UNIVERSIDADE FEDERAL DO MARANHÃO
PROGRMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE
ÁREA: CIÊNCIA DA COMPUTAÇÃO

UMA SOLUÇÃO EFETIVA PARA A APRENDIZAGEM DE RELACIONAMENTOS NÃO TAXONÔMICOS DE ONTOLOGIAS

Ivo José da Cunha Serra

São Luís
2014

UMA SOLUÇÃO EFETIVA PARA A APRENDIZAGEM DE RELACIONAMENTOS NÃO-TAXONÔMICOS DE ONTOLOGIAS

Ivo José da Cunha Serra

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, para a obtenção do título de Doutor em Engenharia Elétrica na área de Ciência da Computação.

Orientadora: Profª. Drª. Rosario Girardi

São Luís
2014

Serra, Ivo José da Cunha

Uma Solução Efetiva para a Aprendizagem de Relacionamentos Não-Taxonômicos de Ontologias / Ivo José da Cunha Serra. – 2014.

212 f.

Impresso por computador (Fotocópia).

Orientadora: Rosario Girardi.

Tese (Doutorado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2014.

1. Ontologias 2. Relacionamentos não-taxonômicos_Aprendizagem 3. Linguagem natual_Processamento 4. Aprendizagem de máquina I. Título

CDU 004.775

UMA SOLUÇÃO EFETIVA PARA A APRENDIZAGEM DE RELACIONAMENTOS NÃO-TAXONÔMICOS DE ONTOLOGIAS

Ivo José da Cunha Serra.

Tese aprovada em 28 de março de 2014.

Profa. Maria del Rosario Girardi, Ph.D.
(Orientadora)

Prof. Paulo Jorge Freitas de Oliveira Novais, Dr.
(Membro da Banca Examinadora)

Prof. Pedro Manuel Rangel Santos Henriques, Dr.
(Membro da Banca Examinadora)

Profa. Renata Vieira, Ph.D.
(Membro da Banca Examinadora)

Prof. Francisco José da Silva e Silva, Dr.
(Membro da Banca Examinadora)

À minha família!

AGRADECIMENTOS

Agradeço a todos que contribuíram direta ou indiretamente para elaboração desta Tese, em especial:

À minha família, pelo o amor, incentivo, dedicação e apoio. À Professora Rosario Girardi, pela orientação segura e presente, pelos ensinamentos e dedicação imprescindíveis para a realização deste trabalho. Aos professores Paulo Novais e Pedro Henriques, por todas as sugestões e contribuições, além do agradável convívio durante minha estada em Portugal.

Agradeço também aos meus companheiros de laboratório, em especial, Carla Faria e Luís Eduardo, pela amizade e apoio no decorrer dessa trajetória e a toda Coordenação da Pós-Graduação, pelos bons serviços oferecidos.

RESUMO

A Aprendizagem de Relacionamentos Não-Taxonômicos é um sub-campo da Aprendizagem de ontologia e constitui uma abordagem para automatizar a extração desses relacionamentos a partir de fontes de informação textuais. As técnicas de aprendizagem de relacionamentos não taxonômicos, da mesma forma que outras na área de Aprendizagem de Ontologias estão sujeitas a uma grande quantidade de ruído uma vez que a fonte de informação da qual extraem os relacionamentos ser desestruturada. Portanto, soluções customizáveis são necessárias para que essas técnicas sejam aplicáveis a maior variedade possível de situações. O presente trabalho apresentou TARNT, uma Técnica para a Aprendizagem de Relacionamentos Não-Taxonômicos de ontologias a partir de textos na língua inglesa que emprega técnicas de Processamento de Linguagem Natural e estatísticas para etiquetar o texto e selecionar os relacionamentos a serem recomendados. O controle sobre a execução de suas regras de extração e consequentemente sobre o *recall* e precisão na fase "Extração de relacionamentos candidatos"; a "regra de apóstrofo", que confere tratamento particular às extrações que tem maior probabilidade de serem relacionamentos válidos e *Bag of labels*, solução para a fase de "Refinamento" que apresenta o potencial de obter maior efetividade que as que operam sobre relacionamentos compostos por um par de conceitos e um rótulo, estão entre seus aspectos positivos. Avaliações experimentais de TARNT foram realizadas conforme dois procedimentos baseados no princípio de comparação dos relacionamentos aprendidos com os de referência. Esses experimentos consistiram em mensurar com as medidas de avaliação *recall* e precisão, a efetividade da técnica na aprendizagem de relacionamentos não-taxonômicos a partir de dois corpora nos domínios da biologia e do direito da família. Os resultados obtidos foram ainda comparados aos de outra abordagem que utiliza o algoritmo de extração de regras de associação na fase de "Refinamento". Esse trabalho demonstrou ainda a hipótese de pesquisa de que: soluções para a fase de "Refinamento" que utilizam relacionamentos compostos por dois conceitos de uma ontologia e um rótulo são menos efetivas que as que refinam relacionamentos compostos apenas por dois conceitos, uma vez que esses tendem a apresentar menores valores para as medidas de avaliação quando considerados os mesmos corpus e ontologia de referência. A demonstração foi realizada por meio de uma exposição teórica que consistiu na generalização das observações realizadas sobre os resultados obtidos por duas técnicas que refinam relacionamentos dos dois tipos considerados.

Palavras-chave: Aprendizagem de relacionamentos não-taxonômicos. Ontologias. Processamento da linguagem natural. Aprendizagem de máquina.

ABSTRACT

Learning Non-Taxonomic Relationships is a sub-field of ontology learning and is an approach to automate the extraction of these relationships from textual information sources. Techniques for learning non-taxonomic relationships just like others in the area of Ontology Learning are subject to a great amount of noise since the source of information from which the relationships are extract is unstructured. Therefore, customizable solutions are needed for these techniques to be applicable to the widest variety of situations. This Thesis presents TARNT, a Technique for Learning for Non-Taxonomic Relationships of ontologies from texts in English that employs techniques from Natural Language Processing and statistics to structure text and to select relationships that should be recommended. The control over the execution of its extraction rules and consequently on the recall and precision in the phase "Extraction of candidate relationships", the "apostrophe rule", which gives particular treatment to extractions that have greater probability to be valid ones and “Bag of labels”, a refinement technique that has the potential to achieve greater effectiveness than those that operate on relationships consisting of a pair of concepts and a label, are among its positive aspects. Experimental evaluations of TARNT were performed according to two procedures based on the principle of comparing the learned relationships with reference ones. These experiments consisted in measuring with recall and precision, the effectiveness of the technique in learning non-taxonomic relationships from two corpora in the domains of biology and family law. The results were compared to that of another approach that uses an algorithm for the extraction of association rules in the Refinement phase. This Thesis also demonstrate the hypothesis that solutions to the Refinement phase that use relationships composed of two ontology concepts and a label are less effective than those that refine relationships composed of only two concepts, since they tend to have lower values for the evaluation measures when considering the same corpus and reference ontology. The demonstration was conducted by a theoretical exposition that consisted of the generalization of the observations made on the results obtained by two techniques that refine relationships of the two types considered.

Keywords: Learning non-taxonomic relationships of ontologies. Ontology. Natural language processing. Machine learning.

ABREVIATURAS E SÍMBOLOS

AM	Aprendizagem de Máquina
AO	Aprendizagem de Ontologias
ARNT	Aprendizagem de Relacionamentos Não-Taxonômicos de ontologias
CREOLE	*Collection of Reusable Objects for Language Engineering*
GATE	General Architecture for Text Engineering
LR	*Language Resources*
OWL	Web Ontology Language
PLN	Processamento da Linguagem Natural
POS	Part of Speech
PR	Processing Resources
REN	Reconhecimento de Entidades Nomeadas
RI	Recuperação de Informação
IA	Inteligência Artificial
TARNT	Técnica de Aprendizagem de Relacionamentos Não-Taxonômicos

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

1.1. Motivação

A demanda por sistemas baseados em conhecimento [17] [40] [54] é crescente considerando suas aptidões para a solução de problemas complexos e para o suporte à tomada de decisão. Estes sistemas têm como principais componentes uma base de conhecimento e um mecanismo de raciocínio capaz de realizar inferências sobre esta base e obter conclusões a partir desse conhecimento. As ontologias [35] [36] [66] são formalismos para a representação de conhecimento, usadas pelos modernos sistemas baseados em conhecimento para representar e compartilhar a informação de um determinado domínio de aplicação. Estes formalismos permitem expressar um conjunto de entidades, seus relacionamentos, restrições, axiomas e o vocabulário de um dado domínio. Através do processamento semântico das informações contidas em uma ontologia, os sistemas baseados em conhecimento podem realizar interpretações mais precisas das informações e, assim, demonstrar maior efetividade e usabilidade que os sistemas de informação tradicionais.

A aquisição de conhecimento para construção de ontologias é um processo custoso e sujeito a erros. Tradicionalmente, as ontologias são desenvolvidas por especialistas de domínio e engenheiros de conhecimento em um trabalho complexo e lento. Esta dificuldade na captura do conhecimento requerido pelos sistemas baseados em conhecimento é conhecida como gargalo da aquisição de conhecimento. Por isso, é desejável a semi-automatização ou automatização desse processo, o que é conhecido como Aprendizagem de ontologias (AO) [10] [11] [15] [16] [17] [33] [37] [64].

A Aprendizagem de Relacionamentos Não-Taxonômicos (ARNT) é um sub-campo da Aprendizagem de ontologia (AO) [10] [11] [15] [16] [17] [33] [37] [64] e constitui uma abordagem para automatizar ou semi-automatizar a extração desses relacionamentos a partir de fontes de informação textuais [59] [60] [61].

Algumas técnicas para a ARNT [29] [42] [48] [55] [56] [69] já foram propostas. Todas elas utilizam técnicas de Processamento de Linguagem Natual (PLN) [4] [8] [19] [22] [23] [68] para anotar o corpus com informação necessária ao processamento subsequente. Por exemplo, a técnica "ARNT baseada em regressão logística" [29] e "ARNT baseada na extração de regras de associação" [69] utilizam *vp chunk* (seção 2.5.10) para identificar as frases verbais que são sugeridas como os rótulos dos relacionamentos, já a técnica "ARNT baseada na extração de regras de associação generalizadas" [42], por não recomendar rótulos aos relacionamentos, não a utiliza. Além dessas, são também utilizadas técnicas de Aprendizagem de Máquina (AM) [9] [46] [49] [50] ou estatísticas [56] [61] para refinar os relacionamentos antes de serem sugeridos ao engenheiro de conhecimento. Entretanto, várias dessas técnicas de ARNT apresentam pelo menos um dos dois aspectos negativos: O primeiro é a exigencia de características específica do corpus, o que consequentemente limita os tipos de documentos aos quais essas técnicas podem ser aplicadas. Esse é o caso de "ARNT baseada na extração de regras de associação generalizadas" [42] (seção 3.2) que utiliza corpora compostos por documentos estruturados em títulos e corpo de texto, para maximizar a quantidade de relacionamentos candidatos obtidos.

O segundo aspecto é a adoção de soluções que não sejam as mais adequadas para cada uma das fases do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3). Esse fato pode acarretar perda de efetividade na aprendizagem de relacionamentos, a qual pode ser formalmente verificada com o uso de procedimentos e medidas de avaliação. Um caso concreto dessa situação, com relação a duas técnicas para a fase de refinamento de ARNT, *Bag of labels* [61] e "Extração de regras de associação" [1] [2] será apresentado e detalhadamente discutido no capítulo 5 dessa Tese. Outra consequência diz respeito à perda no grau de usabilidade da técnica de ARNT devido a informações não sugeridas ao engenheiro de conhecimento. Por exemplo, a técnica "ARNT baseada em extração de regras de associação generalizadas" [42] (seção 3.2) possui a funcionalidade de sugerir o melhor nível hierárquico no qual o relacionamento deva ser acrescentado, entretanto não sugere rótulos a eles. Já a "ARNT baseada na regressão logística" [29]

(seção 3.4) sugere nomes de conceitos da ontologia como frases nominais obtidas do corpus, o que de forma geral não é muito adequado, pois estas não são as de fato normalmente utilizadas para essa finalidade. Todas essas questões serão detalhadamente discutidas na seção 3.6.

Nas seções seguintes são apresentadas respectivamente, uma introdução à problemática de ARNT (seção 1.2), a hipótese de trabalho juntamente com os objetivos que essa pesquisa se propõe a alcançar (seção 1.3) e a organização deste manuscrito (seção 1.4).

1.2.Caracterização do problema

A Aprendizagem de relacionamentos não-taxonômicos de ontologias é um problema de mineração de texto e pode ser decomposto em dois subproblemas: extração inicial de relacionamentos e refinamento dos relacionamentos.

Na extração inicial de relacionamentos, os principais desafios são trabalhar com textos em linguagem natural e a definição de regras que permitam extrair do texto anotado os relacionamentos candidatos. Os textos em linguagem natural são desestruturados. Os sistemas de computação estão aptos a compreender instruções escritas em linguagens de programação, entretanto não o fazem perfeitamente com relação a instruções escritas em linguagem natural. Isso se deve ao fato das linguagens de programação serem formais, o que permitem ao sistema de computação saber exatamente como deve proceder a cada comando. Em linguagens naturais, uma simples frase normalmente contém ambigüidades, nuances e interpretações que dependem do contexto, do conhecimento de mundo, de regras gramaticais, culturais e de conceitos abstratos. Dessa forma é necessário que os textos passem por um processo de estruturação, com a finalidade de identificação dos elementos que compõem um relacionamento não-taxonômico, que são geralmente *strings* correspondentes a conceitos, caso esses sejam dados como *input* para a técnica; frases nominais, caso os conceitos não sejam informados a técnica e frases verbais, usadas como rótulos dos relacionamentos por técnicas que disponibilizam essa funcionalidade.

Já as regras de extração são regras para extrair do texto anotado os relacionamentos candidatos. Por exemplo, a regra de sentença extrai pares de conceitos que estejam em uma mesma sentença e a regra de sentença com frase verbal extrai pares de conceitos e a frase verbal entre eles como seu rótulo. Essas duas regras são formalmente definidas na seção 4.2.3, que descreve as regras de extração de TARNT [60] [61].

Os relacionamentos provenientes da etapa anterior não devem ser todos recomendados ao usuário, pois geralmente há uma quantidade substancial deles que não corresponde a boas sugestões. Por esse motivo, na fase de "Refinamento" podem ser utilizadas técnicas de AM [9] [46] [49] [50] ou estatísticas [56] [61] para sugerir ao especialista os relacionamentos mais prováveis.

1.3.Hipótese de pesquisa e objetivos

1.3.1.Hipótese de pesquisa

Soluções para a fase de refinamento de ARNT que recomendam relacionamentos compostos por dois conceitos de uma ontologia e um rótulo são menos efetivas que as que recomendam relacionamentos compostos apenas por dois conceitos. Formalmente, pretende-se demonstrar que soluções de refinamento que recomendam relacionamentos do tipo $<c_1, fv, c_2>$[1] são menos efetivas que as que recomendam relacionamentos do tipo $<c_1, c_2>$, uma vez que tendem a apresentar menores valores para as medidas de avaliação quando executados no mesmo contexto.

1.3.2.Objetivo geral

Contribuir para aliviar o problema do gargalo de aquisição de conhecimento na aprendizagem de ontologias através da automatização da aquisição de relacionamentos não-taxonômicos a partir de fontes de informação textuais.

[1] "c1" e "c2" são conceitos da ontologia e "fv" é uma frase verbal obtida do corpus

1.3.3. Objetivos Específicos

a) Análise do estado da arte em Aprendizagem de relacionamentos não-taxonômicos para propôr um processo genérico que possa ser utilizado na avaliação de diferentes técnicas quanto às soluções por elas adotadas e também para guiar a proposição de novas técnicas.

b) Desenvolvimento de uma técnica para a aprendizagem de relacionamentos não-taxonômicos de ontologias a partir de fontes textuais que tenha o potencial de apresentar boa efetividade quando comparada a outras para a mesma tarefa.

c) Disponibilização de uma ferramenta de software que dê suporte a aplicação da técnica de ARNT proposta. Essa ferramenta foi utilizada para a realização de experimentos com o objetivo principal de demonstrar a hipótese de pesquisa (seção 1.3.1).

d) Avaliação quantitativa e qualitativa da técnica desenvolvida por meio de sua aplicação na aprendizagem de relacionamentos não-taxonômicos em diferentes domínios como o da biologia e o jurídico e comparação com técnicas do estado da arte.

1.4. Organização da Tese

Esta Tese está organizada em seis capítulos. O segundo apresenta assuntos que correspondem à fundamentação teórica de ARNT, particularmente o Processamento da Linguagem Natural (PLN) [4] [8] [19] [22] [23] [68] e a Aprendizagem de Máquina (AM) [9] [46] [49] [50]; além disso, é definido um processo genérico para ARNT [57] [58] [59] [60] que permite propor novas soluções para a realização dessa tarefa e avaliar comparativamente técnicas já propostas em termos das soluções por elas adotadas para cada uma de suas fases.

O capítulo três apresenta e discute comparativamente algumas propostas representativas do estado da arte de ARNT [29] [42] [48] [55] [56] [69] destacando as soluções por elas empregadas para as fases do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3). O objetivo foi o de identificar

aspectos positivos e limitações nessas técnicas que motivassem a proposição de novas soluções.

O quarto capítulo apresenta TARNT (Técnica para a Aprendizagem de Relacionamentos Não-Taxonômicos de Ontologias) [60] [61], a solução para ARNT proposta nesse trabalho. TARNT [60] [61] obtém esses relacionamentos entre conceitos de uma ontologia conhecidos aprioi a partir de corpora na língua inglesa. É também apresentada TARNTool, a ferramenta de software desenvolvida para dar suporte a aplicação de TARNT [60] [61] e utilizada nos experimentos de ARNT realizados no capítulo 5. Por fim, é apresentada uma discussão sobre as características de TARNT [60] [61] em relação às técnicas de ARNT apresentadas no capítulo 3.

No quinto capítulo são apresentas quatro avaliações realizadas com TARNT [60] [61] e particularmente *Bag of labels* [61] com o objetivo de mensurar sua efetividade na tarefa de ARNT e de demonstrar a hipótese postulada nesse trabalho (seção 1.3). Para tanto foram desenvolvidos dois procedimentos de avaliação de técnicas de ARNT que realizam a comparação dos relacionamentos aprendidos com os de uma ontologia de referência. Os resultados obtidos são apresentados, discutidos e por fim generalizados para quaisquer técnicas de refinamento que trabalhem com relacionamentos dos dois tipos considerados na hipótese: $<c_1, fv, c_2>$ e $<c_1, c_2>$ (seção 1.3).

No sexto e último capítulo são apresentadas e discutidas as conclusões, nos âmbitos científico e tecnológico, obtidas a partir da execução desse trabalho, além de sugeridos trabalhos futuros.

2. Fundamentação teórica sobre ARNT

Nesse capítulo são apresentadas as áreas de conhecimento que fornecem a fundamentação teórica sobre ARNT. As seções 2.1 e 2.2 discorrem respectivamente sobre a definição de ontologia [33] adotada no contexto desse trabalho e a recente área de pesquisa Aprendizagem de ontologias [10] [11] [15] [16] [17] [33] [37]. A seção 2.3 discorre sobre o processo genérico para a aprendizagem de relacionamentos não-taxonômicos proposto nesse trabalho [57] [58] [59] [60]. Ele define os problemas gerais a serem solucionados por qualquer técnica para essa tarefa. Na seção 2.4 são apresentadas algumas das principais formas de realização textual dos relacionamentos não-taxonômicos e as representações desses relacionamentos normalmente adotadas pelas técnicas de ARNT. Na seção 2.5 é feita uma breve apresentação de Processamento da Linguagem Natural [4] [8] [19] [22] [23] [68], subcampo de Inteligência Artificial [54] cujas técnicas são utilizadas para estruturar o texto do qual são extraídos os relacionamentos não-taxonômicos. A seção 2.6 apresenta a Aprendizagem de máquina [9] [46] [49] [50], subcampo de Inteligência artificial [54] cujas técnicas tem sido empregadas para realizar um refinamento dos relacionamentos não-taxonômicos antes de serem definitivamente recomendados ao engenheiro de conhecimento. Por fim a seção 2.7 apresenta um resumo do conteúdo desse capítulo.

2.1. Definição de ontologia

As ontologias [35] [36] [66] são estruturas de representação de conhecimento que ganharam importância na última década. Atualmente, elas são aplicadas, por exemplo, na comunicação de agentes de software [31], na integração da informação [3] [71], na composição de Web Services [65], na descrição do conteúdo para facilitar a sua recuperação [36] [71], no processamento da linguagem natural [4] [8], na Web Semântica [36] [37] [44] e na construção de sistemas baseados em conhecimento [17] [40] [54].

Uma ontologia é uma especificação formal e explícita de uma conceituação compartilhada de um domínio de interesse [35]. Conceituação refere-se a um modelo abstrato de algum fenômeno do mundo. Explícito

significa que o tipo de conceitos utilizados e as limitações do seu uso, são explicitamente definidos. Formal refere-se ao fato de que a ontologia deve ser legível por máquina. Compartilhada reflete a noção de que uma ontologia captura o conhecimento consensual, isto é, não é privada de algum indivíduo, mas aceita por um grupo.

Em [51] define-se ontologia como os termos básicos e os relacionamentos que constituem o vocabulário de uma área temática, bem como as regras para combinar termos e relacionamentos para definir extensões ao vocabulário. Formalmente, uma ontologia pode ser representada por uma 6-tupla (1) [33]:

$$O = (C, H, I, R, P, A) \quad (1)$$

a) $C = C^C \cup C^I$ é o conjunto de entidades da ontologia, ou seja, ele representa as entidades do domínio sendo modelado. São designados por um ou mais termos em linguagem natural. O conjunto C^C é formado por classes, ou seja, conceitos que representam entidades que descrevem um conjunto de objetos (por exemplo, “Pessoa” $\in C^C$) enquanto o conjunto C^I é formado por instâncias, ou seja, entidades únicas no domínio (por exemplo, “Anne” $\in C^I$);
b) $H = \{\text{tipo\_de } (c_1,c_2) \mid c_1 \in C^C \wedge c_2 \in C^C \}$ é o conjunto das relações taxonômicas entre os conceitos. Tais relações definem a hierarquia de conceitos e são denotadas por "*tipo_de"*(c_1,c_2) significando que c_1 é um tipo de c_2. Um exemplo desse relacionamento é *"tipo_de"(Cliente,Pessoa)*;
c) $I = \{\text{é\_um } (c_1,c_2) \mid c_1 \in C^I \wedge c_2 \in C^C \}$ é o conjunto de relacionamentos entre classes e instâncias (relacionamento “é um”) de uma ontologia, por exemplo *"é_um"(Erick,Advogado)*;
d) $R = \{rel_k (c_1,c_2,..., c_n) \mid \forall i, c_i \in C\}$ é o conjunto de relacionamentos que não são nem taxonômicos nem de instanciação entre classes e instâncias de uma ontologia. Alguns exemplos são *“representa”(Advogado, Cliente)* e “*representa”(Erick, Anne)*;

e) $P = \{prop^C (c_k, tipo) \mid c_k \in C^C\} \cup \{prop^I (c_k, valor) \mid c_k \in C^I\}$ é o conjunto de propriedades das entidades de uma ontologia. O relacionamento $prop^C$ define o tipo básico de dado de uma propriedade de classe, enquanto o relacionamento $prop^I$ define os valores das instâncias. Por exemplo, *"assunto" (Caso, String)* é uma exemplo de uma propriedade do tipo $prop^C$ e *"assunto" (Caso12, adoção)* é um exemplo de propriedade $prop^I$.

f) $A = \{condição_x \Rightarrow conclusão_y (c_1, c_2, ..., c_n) \mid \forall j, c_j \in C^C\}$ é um conjunto de axiomas, regras que permitem checar a consistência da ontologia e deduzir novos conhecimentos através de algum mecanismo de inferência. O termo $Condição_x$ é dado por: $Condição_x = \{(cond_1, cond_2, \ldots, cond_n) \mid \forall_z, cond_z \in H \cup I \cup R\}$. Por exemplo: *"aplicada"(Argumento_Defesa22, Caso12)* ∧ *"similar"(Caso12, Caso13)* ⇒ *"aplicada"(Argumento_Defesa22, Caso13)* é uma regra que indica que esses dois casos legais são similares e, conseqüentemente, o argumento de defesa usado em um caso pode ser aplicado ao outro.

Para esclarecer esta definição, tomemos como exemplo uma ontologia simples que descreve o domínio de um escritório de advocacia mostrada na figura 1. Cada advogado é responsável pelos casos dos clientes que ele atende.

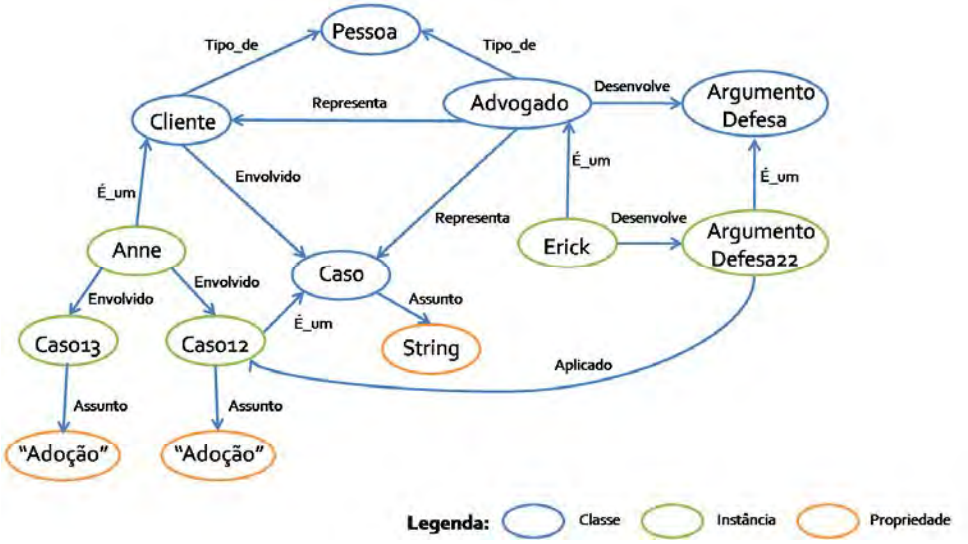


Figura 01: Exemplo de ontologia de um escritório de advocacia

Dessa forma, têm-se:

a) $C^C = \{\text{pessoa}, \text{cliente}, \text{advogado}, \text{caso}, \text{argumento de defesa}\}$
b) $C^I = \{\text{Erick}, \text{Anne}, \text{Caso12}, \text{Caso13}, \text{Argumento\_Defesa22}\}$
c) $H = \{\text{tipo\_de}(\text{cliente}, \text{pessoa}), \text{tipo\_de}(\text{advogado}, \text{pessoa})\}$
d) $I = \{\text{é\_um}(\text{Erick}, \text{advogado}),$
$\text{é\_um}(\text{Anne}, \text{cliente}), \text{é\_um}(\text{Caso12}, \text{caso}), \text{é\_um}(\text{Caso13}, \text{caso}),$
$\text{é\_um}(\text{Argumento\_Defesa22}, \text{argumento\_defesa})\}$
e) $R = \{\text{representa}(\text{advogado}, \text{cliente}), \text{representa}(\text{advogado}, \text{caso}),$
$\text{aplicado}(\text{argumento\_defesa}, \text{caso}),$
$\text{desenvolve}(\text{advogado}, \text{argumento\_defesa}), \text{envolvido}(\text{cliente}, \text{caso})\}$
f) $P = \{\text{assunto}(\text{caso}, \text{String}), \text{assunto}(\text{Caso12}, \text{"Adoção"}),$
$\text{assunto}(\text{Caso13}, \text{"Adoção"})\}$
g) $A =$
$\{\text{aplicado}(\text{Argumento\_Defesa22}, \text{Caso12}) \wedge$
$\text{similar}(\text{Caso12}, \text{Caso13}) \rightarrow$
$\text{aplicado}(\text{Argumento\_Defesa22}, \text{Caso13})\}$

2.2.Aprendizagem de ontologias

O termo aprendizagem de ontologias refere-se ao suporte automatizado ou semi-automatizado à construção de ontologia, enquanto o suporte automatizado ou semi-automatizado à instanciação de uma dada ontologia é chamado de povoamento de ontologia [10] [11] [64]. Juntos, a aprendizagem e o povoamento de ontologias constituem uma abordagem para automatizar a aquisição de conhecimento através da descoberta de conhecimento em diferentes fontes de dados e representando-o através de ontologias.

De acordo com Benz [6] há dois aspectos fundamentais na aprendizagem de ontologias. O primeiro é a disponibilidade de conhecimento prévio, que pode ser na forma de uma ontologia a ser estendida ou pode ser transformado na primeira versão da ontologia. Tal versão é então estendida automaticamente por procedimentos de aprendizagem ou manualmente pelo engenheiro de conhecimento [24].

O outro aspecto é o formato das fontes de dados a partir das quais se deseja extrair conhecimento. Existem três diferentes tipos de fontes de dados [6]:

a) fontes desestruturadas: documentos em linguagem natural, como documentos PDF, Word e como a maioria das páginas da Web tradicional;

b) fontes semi-estruturadas: dicionários e folksonomias;

c) fontes estruturadas: esquemas de bancos de dados.

Algumas abordagens para a aprendizagem de ontologias a partir de fontes estruturadas [41] e semi-estruturadas [6] [44] foram propostas e apresentaram bons resultados. Contudo, apesar de tais abordagens proverem um determinado suporte ao desenvolvimento de ontologias, a maior parte do conhecimento disponível, especialmente na Web, está na forma de textos em linguagem natural [42]. Por isso, a aprendizagem de ontologias a partir de fontes textuais [15] [16] é um ponto central para algumas áreas como o estabelecimento da Web Semântica.

2.3.O processo genérico de ARNT

A Aprendizagem de Relacionamentos Não-Taxonômicos (ARNT) constitui uma abordagem para automatizar ou semi-automatizar a extração desses relacionamentos a partir de fontes de informação textuais [60] [61]. Os relacionamentos não-taxonômicos correspondem ao conjunto “R” da definição de ontologia (seção 2.1). Por exemplo, “representa” é um relacionamento não-taxonômico entre as classes advogado e cliente.

A ARNT pode ser realizada genericamente através das cinco atividades [57] [58] [59] [60]: “Construção do corpus”; “Extração de relacionamentos candidatos”, composta por “Anotação do corpus” e “Extração de relacionamentos”; “Refinamento”; “Avaliação do especialista” e “Atualização da ontologia” (figura 2).

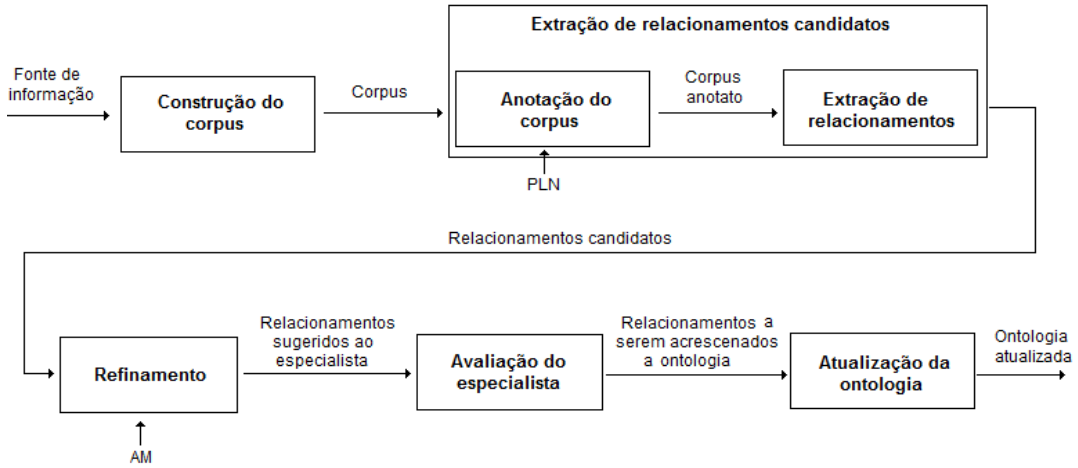


Figura 02: Processo genérico para ARNT [57] [58] [59] [60]

A tarefa de construção do corpus consiste em selecionar documentos sobre o domínio no qual se deseja extrair os relacionamentos. Essa é normalmente uma tarefa custosa e o resultado de qualquer técnica de ARNT depende da qualidade do corpus.

A "Extração de relacionamentos candidatos" tem como finalidade apresentar um conjunto de prováveis relacionamentos, tem como input o corpus construído na fase anterior e como produto relacionamentos candidatos; além disso, é composta pelas sub-atividades: "Anotação do corpus" e "Extração de relacionamentos". A "Anotação do corpus" consiste em anotar o corpus com as técnicas de PLN que forem necessárias à continuação do processo de extração dos relacionamentos. Técnicas de ARNT que não recebem os conceitos da ontologia com *input*, normalmente consideram as frases nominais como tais e, portanto necessitam da execução do *NP chunk*. As técnicas que recomendam frases verbais como rótulos dos relacionamentos necessitam da execução do *VP chunk*. Já as técnicas que não recebem os conceitos da ontologia como input e não recomendam rótulos verbais necessitam obrigatoriamente apenas da tokenização e separação de sentenças e opcionalmente da lematização.

A Extração de relacionamentos candidatos consiste em buscar no corpus anotado as evidencias que sugiram a existência de um relacionamento. Por exemplo, para Maedech e Staab [42] a existência de duas instancias de

conceitos da ontologia em uma sentença é uma evidencia de que estes estão relacionados, já para Villaverde et al. [69] um relacionamento é representado pela presença de dois conceitos da ontologia em uma mesma sentença com uma frase verbal entre eles.

Os relacionamentos provenientes da etapa anterior não devem ser todos recomendados ao especialista, pois geralmente há uma quantidade substancial desses que não correspondem a boas sugestões. Por esse motivo, na fase de "Refinamento", técnicas de aprendizagem de máquina [9] [46] [49] [50] são normalmente utilizadas para refinar aqueles a serem recomendados ao especialista.

Na fase "Avaliação pelo especialista" o especialista, seleciona e possivelmente edita dentre os relacionamentos disponibilizados pela fase anterior, aqueles a serem de fato acrescentados a ontologia.

Por fim, na fase "Atualização da ontologia" é executado um procedimento para atualizar o arquivo da ontologia com os relacionamentos que foram definitivamente selecionados pelo especialista. A tabela 1 resume as fases do processo genérico para ARNT [57] [58] [59] [60].

<table>
<tr><th colspan="2">Fase</th><th colspan="2">Descrição</th></tr>
<tr><td colspan="2">Construção do corpus</td><td colspan="2">Seleção de documentos em quantidade e qualidade necessários à ARNT</td></tr>
<tr><td rowspan="2">Extração de Relacionamentos Candidatos</td><td>Anotação do corpus</td><td>Anotação do corpus com o uso de técnicas de PLN necessárias à continuidade da ARNT</td><td rowspan="2">Extração de um conjunto inicial de relacionamentos</td></tr>
<tr><td>Extração de Relacionamentos</td><td>Aplicação do algoritmo de extração dos relacionamentos a partir do corpus anotado</td></tr>
<tr><td colspan="2">Refinamento</td><td colspan="2">Aplicação de AM para sugerir os relacionamentos mais prováveis</td></tr>
<tr><td colspan="2">Avaliação pelo especialista</td><td colspan="2">Seleção e possível edição dos relacionamentos filtrados</td></tr>
<tr><td colspan="2">Atualização da ontologia</td><td colspan="2">Atualização do arquivo da ontologia com os relacionamentos selecionados pelo especialista</td></tr>
</table>

Tabela 01: Descrição das fases do Processo genérico de ARNT [57] [58] [59] [60]

Para realizar a comparação dos relacionamentos não-taxonômicos obtidos com as técnicas de ARNT com os relacionamentos não-taxonômicos das ontologias de referência são utilizadas as medidas definidas por Dellschaft e Staab [24], o *recall*, a precisão e a medida-F, adaptadas das versões já conhecidas da área de Recuperação de Informação [5]. A adoção dessas medidas para avaliação da efetividade de técnicas de ARNT se deve ao fato de serem as utilizadas pela grande maioria dos trabalhos do gênero [29] [42] [48] [55] [56] [69].

Sejam R_R o conjunto de relacionamentos não-taxonômicos na ontologia de referência (o conjunto de conceitos que deveriam ser extraídos do corpus) e R_L o conjunto de relacionamentos aprendido (o conjunto de relacionamentos extraídos pelo procedimento de aprendizagem) então, o *recall*, a precisão e a medida-F são definidos nas equações 2, 3 e 4, respectivamente.

$$Recall = \frac{|R_R \cap R_L|}{|R_R|} \tag{2}$$

$$\text{Precisão} = \frac{|R_R \cap R_L|}{|R_L|} \tag{3}$$

$$\text{Medida} - \text{F} = \frac{(2 * precisão * recall)}{(precisão + recall)} \tag{4}$$

2.4. Realizações textuais e representações dos relacionamentos não-taxonômicos

Uma vez que a fonte de informação das técnicas de ARNT é textual é necessária uma discussão sobre como os relacionamentos não-taxonômicos podem estar representados no texto. Há diferentes possibilidades, entretanto citamos aqui três dentre as mais comuns.

A primeira consiste em: conceitos da ontologia com uma frase verbal entre eles. Essa é a representação tipicamente considerada por técnicas de ARNT que sugerem rótulos aos relacionamentos. Esse é o caso das técnicas

“ARNT baseada em consultas a Web” [56], “ARNT baseada em regressão logística” [29], “ARNT baseada na extração de regras de associação” [69] e TARNT [60] [61] que serão apresentadas e discutidas nos capítulos 3 e 4. Um exemplo desse tipo de representação é o do relacionamento "<decree, protect, property>" entre os conceitos "decree" e "property" que pode ser extraído a partir da sentença: “The court decree protects the property rights of the parties and provides support for the children”.

A segunda forma de representação textual consiste em: conceitos da ontologia em uma mesma sentença. Essa é uma das representações normalmente levada em consideração por técnicas de ARNT que não sugerem rótulos aos relacionamentos. Esse é o caso da técnica “ARNT baseada na extração de regras de associação generalizadas” [42] apresentada e discutida no capítulo 3. Um exemplo desse tipo de representação é o do relacionamento "<court, decree>" que pode ser extraído a partir da sentença: “The court decree protects the property rights of the parties and provides support for the children”.

Por último mencionamos a representação textual que consiste em: conceitos da ontologia na mesma sentença com “” ou “'s” entre eles. Essa representação indica maior probabilidade de haver relacionamento entre os conceitos que as duas apresentadas anteriormente e é normalmente utilizada por técnicas que não sugerem rótulos aos relacionamentos. Das técnicas discutidas nesse trabalho TARNT [60] [61] é capaz de extrair relacionamentos com essa forma de representação textual. Um exemplo desse tipo de representação é o do relacionamento "<party, lawyer>" que pode ser extraído a partir da sentença “The court can order one party to pay the fee for the other party’s lawyer and for all costs closely related to bringing or enforcing an action for divorce”.

Além das formas de realização textual dos relacionamentos convêm apresentar as formas mais comuns de representação dos relacionamentos não-taxonômicos utilizadas pelas técnicas de ARNT.

A primeira delas é a que recomenda rótulos aos relacionamentos. Os rótulos são tipicamente frases verbais presentes na sentença entre a ocorrência de dois conceitos. Há dois subtipos para essa representação: a utilizada por técnicas que recebem os conceitos da ontologia (conjunto C,

seção 2.1) como *input* e a utilizada pelas técnicas que não o fazem. Para o primeiro subtipo, a representação interna é <c_1, fv, c_2> onde "c_1" e "c_2" são conceitos da ontologia e "fv" é uma frase verbal. Por exemplo, se considerarmos a sentença "The court decree protects the property rights of the parties and provides support for the children" e "decree" e "property" como conceitos de uma ontologia o relacionamento "<decree, protect, property>" seria extraído. Exemplos de técnicas que utilizam essa representação são a "ARNT baseada na extração de regras de associação" [69] e TARNT [60] [61], que serão apresentadas e discutidas nos capítulos 3 e 4 respectivamente.

Técnicas que não recebem os conceitos da ontologia como input utilizam frases nominais extraídas do corpus para tal finalidade. Nesse caso a representação interna é <fn_1, fv, fn_2> onde: fn_1 e fn_2 são frases nominais e fv é uma frase verbal. Por exemplo, a partir da sentença "The judge granted the custody of the child to his grandmother." o relacionamento <the judge, granted, the custody> seria extraído. Exemplos de técnicas que utilizam essa representação são a "ARNT baseada em consultas a Web" [56] e "ARNT baseada na regressão logística" [29], apresentadas e discutidas no capítulo 3.

A segunda representação é a que não recomenda rótulos aos relacionamentos. Aqui novamente há dois subtipos, dependendo do recebimento ou não dos conceitos da ontologia como *input*.

O primeiro subtipo é o utilizado por técnicas que recebem os conceitos da ontologia (conjunto C, seção 2.1) como *input*. Sua representação tem a forma <c_1, c_2> onde c_1 e c_2 são conceitos da ontologia. Por exemplo, considerando "court" e "decree" como conceitos de uma ontologia e a sentença "The court decree protects the property rights of the parties and provides support for the children" o relacionamento "<court, decree>" seria extraído. Um exemplo de técnica que utiliza essa representação é a "ARNT baseada na extração de regras de associação generalizadas" [42], apresentada e discutida no capítulo 3.

O segundo subtipo tem o formato <fn_1, fn_2>; onde fn_1 e fn_2 são frases nominais. Esse formato é utilizado quando os conceitos da ontologia não são dados como input para a técnica de ARNT. Das técnicas do estado da arte discutidas no capítulo 3, nenhuma utiliza esse tipo de representação. A tabela 2

resume os quatro tipos de representações de relacionamentos não-taxonômicos apresentados.

Recomendação de rótulos	Conceitos como *input*	Representações dos relacionamentos não-taxonômicos
Sim	Sim	$<c_1, fv, c_2>$
	Não	$<fn_1, fv, fn_2>$
Não	Sim	$<c_1, c_2>$
	Não	$<fn_1, fn_2>$

Tabela 02: Representações dos relacionamentos não-taxonômicos

2.5. Áreas de conhecimento aplicadas a ARNT: Processamento da linguagem natural

Segundo Drake [27], Processamento da Linguagem Natural (PLN) é um conjunto de técnicas computacionais teoricamente motivadas para analisar e representar textos que ocorrem naturalmente, em um ou mais níveis de análise lingüística para o fim de atingir a maneira humana de processamento da linguagem para uso em uma gama de tarefas ou aplicações.

Dessa forma, o objetivo da PLN é realizar o processamento da linguagem de modo humano. Há mais objetivos práticos para PLN, muitos relacionados com aplicações específicas para as quais ele está sendo usado. Por exemplo, um sistema de recuperação de informação [5] baseado em PLN tem o objetivo de fornecer a informação mais precisa e completa em resposta a real necessidade de informação dos usuários.

A noção de níveis de análise lingüística refere-se ao fato que há múltiplos tipos de conhecimento lingüístico no processamento da linguagem natural trabalhados quando humanos produzem ou compreendem uma linguagem. Pensa-se que os seres humanos normalmente usam todos esses níveis, porque cada nível transmite diferentes tipos de significado. Os sistemas de PLN usam diferentes níveis ou uma combinação de níveis de análise

lingüística de acordo com a tarefa que desempenham. Há uma variedade de métodos e técnicas para a realização de um determinado tipo de análise de linguagem.

Técnicas de PLN [4] [8] [19] [22] [23] [68] são utilizadas na fase de "Anotação do corpus", do processo genérico de ARNT [57] [58] [59] [60] com o objetivo de estruturar o texto. A aplicação dessas técnicas é pré-requisito para todo o procesamento subsequente na aprendizagem de relacionamentos não-taxonômicos.

2.5.1.Conhecimentos Lingüísticos

Qualquer abordagem para o processamento da linguagem natural requer maior ou menor grau de conhecimento lingüístico. Sistemas de linguagem natural usam certos conhecimentos sobre a estrutura da linguagem, incluindo o que as palavras são, como elas são combinadas para formar as sentenças, o que elas significam e como elas contribuem para o significado das sentenças [4]. Os níveis de conhecimento lingüístico são: fonético e fonológico, léxico, morfológico, sintático, semântico, de discurso e pragmático [27].

Nas seções seguintes são apresentados cada um desses níveis acompanhados de alguns exemplos na língua Inglesa.

2.5.2.Nível Fonético e Fonológico

O conhecimento fonético e fonológico está relacionado ao sistema de sons de uma língua. Segundo Vieira e Lima [68]: "A fonética está relacionada ao estudo da produção da fala humana, considerando as questões fisiológicas envolvidas, tais como a estrutura do aparelho fonador: mandíbula, laringe, boca, dentes e língua. [...] A fonologia é o estudo das regras abstratas e princípios envolvidos na organização, estrutura e distribuição dos sistemas de sons de uma determinada língua".

É necessário um conjunto de símbolos que representem os sons de uma língua, pois, em vários idiomas, sons diferentes podem estar associados a uma mesma grafia e várias grafias podem representar um mesmo som. Esse

nível de conhecimento é a base para o funcionamento de sistemas de reconhecimento e síntese de voz.

Em um sistema de PLN de reconhecimento de voz, ou seja, que aceita fala como entrada, as ondas de som são analisadas e codificadas em sinais digitais para posterior interpretação por meio de regras ou de comparação utilizando um modelo particular de linguagem [27]. Já no processo da síntese da fala em sistemas de PLN ocorre o inverso: uma representação digital é convertida em fala.

2.5.3.Nível Léxico

O Léxico pode ser definido como o acervo de palavras de um determinado idioma: todo o universo de palavras que as pessoas de uma determinada língua têm à sua disposição para expressar-se, oralmente ou por escrito. Podemos dizer que uma característica básica do léxico é sua mutabilidade, já que ele está em constante evolução. Algumas palavras se tornam arcaicas, outras são incorporadas, outras mudam seu sentido, e tudo isso ocorre de forma gradual e quase imperceptível. O sistema léxico de uma língua traduz a experiência cultural acumulada por uma sociedade através do tempo, ou seja, o léxico pode ser considerado como o patrimônio vocabular de uma comunidade linguística através de sua história, um acervo que é transmitido de uma geração para a geração seguinte. O usuário da língua utiliza o léxico, esse inventário aberto de palavras disponíveis no seu idioma, para a formação do seu vocabulário, para sua própria expressão no momento da fala ou escrita para a efetivação do processo comunicativo. Assim, o vocabulário de um indivíduo caracteriza-se pela seleção e pelo emprego pessoal que ele faz do léxico. Quanto maior for o vocabulário do usuário, maior a possibilidade de escolha da palavra mais adequada ao seu intento expressivo.

No contexto de PLN o léxico é a estrutura de dados que contém os itens lexicais e informações correspondentes a esses itens. Os itens que constituem as entradas em um léxico podem ser palavras isoladas, como: “*bus*” e “*stop*”, ou composições de palavras, as quais reunidas apresentam um significado específico, por exemplo: “*bus stop*”. Entre as informações

associadas aos itens lexicais estão a categoria gramatical (exemplo: substantivo, advérbio), gênero, número, grau, pessoa, tempo e regência verbal e nominal.

2.5.4.Nível Morfológico

O nível morfológico estuda a estrutura e a formação das palavras. A peculiaridade da morfologia é estudar as palavras olhando para elas isoladamente e não dentro da sua participação na frase.

O nível morfológico consiste em analisar como as palavras são construídas a partir de unidades básicas de significado chamadas morfemas [27]. Algumas palavras, como árvore, não podem ser quebradas em unidades menores, mas isso pode ocorrer com palavras como árvores ou arvorezinhas, por exemplo. Ou ainda palavras como impossível, ou sobremesa. Os morfemas podem ser independentes, como em árvore ou dependentes como no caso dos sufixos (“s” em árvores) e prefixos (“im” em impossível) [27].

Por exemplo, a palavra “pré-registration” pode ser morfologicamente analisada em três morfemas separados: o prefixo “pré”, a raiz “registra”, e o sufixo “tion”. Podemos dividir uma palavra desconhecida em seus morfemas constituintes para compreender o seu significado. Por exemplo, adicionar o sufixo “ed” ao verbo em língua inglesa, faz saber que a ação do verbo teve lugar no passado.

Além de estudar a estrutura das palavras, em morfologia estuda-se a classificação das palavras em diferentes categorias, ou, conforme o termo popularmente conhecido na área, as palavras são classificadas em partes do discurso (part-of-speech ou POS). Entre tais categorias encontramos substantivos (cachorro), verbos (correr), adjetivos (grande), preposições (em), e advérbios (rapidamente). As palavras de uma mesma categoria compartilham várias propriedades em comum como, por exemplo, o tipo de plural (+ s) ou o tipo de diminutivo (+ “inho”). Os verbos e suas conjugações podem apresentar modificações regulares em vários casos. Na língua inglesa, os adjetivos podem ser acompanhados dos sufixos “er” e “est”, como em *big*, *bigger*, *biggest*, significando uma troca de adjetivo comum para um adjetivo comparativo ou

superlativo. As categorias de palavras podem ainda ser divididas em classes abertas ou fechadas. As classes abertas são compostas por categorias que abrangem um grande número de palavras e podem, ainda, abrigar o surgimento de novas palavras. Classes dessas naturezas são os substantivos, verbos e adjetivos. As classes fechadas são aquelas que têm funções gramaticais bem definidas, tais como artigos, demonstrativos, quantificadores, conjunções e preposições [27].

Outra característica compartilhada entre as palavras de uma mesma categoria é a contribuição da palavra para o significado da frase que a contém. Por exemplo, substantivos podem ser usados para identificar um objeto ou conceito determinado, e adjetivos são usados para qualificar esse objeto ou conceito. Ainda a categoria pode dizer algo sobre a posição que as palavras podem ocupar nas frases. As palavras de determinada categoria podem ser usadas como base de um determinado grupo (ou sintagma). Tais palavras são chamadas de núcleo e identificam o tipo de objeto ou conceito que o sintagma descreve. Por exemplo, os sintagmas nominais possuem por núcleo um substantivo (ou nome); em o cachorro, o cachorro raivoso ou em o cachorro raivoso do canil, temos sintagmas nominais que descrevem o mesmo tipo de objeto. Da mesma forma, os sintagmas adjetivais faminto, muito faminto, faminto como um cavalo, descrevem um mesmo tipo de qualidade [27].

2.5.5.Nível Sintático

O nível sintático estuda como as palavras podem ser justapostas para formar frases corretas e determina o papel estrutural que cada palavra desempenha na frase [27] [45]. Assim, esse nível foca na análise das palavras em uma sentença de maneira a descobrir a estrutura gramatical da sentença. Isso requer tanto uma gramática, que é um conjunto finito de regras e princípios, quanto um analisador sintático (comumente conhecidos por sua denominação em inglês, “parser”) e cada sentença pode ser armazenada em uma árvore sintática.

Uma gramática define regras válidas para organização das palavras em frases. A ordem em que as palavras aparecem em uma frase é usada para

identificar a composição de constituintes que têm funções bem definidas na frase, como, por exemplo, sujeito e predicado.

As árvores sintáticas são usadas para representar a constituição das frases de acordo com as regras estabelecidas pela gramática. Por exemplo, na Figura 03 é apresentada a árvore sintática para a frase "*The dog chased the cat*" (O cão perseguiu o gato). A sentença "S" é formada pelo sintagma nominal NP ("*The dog*") e pelo sintagma verbal VP ("*chased the cat*"). Os sintagmas nominal e verbal, por sua vez, são formados por outros elementos como artigo (ART), verbo(V) e nome(N).

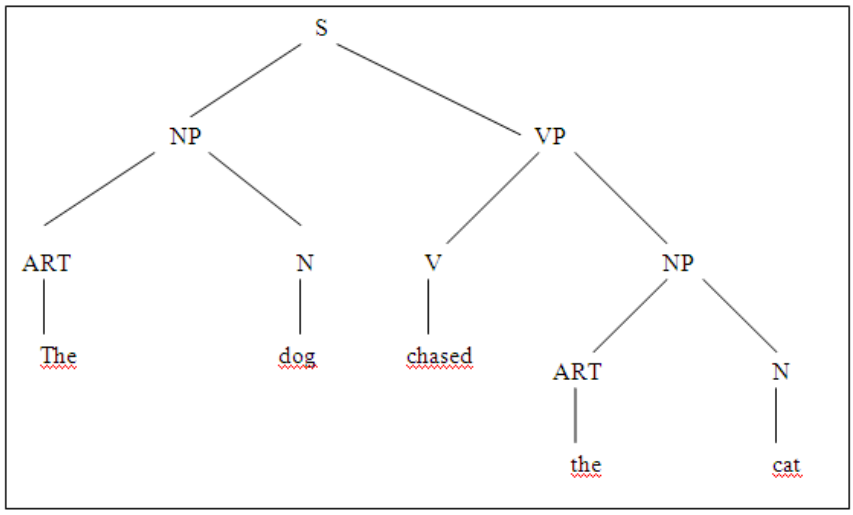


Figura 03: Exemplo de árvore sintática

O uso de sintagmas é fundamental para descrição de cadeias permitidas da linguagem, e serve, por exemplo, para identificar objetos do mundo, que em geral são representados por sintagmas nominais. Exemplos: "*the dog*" e "*the cat*". Nem todas as aplicações de PLN requerem uma análise sintática completa das sentenças, algumas utilizam apenas uma análise superficial ou parcial.

A sintaxe transmite significado [27] já que a mudança na ordem dos elementos de uma sentença pode alterar o sentido da frase. Por exemplo, as duas sentenças "*The dog chased the cat*" (o cão perseguiu o gato) e "*The cat chased the dog*" (o gato perseguiu o cão), diferem somente em termos de sintaxe e transmitem significados completamente diferentes. Nesse exemplo, percebe-se claramente que a troca na ordem dos termos "*dog*" e "*cat*" provocou uma alteração na determinação do agente que sofreu e no que executou a

ação. Na primeira frase o cão realiza a ação, enquanto que na segunda o gato é o executor.

2.5.6.Nível Semântico

A semântica se ocupa com a identificação do significado em nível de sentença. Nela é feito o levantamento dos possíveis significados de uma frase com base na relação entre as palavras que a constituem.

Segundo Gonzalez e Lima [34] "a semântica está relacionada ao significado, não só das palavras, mas também do conjunto resultante delas". Assim, o processamento semântico determina os significados possíveis de uma sentença com base nas interações entre os significados das palavras nela contidas.

A área da semântica é uma área de estudo mais complexa que a área da sintaxe, por apresentar questões que são difíceis de tratar de maneira exata e completa. Isso se deve ao fato de que a mesma frase pode ter significados diversos, o que somente pode ser resolvido por meio da análise situacional feita na etapa de interpretação contextual.

A Interpretação contextual leva em conta o fato de que as mesmas palavras podem ter significados diferentes em situações diferentes, ou seja, leva em consideração o contexto no qual as palavras estão inseridas para descobrir seu real significado.

Vieira e Lima [68] expressam a dificuldade de se trabalhar com interpretação contextual fazendo o seguinte comentário entre o processamento de textos em nível de discurso (contexto lingüístico) e em nível pragmático (contexto situacional): "O contexto lingüístico é o mais fácil de tratar computacionalmente, pois refere-se ao que é explicitado no texto. [...] É mais difícil tratar computacionalmente o contexto imediato, ou contexto situacional de uma expressão, devido à dificuldade de se chegar a uma representação adequada do conhecimento compartilhado entre os participantes de uma conversação ou comunicação". Por exemplo, considerando a sentença "*Edward is dead.*" (Edward está morto.), um analisador semântico poderia apresentar os significados que Edward está *sem vida* ou *cansado*.

2.5.7.Nível de Discurso

A sintaxe e a semântica trabalham com unidades da sentença, enquanto o nível de discurso de PLN trabalha com o texto como um todo, ou seja, não interpreta sentenças isoladamente. A análise de discurso faz conexões entre as frases componentes do texto. A análise do Discurso trata do fato de que sentenças precedentes afetam a interpretação da próxima sentença [4]. Existem diversos tipos de processamento de discurso; dois dos mais comuns é a resolução de anáfora e reconhecimento de estrutura discurso/texto [27]. A resolução de anáfora é a substituição de palavras tais como pronomes, que são semanticamente vagos, pelas apropriadas entidades as quais se referem. Por exemplo, na sentença "*John saw Mike with his ball.*" (John viu Mike com sua bola.), não se sabe ao certo de quem é a bola, ou seja, não se pode afirmar se o pronome possessivo "*his"* faz referência ao termo *"John"* ou *"Mike"* como proprietário da bola. Entretanto, é possível identificar a quem pertence a bola caso a sentença esteja precedida por uma frase esclarecedora, como por exemplo: "*John was looking for his ball*" (John estava procurando sua bola.). Dessa forma, pode-se saber que a bola é de John.

O reconhecimento de estrutura de discurso/texto determina as funções das sentenças no texto, que por sua vez, adicionam ao texto representações significativas. Por exemplo, artigos de jornais podem ser desconstruídos em componentes do discurso tais como história principal, eventos anteriores, avaliações por parte do escritor, citações, dentre outros.

2.5.8.Nível Pragmático

Este nível está relacionado com o uso intencional da linguagem. Nele procura-se obter o significado não literal fazendo uso do conteúdo e contexto do texto, e de outros tipos de conhecimentos mais amplos para a compreensão da mensagem que está no documento.

A análise pragmática ocupa-se com o uso das sentenças em diferentes situações e como isso afeta a interpretação da sentença [4]. Ela se aproximar do modo como as pessoas interpretam as entrelinhas do que lêem e escutam. O objetivo é explicar como o significado extra é lido em textos sem

realmente estar codificados neles. Isto requer muito conhecimento de mundo, incluindo a compreensão de intenções, planos e objetivos. Algumas aplicações de PLN podem usar bases de conhecimento e módulos de inferência que tentam simular esse comportamento.

Nessa fase, também conhecida como interpretação contextual, é feita a ligação das frases do texto entre si e a interpretação da mensagem transmitida, de acordo com a situação e com as condições do enunciado.

Por exemplo, a sentença "*The city councilors refused the demonstrators a permit because they advocated revolution.*" (Os vereadores recusaram aos manifestantes uma autorização, porque eles defendiam a revolução) requer a resolução do termo anafórico *"they*", mas essa resolução exige conhecimento pragmático ou de mundo para ser feita caso o texto de onde foi extraída a sentença não ofereça pistas para a resolução por meio de análise de discurso. Nesse exemplo, não se sabe ao certo se o pronome "*they"* se refere a "*city councilors" ou a "demonstrators".* Somente as pessoas que tenham o conhecimento da situação em que a frase foi formulada e usada podem determinar a que termo o pronome se refere.

2.5.9.Aplicações que usam PLN

A seguir temos alguns tipos de aplicações citados por que usam PLN:

a) Recuperação de informação (RI) – Sistemas de RI fornecem uma lista de documentos potencialmente relevantes em resposta a uma consulta do usuário [5].
b) Extração de informação (EI) – É a mais recente área de aplicação e focaliza o reconhecimento, rotulação e extração de elementos de informação (pessoas, companhias, localizações, organizações, etc.) a partir de grandes coleções de textos [12] [18] [28].
c) Pergunta-resposta – Sistemas de perguntas e respostas fornecem ao usuário apenas o texto da própria resposta ou passagens de respostas disponíveis [70].

d) Resumo – Os altos níveis de PLN, particularmente o nível de discurso, podem ser utilizados na implementação de aplicações que produzem a partir de um texto, outro menor constituído de uma representação narrativa abreviada do documento original.
e) Máquina de tradução – Talvez a mais antiga de todas as aplicações que usam PLN. Os vários níveis da PLN têm sido usados em tradutores, que vão desde as abordagens baseadas em palavras (nível léxico) até aplicações que incluem altos níveis de análise.
f) Sistema de Diálogo – Talvez esse tipo de sistema seja a aplicação do futuro. Normalmente se concentram em tarefas estritamente definidas (por exemplo, sistemas de reconhecimento de voz usados em portais de voz e agendas de telefones celulares) e usam predominantemente os níveis fonéticos e lexicais de processamento de linguagem.

2.5.10.Técnicas de Processamento da Linguagem Natural

O processamento da linguagem natural consiste de uma aplicação seqüencial de diferentes componentes de análise em uma arquitetura *pipeline*. Um *pipeline* é um conjunto de processos encadeados através das suas saídas padrão, de forma que a saída de um processo é utilizada como entrada do processo seguinte.

A definição de etapas ou fases de processamento da linguagem natural se baseia nos conhecimentos lingüísticos necessários à compreensão da linguagem natural que foram vistos anteriormente (fonético, léxico, morfológico, sintático, semântico). Nas seções seguintes comentaremos sobre algumas das principais técnicas de PLN que tem aplicação em ARNT: Tokenização, Separação de sentenças, POS Tag, Lematização e *Chunk*.

2.5.10.1.Tokenização

A técnica de tokenização divide o texto em tokens ou marcas, que são unidades mais simples, como números, pontuação e palavras [23]. Em muitas linguagens que utilizam o alfabeto latino as palavras são separadas por

espaços em branco. Entretanto há questões ambíguas. A maior parte da ambiguidade na tokenização advem do uso das marcas de pontuação, tais como: vírgula, ponto e vírgula, ponto, exclamação, aspas e apóstrofo; uma vez que a mesma marca de pontuação pode ter diferentes funcionalidades. Por exemplo, consideremos a sentença (5):

Clairson International Corp. said it expects to report a net loss for its second quarter ended March 26 and doesn't expect to meet analysts' profit estimates of $3.9 to $4 million, or 76 cents a share to 79 cents a share, for its year ending Sept. 24. (5)

Essa sentença possui algumas situações de ambiguidade. Primeiro, o ponto é utilizado com três funcionalidades diferentes: entre números, como um ponto decimal ($3.9) em abreviações como "*Corp.*" e "*Sept.*" e para representar o fim da sentença, sendo que nesse caso o ponto após o número "24" não representa um ponto decimal. O apóstrofo é utilizado com duas funções. Para representar o caso genitivo (quando o apóstrofo representa possessão) em "*analysts'*" e para representar contrações como em "*doesn't*". Além disso, o "*'s*" serve como forma contracta do verbo "*to be*" como em "*he's*", "*it's*" e também como a forma plural de algumas palavras, por exemplo, "*I.D's*" e "*1980's*". A decisão sobre a forma de tokenização nesses caso é dependente do contexto.

Uma outra situação comum são as abreviações, utilizadas para denotar uma forma mais curta de uma dada palavra. Em muitos casos abriações são escritas como uma sequencia de caracteres terminada com um ponto. Portanto reconhece-las é essencial tanto para a tokenização (seção 2.5.10.1) quanto para a separação de sentenças (seção 2.5.10.2), já que essas podem ocorrer no final de uma sentença, situação na qual o ponto tem duas funções. Criar uma lista de abreviações pode ajudar a reconhecer essas estruturas, entretanto não é possível construir uma lista completa tão pouco definitiva de tais termos. Além disso, abriviações podem aparecer como palavras completas no texto, por exemplo, a palavra "*mass*" é também a abreviação de "*Massachusetts*". Uma abreviação pode ainda representar diferentes paravras, esse é o caso de "*St.*" que pode representar "*Saint*", "*Street*" ou "*State*".

As sentenças 6 e 7 demonstram a dificuldade proveniente desses casos ambíguos, nos quais a mesma abreviação pode representar diferentes palavras e podem ocorrer tanto no interior quanto no final de uma sentença.

The contemporary viewer may simply ogle the vast wooded vistas rising up from the Saguenay River and Lac St. Jean, standing in for the St. Lawrence river. (6)

The firm said it plans to sublease its current headquarters at 55 Water St. A spokesman declined to elaborate. (7)

Uma outra questão relevante é que a convenção de espaço nem sempre corresponde a tokenização desejada. Por exemplo, a expressão "*in spite of*" formada por três palavras é equivalente a "*despite*" e ambas deveriam ser tratadas como uma única marca. Algoritmos de tokenização estão disponíveis em ferramentas para processamento da linguagem natural como o Gate (General Architecture for Text Engineering – Arquitetura Genérica para Engenharia de Texto) [20] [21] e o NLTK (Natural Language ToolKit – Ferramenta de Linguagem Natural) [8], ambas para a língua inglesa.

A tokenização é utilizada na fase "Anotação do corpus" do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3) uma vez que sua execução é pré-requisito para a aplicação das demais técnicas de PLN necessárias no contexto de ARNT, que são tipicamente: Separação de sentenças, Análise morfológica e *Chunk.*

2.5.10.2.Separação de sentenças

Sentenças, na maioria das linguagens, são delimitadas por marcas de pontuação, apesar de as regras específicas utilizadas não serem sempre definidas de forma tão coerente [19]. Mesmo quando um conjunto de regras existe a aderência a essas regras pode variar bastante. Adicionalmente, em diferentes linguagens, sentenças e subsentenças são frequentemente delimitadas por diferentes marcas de pontuação. O sucesso na identificação de sentenças para uma dada linguagem exige o entendimento dos diferentes usos dos caracteres de pontuação naquela linguagem. Na maioria das linguagens o

problema de separação de sentenças consiste em desambiguar todas as instancias de caracteres de pontuação que podem delimitar sentenças.

Linguagens que não possuem muitas marcas de pontuação apresentam uma grande dificuldade para o reconhecimento dos limites das sentenças. O idioma Tailandês, por exemplo, não utiliza um ponto final para delimitar sentenças sendo esse, em alguns casos, representado por um espaço ou mesmo pela ausência dele. Há ainda situações nas quais um espaço é utilizado para denotar o que em inglês seria representado como uma vírgula. Mesmo linguagens com relativa riquesa em sinais de pontuação, como o Inglês apresentam problemas. Reconhecer os limites de sentenças envolve determinar a função de cada marca de pontuação que podem denotar delimitação de sentenças. As marcas: ponto, interrogação, exclamação, redicencia e em algumas situações dois pontos, ponto e vírgula, vírgula e barras (*dashes*) podem ter difererentes funcionalidades além da de demarcar limite de sentenças. Um ponto, por exemplo, pode denotar um ponto decimal, um marcador de milhar, uma abreviação, o final de uma sentença ou ainda uma abreviação no final de uma sentença. O mesmo ocorre com a marca de redicencia, que pode estar tanto no final quanto no meio de uma sentença.

Assim como a definição de o que constitui uma sentença é de certa forma arbitrária, o uso de certas marcas de pontuação para separar sentenças depende muito da aderência do autor a certas convenções de estilo de escrita. Na maioria das aplicações de PLN as únicas pontuações de delimitação de sentenças são o ponto, interrogação e exclamação, como descrito por Nunberg [52]. Entretanto, sentenças gramaticais podem ser delimitadas por outras marcações e restringí-las as essas três pode ter como consequência o desprezo de sentenças significativas. Por exemplo, consideremos as sentenças (8) e (9) que possuem a mesma semântica. A primeira, num sentido mais tradicional, seria considerada como duas sentenças, já a segunda como apenas uma.

Here is a sentence. Here is another. (8)

Here is a sentence; here is another. (9)

A sentença é a unidade de texto da qual são extraídos os relacionamentos não-taxonômicos. Virtualmente todas as técnicas do estado da arte [29] [42] [48] [55] [56] [69] trabalham com essa heurística. Por exemplo, considerando a regra de sentença (definida na seção 4.5.1), os conceitos "*marriage*" e "*divorce*" e a sentença "*An absolute divorce dissolves the marriage.*", esses dois conceitos seriam considerados relacionados de forma não-taxonômica.

2.5.10.3.POS Tag

O etiquetador gramatical (*Part of speach tagger - Pos tagger*) é um sistema responsável por identificar em uma sentença, para cada um dos itens lexicais (*tokens*), a categoria a qual esse item pertence. Por exemplo, para a palavra "*a*" o analisador deverá decidir qual a categoria correta, de acordo com a posição que ocupa na frase. As etiquetas de partes do discurso costumam incluir: substantivo, verbo, pronome, preposição, advérbio, conjunção e artigo. Há uma variedade de conjuntos de marcação que podem ser utilizadas no *POS Tag*; um dos mais utilizados é o conjunto PennTreebank [43] que possui 45 *tags*. Por exemplo, o resultado do *POS tag* à frase: "*Bela wrote a sad letter to Edward.*", considerando as marcações PennTreebank [43] é dada na tabela 3.

Token	***POS Tag***	***Categoria gramatical***
Bela	*NNP*	*Frase nominal*
wrote	*VBD*	*Verbo no passado*
a	*DT*	*Determinante*
sad	*JJ*	*Adjetivo*
letter	*NN*	*Substantivo*
to	*TO*	*"to"*
Edward	*NNP*	*Frase nominal*

Tabela 03: Sentença anotada com *tags* Penntreebank

A etiquetagem é o processo de assinalamento de um marcador de classe gramatical (ou outro "maracador" ou etiqueta de interesse) a cada palavra num corpus. A entrada para a etiquetagem é uma cadeia de itens lexicais e um conjunto específico de etiquetas. A saída é o conjunto de itens

lexicais com a melhor etiqueta associada a cada item. O desafio do processo de etiquetagem consiste em resolver ambiguidades.

Os algoritmos para etiquetagem estão fundamntados em dois modelos mais conhecidos: os baseados em regras e os estocáticos. Os algoritmos baseado em regras, fazem uso de bases de regras para identificar a categoria de um certo item lexical. Nesse caso, novas regras são acrescentadas a base à medida que novas situações de uso do item são encontradas. Os algoritmos baseados em métodos estocásticos costumam resolver as abiguidades com o uso de corpora de treino marcados corretamente, calculando a probabilidade que um certo item lexical terá de ser associado a uma certa etiqueta em certo contexto.

A execução do *POS tag* é necessária para permitir a posterior identificação de frases verbais, que são utilizadas como rótulos dos relacionamentos não-taxonômicos. Exemplos de técnicas de ARNT que possuem essa caracteristica são: ARNT baseada na extração de regras de associação [69], a ARNT baseada em consultas a Web [56] e TARNT [60] [61]. Essa técnica de PLN é também utilizada pelas técnicas de ARNT que não recebem os conceitos da ontologia como *input* e que, portanto utilizam frases nominais em sua substituição. Esse é o caso das técnicas: ARNT baseada em consultas à Web [56] e ARNT baseada na regressão logística [29].

2.5.10.4.Lematização

A Lematização é um tipo de normalização morfológica que deriva o lema (forma lematizada) da palavra original [19]. Assim, o lema de um verbo é sua forma infinitiva. O lema de uma palavra variável (que não seja verbo) é sua forma singular e (quando existe) masculina. Por exemplo, as formas verbais "*walk*", "*walked*", "*walks*" e "*walking*" possuem como lema a forma verbal infinitiva "*walk*"; já o lema das palavras "*cat*" e "*cats*" é o mesmo, "*cat*".

Em recuperação de informação [5], a lematização é utilizada para expansão da consulta. Com esse artifício a busca é feita por mais elementos que o exatamente informado pelo usuário. Por exemplo, se um usuário informar "*cosmology uruguay*" um sistema de informação utilizando lematização poderia

incluuir "*cosmologist*" a consulta. O aspecto negativo refernte a essa técnica é que a precisão pode ser reduzida pela inclusão de termos indesejáveis.

Essa técnica é utilizada com o objetivo de possivelmente incrementar o *recall* da busca por conceitos quando esses são dados como *input* para a técnica de ARNT. Por exemplo, o casamento entre o conceito de uma ontologia, "*lawyer*" e o termo "*lawyers*" do corpus não ocorrerá caso esse último não esteja lematizado.

2.5.10.5.Chunk

Uma tarefa fundamental em PLN é a segmentação e rotulação (etiquetagem) de texto. Em um passo incial o texto é dividido em unidades lexicais (*tokens*) e em sentenças. O *Chunk* atribui rótulos a conjuntos de *tokens* que representam uma estrutura parcial de uma sentença [23].

O *Chunk* se originou de uma motivação psicolinguística [23]. Ele é baseado na intuição de que quando lemos uma sentença, o fazemos um *chunk* por vez. Por exemplo, a sentença "*I begin with an intuition: When I read a sentence, I read it a chunk at a time.*" pode ser decomposta como: "[I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time]".

No contexto do processamento de linguagem natural o *Chunk* é considerado uma abordagem efeciente em relação ao *parsing,* uma vez que não tem que lidar com todas as questões sintáticas de uma linguagem.

Há dois tipos de *chunk*: *VP chunk* e *NP chunk*. O *VP chunk* faz a anotação de frases verbais e pode ser utilizado por técnicas de ARNT que as usem como rótulos dos relacionamentos. Por exemplo, na sentença "*The decree may also award alimony.*", os termos "*may also award*" constituem a frase verbal. Algumas técnicas de ARNT que recomendam rótulos aos relacionamentos como frases verbais são a ARNT baseada em consultas à Web [56], a ARNT baseada na regressão logística [29] e TARNT [60] [61], apresentadas respectivamente nos capítulos 3 e 4.

O *NP chunk* faz a anotação de frases nominais e é normalmente utilizado por técnicas de ARNT que não recebem os conceitos (conjunto C, seção 2.1) ou a hierarquia (conjunto H, seção 2.1) como *input*. Por exemplo, na

sentença "*An absolute divorce dissolves the marriage.*", os termos "*An absolute divorce*" e "*the marriage*" constituem as frases nominais. Algumas técnicas de ARNT que utilizam frases nominais como conceitos são a "ARNT baseada em consultas à Web" [56] e a "ARNT baseada na regressão logística" [29], apresentadas no capítulo 3. Para que possam ser executados, tanto o *VP chunk* quanto o *NP chunk* necessitam que as classes gramaticais das palavras tenham sido previamente identificadas com o *POS Tag (seção 2.5.10.3)*.

2.6. Áreas de conhecimento aplicadas a ARNT: Aprendizagem de máquina

A maioria das técnicas de ARNT não recomenda ao especialista do domínio todos os relacionamentos extraídos do texto na fase "Extração de relacionamentos candidatos". Isso se deve a dois motivos: o primeiro é diminuir a quantidade de recomendações feitas ao especialista. Em soluções que utilizam grandes corpora a quantidade de tuplas extraídas é potencialmente grande. Por exemplo, para o experimento realizado na seção 5.6 com o corpus Genia [53], que possui 2.000 documentos e 18.545 sentenças, a abordagem de ARNT AERA (seção 5.2) extraiu 3.609 relacionamentos. Disponibilizar ao especialista um conjunto tão grande de possíveis relacionamentos é inapropriado. O segundo motivo é filtrar as tuplas que possuem maior probabilidade de serem relacionamentos das que possuem menor probabilidade. Alcançar esses dois objetivos é a finalidade da fase de "Refinamento" do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3).

As soluções normalmente utilizadas pelas técnicas de ARNT para a fase de "Refinamento" são do campo de AM [9] [46] [49] [50]. Por exemplo, as técnicas "ARNT baseada na Extração de regras de associação" [69], "ARNT baseada na Extração de regras de associação generalizadas" [42], "ARNT baseada na regressão logística" [29] e "ARNT baseada na classificação de relacionamentos" [48], apresentadas e discutidas no capítulo 3 utilizam respectivamente os algoritmos de extração de regras de associação [1] [2], extração de regras de associação generalizadas [67], regressão logística [46] e um algoritmo de agrupamento [46].

A AM [9] [46] [49] [50] é uma área da Inteligência Artificial [54] cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo do aprendizado [9] [46] [49] [50]. Em outras palavras, pode-se dizer que a AM trata do desenvolvimento de técnicas que possibilitem ao computador melhorar o seu desempenho em determinada tarefa. Além disso, é objetivo da AM a construção de sistemas que sejam capazes de adquirir conhecimento de maneira automática [49].

Cabe ressaltar que aquisição de conhecimento e aprendizado são processos distintos e que aquele não é condição suficiente para este. Para que haja aprendizado é necessário que o conhecimento adquirido por um sistema interfira positivamente no seu desempenho frente à execução de um determinado conjunto de tarefas. Referente a isto, Mitchell [46] afirma que "um programa aprende, a partir da experiência *E*, em relação a uma classe de tarefas *T*, com medida de desempenho *P*, se seu desempenho em *T*, medido por *P*, melhora com *E*". Os algoritmos de AM podem ser categorizados em três classes conforme a tabela 4.

Tipo de Aprendizagem	Caracterização	Aplicação
Supervisionado	Envolve a aprendizagem de uma função a partir de um conjunto de exemplos de suas entradas e saídas.	É tipicamente utilizado para predizer categorias apropriadas de um exemplo a partir de um conjunto de categorias representadas.
Não supervisionado	Envolve a aprendizagem de padrões de entrada, visto que não são fornecidos ao algoritmo de aprendizagem os valores de saída específicos.	É utilizado para encontrar aglomerados de conjuntos de dados semelhantes entre si (*clusters*).
Por Reforço	A aprendizagem se dá a partir da observação das consequências das ações no ambiente.	Utilizado em tarefas de controle e robótica.

Tabela 04: Classificação das técnicas de AM [46]

2.6.1. Aprendizagem de Máquina Supervisionada

A Aprendizagem de Máquina Supervisionada envolve a aprendizagem de uma função a partir de exemplos de suas entradas e suas saídas [46]. Um exemplo é um par (x, f(x)), onde x é a entrada e f(x) é a saída da função aplicada a x. O objetivo é aprender uma função matemática que aproxime x de f(x) (figura 4). Sempre é fornecida uma referência do objetivo a ser alcançado, ou seja, o algoritmo de aprendizagem recebe o valor de saída desejado para cada conjunto de dados de entrada apresentado.

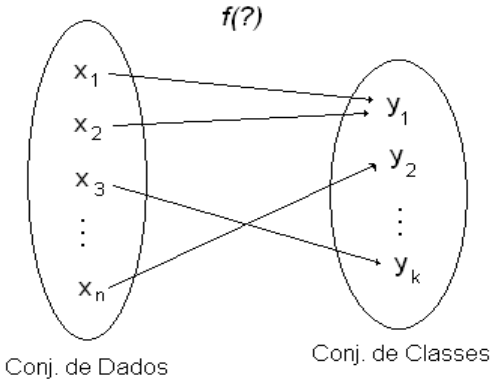


Figura 04: aprendizagem de máquina supervisionada

No caso da aprendizagem supervisionada, este conjunto de exemplos é composto geralmente por um vetor de características e pelo rótulo de uma classe associada. Dessa forma, pode-se afirmar que o objetivo do algoritmo de aprendizagem é gerar um classificador capaz de definir corretamente a classe de novos exemplos ainda não rotulados.

Pode-se falar em duas categorias para aprendizagem supervisionada: a classificação e a regressão. Chama-se classificação o processo de aprendizado que envolve rótulos de classes discretos. Quando os rótulos de classe são contínuos, o processo é chamado de regressão. Neste trabalho é dado ênfase ao processo de regressão e em particular à regressão logística [46], utilizada pela técnica de ARNT baseada na regressão logística [29].

Regressão logística [46] é um tipo de regressão na qual a função *f(z)* aprendida é uma função logística. Esse tipo de função retorna a probabilidade *p* de um evento, dado que esse é afetado por uma ou mais variáveis

explicativas. A função logística *f(z)* é dada pela equação (10). Já a variável "z" é a medida da contribuição de todas as variáveis explicativas usadas no modelo para o calculo da probabilidade do evento considerado (11).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{10}$$

$$z = a_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{11}$$

A aplicação da regressão logística pode ser ilustrada por meio de um exemplo de cálculo da probabilidade de morte por doença cardíaca. Este modelo usa apenas três fatores de risco (variáveis explicativas): idade, sexo e nível de colesterol no sangue. Os pesos e suas respectivas variáveis explicativas são:

- β_0 = -5.0
- β_1 = +2.0
- β_2 = -1.0
- β_3 = + 1.2
- x1 = idade em anos, acima de 50
- x2 = sexo, 0 é masculino 1 feminino
- x3 = nível d colesterol em mmol/L, acima de 5,0

O modelo pode ser expresso por (12) e (13):

$$risco\ de\ morte = \frac{1}{1 + e^{-z}} \tag{12}$$

$$z = -5{,}0 + 2{,}0 * x_1 - 1{,}0 * x_2 + 1{,}2 * x_3 \tag{13}$$

Nesse modelo, aumentar a idade está associado ao aumento do risco de morte por doenças cardíacas (*z* aumenta 2.0 por cada ano depois dos 50 anos de idade), o sexo feminino é associado a um menor risco de doenças cardíacas (*z* diminui 1.0 se o paciente for do sexo feminino), e o aumento do colesterol está associado a um aumento do risco de morte (*z* aumenta 1.2 para cada 1mmol/L aumento no colesterol acima de 5mmol/L).

Para um homem de 50 anos de idade e nível de colesterol é 7,0 mmol/L pode-se calcular o risco de morte conforme (14) e (15).

$$risco\ de\ morte = \frac{1}{1 + e^{-z}} \tag{14}$$

$$z = -5{,}0 + 2{,}0 * 0 - 1{,}0 * 0 + 1{,}2 * 2 \tag{15}$$

Isso significa que por esse modelo, o risco de morte do referido individuo por doença do coração é 0,07 (ou 7%). A regressão logística é utilizada na técnica de ARNT baseada em regressão logística [29] (seção 3.4) para estimar a probabilidade de um relacionamento extraído ser de fato um relacionamento não-taxonômico.

2.6.2.Aprendizagem de Máquina Não Supervisionada

A Aprendizagem de Máquina Não Supervisionada é tipicamente aplicada na exploração de dados [46]. O problema da AMNS envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saída específicos. Essas técnicas são utilizadas para encontrar aglomerados de conjuntos de dados semelhantes entre si (*clusters*) [46] ou regras de associação [1] [2] entre esses dados. As seções 2.6.2.1 e 2.6.2.2 apresentam a técnica de extração de regras de associação [1] [2] e a de agrupamento [46], utilizadas pelas técnicas de ARNT propostas por Villaverde et al. [69] (seção 3.1) e Mohamed et al. [48] (seção 3.5) respectivamente.

2.6.2.1.Extração de Regras de associação

Formalmente, uma regra de associação é uma implicação da forma X→Y, onde X e Y são conjuntos de itens tais que $X \cap Y = \varnothing$ [1] [2]. Convém destacar que a interseção vazia entre antecedente e conseqüente das regras assegura que não sejam extraídas regras óbvias que indicam que um item está associado a ele próprio. A seguir estão indicados dois exemplos de regras de associação no contexto de transações de venda em um supermercado. A regra (16) indica que a compra de leite pode levar a compra de pão. Segundo a regra (17) a compra de pão e manteiga pode induzir a compra de café.

$$\text{Leite} \rightarrow \text{Pão} \tag{16}$$

$$\text{Pão} \wedge \text{Manteiga} \rightarrow \text{Café} \tag{17}$$

Uma regra de associação é considerada "frequente" se o número de vezes em que a união de conjunto de itens ($X \cup Y$) ocorrer em relação ao número total de transações do banco de dados for superior a uma freqüência mínima (denominada *suporte mínimo*) que é estabelecida em cada aplicação. Busca-se, por meio do suporte, identificar quais associações surgem em uma quantidade expressiva a ponto de se destacarem das demais existentes. Considerando a tabela 5 que possui 10 transações de venda e indica quais produtos estão presentes em cada transação, as regras (8) e (9) possuem suporte 20% e 30%, respectivamente.

Transações	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

Tabela 05: Exemplos de transações de venda em um supermercado

Uma regra de associação é considerada *válida* se o número de vezes em que X ∪ Y ocorrer em relação ao número de vezes que "X" ocorrer for superior a um valor denominado "confiança mínima", também estabelecido em cada aplicação. A medida de confiança expressa a qualidade de uma regra, indicando o quanto a ocorrência do antecedente da regra pode assegurar a ocorrência do conseqüente desta regra. Considerando a tabela 5, as regras (15) e (16) possuem confiança 100% e 75%, respectivamente.

Assim sendo, a tarefa de Descoberta de Regras de Associação pode ser definida formalmente como a busca por regras de associação freqüentes e válidas em um banco de dados, a partir da especificação dos parâmetros de suporte e confiança mínimos [1] [2]. É importante destacar que os valores dos parâmetros de suporte e confiança mínimos são definidos experimentalmente.

O conceito de *K-itemset* refere-se a todo conjunto de itens com exatamente "K" elementos. As regras (15) e (16) correspondem, respectivamente, a 2-*itemset* e 3-*itemset*.

O *Apriori* [9] é um algoritmo clássico de Extração de Regras de Associação [1] [2] e se baseia no princípio d*a* "antimonotonicidade do suporte". Segundo esse princípio "Um *k-itemset* somente pode ser freqüente se todos os seus (k-1)-*itemsets* forem freqüentes". Dessa forma, a combinação de *itemsets* para gerar um novo *itemset* somente ocorre quando estes são freqüentes. O suporte de um conjunto de itens nunca pode crescer quando este é expandido para um conjunto com mais itens; ele pode, na melhor hipótese, permanecer igual, ou simplesmente diminuir.

O algoritmo *Apriori* pode ser decompostos em duas etapas:

a) Encontrar todos os *k-itemsets* freqüentes (que satisfazem a condição de suporte mínimo).
b) Utilizar os *k-itemsets* freqüentes, com k ≥ 2, para gerar as regras de associação (que satisfazem à condição de confiança mínima).

A complexidade de tempo do algoritmo *Apriori* é linear em relação ao número de transações. Entretanto, é relatado que o *Apriori* apresenta bom desempenho quando o suporte mínimo é considerado grande. Caso contrário, muito conjuntos de itens são gerados, degradando sensivelmente o tempo de

execução. Na maioria das aplicações o suporte mínimo deve ser grande, para que os conjuntos freqüentes sejam significativos.

Consideremos como exemplo a tabela 5. A seguir são descritas as etapas de funcionamento de algoritmo *Apriori* sobre este exemplo.

Primeiramente, o usuário deve definir os valores de suporte e confiança mínimos a serem considerados pelo algoritmo. Consideramos para este exemplo que tenham sido definidos MinSup = 0,3 e MinConf = 0,8.

A etapa (a) é iterativa. Em cada iteração são identificados os *itemsets* freqüentes. No exemplo:

1ª Iteração – São identificados os 1-itemsets com suporte maior que MinSup (tabela 6).

1-itemsets	Suportes
Leite	0,2
Café	0,3
Cerveja	0,2
Pão	0,5
Manteiga	0,5
Arroz	0,2
Feijão	0,2

Tabela 06: Ilustração do Apriori: tabela de 1-itemsets

2ª Iteração – São combinados os 1-itemsets freqüentes para gerar os 2-itemsets (tabela 7). Destes, são selecionados aqueles cujo suporte seja superior ao mínimo definido.

2-itemsets	Suportes
Café, Pão	0,3
Café, Manteiga	0,3
Pão, Manteiga	0,4

Tabela 07: Ilustração do Apriori: tabela de 2-itemsets

3ª Iteração - São combinados os 2-itemsets freqüentes para gerar os 3-itemsets (tabela 8). Destes, são selecionados aqueles cujo suporte seja superior ao mínimo definido.

3-itemsets	Suportes
Café, Pão, Manteiga	0,3

Tabela 08: Ilustração do Apriori: tabela de 3-itemsets

Assim sendo, os k-itemsets com $k \geq 2$ resultantes da etapa (a) foram:

Café e Pão

Café e Manteiga

Pão e Manteiga

Café, Pão e Manteiga

Esta lista é informada à etapa (b) que deve identificar as regras válidas, ou seja, aquelas regras cuja confiança seja superior à confiança mínima definida pelo usuário. Para isso, são geradas as regras possíveis a partir de cada *itemset* freqüente, assim como sua confiança.

1. Itemset {Café, Pão}

   SE café ENTÃO pão.Conf. = 1,0.

   SE pão ENTÃO café.Conf. = 0,6.

2. Itemset {Café, Manteiga}

   SE café ENTÃO manteiga.Conf. = 1,0.

   SE manteiga ENTÃO café.Conf. = 0,6.

3. Itemset {Pão, Manteiga}

   SE manteiga ENTÃO pão.Conf. = 0,8.

   SE pão ENTÃO manteiga.Conf. = 0,8.

4. Itemset {Café, Pão, Manteiga}

   SE café, pão ENTÃO manteiga.Conf. = 1,0.

   SE café, manteiga ENTÃO pão.Conf. = 1,0.

   SE manteiga, pão ENTÃO café.Conf. = 0,75.

   SE café ENTÃO pão, manteiga.Conf. = 1,0.

SE pão ENTÃO café, manteiga.Conf. = 0,6.

SE manteiga ENTÃO café, pão.Conf. = 0,6.

Após a filtragem das regras com confiança igual ou superior à confiança mínima, o *Apriori* produz como saída as seguintes regras de associação:

SE café ENTÃO pão.

SE café ENTÃO manteiga.

SE manteiga ENTÃO pão.

SE pão ENTÃO manteiga.

SE café, pão ENTÃO manteiga.

SE café, manteiga ENTÃO pão.

SE café ENTÃO pão, manteiga.

2.6.2.2.Agrupamento

O Agrupamento é a tarefa de encontrar grupos ou "*clusters*" de objetos similares nos dados [46]. Um bom método de agrupamento deve produzir "*clusters"* com qualidade, ou seja, alta similaridade intra-grupo e baixa similaridade inter-grupo. A qualidade do resultado de um processo de agrupamento depende da medida de similaridade do método utilizado e sua implementação. Para descrever mais formalmente o agrupamento, apresentamos em linhas gerais o K-Means [54], um algoritmo bem conhecido e que posui os seguintes passos:

1. Toma-se randomicamente "K" pontos de dados (dados numéricos) como sendo os centróides dos clusters.
2. Cada ponto (ou registro da base dados) é atribuído ao *cluster* cuja distância deste ponto em relação ao centróide de cada cluster é a menor dentre todas as distâncias calculadas.
3. Um novo centróide para cada cluster é calculado pela média dos pontos do cluster, caracterizando a configuração dos clusters para a interação seguinte.

4. O processo termina quando os centróides dos clusters param de se modificar ou após um número limitado de interações que tenha sido especificado pelo usuário.

O K-Means [54] toma um parâmetro de entrada “K” e divide um conjunto de “n” objetos em “K” *clusters* tal que a similaridade intracluster resultante seja alta e a similaridade intercluster seja baixa. A similaridade em um cluster é medida em respeito ao valor médio dos objetos neste cluster (centro de gravidade do cluster). A execução de K-Means consiste em, primeiro, selecionar aleatoriamente “K” objetos, que inicialmente representam cada um a média de um cluster. Para cada um dos objetos remanescentes é feita a atribuição ao cluster ao qual o objeto é mais similar, baseado na distância entre o objeto e a média do *cluster*. A partir de então o algoritmo computa as novas médias para cada *cluster*. Este processo se repete até que uma condição de parada seja atingida. A figura 5 ilustra a aplicação do K-Means a um arquivo com vinte registros de dados, considerando K = 3.

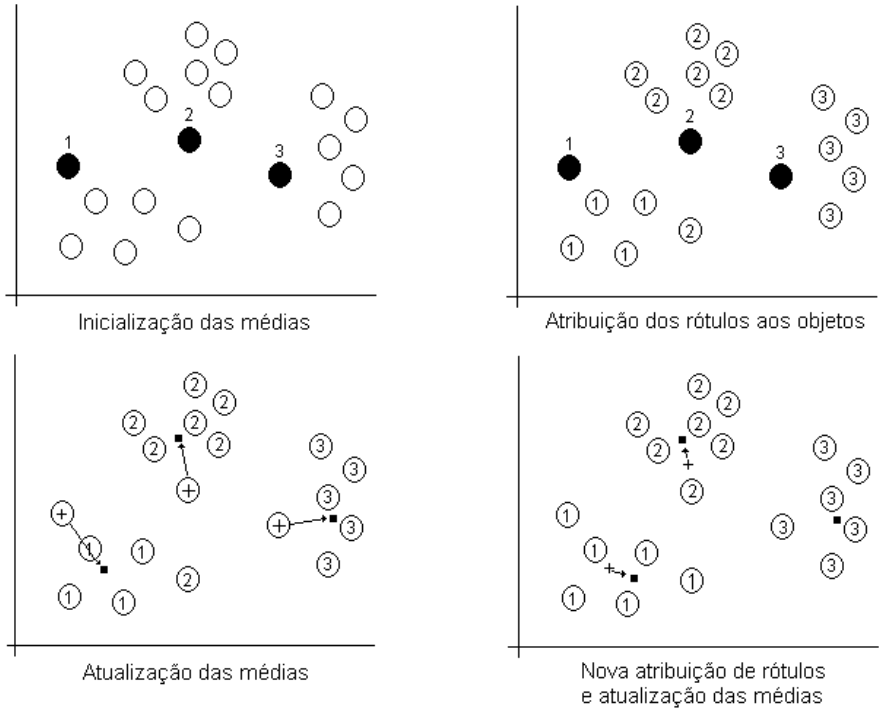


Figura 05: Exemplo de aplicação do algoritmo de agrupamento K-Means

As principais abordagens de agrupamento são [46]: algoritmos de partição, algoritmos hierárquicos, algoritmos fundamentados na densidade, algoritmos baseados em *grids* e os algoritmos baseados em modelos.

Os algoritmos hierárquicos por sua vez são divididos em: aglomerativo e divisivo. O *clustering* hierárquico aglomerativo é uma abordagem *bottom-up*. Na inicialização é criado um cluster próprio para cada elemento. Nas iterações seguintes, os *clusters* mais similares são fundidos. O cálculo da similaridade é feito entre os *clusters*, através do método de links, que é calculado através da distância. Há três tipos de *link*: o link mínimo, que é a distância mínima entre os *clusters*; o link completo, que é a máxima distância entre os *clusters*; e o link médio, que é a média das distâncias entre os *clusters*.

O *clustering* hierárquico divisivo é uma abordagem *top-down* que particiona um *cluster* universal contendo todos os elementos. O cálculo da similaridade aborda duas questões: como selecionar o próximo *cluster* a ser dividido e como realizar a divisão. Essas questões são abordadas através da função de coerência, onde os elementos menos coerentes são candidatos a deixarem o cluster e a função baseada na variância selecionando o cluster com maior valor para ser dividido.

Um exemplo de técnica de ARNT que utiliza um algoritmo de agrupamento é a “ARNT baseada na classificação de relacionamentos” [48]. Para cada par de conceitos, presente em uma sentença, extraído do corpus, é criada uma matriz cujas linhas e colunas correspondem a frases verbais (contextos) que ocorrem entre os conceitos nas sentenças e cada célula da matriz corresponde a frequência na qual o par de conceitos co-ocorre com ambas as frases verbais. Sobre essa matriz é realizado um agrupamento sobre os valores de frequencia de co-ocorrência e cada agrupamento é então utilizado para propor um possível novo relacionamento. O centroide de cada agrupamento é utilizado para sugerir o nome do novo relacionamento. Essa técnica é apresentada na seção 3.5.

2.7.Considerações finais

Nesse capítulo foi apresentada a definição formal de ontologia adotada nesse trabalho, algumas de suas aplicações e a relevância da emergente área de pesquisa Aprendizagem de Ontologias [10] [11] [15] [16] [17] [33] [37] [64] e em particular da Aprendizagem de Relacionamentos Não-Taxonômicos [29] [42] [48] [55] [56] [60] [69] . A ARNT foi ainda sistematizada com a definição de um processo genérico [57] [58] [59] [60] composto de cinco fases que pode ser utilizado tanto para a realização de avaliações qualitativas entre diferentes técnicas com relação às soluções adotadas para cada fase quanto para guiar a proposição de novas técnicas.

Em seguida, os quatro principais tipos de relacionamentos não-taxonômicos tipicamente utilizados pelas técnicas de ARNT (seção 2.4) foram formalmente definidos assim como as medidas *recall* e precisão, que são utilizadas nas avaliações realizadas sobre TARNT no capitulo 5.

Foram ainda apresentadas as principais áreas de conhecimento que têm aplicação em ARNT: Processamento de Linguagem Natural e Aprendizagem de Máquina. O Processamento de Linguagem Natural é um subcampo da Inteligência Artificial cujas técnicas são utilizadas na fase "Anotação do corpus" do processo genérico de ARNT [57] [58] [59] [60] para estruturar o texto do qual são extraídos os relacionamentos não-taxonômicos. A Aprendizagem de Máquina é um subcampo da Inteligência Artificial cujas técnicas são utilizadas para refinar os relacionamentos não-taxonômicos, em uma fase de mesmo nome, antes de serem definitivamente recomendados ao engenheriro de conhecimento.

No próximo capítulo serão apresentadas e discutidas técnicas do estado da arte em ARNT enfatizando as soluções por cada uma adotada para as fases do processo genérico de ARNT [57] [58] [59] [60] apresentado no presente capítulo.

3. Técnicas de Aprendizagem de Relacionamentos Não-Taxonômicos de Ontologias

Nesse capítulo algumas técnicas representativas do estado da arte em de ARNT [29] [42] [48] [55] [56] [69] são apresentadas e analisadas comparativamente. As soluções por cada uma delas adotadas para as fases do processo genérico de ARNT [57] [58] [59] [60] são destacadas e seus aspectos positivos e limitações discutidos.

A técnica baseada na extração de regra as de associação [69], apresentada na seção 3.1, utiliza técnicas de PLN para extrair do corpus conceitos da ontologia dados como *input* e verbos que estejam entre esses conceitos para gerar tuplas formadas por dois conceitos e um verbo, que representam relacionamentos candidatos. Esses relacionamentos são submetidos ao algoritmo de mineração de regra de associação [1] [2] que sugere relacionamentos na forma de regras de associação [1] [2]. Essas regras tem a forma "X→Y", o que significa que a ocorrência de "X" implica na ocorrência de "Y".

A técnica baseada na extração de regras de associação generalizadas [42], apresentada na seção 3.2, difere da proposta por Villaverde et al. [69] (seção 3.1), principalmente pela adoção do algoritmo para extração de regras de associação generalizadas [67] para refinar os relacionamentos candidatos e recomendá-los na forma ($conceito_1 \rightarrow conceito_2$). Esse algoritmo sugere o melhor nível da hierarquia de conceitos da ontologia no qual o relacionamento deva ser acrescentado. Por exemplo, no caso de um supermercado o algoritmo poderia sugerir que "*snacks* são comprados com *drinks*" e não que "*chips* são comprados com *beer*" ou "*peanuts* são comprados com *soda*".

A seção 3.3 apresenta uma técnica que recomenda relacionamentos não-taxonômicos baseado em estatísticas calculadas sobre os resultados de consultas a um mecanismo de busca Web [55] [56]. A partir de uma palavra chave inicial (ex.: hipertensão), representando um conceito da ontologia, a técnica sugere relacionamentos do domínio (ex.: *hipertensão* é causada por *problemas hormonais*).

Na seção 3.4 é apresentada a proposta de Fader et al. [29], que utiliza um classificador de regressão logística [46] para ranquear os relacionamentos não-taxonômicos conforme a probabilidade de serem válidos.

A última técnica discutida é uma proposição de Mohamed et al. [48] (seção 3.5) que realiza tanto a aprendizagem quanto o povoamento de relacionamentos não-taxonômicos. Essa técnica recomenda relacionamentos não-taxônomicos a partir de um corpus em lingua inglesa e infere novos relacionamentos a partir desses, além de identificar suas instancias.

3.1.ARNT baseada na Extração de regras de associação

Villaverde, et al. [69] propõem uma técnica independente de domínio baseada em técnicas de PLN e AM para a extração de relacionamentos não-taxonômicos binários, com rótulos em forma de frases verbais, a partir de textos na língua inglesa. A figura 6 apresenta um esquema da referida técnica.

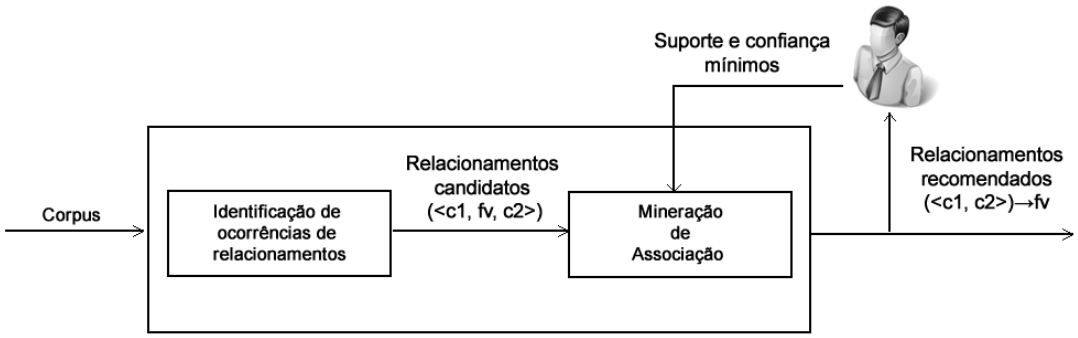


Figura 06: ARNT baseada na Extração de regras de associação

A fase "Identificação de ocorrências de relacionamentos" recebe um corpus e o conjunto de conceitos da ontologia e tem como produto um conjunto de tuplas do tipo <c_1, fv, c_2>, onde "c_1" e "c_2" são conceitos da ontologia e "fv" é uma frase verbal. Esse conjunto é obtido através do procedimento descrito a seguir. Inicialmente cada conceito é expandido com seus sinônimos para aumentar o *recall* da busca sendo o *Wordnet* [13] [30] utilizado para tal finalidade. Em seguida são identificadas sentenças que possuem exatamente dois conceitos, ou algum de seus sinônimos. A essas sentenças é realizado o *Chunk* (seção 2.5.10.5) com o objetivo de identificar as frases verbais. Para cada sentença que satisfaça a condição de apresentar dois conceitos da

ontologia ou algum de seus sinônimos (retornados do *Wordnet* [13] [30]) e uma frase verbal entre eles é gerada uma tupla "<c_1, fv, c_2>".

Fase de "Mineração de associações": uma vez que se tenha um conjunto de relacionamentos candidatos (conjunto de tuplas obtido da fase anterior), é realizada a segunda fase "Mineração de associações" que tem o objetivo de realizar um refinamento nos resultados da fase anterior antes que esses sejam sugeridos ao engenheiro de conhecimento. Para tanto é utilizado um algoritmo de extração de regras de associação [1] [2] (seção 2.6.2.1) que sugere relacionamentos não-taxonômicos como regras no formato "<c_1, c_2> → fv" que possuam valores para os parâmetros de corte, suporte e confiança mínimos maiores que os experimentalmente informados pelo usuário.

Por exemplo, consideremos a seguinte sentença: " The *court decree protects the property rights of the parties and provides support for the children.*" e os conceitos de uma ontologia "*decree*" e "*property*" dados como entrada para a técnica. Caso a regra "<*decree*, *property* > → <*protect*>" possua valores para as medidas de suporte e confiança maiores ou iguais ao suporte e confiança mínimos, ela é recomendada ao engenheiro de conhecimento. A tabela 9 apresenta as soluções em particular adotadas para cada uma das tarefas genéricas para a aprendizagem de relacionamentos não-taxonômicos [57] [58] [59] [60] (seção 2.3), bem como o tipo de relacionamento adotado (2.4).

<table>
<tr><th colspan="2">Tarefa genérica</th><th>Solução adotada</th><th>Tipo de rel.</th></tr>
<tr><td colspan="2">Construção do corpus</td><td>Não abordada.</td><td rowspan="5">$<c_1, fv, c_2>$</td></tr>
<tr><td rowspan="2">Extração de Rel. Candidatos</td><td>Anotação do corpus</td><td>Chunk.</td></tr>
<tr><td>Extração de Rel.</td><td>Utiliza o algoritmo: seleciona sentenças que tenham dois conceitos ou sinônimos. Nessas sentenças realiza Chunk. Verifica se há uma frase verbal entre os conceitos. Para cada sentença que satisfaça essa condição cria a tupla: $<c_1, fv, c_2>$. O produto dessa fase é um conjunto de tuplas: $<c_1, fv, c_2>$ que são os relacionamentos candidatos.</td></tr>
<tr><td colspan="2">Refinamento</td><td>Utiliza a técnica de “Extração de regras de associação” [1] [2]. O produto são os relacionamentos sugeridos ao usuário na forma “$<c_1, c_2> \rightarrow fv$”.</td></tr>
<tr><td colspan="2">Avaliação pelo especialista</td><td>Não abordada.</td></tr>
<tr><td colspan="2">Atualização da ontologia</td><td>Não abordada.</td></tr>
</table>

Tabela 09: Soluções adotadas por ARNT baseada na Extração de regras de associação [69] para as tarefas genéricas de ARNT

Um aspecto positivo a ser resaltado nessa proposta é que os relacionamentos recebem rótulos na forma de frases verbais encontradas entre dois conceitos da ontologia presentes em cada sentença do corpus. Além disso, uma vez que recebe os conceitos da ontologia como *input*, essa técnica tem maior potencial de apresentar melhores valores de precisão (seção 2.3) que técnicas que não o fazem, como a “ARNT baseada em consultas a Web” [56] (seção 3.3). Com relação ao corpus dado como entrada para a técnica, esse deve possuir o maior número de conceitos possível em detrimento do número de instancias desses, já que relacionamentos serão identificados somente mediante a ocorrencia de conceitos nas sentenças. A ocorrencia de instancias das classes da ontologia são simplesmente ignoradas pela técnica. Já um aspecto negativo é o fato de que nenhum tratamento é dado a forma possesiva contracta (“’s”) que tem grande probabilidade de representar um relacionamento não-taxonômico e que pode existir com razoável frequência no corpus. Além disso, de forma não muito apropriada para o contexto de ARNT,

esse algoritmo valoriza regras de associação que possuem altos valores de confiança, medida que representa a informação de quanto um rótulo (frase verbal) está associado a um par de conceitos e em contrapartida desprestigia uma informação mais relevante, que consiste no quanto um conceito "c_1" da ontologia está associado a um conceito "c_2". Essa informação é melhor representada pela taxa de ocorrência do par de conceitos no corpus. Uma implementação dessa medida é apresentada no algoritmo de refinamento *Bag of labels* [61] (seção 4.2.4.2), adotado por TARNT [60] [61].

3.2. ARNT baseada na Extração de regras de associação generalizadas

Maedche e Staab [42] propõem um processo para extração de relacionamentos binários do tipo <c_1, c_2> (seção 2.4) a partir de textos na língua alemã em domínios específicos. São utilizadas técnicas de PLN e heurísticas, tanto específicas quanto independentes de domínio, além de uma técnica de AM denominada extração de regras de associação generalizadas [67]. É utilizado um ambiente de PLN para a língua alemã denominado SMES [42], que além de disponibilizar as técnicas de PLN normalmente conhecidas possui um Lexicon, o qual associa instancias a conceitos de uma ontologia. Os autores definiram 400 conceitos e suas respectivas instancias associadas no domínio turístico. A figura 7 mostra um esquema do referido processo.

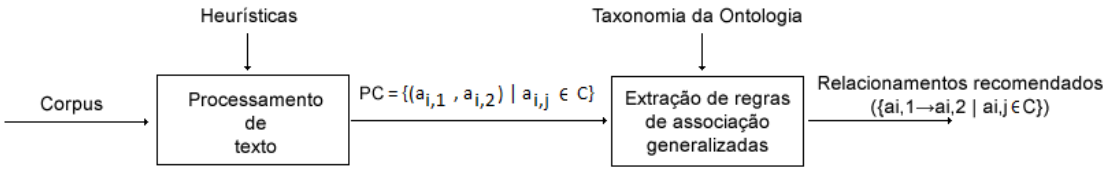


Figura 07: ARNT baseada na extração de regras de associação generalizadas

O objetivo da fase "Processamento de texto" é identificar no texto termos que representem os conceitos da ontologia e formar pares de conceitos (PC), PC = {($a_{i,1}$, $a_{i,2}$) | $a_{i,j} \in$ C}), a partir desses. Duas heurísticas são utilizadas para considerar que dois conceitos sejam relacionados de forma não-taxonômica:

a) Heurística de sentença: todos os conceitos contidos em uma sentença formam pares de conceitos relacionados.
b) Heurística de título: essa heurística é especifica de domínio. Ela forma pares de conceitos relacionados a partir dos conceitos presentes no título do documento com todos os conceitos presentes no corpo do documento.

Os conceitos são identificados no corpus por meio de suas realizações lingüísticas que podem ser tanto instancias das classes quanto termos flexionados que representam conceitos. Para o primeiro caso é utilizado Reconhecimento de entidades nomeadas (REN), já para o segundo a lematização. Por exemplo, os termos em negrito nas sentenças representam conceitos relacionados. A tabela 10 resume pares de conceitos formados e os termos a partir dos quais foram identificados.

Mecklenburg's most beautiful **hotel** is located in Rostock.

A **hairdresser** in our **hotel** is a special service for our guests.

The hotel Mercure offers **balconies** with direct **access** to the beach.

All **rooms** have **TV**, telephone, modem and minibar.

$Termo_1$	$Conceito_1$	$Termo_2$	$Conceito_2$
Mecklenburgs	Area	hotel	hotel
hairdresser	Hairdresser	hotel	hotel
balconies	Balcony	Access	Access
Room	Room	TV	television

Tabela 10: Conceitos identificados pela técnica a partir de sentenças no domínio turístico

Na segunda fase, os pares de conceitos extraídos na fase anterior são submetidos à técnica de extração de regras de associação generalizadas [67], que é uma especialização da técnica de extração de regras de associação [1] [2]. Esse algoritmo tem como insumos a taxonomia da ontolologia (conjunto H, seção 2.1) e o conjunto de pares de conceitos da fase anterior e cumpre dois objetivos: primeiro, baseado nas medidas de suporte e confiança [1] [2]

(seção 2.6.2.1) extrai regras que representam os relacionamentos mais prováveis e em segundo lugar sugere o nível de generalização mais provável para o relacionamento.

Uma aplicação bem conhecida desse algoritmo é a de encontrar regras de associação entre produtos vendidos em supermercados e descreve-las no nível hierárquico mais apropriado. Por exemplo, uma associação válida poderia ser expressa pela sentença "*snacks are purchased together with drinks*" ao invés de "*chips are purchased with beer*" e "*peanuts are purchased with soda*" (figura 8).

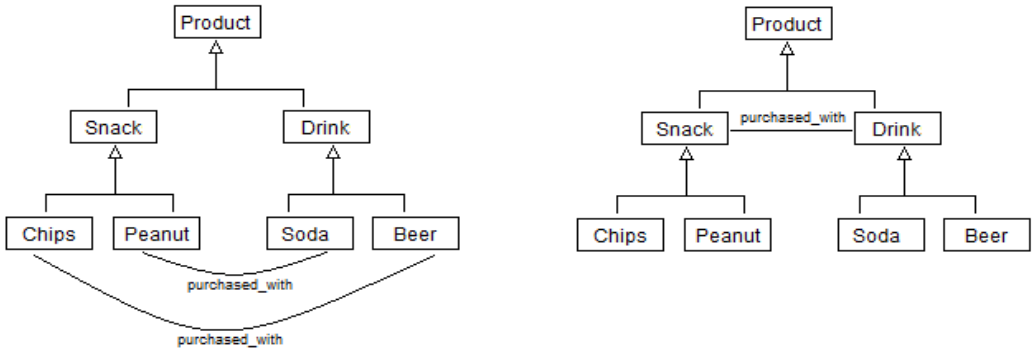


Figura 08: Exemplo de relacionamentos não-taxonômicos em diferentes níveis hierárquicos

O algoritmo básico para extração de regras de associação [1] [2] usa um conjunto de transações $T = \{t_i \mid i = 1...n\}$, cada transação t_i consiste de um conjunto de itens, $t_i = \{a_{i,j} \mid j = 1 \ldots m_i, a_{i,j} \in C\}$ e cada item "$a_{i,j}$" é um elmento de um conjunto de conceitos C. O algoritmo computa regras de associação "$X_k \rightarrow Y_k$" ($X_k, Y_k \subset C$, $X_k \cap Y_k = \varnothing$) que tem valores para as medidas de suporte e confiança acima de determinado limiar. O suporte de uma regra "$X_k \rightarrow Y_k$" representa a porcentagem de transações que têm "$X_k \cup Y_k$" como subconjunto; a confiança é definida como a porcentagem de transações que têm "Y_k" como consequente quando "X_k" for o precedente da regra. Formalmente o suporte e a confiança são dados pelas fórmulas: Suporte $(X_K \rightarrow Y_K) = | \{t_i \mid X_K \cup Y_K \subseteq t_i\} | / n$ e Confiança $(X_K \rightarrow Y_K) = | \{t_i \mid X_K \cup Y_K \subseteq t_i\} | / | \{t_i \mid X_K \subseteq t_i\}|$.

Para extrair associações entre conceitos no nível correto da hierarquia, toda transação t_i é estendida para incluir os ancestrais de cada item "$a_{i,j}$", por exemplo, $t_i' := t_i \cup \{a_{i,l} \mid (a_{i,j}, a_{i,l}) \in H\}$. Então o suporte e a confiança são

calculados para todas as possíveis regras de associação "$X_k \rightarrow Y_k$", tal que "Y_k" não tenha um ancestral de "X_k" uma vez que essa seria uma associação trivial. Finalmente, são excluídas todas as regras de associação "$X_k \rightarrow Y_k$" que tenham valores inferiores para suporte e confiança que uma regra ancestral "$\underline{X}_k \rightarrow \underline{Y}_k$". *Itemsets* "$\underline{X}_k$" e "$\underline{Y}_k$" contêm apenas ancestrais ou itens presentes em *itemsets* da regra "$X_k \rightarrow Y_k$". A figura 9 mostra parcialmente a hierarquia de conceitos para o domínio de informações turísticas.

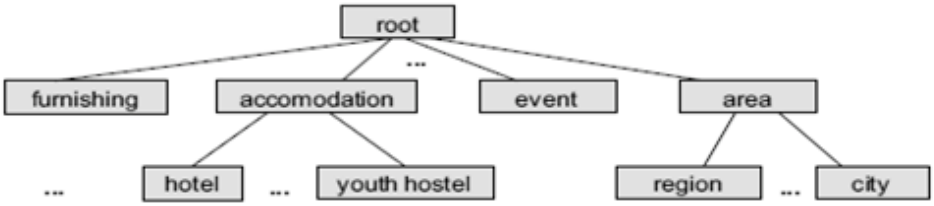


Figura 09: Taxonomia do domínio turístico [42]

O algoritmo de aprendizagem de regras de associação generalizadas [67] utiliza tanto a taxonomia do domínio quanto os pares de conceitos extraídos do corpus. A tabela 11 apresenta os pares de conceitos que representam as relações não-taxonômicas extraídas. Observe que a regra *(room → television)* não foi recomendada ao usuário, uma vez que uma regra ancestral (*room → furnishing*) possui valores de suporte e confiança maiores. O mesmo vale para a regra (*area → hotel*).

Discovered relation	Confidence	Support
(area, accomodation)	0.38	0.04
~~(area, hotel)~~	~~0.1~~	~~0.03~~
(room, furnishing)	0.39	0.03
~~(room, television)~~	~~0.29~~	~~0.02~~
(accomodation, address)	0.34	0.05
(restaurant, accomodation)	0.33	0.02

Tabela 11: Relacionamentos não-taxonômicos extraídos do domínio turístico [42]

A tabela 12 apresenta as soluções em particular adotadas para cada uma das tarefas genéricas para a aprendizagem de relacionamentos não-taxonômicos [57] [58] [59] [60] (seção 2.3), além do tipo de relacionamento adotado (2.4).

Tarefa genérica		Solução adotada	Tipo de rel.
Construção do corpus		Não abordada.	$<c_1, c_2>$
Extração de Rel. Candidatos	Anotação do corpus	REN e *Chunk*.	
	Extração de Rel.	Utiliza duas regras: "sentença" e "título"; para obter um conjunto de pares de conceitos do tipo $<c_1, c_2>$.	
Refinamento de Rel.		Utiliza a extração de regras de associação generalizadas [67]. O produto são os relacionamentos sugeridos ao usuário na forma de regras ($c_1 \rightarrow c_2$).	
Avaliação pelo especialista		Não abordada.	
Atualização da ontologia		Não abordada.	

Tabela 12: Soluções adotadas por ARNT baseada na extração de regras de associação generalizadas [42] para as tarefas genéricas de ARNT

Um aspectos positivo a ser resaltado nessa proposta é o uso da técnica de "Extração de regras de associação generalizadas" [67], na fase de "Refinamento", que além de realizar a recomendação dos relacionamentos mais prováveis permite sugerir o possível melhor nível na hieraquia de conceitos da ontologia onde esses devam ser acrescentados. Por outro lado, alguns aspectos negativos são o fato de a técnica não sugerir um rótulo ao relacionamento e ter seu desempenho dependente de uma heurística associada a estrutura do corpus (heurística de título [42]). Além disso, são utilizadas listas *gazzetiers* para fazer a associação das instancias encontradas no corpus às classes da ontologia. Esse fato deixa a efetividade da técnica dependente da abrangencias dessas listas.

3.3.ARNT baseada em consultas a Web

Sánchez e Moreno [55] [56] propõem um processo independente de domínio para a extração de relacionamentos binários, do tipo $<fn_1, fv, fn_2>$ (seção 2.4), a partir de textos na língua inglesa que usa como corpus páginas indexadas por um mecanísmo de busca na Web e estatísticas calculadas utilizando os resultados de pesquisas no mesmo sistema de busca.

Esse processo está baseado na premissa de que a redundância da informação na Web representa uma medida de sua relevância e veracidade para um domínio. Além disso, considera que os relacionamentos são expressos por frases verbais e pares de conceitos (na forma de frases nominais) a serem extraídos de cada sentença do corpus. O processo (figura 10) é iterativo e consiste em descobrir frases verbais relevantes a um domínio, sendo que um domínio é inicialmente representado por uma palavra chave. Um exemplo é a palavra chave "*hypertension*" (hipertensão), domínio que será usado de agora em diante para ilustrar a aplicação da técnica.

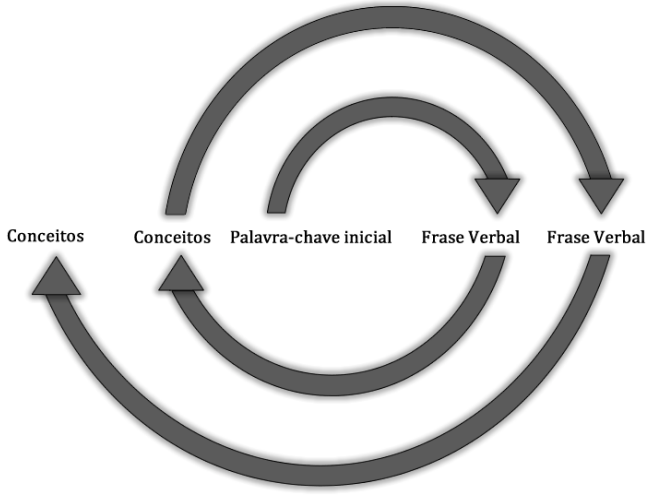


Figura 10: ARNT baseada em consultas a Web

A primeira etapa é utilizar a palavra chave em um buscador para obter um conjunto de documentos Web relacionados ao domínio. Para cada documento recuperado é realizado o *Chunk* (seção 2.5.10.5) para identificar as frases verbais (verbos conjugados e opcionalmente preposições) que vão compor uma lista de relacionamentos candidatos a serem considerados "característicos do domínio". Essas frases verbais candidatas são classificadas em função de sua posição em relação a palavra chave inicial em: predecessora (ex: "*is associated with hypertension*") ou sucessora (ex: "*hypertension is treated with*").

O próximo passo é coletar evidencias de que a frase verbal é característica do domínio. Para tanto, para cada uma delas que tenha sido extraída como predecessora da palavra chave inicial faz-se um cálculo

baseado em consultas feitas sobre um mecanismo de busca Web. Especificamente é calculada a razão entre o número de documentos que apresentam os dois termos: frase verbal (*verbPhrase*) e palavra chave inicial (*initKey*) e o número de documentos que apresentam a frase verbal (*verbPhrase*) (18). Caso a frase verbal candidata seja um sucessor, usa-se a fórmula (19):

$$Score\left(\frac{verbPhrase}{initKey}\right) = \frac{hits("verbPhrase\ initKey")}{hits("verbPhrase")} \quad (18)$$

$$Score\left(\frac{verbPhrase}{initKey}\right) = \frac{hits("initKey\ verbPhrase\ ")}{hits("verbPhrase")} \quad (19)$$

As frases verbais são então ranqueadas segundo os valores calculados. A tabela 13 mostra cada frase verbal, se é predecessora ou sucessora e seu grau de relacionamento com o domínio.

Verb phrase	Position	Relatedness	Verb phrase	Position	Relatedness
is diagnosed in	SUC	0.12	is indicated for	PRE	0.11
are diagnosed as	PRE	0.10	is diagnosed as	PRE	0.08
is associated with	PRE	0.06	are associated with	PRE	0.06
is aggravated by	SUC	0.05	is cured by	SUC	0.03
is caused by	SUC	0.03	occurs during	SUC	0.03
is influenced by	SUC	0.03	suffer from	PRE	0.02
is treated with	SUC	0.02	accelerates	SUC	0.02

Tabela 13: Frases verbais extraídas sobre "*hypertension*"

Na próxima etapa, essas frases verbais são utilizadas para descobrir conceitos que sejam relacionados de forma não-taxonômica à palavra chave inicial. Para tanto, faz-se uma pesquisa em um mecanismo de busca por um dos padrões: "frase verbal palavra_chave" ou "palavra_chave frase verbal". Por exemplo, "*hypertension accelerates*" ou "*is_indicated_for hypertension*". O objetivo é obter conceitos que precedem ou sucedem a *string* utilizada na consulta. Esses novos conceitos são candidatos a serem relacionados de forma não-taxonômica à palavra chave, e o rótulo do relacionamento é a frase verbal. Entretanto, é necessário avaliar quais conceitos extraídos são mais característicos do domínio. Para tanto utiliza-se em um mecanismo de busca para calcular o grau de relacionamento entre o domínio (representado pela

palavra chave) e cada conceito. Especificamente, para cada conceito candidato faz-se o cálculo (20).

$$Score(\frac{Concept}{initKey}) = \frac{hits("initKey"\ AND\ "Concept")}{hits("Concept")} \quad (20)$$

A tabela 14 apresenta os conceitos e seus respectivos valores de relacionamento com o domínio. Aqueles que possuem valores de relacionamento maiores que determinado limiar definido pelo usuário são incorporados à ontologia assim como os relacionamentos.

Subject (NP)	Verb (VP)	Object (NP)	Relat.
diuretic therapy	is indicated for	hypertension	0.61
salt intake	is associated with	hypertension	0.45
latter factors	*are associated with*	*hypertension*	*0.08*
hypertension	*is diagnosed in*	*individuals*	*0.006*
hypertension	is aggravated by	obesity	0.12
hypertension	*is aggravated by*	*the increase*	*0.01*
hypertension	is influenced by	sodium retention	0.65
hypertension	*is influenced by*	*some factors*	*0.02*
hypertension	is treated with	antihypertensives	0.55
hypertension	accelerates	renal disease	0.49
hypertension	*accelerates*	*the development*	*0.003*

Tabela 14: Relacionamentos extraídos sobre "*hypertension*"

A tabela 15 apresenta as soluções em particular adotadas para cada uma das tarefas genéricas para a aprendizagem de relacionamentos não-taxonômicos [57] [58] [59] [60] (seção 2.3), bem como o tipo de relacionamento adotado (2.4).

Tarefa genérica		Solução adotada	Tipo de rel.
Construção do corpus		O corpus consiste em documentos retornados por um mecanismo de busca Web em resposta a consultas com formatos pré-definidos.	<fn_1, fv, fn_2>
Extração de Rel. Candidatos	Anotação do corpus	*Chunk*	
	Extração de Rel.	Realiza busca em cada sentença por frases verbais que precedam ou sucedam a palavra chave inicial.	
Refinamento		É um processamento estatístico: Inicialmente faz uma consulta segundo um padrão pré-estabelecido para verificar o grau de relacionamento da frase verbal com o domínio, em seguida os relacionamentos são ranqueados pelo grau de relacionamento. Aqueles que estiverem acima de um limiar definido pelo usuário são recomendados.	
Avaliação pelo		Não abordada.	
Atualização da ontologia		Não abordada.	

Tabela 15: Soluções adotadas por ARNT baseada em consultas a Web [55] [56] para as tarefas genéricas de ARNT

Um aspecto positivo a ser resaltado nessa proposta é que a aplicação da técnica é facilitada uma vez que os usuários não precisam lidar com a construção ou seleção do corpus; tarefa que pode exigir grande esforço. O corpus é criado automaticamente por um procedimento que realiza consultas a mecanísmos de busca na Web, além disso, o processo é independente de domínio. Por outro lado, um aspecto negativo é o fato de a prendizagem dos relacionamentos ser dependente da aprendizagem de conceitos e vice-versa, o que limita sua aplicabilidade, uma vez que os conceitos da ontologia podem já estar disponíveis e o que é desejado é apenas acrescentar a eles relacionamentos não-taxonômicos. Além disso, os conceitos da ontologia são representados por frases nominais, que normalmente não são termos de fato utilizados para essa finalidade.

3.4.ARNT baseada em Regressão logística

Mohamed et al. [29] desenvolveram uma técnica independente de domínio que extrai relacionamentos não-taxonômicos binários a partir de corpora na língua inglesa do tipo <fn_1, fv, fn_2>. Para tanto são utilizadas duas restrições lingüísticas, uma léxica e outra sintática, conforme explicado a seguir.

Restrição sintática: exige que as frases verbais casem com algum dos padrões expecificados por expressões regulares (figura 11). Esses padrões limitam as frases verbais a serem consideradas relacionamentos a um dos tipos: um verbo (por exemplo, "*invented*"), um verbo seguido imediatamente por uma preposição (por exemplo, "*located in*"), ou um verbo seguido por substantivos, adjetivos ou advérbios que terminam com uma preposição (por exemplo, "*has atomic weight of*").

$V \mid VP \mid VW^{+}P$

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

Figura 11: Padrões sintáticos de ARNT baseada em regressão logística [29]

A restrição sintática serve a dois propósitos. Em primeiro lugar, elimina as extrações "incoerentes", normalmente retornadas por técnicas que não trabalham com esse tipo de restrição. Por exemplo, dada a sentença (21):

Extendicare agreed to buy Arbor Health Care for about US $432 million in cash and assumed debt. (21)

O sistema TextRunner [72] retornaria: <*Arbor Health Care*, *for assumed*, *debt*>. A frase "*for assumed*" claramente não é uma relação válida: ela começa com uma preposição e relaciona duas palavras distantes na sentença. A restrição sintática impede esse tipo de erro, simplesmente restringindo-se a frases relacionais que correspondem aos padrões sintáticos da figura 11.

Em segundo lugar, essa restrição reduz extrações "não informativas", por exemplo, na sentença "*Faust made a deal with devil*" extrair <*Faust*, *made*, *devil*> corresponde a uma relação não informativa. A relação extraída utilizando os padrões sintáticos seria <*Faust*, *made a deal with*, *devil*>, que é um relacionamento válido.

Padrão léxico: se por um lado as restrições sintáticas reduzem "extrações não informativas", por outro permitem que por vezes sejam extraídas relações que por serem demasiadamente específicas possuem muito poucas instancias, até mesmo em grandes corpora extraídos da Web. Por exemplo, dada a sentença (22):

The Obama administration is offering only modest greenhouse gas reduction targets at the conference. (22)

O padrão sintático irá coincidir com a frase: "*is offering only modest greenhouse gas reduction targets at*". Assim há frases que satisfazem a restrição sintática, mas que não são relacionamentos. Para superar essa limitação, é incluída uma restrição léxica que é usada para separar frases relacionais válidas de frases de relacionamentos muito específicos, como o do exemplo (21). A restrição é baseada na intuição de que uma frase relacional válida deve ter muitos argumentos distintos em um corpus de grande porte. A frase na sentença (21) é específica para o par de argumentos (*Obama administration*, *conference*), por isso é pouco provável que represente um relacionamento. A restrição léxica é implementada por um repositório de frases verbais que são consideradas suficientemente genéricas (possuem muitos argumentos distintos). O repositório é construído manualmente e sempre que uma frase verbal que satisfaça algum dos padrões sintáticos for extraída do texto é feita uma busca no repositório. Frases verbais não presentes no repositório não são recomendadas como relacionamentos.

Essa técnica tem como entrada uma sentença anotada por *POS Tag* e *NP chunk* e retorna um conjunto de tuplas extraídas (<fn_1, fv, fn_2>). O processo (figura 12) consiste em duas fases conforme descrito a seguir.

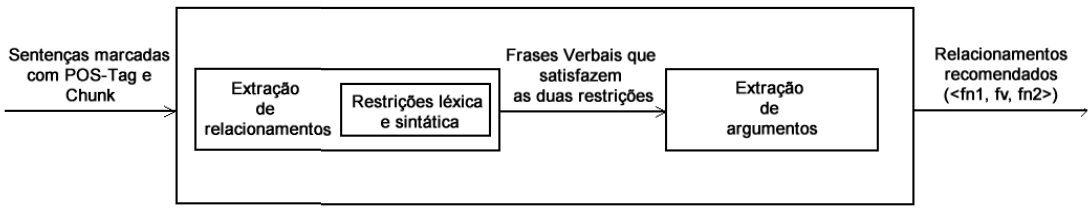


Figura 12: A técnica “ARNT baseada em regressão logística”

Na primeira fase, “Extração de relacionamentos”, para cada verbo “v” em cada sentença “s” do corpus, é identificada a sequencia mais longa de palavras “fv” tal que (1) “fv” é iniciada por “v”, (2) “fv” satisfaz a restrição sintática e (3) “fv” satsifaz a restrição léxica.

Na segunda fase, “Extração de argumentos”, para cada frase verbal “fv” identificada no passo anterior, é encontrada a frase nominal “fn_1” à esquerda de “fv” na sentença de forma que “fn_1” não seja um pronome relativo, advérbio "who" ou o existencial "there". Em seguida, é encontrada a frase nominal mais próxima “fn_2” à direita de “fv” em “s”. Se um par (fn_1, fn_2) tiver sido encontrado, então é retornado “<fn_1, fv, fn_2>” como um relacionamento extraído.

As fases de “Extração de relacionamentos” e “Extração de argumentos” apresentam alto *recall* mas baixa precisão. Dessa forma é necessário um refinamento no intuito de revelar os mais prováveis relacionamentos dentre todos os extraídos com a aplicação dos padrões lingüísticos (restrições léxica e sintática). Para tanto é utilizado um classificador de regressão logística [46] que atribui uma probabilidade a cada relacionamento e permite criar um ranque segundo a probabilidade que cada um tem de representar um relacionamento "válido". As variáveis características da função logística e seus respectivos pesos são dados na tabela 16.

Weight	Feature
1.16	(x, r, y) covers all words in s
0.50	The last preposition in r is *for*
0.49	The last preposition in r is *on*
0.46	The last preposition in r is *of*
0.43	$len(s) \leq 10$ words
0.43	There is a WH-word to the left of r
0.42	r matches VW*P from Figure 1
0.39	The last preposition in r is *to*
0.25	The last preposition in r is *in*
0.23	10 words $< len(s) \leq 20$ words
0.21	s begins with x
0.16	y is a proper noun
0.01	x is a proper noun
-0.30	There is an NP to the left of x in s
-0.43	20 words $< len(s)$
-0.61	r matches V from Figure 1
0.65	There is a preposition to the left of x in s
-0.81	There is an NP to the right of y in s
-0.93	Coord. conjunction to the left of r in s

Tabela 16: Variáveis características do classificador de regressão logística [48]

A tabela 17 apresenta as soluções em particular adotadas para cada uma das tarefas genéricas para a aprendizagem de relacionamentos não-taxonômicos [57] [58] [59] [60] (seção 2.3), bem como o tipo de relacionamento adotado (2.4).

Tarefa genérica		Solução adotada	Tipo de rel.
Construção do corpus		Não abordada.	<fn_1, fv, fn_2>
Extração de Rel. Candidatos	Anotação do corpus	*Chunk*.	
	Extração de Rel.	Utiliza o algoritmo descrito na técnica. Obtém como resultado um conjunto de relacionamentos candidatos: <fn_1, fv, fn_2> que estejam em conformidade com as restrições léxica e sintática da técnica.	
Refinamento		Utiliza a técnica de AM, regressão logística. O produto são os relacionamentos <fn_1, fv, fn_2> associados a uma probabilidade de serem reais relacionamentos não-taxonômicos.	
Avaliação pelo especialista		Não abordada.	
Atualização da ontologia		Não abordada.	

Tabela 17: Soluções adotadas por ARNT baseada em regressão logística [29] para as tarefas genéricas de ARNT

A ARNT baseada em regressão logística [29] utiliza um conhecimento linguistico (língua inglesa) ignorado em maior ou menor grau pelas demais abordagens avaliadas. Esse conhecimento é representado na forma de padões linguísticos formalizados como expressões regulares e de um filtro que verifica a coicidencia de frases verbais obtidas do corpus com as de uma repositório com o objetivo de aumentar a precisão da técnica. Além disso, a técnica é capaz de extrair relacionamentos de corpus muitos pequenos, inclusive de uma única sentença e de qualquer área de conhecimento (independente de domínio).

O aspecto negativo dessa abordagem é que ela utiliza um repositório de frases verbais, construído manualmente, que contém aquelas que são consideradas suficientemente genéricas de forma que possam estar presentes em sentenças relacionando conceitos diferentes. Caso um relacionamento potencialmente "válido" seja representado por uma frase verbal que não esteja no repositório, este não será recomendado pela técnica.

3.5.ARNT baseada na classificação de relacionamentos

Mohamed, Hruschka e Mitchell [48] propõem uma técnica de ARNT que realiza tanto a aprendizagem quanto o povoamento de relacionamentos não-taxonômicos do tipo <c_1, fv, c_2> (seção 2.4). Essa técnica recomenda relacionamentos não-taxônomicos a partir de um corpus em lingua inglesa e infere novos relacionamentos a partir desses, além de identificar suas instancias.

A técnica “ARNT baseada na classificação de relacionamentos” [48] tem como insumos os conceitos de uma ontologia (conjunto C, seção 2.1), uma lista de instancias associada a cada um desses e um corpus. Os produtos entregues pela técnica são: um conjunto de relacionamentos não-taxonomicos representados por três elementos, dois conceitos da ontologia e um rótulo. (ex.: *RiverFlowsThroughCity*(<*River*>, <*City*>)); um conjunto de instancias (ex.: *RiverFlowsThroughCity*(<*Nile*>, <*Cairo*>)) para cada relacionamento e um conjunto de padrões léxicos de extração de texto que podem ser utilizados para extrair novas instancias desse relacionamento, por exemplo "*X in the heart of*

Y" com o qual poderia ser identificado o relacionamento entre “X” e “Y” na sentença "*Tamisa in the hart of London*").

Na fase de "Pré-processamento", para cada par de conceitos é criado um conjunto “S” formado por todas as sentenças contendo instancias conhecidas de ambos.

Na fase "Geração de relacionamentos", uma matriz de co-ocorrencia de contextos é gerada para cada par de conceitos da ontologia (figura 13). Nessa matriz cada célula corresponde a frequencia de instancias de pares de conceitos com os quais ambos os contextos co-ocorrem. Por exemplo, para as sentenças "*Vioxx can cure Arthritis*" e "*Vioxx is a treatment for Arthritis*" os contextos "*can cure*" e "*is a treatment for*" co-ocorrem com um par de instancias "*Vioxx*" e "*Arthritis*".

Para ilustrar a aplicação da técnica considere que como resultado do pré-processamento, para o par de conceitos <*drug*, *disease*> tenham sido obtidos 122 contextos. Contextos como "*to tread*", "*for treatment of*" e "*medication*" que indicam o mesmo relacionamento (*drug-to treat-disease*) tem altos valores de co-ocorrencia (figura 13). O mesmo ocorre para os contextos "*can cause*", "*may cause*" e "*can lead to*", que indicam o relacionamento “*drug-can cause-disease*” (figura 13).

Contexts / Contexts	may cause	can cause	can lead to	to treat	for treatment of	medication
may cause	0.176	0.074	0.030	0.015	0.011	0.000
can cause	0.051	0.150	0.039	0.018	0.013	0.010
can lead to	0.034	0.064	0.189	0.019	0.021	0.018
to treat	0.006	0.011	0.007	0.109	0.043	0.015
for treatment of	0.005	0.008	0.008	0.045	0.086	0.023
medication	0.000	0.011	0.009	0.030	0.036	0.111

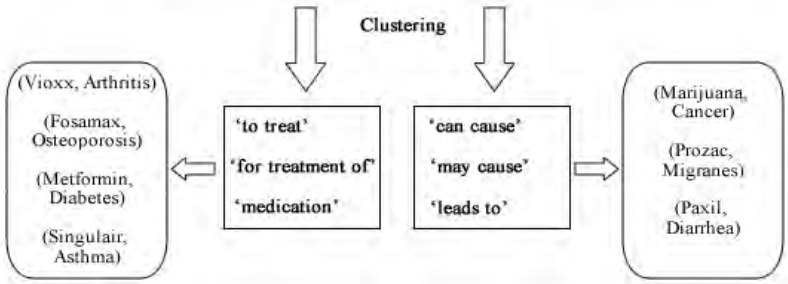


Figura 13: ARNT baseada na classificação de relacionamentos [48]

Sobre essa matriz é realizado um agrupamento sobre os valores de co-ocorrência. Cada agrupamento é então utilizado para propor um possível novo relacionamento. O centroide de cada agrupamento é utilizado para sugerir o nome do novo relacionamento. Por exemplo, se o centróide de um agrupamento for "*for treatment of*" então o nome do relacionamento é "*drug-for-treatment-of-disease*". Em seguida são geradas as intancias iniciais (*seed instances*) para os novos relacionamentos. As instancias de relacionamentos (pares de conceitos) que correspondem ao centróide (ou que estejam próximas ao centróide) são as mais representativas do relacionamento. Cada instancia inicial (*seed instance*) de relacionamento é pesada com a fórmula (23):

$$\text{weight}(seed\ instance) = \sum_{c \in \text{PatternCluster}} \text{Occ}(c, s) / (1 + \text{sd}(c)) \qquad (23)$$

Na formúla 19, "*pattern cluster*" é o agrupamento de padrões de contexto para o relacionamento considerado. “Occ(c,s)” é o número de vezes que a instancia de relacionamento (“s”) co-ocorre com o padrão de contexto (“c”). “sd(c)” é o desvio padrão do contexto (“c”) em relação ao centroide do seu agrupamento. As instancias são ranqueadas por essa medida e as 50 primeiras são consideradas as instancias iniciais (*seed instances*) do relacionamento proposto.

Muitos relacionamentos extraidos na fase “Geração de relacionamentos” não correspondem a relacionamentos válidos. Por esse motivo, na fase "Classificando relacionamentos semanticamente válidos" algumas heurísticas [48] são utilizadas como critérios de classificação. Uma delas diz respeito a quão específico é o padrão de contexto em relação a um dado relacionamento. Por exemplo, seja “<c_1, c_2>” um par de conceitos da ontologia. Se o mesmo contexto conectar instancias de “c_1” a um número grande de instancias que não as de “c_2” então esse contexto não deve indicar um relacionamento válido.

A tabela 18 apresenta as soluções em particular adotadas para cada uma das tarefas genéricas para a aprendizagem de relacionamentos não-taxonômicos [57] [58] [59] [60] (seção 2.3), bem como o tipo de relacionamento adotado (2.4).

Tarefa genérica		Solução adotada	Tipo de rel.
Construção do corpus		Não abordada.	$<c_1, fv, c_2>$
Extração de Rel. Candidatos	Anotação do corpus	Tokenização, separação de sentenças e REN.	
	Extração de Rel.	Identifica relacionamentos de duas formas: 1-extração de tuplas "$<c_1, fv, c_2>$" a partir do corpus. 2-geração de novas tuplas, a partir do agrupamento de contextos, conforme descrito na técnica.	
Refinamento		Usa um classificador que utiliza heurísticas como critérios de classificação para recomendar apenas relacionamentos considerados válidos.	
Avaliação pelo especialista		Não abordada.	
Atualização da ontologia		Não abordada.	

Tabela 18: Soluções adotadas por ARNT baseada na classificação de relacionamentos [48] para as tarefas genéricas de ARNT

3.6.Avaliação das técnicas de ARNT

As técnicas de ARNT assim como quaisquer outras na área de AO estão sujeitas a grande quantidade de ruído uma vez que a fonte de informação da qual os relacionamentos são extraídos é desestruturada. Portanto, soluções customizáveis são necessárias para que essas técnicas possam ser aplicáveis a maior gama de situações. Entretanto, conforme discutido nessa seção, as técnicas de ARNT não tem contemplado da forma mais adequada possível esse aspecto.

Villaverde et al. [69] propuseram uma técnica de ARNT que utiliza *Chunk* na fase de "Anotação do corpus" com o objetivo de identificar as frases verbais presentes no texto. Na fase "Extração de relacionamentos" uma tupla na forma $<c_1, fv, c_2>$ ("c_1" e "c_2" são conceitos da ontologia e "fv" é uma frase verbal) é gerada para cada ocorrência de dois conceitos consecutivos com uma frase verbal entre eles.

Por exemplo, na sentença "*The young couple have two children.*" A tupla gerada "*<couple, have, child>*" é um relacionamento candidato (produto da fase "Extração de relacionamentos"). Para possivelmente elevar o *recall*, os

sinônimos dos conceitos da ontologia são incluídos na busca por conceitos no corpus. Na fase de "Refinamento" os relacionamentos candidatos são submetidos a um algoritmo para extração de regras de associação [46] que recomenda relacionamentos não-taxonômicos na forma de regras de associação [1] [2]. Por exemplo, a regra "<*couple*, *child*> → *have*" expressa que há um relacionamento não-taxonômico entre os dois conceitos "*couple*" e "*child*" e que o rótulo é o verbo "*have*". O suporte e a confiança [1] [2] (seção 2.6.2) são dois parâmetros de corte sobre os relacionamentos recomendados e tem seus valores definidos experimentalmente pelo especialista.

Em sua primeira fase "Cálculo do suporte das regras", o algoritmo para extração de regras de associação [1] [2] agrupa tuplas para calcular sua frequência, sempre que os três elementos ("c_1", "c_2" e "fv") de tuplas diferentes coincidem. Essa abordagem não é muito adequada, uma vez que torna mais difícil a separação de tuplas possivelmente mais relevantes das menos relevantes por meio do parâmetro de corte suporte mínimo. Isso ocorre, pois a tendência é que haja um grande número de regras de associação com os mesmos valores para esse parâmetro.

Além disso, de forma não muito apropriada para o contexto de ARNT, esse algoritmo valoriza regras de associação que possuem altos valores de confiança (seção 2.6.2.1), medida que representa a informação de quanto um rótulo (frase verbal) está associado a um par de conceitos. Entretanto, e em contrapartida perde efetividade na realização da tarefa mais relevante em ARNT, que é a identificação de quais pares de conceitos da ontologia estão relacionados de forma não-taxonômica, já que a relevância de um relacionamento recomendado pela técnica é melhor representada pela taxa de ocorrência do par de conceitos no corpus. Uma implementação dessa medida é apresentada no algoritmo de refinamento *Bag of labels* [61] (seção 4.2.4.2), adotado por TARNT [60] [61].

Merece também menção o fato de o especialista ter que lidar com a tarefa custosa de ajustar experimentalmente os valores de dois parâmetros de corte (suporte e confiança mínimos) do algoritmo de extração de regras de associação [1] [2].

Maedche e Staab [42] propõem uma técnica que extrai relacionamentos não-taxonômicos de corpora com instancias, e não com classes como realizado por Villaverde et al. [69]. Para tanto é utilizada REN na fase de "Anotação do corpus" para associar instancias encontradas no texto às suas respectivas classes da ontologia. Na fase de "Extração de relacionamentos" uma tupla na forma "<c_1, c_2>" ("c_1" e "c_2" são conceitos da ontologia) é gerada para quaisquer das duas situações: para cada par de conceitos em uma mesma sentença do corpus e para cada instancia em um título com cada instancia em seu correspondente corpo de texto. Na fase de "Refinamento" os relacionamentos candidatos são submetidos ao algoritmo de mineração de regras de associação generalizadas [67] que sugere regras de associação na forma "$c_1 \rightarrow c_2$", significando que há um relacionamento não-taxonômico entre os conceitos "c_1" e "c_2". Esse algoritmo sugere o possível melhor nível na hierarquia de conceitos da ontologia no qual o relacionamento deva ser inserido. Por exemplo, no caso de um supermercado, o algoritmo poderia sugerir que "*snacks are purchased together with drinks*" ao invés de "*chips are purchased with beer*" e "*peanuts are purchased with soda*".

Na fase "Extração de relacionamentos", Maedche e Staab [42] extraem uma tupla "<c_1, c_2>" para cada instancia de conceito da ontologia em um título com cada instancia em seu correspondente corpo de texto (regra de título). Essa regra aumenta consideravelmente o número de tuplas geradas nessa fase, o que de forma geral é útil para técnicas que utilizam extração de regras de associação e especialmente para esse caso, uma vez que os conceitos nos títulos tendem a ser mais genérico que os do corpo do texto. Entretanto, há a exigência de que o corpus dado como entrada para a técnica possua anotações especificas indicando quais partes do texto correspondem a títulos e quais não. Esse fato limita os textos que podem ser utilizados como corpus e requer adaptações na fase de "Anotação do corpus". Além disso, a técnica proposta por Maedche e Staab [42] não sugere rótulos aos relacionamentos extraídos. Merece também menção o fato de o especialista ter que lidar com a tarefa custosa de ajustar experimentalmente os valores de dois parâmetros de corte (suporte e confiança mínimos) do algoritmo de extração de regras de associação generalizadas [67].

Sanchez e Moreno [55] [56] desenvolveram uma técnica de ARNT que automatiza a fase de "Construção do corpus" por meio de consultas a mecanismos de busca Web. A técnica de PLN POS tag é utilizada na fase de "Anotação do corpus" para identificar as frases verbais em cada sentença. Na fase "Extração de relacionamentos", para cada sentença, relacionamentos do tipo <fn_1, fv, fn_2> (seção 2.4) são extraídos para cada frase verbal ("fv") com as primeiras frases nominais a sua esquerda ("fn_1") e direita ("fn_2"). Na fase de "Refinamento" é utilizada uma solução estatística que consiste na execução de fórmulas predefinidas para calcular o grau de relacionamento das tuplas extraídas (<fn_1, fv, fn_2>) com o domínio.

A proposição de Sanchez e Moreno [55] [56] automatiza a tarefa custosa de construção do corpus com o uso de consultas a mecanismos de busca Web, o que é um aspecto positivo. Entretanto, ela realiza a aprendizagem de conceitos e relacionamentos conjuntamente, ignorando dessa forma a situação realista na qual os conceitos da ontologia são conhecidos a priori e para os quais se deseje extrair relacionamentos não-taxonômicos com um procedimento desenvolvido especificamente para essa finalidade. Além disso, os conceitos da ontologia são representados por frases nominais, que normalmente não são termos de fato utilizados para essa função.

A técnica proposta por Fader et al. [29] utiliza a técnica de PLN *VP chunk* na fase de "Anotação do corpus" para identificar frases verbais em cada sentença. Na fase "Extração de relacionamentos", para cada sentença do corpus, os relacionamentos não-taxonômicos são representados por tuplas do tipo <fn_1, fv, fn_2> (seção 2.4), sendo "fv" uma frase verbal e "fn_1" e "fn_2" as primeiras frases nominais a sua esquerda e direita. Na fase de "Refinamento", um classificador de regressão logística [29] é utilizado para ranquear as tuplas extraídas segundo suas probabilidades de serem relacionamentos válidos. Dois exemplos de variáveis características utilizadas pela função logística são: "a sentença tem menos que dez palavras" e "as frases nominais e verbais correspondem à sentença completa".

Fader et al. [29] utilizam um classificador de regressão logística para ranquear os relacionamentos não-taxonômicos segundo sua probabilidade de

serem válidos o que torna a acurácia da classificação entre relacionamentos válidos e inválidos independente do tamanho do corpus, além disso, recomenda rótulos aos relacionamentos extraídos. Entretanto, uma vez que diferentemente de Villaverde et al. [69] e Maedche e Staab [42] não recebe os conceitos da ontologia como *input,* a proposição de Fader et al. [29] considera as frases nominais como conceitos, as quais normalmente não corespondem a nomes de classes utilizados na prática.

A técnica proposta por Mohamed, Hruschka e Mitchell [48], na fase "Anotação do corpus" utiliza a técnica de PLN *Chunk* para identificação das frases nominais e verbais. Na fase "Extração de relacionamentos", para cada sentença que contenha um par de instancias conhecidas de classes da ontologia, é criada uma tupla do tipo $<c_1, fv, c_2>$. O rótulo "fv" corresponde ao texto existente entre as duas instancias, desde que corresponda a algum dos padrões sintáticos predefinidos [29]. Para cada par de conceitos é criada uma matriz de co-ocorrência conforme descrito na seção 3.5. Um algoritmo de agrupamento é utilizado para criar grupos de contextos mais similares. Para cada um desses *clusters* é criado um novo relacionamento. Suas instancias iniciais são as cinquenta primeiras da lista de instancias dos relacionamentos que compõem o *cluster*. Na fase de "Refinamento", um conjunto de heurísticas é utilizado para classificar os relacionamentos provenientes da fase anterior em válidos ou inválidos.

Alguns aspectos positivos dessa proposição são: extrai relacionamentos com rótulos e sugere novos relacionamentos não-taxonômicos a partir dos previamente conhecidos, com a realização de um agrupamento sobre uma matriz de co-ocorrência de contextos conforme descrito na seção 3.5. Algumas das limitações dessa proposição são: utilização de REN para identificar no texto instancias para cada classe da ontologia ficando dessa forma dependente da abrangência dessas listas; necessidade de dispor de uma ontologia povoada, ou seja, esse procedimento só pode ser aplicado quando se dispõe dos conjuntos "C" e de um subconjunto do conjunto "**I**" (seção 2.1). Além dessas, a regra de extração de relacionamentos da fase "Extração de relacionamentos candidatos" extrai uma tupla sempre que houver exatamente um par de instancias de conceitos em uma sentença. Entretanto,

uma sentença pode ter, com razoável frequência, mais de um relacionamento candidato. A tabela 19 apresenta um resumo comparativo das técnicas de ARNT discutidas neste capítulo com relação às soluções por elas adotadas para as fases do processo genérico [57] [58] [59] [60]. A tabela 20 apresenta alguns dos principais aspectos positivos e limitações para essas mesmas técnicas. As técnicas são representadas pelos números de suas seções nesse capítulo.

Além de avaliações qualitativas, como as apresentadas nesta seção, avaliações quantitativas, como as apresentadas no capítulo 5, podem também ser realizadas. Essas avaliações consistem em medir com o uso de procedimentos e métricas bem definidas a efetividade das técnicas de ARNT no cumprimento de sua tarefa. Medidas tipicamente utilizadas para essa finalidade são o *recall*, precisão e medida-F da área de recuperação de informação [34], conforme definidas na seção 2.3.

Técnica de ARNT	**Fases de ARNT**				**Tipo de rel.**
	Construção do corpus	**Anotação do corpus**	**Extração de relacionamentos**	**Refinamento**	
3.1 [69].	Não abordada	*Chunk*.	Extrai relacionamentos de sentenças que possuam frases verbais entre dois conceitos.	Utiliza Extração de regras de associação [1] [2] para recomendar rel. com rótulos.	(c_1,fv,c_2)
3.2 [42].	Não abordada	*Chunk* e REN.	Utiliza as regras de "sentença" e de "título" (seção 3.2)	Utiliza Extração de regras de associação generalizadas [67] para sugerir o melhor nível hierarquico.	(c_1, c_2)
3.3 [55].	Procedimento baseado em consultas a Web que retorna documentos do domínio.	*Chunk.*	Extrai relacionamentos de sentenças que possuam frases verbais entre dois conceitos.	Processamento estatístico (seção 3.3) que verifica a relevância de cada relacionamento no domínio.	(fn_1,fv,fn_2)
3.4 [29].	Não abordada	*Chunk.*	Aplica padrões léxicos e sintáticos para obter relacionamentos a partir de sentenças.	Utiliza regressão logística para obter os rel. mais prováveis.	(fn_1,fv,fn_2)
3.5 [48].	Não abordada	Separação de sentenças e REN.	Extrai relacionamentos de sentenças que possuam frases verbais entre dois conceitos.	Utiliza heurísticas para classificar e recomendar apenas relacionamentos considerados válidos.	(c_1,fv,c_2)

Tabela 19: Comparação das técnicas de ARNT com relação às soluções para as fases do processo genérico [57] [58] [59] [60].

Técnicas de ARNT	Aspectos positivos	Limitações
3.1 [69].	• Sugere rótulos aos relacionamentos. • Nenhuma característica específica é exigida do corpus, como a feita por Maedech e Staab [42] de possuir títulos e corpo de texto.	• Dificuldade em determinar experimentalmente os valores para os parâmetros de corte suporte e confiança minimos. • Valorização da informação de quanto um rótulo está associado a um par de conceitos em prejuízo de quais pares estão relacionados.
3.2 [42].	• Alto *recall* na fase de "Extração de relacionamentos candidatos" devido às regras de extração utilizadas. • Sugere o possível melhor nível na taxonomia onde os relacionamentos devam ser acrescentados.	• Não sugere rótulos aos relacionamentos. • Aplicável somente sobre corpora que apresente marcação que distinga corpo de texto de título.
3.3 [55].	• Solução automatizada para a fase "Construção do corpus". • Sugere rótulos aos relacionamentos.	• Realiza a ARNT e aprendizagem de conceitos conjuntamente, ignorando a situação realista na qual esses últimos são conhecidos à priori. • Conceitos são representados por frases nominais, que normalmente não são nomes de fato utilizados para essa finalidade.
3.4 [29].	• Sugere rótulos aos relacionamentos. • Mantêm a acurácia independentemente do tamanho do corpus.	• Utiliza frases nominais extraídas do corpus como conceitos, as quais normalmente não corespondem a nomes de classes utilizados na prática. • Utiliza um repositório de frases verbais, construído manualmente para melhorar a precisão.
3.5 [48].	• Sugere rótulos aos relacionamentos. • Sugere novos relacionamentos não-taxonômicos a partir de outros previamente conhecidos	• Dependencia em relação REN para a identificação das instancias de classes da ontologia nas sentenças. • Necessidade de dispor de uma ontologia povoada.

Tabela 20: Aspectos positivos e limitações de técnicas do estado da arte em ARNT

3.7.Considerações finais

Nesse capítulo foram apresentadas e discutidas algumas técnicas para ARNT [29] [42] [48] [55] [56] [69] representativas do estado da arte. Foi evidenciado como cada uma delas soluciona as fases do processo genérico de ARNT [57] [58] [59] [60], apresentado na seção 2.4, destacando os aspectos positivos e limitações de cada uma delas.

A "ARNT baseada em consultas a Web" [55] [56], por exemplo, automatiza a tarefa de construção do corpus, a "ARNT baseada em regressão logística" [29] mantém sua acurácia independentemente do tamanho do corpus, a "ARNT baseada na extração de regras de associação" [69] recomenda rótulos aos relacionamentos, a "ARNT baseada na extração de regras de associação generalizadas" [42] sugere o possível nível mais adequado na hieraquia de conceitos da ontologia onde o relacionamento deva ser acrescentado e a "ARNT baseada na classificação de relacionamentos" [48] sugere novos relacionamentos não-taxonômicos a partir de outros previamente conhecidos.

Por fim, foi realizada uma avaliação comparativa das técnicas de ARNT apresentadas (seção 3.6) com o intuito de identificar deficiências para as quais fosse possível desenvolver soluções mais adequadas e também características consideradas positivas e que pudessem ser mantidas na técnica proposta nesse trabalho.

O próximo capítulo apresenta uma técnica para a aprendizagem de relacionamentos não-taxonômicos de ontologias (TARNT) [60] [61]. Por ser parametrizada e incorporar soluções que a diferencia de outras técnicas como a regra de apóstrofo [61] (seção 4.2.3) na fase de "Extração de relacionamentos candidatos" e *Bag of labels* [61] (seção 4.2.4.2) na fase de "Refinamento", TARNT pode ajudar os engenheiros de ontologia a obter bons resultados em uma maior variedade de situações quando comparado à maioria dos trabalhos relacionados.

4. TARNT – Uma técnica para Aprendizagem de Relacionamentos Não-Taxonômicos de Ontologias

Esse capítulo apresenta TARNT [60] [61], uma Técnica para Aprendizagem de Relacionamentos Não-Taxonômicos de ontologias a partir de fontes de informação textuais na língua inglesa que está em conformidade com o processo genérico para ARNT [57] [58] [59] [60] definido no capítulo 2. TARNT utiliza técnicas de PLN e estatísticas para extrair e sugerir ao engenheiro de conhecimento relacionamentos não-taxonômicos sobre conceitos de ontologias conhecidos a priori. Por ser parametrizada e incorporar soluções que a diferencia de outras técnicas como a regra de apóstrofo (seção 4.2.3) na fase de "Extração de relacionamentos candidatos" e *Bag of labels* [61] na fase de "Refinamento", TARNT pode permitir aos engenheiros de conhecimento obter bons resultados em uma maior variedade de situações quando comparado à maioria dos trabalhos relacionados. É também apresentada a ferramenta de software TARNTool desenvolvida para prover suporte a aplicação de TARNT e que no contexto dessa Tese é utilizada para realizar avaliações quantitativas da técnica e a verificação da hipótese de pesquisa, conforme descrita na seção 1.3.

O capítulo está organizado da seguinte forma: A seção 4.1 descreve algumas características gerais de TARNT [60] [61] que serão detalhadas nas seções seguintes deste capítulo. A seção 4.2 apresenta uma descrição detalhada de TARNT e das soluções por ela adotadas para cada uma das fases do processo genérico de ARNT [57] [58] [59] [60]. Na seção 4.3 é apresentada uma discussão sobre as diferentes configurações de TARNT. Na seção 4.4 é apresentada TARNTool, a ferramenta de software de suporte a TARNT. Na seção 4.5 é apresentada uma avaliação qualitativa de TARNT em relação às técnicas de ARNT apresentas no capítulo 3. Por fim, a seção 4.5 faz as considerações finais sobre o capítulo.

4.1. Considerações gerais sobre TARNT

TARNT [60] [61] é uma técnica para aprendizagem de relacionamentos não-taxonômicos de ontologias que está em conformidade

com o processo genérico para ARNT [57] [58] [59] [60] definido na seção 2.3 e ilustrado na figura 2. TARNT possui seis fases que serão detalhadas nas próximas seções. Na primeira delas, “Construção do corpus”, um conjunto de documentos na língua inglesa no domínio de interesse é construído. Na fase “Anotação do corpus”, técnicas de PLN são utilizadas para marcar o corpus com anotações necessárias à extração dos relacionamentos candidatos. Na fase “Extração de relacionamentos”, regras de extração são utilizadas para extrair do corpus previamente anotado os possíveis relacionamentos não-taxonômicos (relacionamentos candidatos). Na fase de “Refinamento” os relacionamentos candidatos são filtrados com o objetivo de sugerir ao especialista os mais prováveis. Na fase de “Avaliação do especialista”, esse seleciona e possivelmente edita os relacionamentos a serem acrescentados a ontologia. Por fim, na fase “Atualização da ontologia”, um procedimento de atualização do arquivo da ontologia com os relacionamentos não-taxonômicos aprendidos é executado.

TARNT é parametrizada. Na fase “Anotação do corpus”, diferentes soluções são disponibilizadas cujas características são discutidas na seção 4.2.2. Para a fase “Extração de relacionamentos” três regras de extração foram definidas (seção 4.2.3). Já na fase de “Refinamento” duas soluções, *Bag of labels* [61] e “Frequência de co-ocorrência” [61], podem ser utilizadas dependendo se os relacionamentos são recomendados respectivamente com ou sem rótulos (seção 4.2.4).

TARNT pode trabalhar tanto com relacionamentos do tipo “$<c_1, c_2>$” quanto “$<c_1, fv, c_2>$”, dependendo da configuração adotada. Na fase “Extração de relacionamentos”, as regras de extração, regra de sentença (seção 4.2.3.1) e regra de apóstrofo (seção 4.2.3.3) utilizam a representação com apenas dois conceitos, já a regra de sentença com frase verbal (seção 4.2.3.2) utiliza a representação que inclui um rótulo. Na fase de “Refinamento”, tanto *Bag of labels* [61] quanto “Frequência de co-ocorrência” [61] trabalham internamente com relacionamentos “$<c_1, c_2>$”, apesar de, conforme será discutido na seção 4.2.4.2, diferentemente de “Frequência de co-ocorrência” [61], *Bag of labels* [61] ter como *input* relacionamentos do tipo “$<c_1, fv, c_2>$”.

4.2. Descrição detalhada dos componentes de TARNT

As soluções adotadas por TARNT para cada fase do processo genérico de ARNT [57] [58] [59] [60] proposto na seção 2.3 são resumidas na tabela 21 e descritas detalhadamente nas seções 4.2.1 a 4.2.6.

Tarefa genérica		Solução adotada
Construção do corpus		Não abordada.
Extração de Relacionamentos Candidatos	Anotação do corpus	Tokenização, separação de sentenças e opcionalmente lematização, análise morfológica e *vp chunk*
	Extração de Relacionamentos	Regra de sentença [61], regra de sentença com frase verbal [61] e regra de apóstrofo [61] .
Refinamento de Relacionamentos		Frequência de co-ocorrência [61] e *Bag of labels* [61]
Avaliação pelo especialista		Seleção e edição manual de relacionamentos não-taxonômicos
Atualização da ontologia		Execução do procedimento para atualização do arquivo owl da ontologia com os relacionamentos não taxonômicos

Tabela 21: Soluções adotadas por TARNT para as fases do processo genérico de ARNT [61]

4.2.1.Construção do corpus

TARNT não define uma solução específica a ser adotada nessa fase sendo o especialista responsável pela escolha daquela que melhor satisfizer a necessidade de obtenção de um corpus que possua os elementos linguisticos necessários a extração dos relacionamentos.

4.2.2. Anotação do corpus

Essa fase tem o objetivo de marcar o corpus com anotações necessárias à aplicação das regras de extração selecionadas pelo especialista na sub-fase “Extração de relacionamentos”. Portanto, TARNT disponibiliza cinco técnicas de PLN que são executadas conforme o diagrama de atividades da figura 14.

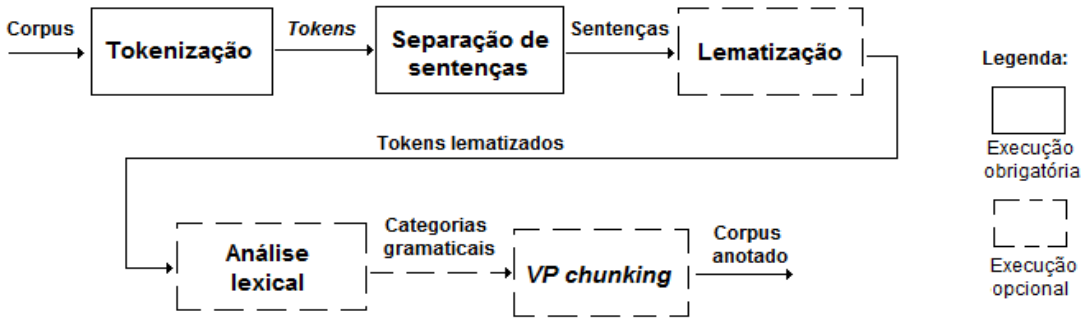


Figura 14: Solução de TARNT para a fase “Anotação do corpus” [61]

A tokenização é uma tarefa básica de PLN e sua execução é pré-requisito para aplicação de qualquer outra técnica. A separação de sentenças é necessária por dois motivos, primeiramente sua execução é pré-requisito a execução das demais técnicas de PLN disponibilizadas por TARNT, em segundo lugar, a sentença é a unidade linguística da qual são extraídos os relacionamentos. A lematização é utilizada visando aumentar o *recall* da busca por conceitos da ontologia no corpus. A análise lexical classifica os termos em classes gramaticais e é necessária para permitir a posterior identificação das frases verbais que são recomendadas como rótulos dos relacionamentos caso TARNT tenha sido configurada para cumprir essa tarefa. *VP chunk* é uma técnica de PLN para identificação de frases verbais que são sugeridas como rótulos dos relacionamentos. Tanto a análise lexical quanto o *VP chunk* são executados apenas quando a regra de sentença com frase verbal (seção 4.2.3.2) é selecionada na fase “Extração de Relacionamentos”.

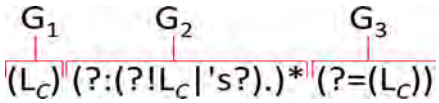
4.2.3.Extração de Relacionamentos

Nessa fase um conjunto de regras de extração selecionadas pelo especialista são utilizadas para extrair do corpus previamente anotado os relacionamentos candidatos. TARNT disponibiliza três regras de extração de relacionamentos: regra de sentença, regra de sentença com frase verbal e regra de apóstrofo, descritas nas seções 4.2.3.1, 4.2.3.2 e 4.2.3.3 respectivamente. Para exemplificar a aplicação das referidas regras são utilizadas sentenças retiradas do corpus *Family law doctrine* (seção 5.4) do

domínio do direito da família e os conceitos: “*agreement*”, “*court*”, “*decree*”, “*divorce*”, “*lawyer*”, “*marriage*”, “*party*”, “*property*” e “*spouse*”.

4.2.3.1.Regra de Sentença

A regra de sentença [61] é baseada na heurística de que dois conceitos consecutivos em uma mesma sentença estão provavelmente relacionados de forma não-taxonômica. Esse é o caso dos conceitos “judge” e “custody” na sentença “The judge granted the custody of the child to his grandmother”. A regra de sentença pode ser formalizada pela expressão regular parametrizada (24) no padrão PCRE (“Perl Compatible Regular Expressions”) [39].



(24)

O parâmetro L_c é uma *string* que corresponde aos conceitos da ontologia sobre os quais deseja-se aprender os relacionamentos não taxonômicos separados pelo operador de disjunção do padrão PCRE [39]. O parâmetro L_c pode ser formalmente definido como a seguinte concatenação: $L_c = c_1$ “|” c_2 “|” ... “|” c_n ; $\forall\, c_i \in C$.

As subexpressões regulares “G_1” e “G_3” realizam o casamento e extração dos conceitos da ontologia respectivamente nas extremidades esquerda e direita do texto casado com a expressão regular completa (24). A subexpressão “G_2” verifica se entre os dois conceitos retornados por “G_1” e “G_3” não há um conceito ou uma das strings “’s” ou “’”. Para cada casamento da expressão regular (19) no corpus, um relacionamento candidato na forma “<c_1, c_2>” (seção 2.4) é gerado. A tabela 22 apresenta sentenças do corpus *Family law doctrine* (seção 5.4) e as correspondentes tuplas extraídas pela regra de sentença.

Sentenças	Tuplas extraídas
The court decree protects the property rights of the parties and provides support for the children.	< court, decree> <decree, property> < property, party>
An absolute divorce dissolves the marriage.	<divorce, marriage>
Either of the parties of the marriage or a third person can bring an action to declare it void at any time.	<parties, marriage>

Tabela 22: Exemplos de relacionamentos candidatos extraídos pela Regra de sentença

4.2.3.2.Regra de sentença com frase verbal

A regra de sentença com frase verbal [61] é baseada na heurística de que dois conceitos consecutivos na mesma sentença com uma frase verbal entre eles estão provavelmente relacionados de forma não taxonômica, sendo a frase verbal o provável rótulo do relacionamento. A regra de sentença com frase verbal pode ser formalizada pela expressão regular parametrizada (25) no padrão PCRE ("Perl Compatible Regular Expressions") [39].

$$\underbrace{(L_C)}_{G_1}(?:(?!L_C|'s?).)*\underbrace{(V)}_{G_2}(?:(?!L_C\ 's?).)*\underbrace{(?=(L_C))}_{G_3} \quad (25)$$

O parâmetro L_c é como definido na seção 4.2.3.1. As subexpressões regulares "G_1" e "G_3" realizam o casamento e extração dos conceitos da ontologia respectivamente nas extremidades esquerda e direita do texto casado com a expressão regular completa (25).

O parâmetro "V" é uma *string* que corresponde às frases verbais presentes no documento do qual deseja-se extrair relacionamentos candidatos, separadas pelo operador de disjunção do padrão PCRE [39]. O parâmetro "V" pode ser formalmente definido como a seguinte concatenação: V = v_1 "|" v_2 "|" ... "|" v_n ; $\forall$ $v_i \in$ F. "F" é o conjunto de frases verbais presentes no documento avaliado.

A subexpressão "G_2" retorna o conjunto de frases verbais (V_{G2}) que houverem na *string* "s" existente entre os dois conceitos retornados por "G_1" e

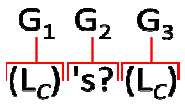
“G_3”, desde que essa não possua “‘s”, “‘” ou um conceito da ontologia como substrings. Formalmente, $V_{G2} = \{v_i \mid substring(s, v_i) \wedge \sim substring(s, ("'s" \vee "'" \vee (c_i \in C)\}$. Para cada casamento da expressão regular (19) no corpus, um relacionamento candidato na forma $<c_1, fv, c_2>$ ($c_1, c_2 \in C$ e $fv \in F$) é gerado. A tabela 23 apresenta sentenças do corpus *Family law doctrine* (seção 5.4) e as correspondentes tuplas extraídas pela regra de sentença com frase verbal.

Sentença	Tuplas extraídas
The court decree protects the property rights of the parties and provides support for the children.	<decree, protect, property>
An absolute divorce dissolves the marriage.	<divorce, dissolves, marriage>
Either of the parties of the marriage or a third person can bring an action to declare it void at any time.	Nenhuma extração

Tabela 23: Exemplos de relacionamentos candidatos extraídos pela Regra de sentença com frase verbal

4.2.3.3.Regra de Apóstrofo

A regra de apóstrofo [61] é baseada na heurística de que dois conceitos consecutivos com as *strings* “‘s” ou “‘” entre eles tem grande probabilidade de estarem relacionados de forma não-taxonômica. A regra de apóstrofo pode ser formalizada pela expressão regular parametrizada (26) no padrão PCRE (Perl Compatible Regular Expressions) [39].



(26)

O parâmetro L_c é como definido na seção 4.2.3.1. As subexpressões regulares “G_1” e “G_3” realizam o casamento e extração dos conceitos da ontologia respectivamente nas extremidades esquerda e direita do texto casado com a expressão regular completa (26). A subexpressão “G_2” verifica se entre os dois conceitos retornados por “G_1” e “G_3” há exclusivamente uma das strings “’s” ou “’”. Para cada casamento da expressão regular (25) no corpus, um relacionamento candidato na forma “$<c_1, c_2>$” (seção 2.4) é gerado. A

tabela 24 apresenta sentenças do corpus *Family law doctrine* e as correspondentes tuplas extraídas pela regra de apóstrofo.

Sentença	Tuplas extraídas
While the court will generally honor the parties' agreements as set forth in the separation agreement, the court may modify provisions affecting the care, custody, education, maintenance and support of the children in order to protect their best interests.	<party, agreement>
Do not rely on the advice of your spouse's lawyer.	<spouse, lawyer >
The court can order one party to pay the fee for the other party's lawyer and for all costs closely related to bringing or enforcing an action for divorce.	<party, lawyer>

Tabela 24: Exemplos de relacionamentos candidatos extraídos pela Regra de apóstrofo

4.2.4.Refinamento

TARNT disponibiliza duas soluções para a fase de refinamento: "Frequência de co-ocorrência" [61] e *Bag of labels* [61]. A "Frequência de co-ocorrência" [61] é utilizada para filtrar tanto os relacionamentos candidatos resultantes da extração feita com a regra de sentença quanto pela regra de apóstrofo. Já *Bag of labels* [61] é utilizada para filtrar os relacionamentos candidatos obtidos da extração realizada com a regra de sentença com frase verbal. As seções seguintes detalham cada uma dessas soluções utilizadas na fase de "Refinamento".

4.2.4.1.Frequência de co-ocorrência

Esse algoritmo calcula a frequência normalizada dos relacionamentos candidatos representados na forma "<c_1, c_2>" (seção 2.4) extraídos tanto pela regra de sentença quanto pela regra de apóstrofo [61]. O parâmetro "frequência mínima" serve como critério de corte dos relacionamentos recomendados por essa solução de refinamento e seu valor é definido experimentalmente.

A figura 15 apresenta o algoritmo "Frequência de co-ocorrência" [61]. Na linha um são informados ao algoritmo os relacionamentos candidatos

(“*setCandidateRel*”) e o parâmetro de corte frequência mínima (“*freqM*”). Na linha dois, à variável “*qtdCandidateRel*” é atribuída a quantidade de relacionamentos candidatos (“*count(setCandidateRel)”*). A linha três corresponde a um laço que percorre cada um dos elementos (“*candidateRel*”) do conjunto de relacionamentos candidatos (“*setCandidateRel*”). Na linha quatro, cada relacionamento candidato (“*candidateRel*”) tem a frequência normalizada (“*freq*”) calculada. A frequência normalizada para determinado par de conceitos (formato “<c_1, c_2>”) é dada pelo quociente entre o número de ocorrências do relacionamento candidato (“*countOccurences(candidateRel, setCandidateRel)*”) e a quantidade de relacionamentos candidatos (“*qtdCandidateRel*”). Das linhas cinco a nove, são armazenados na matriz “*refinedRel*” os relacionamentos (“*candidateRel*”) e correspondentes frequências (“*freq*”) que possuírem frequência maior ou igual a mínima (“*freqM*”). Os que não satisfizerem essa condição tem todas as suas ocorrências excluídas de “*setCandidateRel*”. Por fim, “*refinedRel*” contendo relacionamentos e correspondentes frequências é retornado. Uma implementação desse algoritmo encontra-se na ferramenta TARNTool que suporta a aplicação de TARNT [60] [61] e é descrita na seção 4.4.

```
1. FrequencyOfCoocurrence(setCandidateRel, freqM)
2. qtdCandidateRel := count(setCandidateRel)
3. foreach setCandidateRel as candidateRel do
4.     freq:=countOccurrences(candidateRel,setCandidateRel)
       /qtdCandidateRel
5.     if(freq >= freqM)
6.       refinedRel['setCandidateRel'][] := candidateRel
7.       refinedRel['frequency'][] := freq
8.     else
9.       deleteAllOccurrences(candidateRel,
         setCandidateRel)
10. endForeach
11. return refinedRel
```

Figura 15: O algoritmo “Frequência de co-ocorrência” [61]

4.2.4.2.Bag of labels

O algoritmo *Bag of labels* [61] tem como *input* relacionamentos candidatos do tipo "<c_1, fv, c_2>" obtidos a partir das extrações feitas pela "Regra de sentença com frase verbal" (seção 4.2.3.2) na fase "Extração de relacionamentos".

Bag of labels [61] calcula a frequência normalizada dos relacionamentos candidatos verificando a coincidência apenas dos dois conceitos; a coincidência da frase verbal "fv" não é verificada. A cada nova coincidência do par de conceitos, sua frequência é incrementada e a frase verbal "fv" é acrescentada a um repositório de rótulos (*bag of labels*) associado a esse par de conceitos. Por esse motivo, apesar do algoritmo *Bag of labels* [61] receber relacionamentos candidatos do tipo "<c_1, fv, c_2>" como *input,* para efeito do refinamento, a representação utilizada é "<c_1, c_2>". As frases verbais (rótulos dos relacionamentos) são utilizadas apenas como uma informação adicional associada ao par de conceitos, mas que não é considerada para efeito dos cálculos do parâmetro de corte (frequência de co-ocorrência).

Bag of labels [61] é uma solução considerada de uso simples, uma vez que o especialista tem que informar o valor de apenas um parâmetro de corte (frequência mínima) e não dois (suporte e confiança mínimos) como no caso do algoritmo de "Extração de regras de associação" [1] [2] ou como no de "Extração de regras de associação generalizadas" [67].

A figura 16 apresenta o pseudocódigo do algoritmo *Bag of labels* [61]. Na linha um são informados ao algoritmo os relacionamentos candidatos ("*setCandidateRel*") e o parâmetro de corte frequência mínima ("*freqM*"). A linha dois corresponde a um laço que percorre cada um dos elementos ("*candidateRel*") do conjunto de relacionamentos candidatos ("*setCandidateRel*"). Na linha três, o vetor bidimensional "*setRelBag*" é criado. Na primeira dimensão estão todos os pares de conceitos diferentes, já na segunda dimensão estão todos os rótulos verbais associados a cada par. Os rótulos são obtidos a partir dos relacionamentos candidatos que possuem o par de conceitos em questão.

A linha cinco corresponde a um laço que percorre cada um dos elementos “*relBag*” (par de conceitos e seus correspondentes rótulos) do conjunto de relacionamentos candidatos “*setRelBag*” (todos os pares de conceitos e seus respectivos rótulos). Na linha seis, cada “*relBag*”, elemento do vetor “*setRelBag*”, tem a frequência normalizada (“*freq*”) calculada. A frequência normalizada, para determinado “*relBag*”, é dada pelo quociente entre o número de rótulos verbais presentes no “*relBag*” e a quantidade de rótulos no “*setRelBag*”.

Das linhas sete a onze, todos os “*relBag*” que possuírem frequências (“*freq*”) inferiores a frequência mínima informada ao algoritmo *Bag of labels* [61] como argumento (“freqM”) são excluídas do “setRelBag”. Por fim, o “*setRelBag*” contendo cada par de conceitos, respectivos conjuntos de rótulos e frequências é retornado. Uma implementação desse algoritmo encontra-se na ferramenta TARNTool que provê suporte a aplicação de TARNT e é descrita na seção 4.4.

```
1.  BagOfLabels(setCandidateRel, freqM)
2.  foreach setCandidateRel as candidateRel do
3.      setRelBag[candidateRel]['labels'][] := candidateRel
        ['label'];
4.  EndForeach
5.  foreach setRelBag as relBag do
6.      freq := count(returnLabels(relBag)) /
        count(returnLabels(setRelBag))
7.      if(freq >= freqM)
8.          setRelBag[returnPair(relBag)]['freq'] := freq
9.      else
10.         delete(relBag,setRelBag)
11.     endIf
12. endForeach
13. return setRelBag
```

Figura 16: O algoritmo “*Bag of labels*” [61]

4.2.5.Avaliação pelo especialista

Nenhuma técnica de PLN [4] [8] [19] [22] [23] [68] ou de Aprendizagem de Máquina [9] [46] [49] [50] substitui a decisão do especialista em um ambiente ruidoso como aprender a partir de fontes de informação em linguagem natural [58]. Portanto, o objetivo desta fase é permitir ao especialista selecionar e possivelmente editar relacionamentos dentre os relacionamentos extraídos pelas fases anteriores. Assim, o resultado da técnica deve ser avaliado por um especialista antes que as relações sejam definitivamente adicionadas à ontologia. Questões como o escopo do conhecimento a ser representado, o nível de generalização, a real necessidade da adição de um relacionamento, sua direção e rótulo devem ser avaliados, selecionados e possivelmente ajustados por um especialista.

4.2.6.Atualização da Ontologia

Essa etapa consiste em executar um procedimento de atualização do arquivo da ontologia com os relacionamentos não-taxonômicos selecionados e possivelmente editados na fase anterior. O procedimento a ser executado é fortemente determinado pelo tipo de arquivo e representação da ontologia. A disponibilização dessa funcionalidade por TARNTool (seção 4.4) é sugerida como trabalho futuro.

4.3.Configurações de TARNT

O principal objetivo de uma técnica de ARNT é sugerir relacionamentos não-taxonômicos com o máximo de efetividade. Formalmente, a efetividade pode ser mensurada em termos de medidas de avaliação. Exemplos são *recall*, precisão e medida-F (seção 2.3). O cumprimento deste requisito exige que tanto a fonte de informação quanto a técnica de ARNT sejam adequados a realização da tarefa de aprendizagem.

Para obter boa efetividade, a técnica de ARNT precisa fazer uso de soluções que permitam a extração e posterior refinamento de relacionamentos de forma eficaz. Entretanto, os relacionamentos podem estar presentes no corpus com diferentes realizações textuais, o que pode exigir o uso de

diferentes técnicas para sua obtenção. Por exemplo, há relacionamentos representados por dois conceitos sem frases verbais entre eles. Um caso concreto é o do relacionamento "<decree, protect, property>" entre os conceitos "decree" e "property" que pode ser extraído a partir da sentença: "The court decree protects the property rights of the parties and provides support for the children". Há também relacionamentos representados por dois conceitos sem frases verbais. Um caso concreto é do relacionamento "<court, decree>" que pode ser extraído a partir da sentença: "The court decree protects the property rights of the parties and provides support for the children". Há ainda relacionamentos representados pelo apóstrofo (forma possesiva). Esse é o caso de "<party, lawyer>" que pode ser extraído a partir da sentença "The court can order one party to pay the fee for the other party's lawyer and for all costs closely related to bringing or enforcing an action for divorce". Dessa forma, a efetividade da técnica de ARNT depende de sua capacidade de, por meio do uso de soluções adequadas, extrair do texto e recomendar relacionamentos com a obtenção dos melhores valores para uma medida de avaliação considerada. Por esse motivo, TARNT disponibiliza algumas configurações conforme apresentado na tabela 25.

<table>
<tr><th colspan="2">Anotação do corpus</th><th>Extração de relacionamentos</th><th>Refinamento</th></tr>
<tr><td>Análise morfo-lexical e Chunk</td><td rowspan="2">Lematização</td><td>Regra de sentença com frase verbal</td><td>Bag of labels</td></tr>
<tr><td>Tokenização e Separação de sentença</td><td>Regra de sentença ou Regra de apóstrofo</td><td>Frequência de co-ocorrência</td></tr>
</table>

Tabela 25: Configurações de TARNT

Na fase de "Anotação do corpus" quatro soluções estão disponíveis: A primeira consiste nas técnicas de tokenização e separação de sentenças. Essa configuração permite a aplicação da "Regra de Sentença" (seção 4.2.3.1) e da "Regra de apóstrofo" (seção 4.2.3.3) na fase de "Extração de relacionamentos". A segunda configuração consiste nas técnicas anteriores acrescidas da análise morfo-lexical e VP chunk. Essa configuração permite a

aplicação da “Regra de Sentença com Frase Verbal” (seção 4.2.3.2). Cada uma dessas duas configurações pode ainda ser executada incluindo a lematização.

Na fase “Extração de relacionamentos” três regras de extração são disponibilizadas: a “Regra de Sentença”, a “Regra de Sentença com Frase Verbal”; e a “Regra de Apóstrofo” (seção 4.2.3). Estas regras devem ser utilizadas de forma a maximizar a medida de avaliação dependendo do tipo de realização textual predominante dos relacionamentos de referência no corpus conforme discutido anteriormente nesta seção.

Na fase de “Refinamento” duas soluções são disponibilizadas: *Bag of labels* [61] e “Frequencia de co-ocorrência” [61]. *Bag of labels* deve ser utilizada na situação na qual rótulos são recomandados ao especialista, ou seja, refina extrações realizadas pela “Regra de sentença com frase verbal”. O algoritmo “Frequência de co-ocorrência” é utilizado caso os rótulos dos relacionamentos não sejam recomendados ao especialista. Essa solução realiza o refinamento dos relacionamentos candidatos extraídos pela “Regra de sentença” ou pela “Regra de apóstrofo”. Em ambas as soluções, a alteração da medida de avaliação é realizada por meio do ajuste experimental dos seus parâmetros de refinamento (frequência mínima).

4.4.A ferramenta de software TARNTool

TARNTool é uma ferramenta de software que provê suporte a aplicação de TARNT [60] [61] e foi totalmente desenvolvida em Java. Para a fase de “Anotação do corpus” TARNTool utiliza o GATE (General Architecture for Text Engineering – Arquitetura Genérica para Engenharia de Texto) [20] [21], uma API de PLN desenvolvida na Universidade de Sheffield [20]. A seção 4.4.1 apresenta uma visão geral do GATE [20] [21], além de destacar seus recursos utilizados por TARNTool. Na seção 4.4.2 é apresentada a arquitetura de TARNTool.

4.4.1.Recursos do GATE utilizados em TARTool

O GATE (General Architecture for Text Engineering – Arquitetura Genérica para Engenharia de Texto) [20] [21] é uma infraestrutura para desenvolvimento e implantação de componentes de software que processam a linguagem natural [21]. O GATE foi desenvolvido na linguagem Java pela Universidade de Sheffield. A arquitetura do GATE é baseada em componentes conhecidos como "*Resources*" (recursos). Os recursos do GATE são diferenciados em três categorias:

a) Recursos de Linguagem (*Language Resources* – LR): são entidades linguísticas, tais como documentos e corpora.

b) Recursos de Processamento (*Processing Resources* – PR): são entidades algorítmicas que efetuam algum processamento. Exemplos são: analisadores sintáticos (parsers).

c) Recursos Visuais (*Visual Resources* – VR): são componentes da Interface Gráfica do Usuário (*Graphical User Interface* – GUI), que permitem a visualização ou edição de outros componentes.

A correspondência entre os Recursos de Processamento (*Processing Resources* – PR) do GATE e as tarefas de PLN utilizadas por TARNTool é mostrada na tabela 26.

Tarefas de PLN presentes em TARNT	**Recurso de processamento do GATE**
Tokenização	*ANNIE English Tokeniser*
Divisão em Sentenças	*Sentence Splitter*
Análise morfológica	*ANNIE POS Tagger*
Lematização	*GATE Morphological Analyser*
VPchunking	*ANNIE VP chunker*

Tabela 26: Recursos de processamento do GATE utilizados por TARNTool

4.4.2.Arquitetura de TARNTool

A figura 17 apresenta o diagrama de sequência para TARNTool. A classe "*Application*" (Aplicação) é a que gerencia todo o processo. Inicialmente ela informa o arquivo no formato "OWL" da ontologia à classe "*ClassesExtractor*" (Extrator de classes) que retorna todas as classes da ontologia. Em seguida a classe "*Application*" fornece essas classes à classe "*ClassesLemmatisation*" (Lematização de classes) e tem como retorno essas classes lematizadas.

A classe "*Application*" fornece o corpus (arquivos txt) e a solução de PLN que deseja que seja executada e a classe "*Gate*" retorna o corpus (arquivos txt) anotado. TARNT permite a escolha dentre diferentes soluções de PLN conforme descrito na seção 4.2.2. A classe "*Application*" então informa à classe "*RelExtractor*" (Extrator de relacionamentos) o corpus anotado, as classes originais (não lematizadas), as classes lematizadas e a regra de extração a ser executada e tem como retorno os relacionamentos candidatos. Em seguida, a classe "*Application*" informa a classe "*Refinement*" (Refinamento) que execute a solução de refinamento selecionada sobre os relacionamentos candidatos. A classe "*Refinement*" então retorna os relacionamentos a serem sugeridos ao especialista.

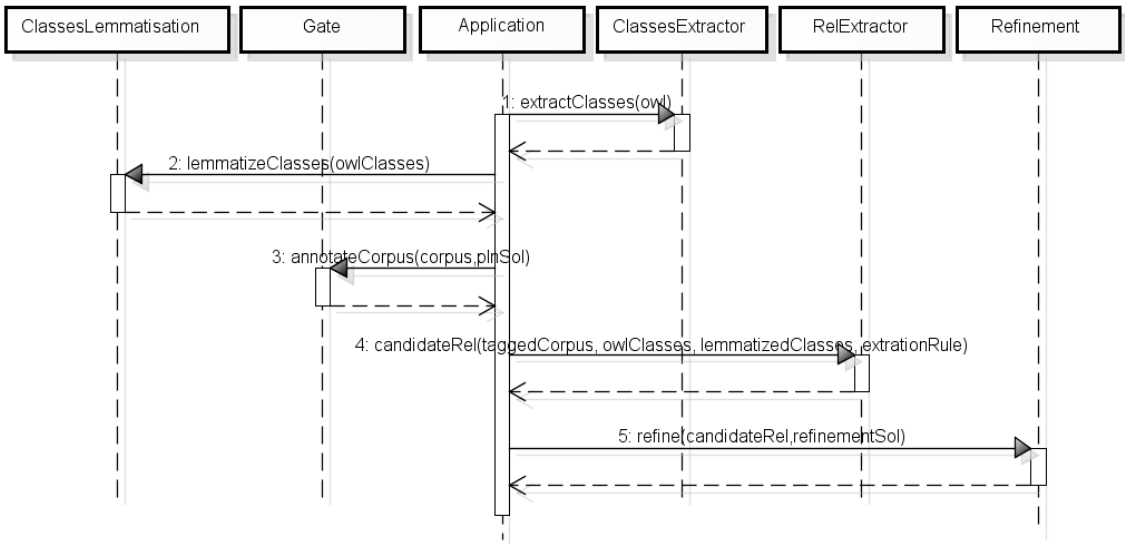


Figura 17: Diagrama de sequência de TARNTool

A tabela 27 apresenta a correspondência entre as classes de TARNTool e as fases do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3). A figura 18 apresenta o código da classe "*Application*", responsável por gerenciar todo o processo de ARNT realizado por TARTool. Seu construtor recebe os seguintes argumentos: O arquivo "OWL" da ontologia da qual são obtidas as classes ("owl"). O corpus do domínio do qual são aprendidos os relacionamentos ("*corpus*"). Uma *string* que informa a regra de extração a ser excutada e assume um dentre os três valores: "SR" (*Sentence Rule*), "AR" (*Apostrophe Rule*) ou "SRVP" (*Sentence Rule with Verb Phrase*). A solução de refinamento a ser executada, "*Bag of labels*" ou "Frequencia de co-ocorrência" e o valor do parâmetro de corte frequencia mínima ("*FreqM*"), como o objeto "*refinementSol*" (Solução de refinamento) e também as técnicas de PLN a serem executadas como o objeto "*nlpSol"* (solução de PLN). O método "*execApplication*" é o que de fato executa a lógica da aplicação e retorna os relacionamentos a serem recomendados ao especialista na forma do objeto "*recommendedRel*".

<table>
<tr><th colspan="2">Fases de TARNT</th><th>Classes de TARNTool</th></tr>
<tr><td colspan="2">Construção do corpus</td><td>-</td></tr>
<tr><td rowspan="2">Extração de Relacionamentos Candidatos</td><td>Anotação do corpus</td><td>Gate, ClassesLemmatisation</td></tr>
<tr><td>Extração de Relacionamentos</td><td>RelExtractor, ClassesExtractor</td></tr>
<tr><td colspan="2">Refinamento</td><td>Refinement</td></tr>
<tr><td colspan="2">Avaliação pelo especialista</td><td>-</td></tr>
<tr><td colspan="2">Atualização da ontologia</td><td>-</td></tr>
</table>

Tabela 27: Fases do processo de ARNT e correspondentes classes de TARNTool

```
class Application {
      private OWL owl;
      private Corpus corpus;
      private String extractionRule;
      private RefinementSol refinementSol;
      private NLPSol nlpSol;
      Application(OWL owl, Corpus corpus, String extractionRule,
      RefinementSol refinementSol, NLPSol nlpSol) {
          this.owl = owl;
          this.corpus = corpus;
          this.extractionRule = extractionRule;
          this.refinementSol = refinementSol;
          this.nlpSol = nlpSol;
}
      public RecommendedRel execApplication() {
          ClassesExtractor classesExtractor = new ClassesExtractor();
          OWLClasses owlClasses = classesExtractor.extractClasses
          (this.owl);
          ClassesLemmatisation classesLemmatisation = new
          ClassesLemmatisation();
          LemmatizedClasses lemmatizedClasses =
          classesLemmatisation.lemmatizeClasses(owlClasses);
          Gate gate = new Gate();
          TaggedCorpus taggedCorpus = gate.annotateCorpus
          (this.corpus,this.nlpSol);
          RelExtractor relExtractor = new RelExtractor();
          CandidateRel candidateRel = relExtractor.candidateRel
          (taggedCorpus,owlClasses,lemmatizedClasse,this.extractionRule);
          Refinement refinement = new Refinement();
          RecommendedRel recommendedRel = refinement.refine
          (candidateRel,this.refinementSol);
          return recommendedRel
}
}
```

Figura 18: A classe "*Application*" de TARNTool

4.5.Avaliação qualitativa de TARNT

TARNT foi desenvolvida levando em consideração a premissa de que a fonte de informação textual, por ser ruidosa, exige uma solução

customizável para a obtenção dos melhores resultados na aprendizagem de relacionamentos não-taxonômicos.

As técnicas apresentadas no capítulo três não possuem o grau de customização apresentado por TARNT. Por exemplo, a "ARNT baseada na extração de regras de associação generalizadas" [42] utiliza duas regras de extração que são executadas simultaneamente, a regra de sentença e a regra de título, entretanto essa última limita os textos sobre os quais a técnica pode ser aplicada a aqueles que possuem títulos e corpo de texto. A "ARNT baseada na extração de regras de associação" [69] e a "ARNT baseada em consultas a Web" [55] [56] utilizam uma única regra de extração de relacionamentos, sendo a de sentença e de sentença com frase verbal respectivamente.

TARNT disponibiliza na fase de "Anotação do corpus" quatro configurações possíveis. A primeira consiste nas técnicas de tokenização e separação de sentenças. Essa configuração permite a aplicação da "Regra de Sentença" [61] e da "Regra de apóstrofo" [61] na fase de "Extração de relacionamentos". A segunda configuração consiste nas técnicas anteriores acrescidas de análise morfo-lexical e *VP chunk*. Essa configuração permite a aplicação da "Regra de sentença com frase verbal" [61]. Cada uma dessas duas configurações pode ainda ser executada incluindo a lematização. A utilização dessa técnica tem o potencial de aumentar o *recall* da fase "Extração de relacionamentos candidatos" e também consequentemente do resultado obtido pela técnica de ARNT como um todo. A contrapartida é que há também o potencial de diminuição da precisão. A decisão sobre o uso desse ou de outros recursos é tomada experimentalmente para cada situação específica.

Na fase de "Extração inicial de relacionamentos", três regras de extração estão disponíveis: a "Regra de sentença" [61] que é baseada na heurística de que dois conceitos consecutivos em uma mesma sentença estão provavelmente relacionados de forma não-taxonômica, a "Regra de sentença com frase verbal" [61], baseada na heurística de que dois conceitos consecutivos na mesma sentença com uma frase verbal entre eles estão provavelmente relacionados de forma não-taxonômica, sendo a frase verbal o provável rótulo do relacionamento; e por fim, a "Regra de apóstrofo" [61], baseada na heurística de que dois conceitos consecutivos com as *strings* "'s"

ou “” entre eles tem grande probabilidade de estarem relacionados de forma não-taxonômica. Essas regras de extração permitem ao especialista maximizar a quantidade de relacionamentos candidatos obtidos do corpus.

Na fase de “Refinamento” todas as técnicas avaliadas no capítulo três disponibilizam uma única solução. A solução adotada por Maedech e Staab [42], “Extração de regras de associação generalizadas” [67], faz a sugestão do melhor nível hierárquico no qual cada relacionamento deva ser incluído, entretanto não permite à técnica recomendar rótulos aos relacionamentos, funcionalidade que é normalmente mais útil.

Villaverde et al. [69] utiliza o algoritmo de “Extração de regras de associação” [1] [2], uma solução que além de recomandar pares de conceitos relacionados também sugere um rótulo. As desvantagens nessa abordagem são principalmente duas: Primeiro, esse algoritmo atribui maior relevância a relacionamentos representados por regras que possuem maior valor de confiança (ver seção 2.6.2.1), medida que expressa o quanto um rótulo está ssociado a um par de conceitos. Entretanto, regras de associação que possuem os maiores valores de confiança podem conter pares de conceitos que apresentam baixa ocorrência no corpus, medida que é normalmente mais representativa da relevância de um relacionamento recomendado pela técnica de refinamento. A segunda desvantagem é que relacionamentos candidatos extraídos pela regra de sentença (em uma possível extensão da técnica) não podem ser processados por esse algoritmo.

A solução “Regressão logística” [46] utilizada por Fader et al. [29] recomenda relacionamentos compostos por frases nominais e frases verbais utilizadas como rótulos. Ela atribui probabilidades aos relacionamentos recomendados baseado em um conjunto de heurísticas como as observadas na tabela 16 (seção 3.4), o que lhe permite, diferentemente dos algoritmos de “Extração de regras de associação” [1] [2], “Extração de regras de associação generalizadas” [67], *Bag of labels* [61] e Frequência de co-ocorrência [61], manter sua acurácia independentemente do tamanho do corpus. O aspecto negativo é que ao utilizar frases nominais extraídas do corpus como conceitos, por não dispor desses a partir de uma ontologia, os valores para as medidas de avaliação (ex.: *recall*, precisão e medida-F) tendem a ser mais baixos,

principalmente levando-se em consideração procedimentos de avaliação automatizados, que são os mais viáveis.

Para a fase de refinamento TARNT disponibiliza duas soluções: *Bag of labels* [61] e “Frequencia de co-ocorrência” [61]. *Bag of labels* [61] deve ser utilizada na situação na qual rótulos são recomandados ao especialista, ou seja, refina extrações realizadas pela “Regra de sentença com frase verbal” [61]. Uma das grandes vantagens desse algoritmo em relação a outros que além de receber os conceitos da ontologia também recomandam rótulos é que apesar de realizar essa recomendação, esses não são de fato utilizados na realização do processo de refinamento, ficando apenas como uma informação adicional dada ao especialista junto ao par de conceitos recomendado. Os rótulos são armazenados nos *bag of labels* associados a cada par de conceitos e para apenas esses os cálculos de refinamento são realizados.

O algoritmo “Frequência de co-ocorrência” [61] é utilizado caso os rótulos dos relacionamentos não sejam recomendados ao especialista. Essa solução realiza o refinamento dos relacionamentos candidatos extraídos pela “Regra de sentença” [61] ou pela “Regra de apóstrofo” [61].

Levando-se em consideração todos os aspectos aqui discutidos pode-se afirmar que TARNT [60] [61] é uma técnica customizável que incorpora soluções adequadas para cada uma das fases do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3) e que, além disso, é capaz de apresentar bons resultados na aprendizagem de relacionamentos não-taxonômicos, como inclusive demonstrado pelos experimentos apresentados no capítulo 5 desta Tese. A tabela 28 apresenta de forma resumida alguns dos principais aspectos positivos e limitações das técnicas de ARNT discutidas nesse trabalho [29] [42] [48] [55] [56] [69], incluindo TARNT.

Técnicas de ARNT	Aspectos positivos	Limitações
3.1 [69].	• Sugere rótulos aos relacionamentos. • Nenhuma característica específica é exigida do corpus, como a feita por Maedech e Staab [42] de possuir títulos e corpo de texto.	• Dificuldade em determinar experimentalmente os valores para os parâmetros de corte suporte e confiança minimos. • Valorização da informação de quanto um rótulo está associado a um par de conceitos em prejuízo de quais pares estão relacionados.
3.2 [42].	• Alto *recall* na fase de "Extração de relacionamentos candidatos" devido às regras de extração utilizadas. • Sugere o possível melhor nível na taxonomia onde os relacionamentos devam ser acrescentados.	• Não sugere rótulos aos relacionamentos. • Aplicável somente sobre corpora que apresente marcação que distinga corpo de texto de título.
3.3 [55].	• Solução automatizada para a fase "Construção do corpus". • Sugere rótulos aos relacionamentos.	• Realiza a ARNT e aprendizagem de conceitos conjuntamente, ignorando a situação realista na qual esses últimos são conhecidos à priori. • Conceitos são representados por frases nominais, que normalmente não são nomes de fato utilizados para essa finalidade.
3.4 [29].	• Sugere rótulos aos relacionamentos. • Mantêm a acurácia independentemente do tamanho do corpus.	• Utiliza frases nominais extraídas do corpus como conceitos, as quais normalmente não corespondem a nomes de classes utilizados na prática. • Utiliza um repositório de frases verbais, construído manualmente para melhorar a precisão.
3.5 [48].	• Sugere rótulos aos relacionamentos. • Sugere novos relacionamentos não-taxonômicos a partir de outros previamente conhecidos	• Dependencia em relação REN para a identificação das instancias de classes da ontologia nas sentenças. • Necessidade de dispor de uma ontologia povoada.
TARNT [60] [61]	• Maior customização em relação a outras técnicas, o que lhe confere o potencial de obtenção de melhores resultados. • Recomenda rótulos sem perder efetividade na identifivação de pares de conceitos mais relevantes (relacionamentos mais prováveis), diferentemente do que ocorre com técnicas que recomendam relacionamentos do tipo "<c_1, fv, c_2>".	• Não sugere o nível taxonômico no qual o relacionamento deva ser acrescentado. • Necessidade de quantidade substancial de texto no corpus devido às técnicas estatísticas que utiliza.

Tabela 28: Aspectos positivos e limitações de técnicas de ARNT

4.6.Considerações finais

Nesse capítulo foi apresentado TARNT [60] [61], uma técnica semiautomática para aprendizagem de relacionamentos não-taxonômicos de ontologias que está em conformidade com o processo genérico de ARNT proposto no presente trabalho [57] [58] [59] [60] (seção 2.3).

TARNT apresenta algumas vantagens em relação às demais técnicas de ARNT discutidas no capítulo 3. Dentre elas pode-se citar o controle sobre a execução de suas regras de extração, que permitem ao especialista configurar a fase de "Extração de relacionamentos" de forma a obter os melhores resultados em termos de *recall,* precisão e medida-F com relação a um corpus específico; a "Regra de apóstrofo" que permite dar tratamento diferenciado a relacionamentos candidatos que possuem maior probabilidade de serem reais relacionamentos não-taxonômicos; e a solução de refinamento *Bag of labels* [61], algoritmo que recomenda relacionamentos não-taxonômicos indicando quais classes estão relacionadas além de sugerir rótulos a cada relacionamento. Foi também apresentada TARNTool, uma ferramenta de software que dá suporte a aplicação de TARNT e que utiliza a API GATE na fase de "Anotação do corpus".

No próximo capítulo, com o intuito de verificar a hipótese de pesquisa postulada nesse trabalho (seção 1.3), TARNT e, em particular, a solução de refinamento *Bag of labels* são avaliadas comparativamente a outra abordagem que utiliza o algoritmo de "Extração de regras de associação" [1] [2] na fase de "Refinamento". Esse algoritmo teve o uso sugerido pela técnica "ARNT baseada na extração de regras de associação" [69]. Os resultados obtidos são posteriormente generalizados para algoritmos que recomendam relacionamentos dos mesmos tipos desses. Para tornar possível a avaliação foram ainda formalmente definidos dois procedimentos baseados na comparação dos relacionamentos aprendidos com os de uma ontologia de referência.

5. Avaliação das soluções para a fase de refinamento: *Bag of Labels* e Extração de regras de associação

O objetivo desse capítulo é demonstrar a hipótese postulada nesse trabalho (seção 1.3) de que soluções para a fase de refinamento do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3) que recomendam relacionamentos do tipo "<c_1, fv, c_2>" são menos efetivas que as que recomendam relacionamentos "<c_1, c_2>", uma vez que tendem a apresentar menores valores para as medidas de avaliação, quando considerados o mesmo corpus e ontologia de referência. Para tanto, experimentos de aprendizagem de relacionamentos foram realizados com a ferramenta de software TARNTool (seção 4.4) que provê suporte a TARNT [60] [61] e AERA-Tool, apresentada na seção 5.2 e desenvolvida para permitir a avaliação comparativa dos resultados obtidos por TARNT. TARNT [60] [61] e AERA (Aprendizagem de relacionamentos com Extração de Regras de Associação) foram configuradas com a mesma solução para a fase de "Extração de relacionamentos candidatos" (regra de sentença com frase verbal [61]) e respectivamente com as soluções *Bag of labels* [61] e "Extração de regras de associação" [1] [2] para a fase de "Refinamento". Os resultados obtidos foram avaliados e generalizados para quaisquer algoritmos que utilizem relacionamentos dos dois tipos considerados na hipótese de pesquisa (seção 1.3).

O capítulo está organizado da seguinte forma: A seção 5.1 discorre sobre a avaliação de técnicas de Aprendizagem de ontologias, resaltando qual das soluções propostas foi adotada no presente trabalho. A seção 5.2 apresenta a abordagem AERA, desenvolvida para realizar avaliações comparativas entre as duas soluções de refinamento, *Bag of labels* [61] e "Extração de regras de associação" [1] [2]. As seções 5.3 e 5.4 apresentam os corpora e correspondentes ontologias utilizados nos experimentos. Na seção 5.5 são apresentados os procedimentos de avaliação de técnicas de ARNT utilizados. As seções 5.6 a 5.9 apresentam e discutem os resultados das avaliações realizadas com TARNT e AERA. Na seção 5.10, por meio de uma exposição formal que utiliza esses resultados, a hipótese de pesquisa é verificada. Por fim, a seção 5.11 faz as considerações finais sobre o capítulo.

5.1.Avaliação de técnicas de Aprendizagem de Ontologias

Comparar técnicas para a aprendizagem de ontologias não é uma tarefa trivial. Para um dado domínio não há uma possibilidade única de conceitualização [66] e cada uma das técnicas existentes podem ser mais ou menos úteis para determinadas tarefas e ainda assim serem justificáveis [25].

Apesar da avaliação das técnicas de aprendizagem de ontologias ainda ser um problema em aberto, já existem trabalhos nessa direção. De acordo com Dellschaft e Staab [25], as ontologias resultantes podem ser comparadas avaliando-as em uma aplicação executável, por avaliação a posteriori por especialistas ou pela comparação dos resultados aprendidos com uma ontologia de referência pré-definida ("*goldstandard*").

A comparação de ontologias em uma aplicação executável visa medir a efetividade de um sistema que utiliza as ontologias sendo avaliadas. Uma desvantagem dessa abordagem é que outros fatores podem ter impacto na saída do sistema e, algumas vezes a ontologia é, de fato, uma parte do sistema que pouco interfere em seus resultados [25].

A avaliação manual tem suas vantagens, uma vez que se espera que os especialistas conheçam os conceitos e relacionamentos dos seus domínios de atuação e, portanto, eles são supostamente capazes de dizer se uma dada ontologia representa bem o domínio ou não. Desvantagens dessas abordagens são sua subjetividade e demora. Além disso, esse método não é viável para avaliações em larga escala. Assim, a comparação com uma ontologia de referência é uma alternativa plausível. Trabalhos avaliados através de comparação com ontologias de referência são apresentados em Maedche e Staab [42] e Dellschaft e Staab [25]. Entretanto, como se pode afirmar que uma ontologia é boa o suficiente para ser uma ontologia de referência? A ontologia de referência é uma ontologia feita à mão, desenvolvida pelos mesmos processos custosos e propensos a erros que a aprendizagem automática de ontologias tenta evitar. Se a ontologia de referência apresenta problemas de modelagem, o método de avaliação recompensa ontologias com problemas similares e penaliza ontologias com conceitos ou relacionamentos que não

aparecem na ontologia de referência. A tabela 29 resume as vantagens e desvantagens dessas propostas de avaliação de técnicas de AO.

No contexto do presente trabalho foram desenvolvidos dois procedimentos para avaliação das técnicas de ARNT, apresentados na seção 5.5, baseados na proposta de comparação da ontologia aprendida com uma ontologia de referência.

Propostas para avaliação de técnicas de AO	Aspectos positivos	Aspectos negativos
Avaliação da ontologia aprendida em uma aplicação executável	A ontologia aprendida é avaliada em uma situação real.	Outros fatores podem ter impacto na efetividade nos resultados.
Avaliação da ontologia aprendida a posteriori por especialistas	Os especialistas conhecem o domínio, suas entidades e relacionamentos.	Inviável para avaliações em larga escala.
Comparação da ontologia aprendida com uma ontologia de referência	Pode ser definido um procedimento para automatizar a tarefa.	Necessidade de determinar se uma ontologia pode ser utilizada como referência.

Tabela 29: Propostas para avaliação de técnicas de AO

5.2.Abordagem para Aprendizagem de relacionamentos com Extração de Regras de Associação

AERA (Aprendizagem de relacionamentos com Extração de Regras de Associação) é uma abordagem para a Aprendizagem de relacionamentos não-taxonômicos cujo nome remete ao algoritmo utilizado na fase de "Refinamento"; o algoritmo para "Extração de Regras de Associação" [1] [2]. Essa abordagem foi desenvolvida com a finalidade de permitir a verificação da hipótese postulada nesse trabalho de que soluções para a fase de refinamento do processo de genérico de ARNT [57] [58] [59] [60] (seção 2.3) que utilizam relacionamentos do tipo "$<c_1$, fv, $c_2>$" são menos efetivas que as que utilizam relacionamentos "$<c_1$, $c_2>$", uma vez que tendem a apresentar menores valores para as medidas de avaliação quando considerados o mesmo corpus e ontologia de referência.

O algoritmo de "Extração de regras de associação" [1] [2] teve seu uso na fase de "Refinamento" proposto pela técnica de "ARNT baseada na extração de regras de associação" [69] (seção 3.1) e foi a solução adotada,

para efeito da verificação da hipótese de pesquisa, uma vez que dentre as soluções de refinamento utilizadas pelas técnicas de ARNT avaliadas nesse trabalho [29] [42] [48] [55] [56] [69] (capítulo 3) é a que emprega relacionamentos do tipo "<c_1, fv, c_2>". Assim como TARNTool, a AERA-Tool, ferramenta de suporte a AERA, utiliza a API Gate [20] [21] de processamento de linguagem natural para realizar a tokenização e separação de sentenças na fase de "Anotação do corpus". Para a fase de "Extração de relacionamentos candidatos" é utilizada a regra de sentença com frase verbal, conforme definida na seção 4.2.3.2. Na fase de "Refinamento" é utilizada uma especialização do algoritmo de extração de regras de associação (apresentado na seção 2.6.2.1), batizado "*RulesExtractor*" (Extrator de regras - figura 20). Esse algoritmo sugere relacionamentos não-taxonômicos como regras de associação com o formato "<c_1, c_2> → fv" onde "c_1" e "c_2" são conceitos e "fv" é uma frase verbal, conforme sugerido na técnica de "ARNT baseada na extração de regras de associação" [69].

A figura 19 apresenta o diagrama de sequência de AERA-Tool. A classe "*Application*" (Aplicação) é a que gerencia todo o processo. Inicialmente ela informa o arquivo no formato "OWL" da ontologia à classe "*ClassesExtractor*" (Extrator de classes) que retorna as *strings* correspondentes às classes da ontologia. Em seguida, "*Application*" fornece essas classes à classe "*ClassesLemmatisation*" (Lematiza classes) e tem como retorno essas classes lematizadas. A classe "*Application*" fornece o corpus (arquivos txt) à classe "Gate" que retorna o corpus anotado. A classe "*Application*" informa então à classe "*RelExtractor*" (Extrator de relacionamentos) o corpus anotado, as classes originais (não lematizadas), as classes lematizadas e tem como retorno os relacionamentos candidatos. Em seguida, a classe "*Application*" informa à classe "*RulesExtractor*" (Extrator de Regras) os relacionamentos candidatos que por sua vez retorna os relacionamentos a serem sugeridos ao especialista como regras de associação no formato "<c_1, c_2> → fv", sendo "c_1" e "c_2" conceitos e "fv" uma frase verbal. A figura 20 apresenta o código da classe "*RulesExtractor*" (Extrator de regras), responsável por retornar os relacionamentos refinados no formato de regras de associação (<c_1, c_2> → fv). O método "*extractRules*" (extrai regras) executa o algoritmo Apriori [1] [2]

informando a ele os relacionamentos candidatos ("*candidateRel*") e os valores de suporte e confiança mínimos, "*supportMin*" e "*confidenceMin*" respectivamente. Em seguida, o método "*rulesCCVP*" filtra os relacionamentos que tem o padrão "<c_1, c_2> → fv" e os retorna como resultado final da fase de "Refinamento". A tabela 30 apresenta a correspondência entre as classes de AERA-Tool e as fases do processo genérico de ARNT (seção 2.3).

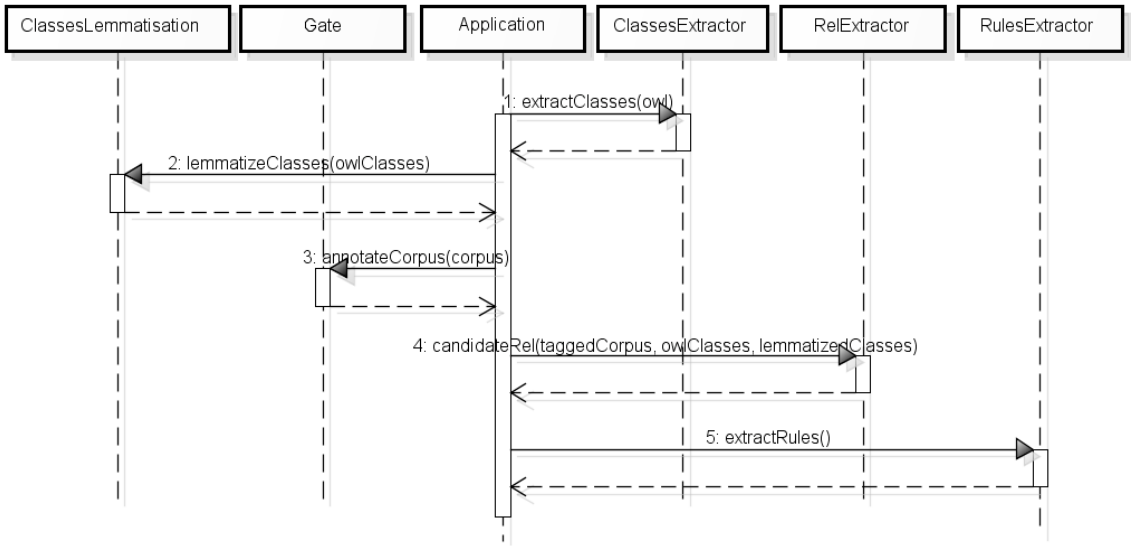


Figura 19: Diagrama de sequência de AERA-Tool

```
import ExecApriori;
class RulesExtractor{
private CandidateRel candidateRel;
private double supportMin;
private double confidenceMin;
// Initializes the properties of RulesExtractor
RulesExtractor(candidateRel, supportMin, confidenceMin){
this.candidateRel = candidateRel;
this.supportMin = supportMin;
this.confidenceMin = confidenceMin;
}
//Execute apriori
//Return rules in the format C1,C2->VP
public ExecApriori extractRules(){
            ExecApriori      execApriori      =      new      ExecApriori
            (this.candidateRel,this.supportMin, this.confidenceMin);
return execApriori.rulesCCVP()
}
}
```

Figura 20: A classe "*RulesExtractor*" da ferramenta AERA-Tool

Fases de ARNT		Classes de AERA-Tool
Construção do corpus		-
Extração de Relacionamentos Candidatos	Anotação do corpus	*Gate, ClassesLemmatisation*
	Extração de Relacionamentos	*RelExtractor, ClassesExtractor*
Refinamento de Relacionamentos		*RulesExtractor*
Avaliação pelo especialista		-
Atualização da ontologia		-

Tabela 30: Fases do processo de ARNT e correspondentes classes da ferramenta AERA-Tool

5.3.O corpus e ontologia Genia

O corpus e ontologia utilizados nos experimentos descritos nas seções 5.6 e 5.7 do presente capítulo fazem parte do projeto Genia [14] [53]. O corpus Genia [14] [53], é uma coleção de documentos sobre biologia molecular que consiste de resumos de artigos extraídos do banco de dados MEDLINE. O objetivo do projeto Genia é criar um recurso para apoiar o desenvolvimento de aplicações de processamento de linguagem natural no domínio da biologia molecular. A ontologia Genia [14] [53] representa o conhecimento desse domínio e seus relacionamentos não-taxonômicos (anexo B) foram utilizados como referência para o cálculo das medidas de avaliação (*recall*, precisão e medida-F) definidas na seção 2.3. As tabelas 31 e 32 apresentam algumas características do corpus e da ontologia Genia.

Corpus	Domínio	Número de documentos	Número de sentenças	Número de palavras
Genia [53]	Biologia	2.000	18.545	436.967

Tabela 31: Características do corpus Genia

Ontologia	Domínio	Número de conceitos	Número de relacionamentos não-taxonômicos
Genia [53]	Biologia	27	38

Tabela 32: Características da ontologia Genia

5.4.O corpus e ontologia *Family law doctrine*

O corpus e ontologia utilizados nos experimentos descritos nas seções 5.8 e 5.9 do presente capítulo fazem parte do projeto *Family law doctrine* [62]. O corpus *Family law doctrine* é uma coleção de documentos sobre a doutrina do Direito de família. O Direito de família é um ramo do Direito civil, que constitui o conjunto de normas que regulam entre outros fatos, a celebração do casamento, sua validade e os efeitos que dele resultam; as relações pessoais e econômicas da sociedade conjugal; a dissolução desta; as relações entre pais e filhos e os vínculos de parentesco. A ontologia *Family law doctrine* representa o conhecimento desse domínio e seus relacionamentos não-taxonômicos (anexo F) foram utilizados como referência para o cálculo das medidas de avaliação (*recall*, precisão e medida-F) definidas na seção 2.3. As tabelas 33 e 34 apresentam algumas características do corpus e da ontologia *Family law doctrine*.

Corpus	Domínio	Número de documentos	Número de sentenças	Número de palavras
Family law doctrine	Direito de Família	926	8.334	93.247

Tabela 33: Características do corpus *Family law doctrine* [62]

Ontologia	Domínio	Número de conceitos	Número de relacionamentos não-taxonômicos
Family law doctrine	Direito de Família	27	42

Tabela 34: Características da ontologia *Family law doctrine* [62]

5.5.Procedimentos de avaliação

Para verificar a efetividade de TARNT no cumprimento de seu objetivo foram definidos dois procedimentos de avaliação de técnicas de ARNT baseados na comparação dos relacionamentos aprendidos com os de referência: os procedimentos RAPR (Recomendação de relacionamentos com Anulação dos Parâmetros de Refinamento) e RMMA (Recomendação de

relacionamentos com Maximização de Medida de Avaliação) os quais são descritos nas seções 5.5.1 e 5.5.2 respectivamente.

5.5.1.Procedimento de avaliação RAPR

O objetivo do procedimento RAPR (Recomendação de relacionamentos com Anulação dos Parâmetros de Refinamento) é avaliar comparativamente técnicas de ARNT em termos de uma medida de avaliação cujo valor é calculado para grupos de relacionamentos por elas sugeridos. Os relacionamentos sugeridos devem estar ordenados por algum dos parâmetros da solução de refinamento e os grupos devem ser considerados em igual aridade. Os parâmetros de corte da solução para a fase de refinamento devem ser anulados. Por exemplo, no caso do algoritmo de "Extração de regras de associação" [1] [2] isso corresponde a ajustar os valores de suporte e confiança mínimos para zero, já no caso de *Bag of labels* [61] corresponde a ajustar para zero o valor da frequência mínima.

O procedimento de avaliação está formalizado no código da figura 21. Na linha um, os oito argumentos, ontologia de referência ("*ontology*"), a técnica a ser avaliada ("*tec*"), os valores dos parâmetros com os quais a técnica deva ser executada ("*paramT*"), o corpus do qual são extraídos os relacionamentos ("*corpus*"), o parâmetro para ordenação dos relacionamentos refinados ("*paramS*"), a quantidade de relacionamentos a ser considerada no resultado ("*max*"), o tamanho dos subgrupos de relacionamentos para os quais a medida de avaliação é calculada ("*inc*") e a medida de avaliação escolhida ("*measure*") são informados ao procedimento RAPR.

Na linha dois, "*ntrOntology*" recebe como argumento uma ontologia de referência ("*ontology*") e retorna seus relacionamentos não-taxonômicos. Esses relacionamentos, assim como os relacionamentos extraídos pela técnica de ARNT, serão utilizados para o cálculo dos valores das medidas de avaliação. Na linha três, "*execTec*" recebe como argumentos a técnica a ser executada ("*tec*"), os parâmetros com os quais deva ser executada ("*paramT*") e o corpus ("*corpus*") e retorna os relacionamentos recomendados pela referida técnica. Na linha quatro, "*sort*" retorna os relacionamentos recomendados pela

técnica (“*recRel*”) ordenados de forma decrescente (“*descending*”) pelo parâmetro da solução de refinamento “*paramS*”. Na linha seguinte têm-se um laço variando de zero até um valor menor que a quantidade de relacionamentos que devam ser considerados para os cálculos (“*max*”). A variável “*inc*” utilizada no incremento é, assim como “*max*”, informada pelo usuário e corresponde ao tamanho dos subgrupos de relacionamentos utilizados para o calculo da medida de avaliação. Por exemplo, nos experimentos realizados nas seções 5.6 e 5.8, “*max*” e “*inc*” receberam respectivamente 100 e 5.

Na linha seis, a variável “*setRel*” recebe os primeiros “*i + inc*” relacionamentos não-taxonômicos recomendados pela técnica (“*sortedRel*”). Na linha seguinte, o vetor “*evalMeasure*” recebe, para cada uma de suas posições, o valor para a medida de avaliação calculada (“*measure*”) para grupos de relacionamentos de tamanho “*inc*” considerados cumulativamente a partir do primeiro (“*setRel*”). Para realizar o cálculo são ainda informados os relacionamentos de referência obtidos no passo dois (“*refRel*”). Por exemplo, se considerarmos *max* = 100 e *inc* = 5, então a posição “0” do vetor conterá o valor da medida de avaliação calculado para as cinco primeiras recomendações, já a posição “1” conterá o valor da medida calculada para as 10 primeiras recomendações. Por fim, o vetor “*evalMeasure*” contendo os valores para a medida de avaliação é retornado.

```
1. RAPR(ontology, tec, paramT, corpus, paramS, max, inc,
   measure)
2. refRel := ntrOntology(ontology)
3. recRel := execTec(tec, paramT, corpus)
4. sortedRel := sort(recRel, paramS, 'descending')
5. for(i:=0 ; i<max; i:=i+inc) do
6.     setRel := returnRel(i+inc, sortedRel)
7.     evalMeasure[] := calcEvalMeasure(setRel, refRel,
       measure)
8. endFor
9. return evalMeasure
```

Figura 21: Procedimento de avaliação RAPR

5.5.2.Procedimento de avaliação RMMA

O objetivo do procedimento RMMA (Recomendação de relacionamentos com Maximização de Medida de Avaliação) é avaliar comparativamente técnicas de ARNT em termos de uma medida de avaliação a ser maximizada. Dessa forma, cada técnica deve ser executada com a configuração que lhe permita obter o maior valor para a medida de avaliação considerada.

O procedimento de avaliação RMMA está formalizado no pseudocódigo da figura 22. Na primeira linha, os quatro argumentos, ontologia de referência ("*ontology*"), a técnica a ser avaliada ("*tec*"), o corpus do qual são extraídos os relacionamentos ("*corpus*") e a medida de avaliação escolhida ("*measure*") são informados ao procedimento RMMA. Na linha dois, "*ntrOntology*" recebe como argumento uma ontologia de referência ("*ontology*") e retorna seus relacionamentos não-taxonômicos, que são atribuídos a "*refRel*". Na linha três, "*paramMax*" recebe como argumentos a técnica de ARNT ("*tec*"), o corpus do qual os relacionamentos devem ser obtidos ("*corpus*") e a medida de avaliação a ser maximizada ("*measure*") e retorna os valores para os parâmetros da técnica ("*tec*") que maximizam o valor da medida de avaliação informada ("*paramM*"). No quarto passo, "*execTec*" executa a técnica de ARNT ("*tec*") com os valores dos parâmetros obtidos no passo três ("*paramM*") sobre o corpus informado como seu terceiro argumento ("*corpus*") e retorna o conjunto de relacionamentos recomendados pela técnica ("*recRel*"). No passo cinco, os relacionamentos de referência obtidos no passo dois ("*refRel*") e os recomendados pela técnica no passo quatro ("*recRel*"), são utilizados para calcular a medida de avaliação ("*measure*") informada como terceiro argumento da função "*calcEvalMeasure*". Por fim, é retornado o valor maximizado da medida de avaliação ("*maxMeasure*").

```
1. RMMA(ontology, tec, corpus, measure)
2.    refRel := ntrOntology(ontology)
3.    paramM := paramMax(tec, corpus, measure)
4.    recRel := execTec(tec, paramM, corpus)
5.    maxMeasure := calcEvalMeasure(refRel, recRel,
      measure)
6. return maxMeasure
```

Figura 22: Procedimento de avaliação RMMA

5.6. Avaliação *Bag of labels* versus "Extração de regras de associação" utilizando RAPR e o corpus Genia

Nesse experimento o procedimento de avaliação RAPR (seção 5.5.1) foi aplicado a AERA e TARNT com as configurações apresentadas nas tabelas 35 e 36 respectivamente. O corpus e ontologia Genia [53] foram utilizados como fonte de extração dos relacionamentos e ontologia de referência respectivamente.

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0	0

Tabela 35: Configuração de AERA para avaliação utilizando RAPR e o corpus Genia [53]

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0

Tabela 36: Configuração de TARNT para avaliação utilizando RAPR e o corpus Genia [53]

A listagem dos relacionamentos recomendados por cada abordagem ordenados pelos parâmetros, "frequência de co-ocorrência" (TARNT) e

“confiança” (AERA) estão nos anexos C e D respectivamente. No anexo B estão os relacionamentos de referência obtidos da ontologia Genia [53]. Considerou-se que para AERA um casamento entre um relacionamento recomendado e um de referência ocorreu sempre que os três elementos “<c_1, fv, c_2>” (“c_1” e “c_2” são conceitos da ontologia e “fv” é uma frase verbal) de um relacionamento de referência coincidiram com os correspondentes elementos de um relacionamento recomendado pela ferramenta na forma de uma regra de associação “<c_1, c_2> → fv”. Já para TARNT um casamento ocorreu sempre que um par de conceitos e sua frase verbal correspondente, no caso de um relacionamento de referência, coincidiram respectivamente com um par de conceitos recomendado por TARNT e uma frase verbal em seu *bag of labels*. Por exemplo, o relacionamento (CELL,VIRUS → host) (anexo D) recomendado por AERA casou com o relacionamento de referência (CELL, host, VIRUS) (anexo B). Houve também o casamento desse relacionamento de referência com uma recomendação de TARNT (anexo C), já que os conceitos "Cell" e "Virus" coincidiram e a frase verbal "host" está presente em seu respectivo *bag of labels*. As tabelas 37, 38 e 39 apresentam as quantidades de relacionamentos válidos para cada grupo de cinco recomendações e também os valores de *recall*, precisão e medida-F para esses grupos considerados cumulativamente a partir do primeiro. As figuras 23, 24 e 25 apresentam os gráfico de *recall*, precisão e medida-F correspondentes as tabelas 37, 38 e 39 para ambas as abordagens.

	TARNT		AERA		
Grupos de relacionamentos	Qtd. rel. válidos	*Recall* (A)	*Recall* (B)	Qtd. rel. válidos	A – B
5	3	0,0789	0,0263	1	0,0526
10	3	0,1578	0,0789	2	0,0789
15	2	0,2105	0,1315	2	0,0789
20	1	0,2368	0,2105	3	0,0263
25	1	0,2631	0,2631	2	0,0000
30	3	0,3421	0,3157	2	0,0263
35	1	0,3684	0,3421	1	0,0263
40	1	0,3947	0,3684	1	0,0263
45	1	0,4210	0,3684	0	0,0526
50	1	0,4473	0,3947	1	0,0526
55	1	0,4736	0,4210	1	0,0526
60	0	0,4736	0,4473	1	0,0263
65	2	0,5263	0,4473	0	0,0789
70	1	0,5526	0,4473	0	0,1053
75	0	0,5526	0,4473	0	0,1053
80	2	0,6052	0,4736	1	0,1316
85	1	0,6315	0,4736	0	0,1579
90	1	0,6578	0,5000	1	0,1579
95	2	0,7105	0,5000	0	0,2105
100	2	0,7631	0,5263	1	0,2368

Tabela 37: *Recall* de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

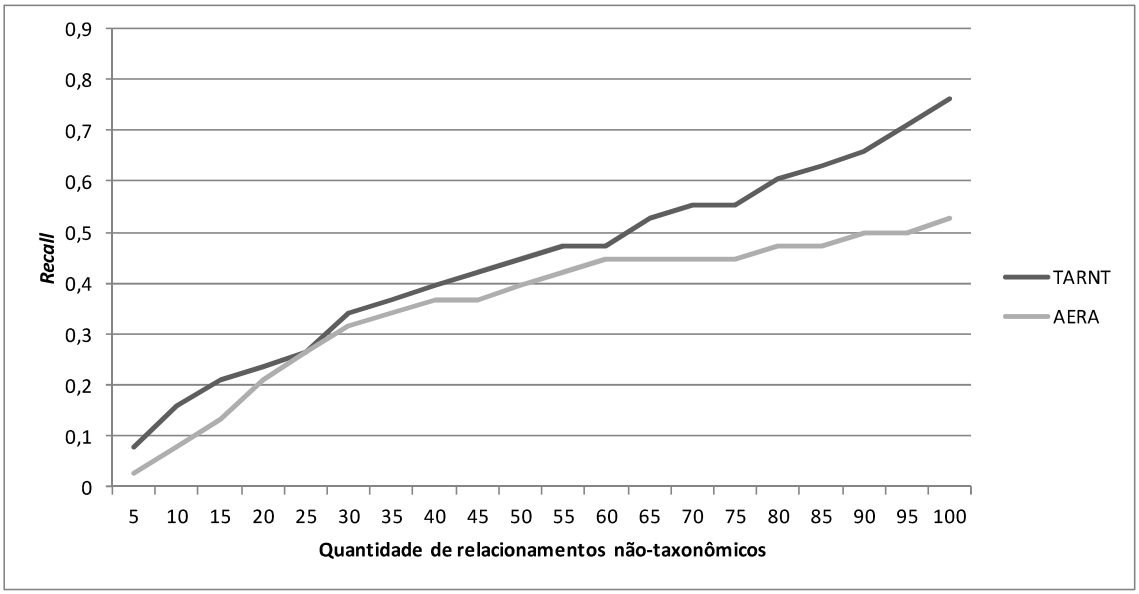


Figura 23: *Recall* de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

	TARNT		AERA		
Grupos de relacionamentos	Qtd. rel. válidos	Precisão (A)	Precisão (B)	Qtd. rel. válidos	A – B
5	3	0,6000	0,2000	1	0,4000
10	3	0,6000	0,3000	2	0,3000
15	2	0,5333	0,3333	2	0,2000
20	1	0,4500	0,4000	3	0,0500
25	1	0,4000	0,4000	2	0,0000
30	3	0,4333	0,4000	2	0,0333
35	1	0,4000	0,3714	1	0,0286
40	1	0,3750	0,3500	1	0,0250
45	1	0,3555	0,3111	0	0,0444
50	1	0,3400	0,3000	1	0,0400
55	1	0,3272	0,2909	1	0,0364
60	0	0,3000	0,2833	1	0,0167
65	2	0,3076	0,2615	0	0,0462
70	1	0,3000	0,2428	0	0,0571
75	0	0,2800	0,2266	0	0,0533
80	2	0,2875	0,2250	1	0,0625
85	1	0,2823	0,2117	0	0,0706
90	1	0,2777	0,2111	1	0,0666
95	2	0,2842	0,2000	0	0,0842
100	2	0,2900	0,2000	1	0,0900

Tabela 38: Precisão de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

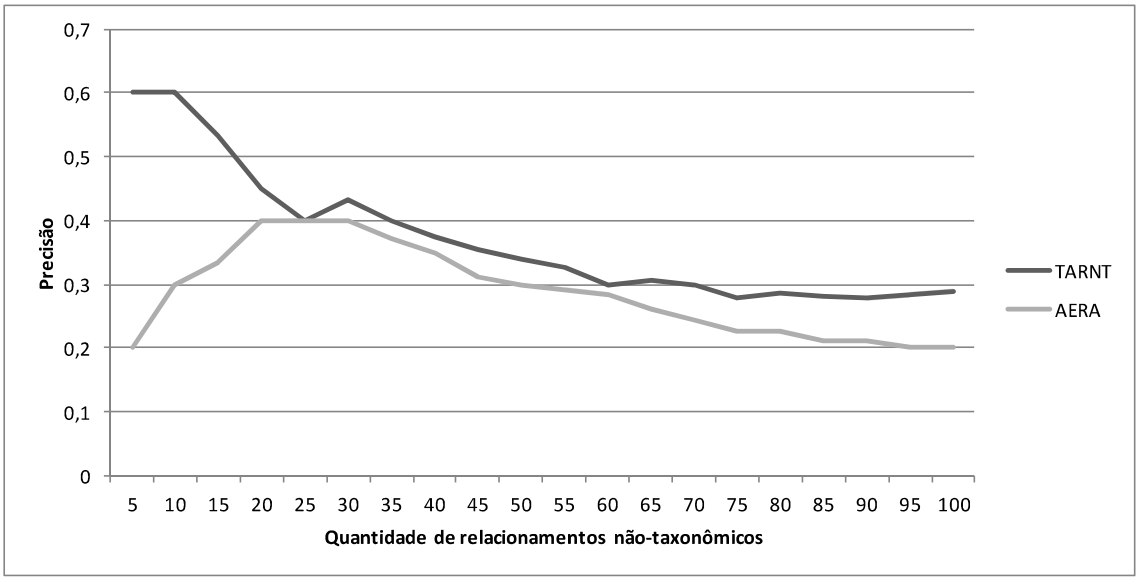


Figura 24: Precisão de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

	TARNT		AERA		
Grupos de relacionamentos	Qtd. rel. válidos	Medida-F (A)	Medida-F (B)	Qtd. rel. válidos	A – B
5	3	0,1395	0,0465	1	0,0930
10	3	0,2500	0,1250	2	0,1250
15	2	0,3018	0,1886	2	0,1132
20	1	0,3103	0,2758	3	0,0345
25	1	0,3174	0,3174	2	0,0000
30	3	0,3823	0,3529	2	0,0294
35	1	0,3835	0,3561	1	0,0274
40	1	0,3852	0,3589	1	0,0263
45	1	0,4180	0,3373	0	0,0807
50	1	0,3863	0,3409	1	0,0454
55	1	0,3870	0,3440	1	0,0430
60	0	0,3673	0,3469	1	0,0204
65	2	0,3883	0,3300	0	0,0583
70	1	0,3888	0,3148	0	0,0740
75	0	0,3716	0,3008	0	0,0708
80	2	0,3898	0,3050	1	0,0848
85	1	0,3902	0,2926	0	0,0976
90	1	0,3906	0,2968	1	0,0938
95	2	0,4060	0,2857	0	0,1203
100	2	0,4096	0,2898	1	0,1198

Tabela 39: Medida-F de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

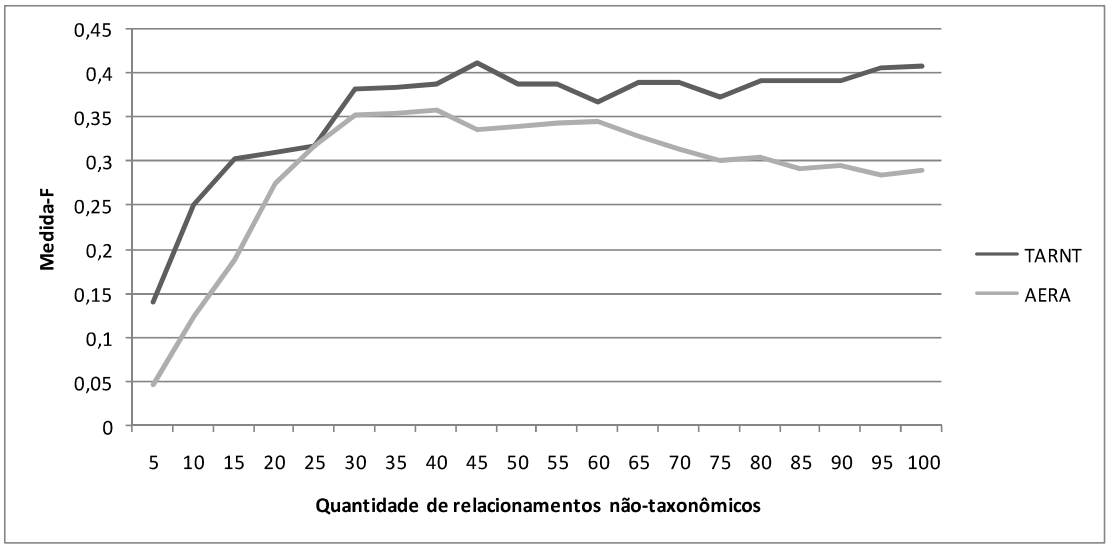


Figura 25: Medida-F de TARNT e AERA para as cem primeiras recomendações a partir do corpus Genia [53]

Em todo o intervalo de recomendações observado (da primeira a centésima), TARNT [60] [61] obteve valores de *recall* iguais ou superiores aos obtidos por AERA, tendo sido de 0,0842 (aproximadamente 8,4%) sua diferença média. Essa observação pode ser explicada pelo fato de a mesma quantidade total de relacionamentos de referência identificados por ambas as ferramentas (37) estarem distribuídos por um número maior de relacionamentos recomendados por AERA (524) que pelos 134 recomendados no caso de TARNT. O número de recomendações feita por AERA foi 3,9 vezes superior ao de TARNT.

De forma geral a tendência é que os valores de *recall* de TARNT quando configurada com a "Regra de sentença com frase verbal" [61] na fase "Extração de relacionamentos" sejam iguais ou maiores que os apresentados por AERA considerando o mesmo corpus e conceitos como entrada, uma vez que o conjunto de tuplas extraídas por AERA (relacionamentos candidatos) será maior ou igual ao de TARNT. A situação de igualdade entre o número de relacionamentos candidatos de TARNT e AERA é possível apenas na situação hipotética na qual todas as frases verbais de cada par de conceitos extraídos do corpus coincidam.

A diferença entre os valores de *recall* obtidos por TARNT e AERA aumentou a medida que aumentou a quantidade de relacionamentos recomendados. A diferença no grupo das cinco primeiras recomendações foi de 0,0526, no grupo das 10 primeiras foi de 0,0789 e nos grupos das 80 e 100 primeiras foi de 0,1316 e 0,2368, respectivamente. Isso significa que AERA foi gradativamente identificando menos relacionamentos válidos que TARNT com o aumento do número de recomendações. Esse fato sugere que *Bag of labels* [61] e a medida "frequência de co-ocorrência" foram mais eficazes na realização da separação entre verdadeiros e falsos relacionamentos não-taxonômicos e que *Bag of labels* [61] é capaz de identificar com o menor número de recomendações um maior número de relacionamentos em relação ao algoritmo de "Extração de regras de associação" [1] [2].

Com relação à precisão (figura 24), da mesma forma que para o *recall*, TARNT, em todo o intervalo de recomendações observado (da primeira a centésima), obteve valores iguais ou superiores aos obtidos por AERA, tendo

sido de 0,0852 (aproximadamente 8,5%) sua diferença média. Essa observação pode ser explicada pelo fato de a mesma quantidade de relacionamentos de referência identificados por ambas as ferramentas (37) estarem distribuídos por um número maior de relacionamentos recomendados por AERA (524) que pelos 134 recomendados no caso de TARNT.

De forma geral a tendência é que os valores de precisão de TARNT quando configurada com quaisquer das regras de extração ("Regra de apóstrofo" [61] ou "Regra de sentença com frase verbal" [61]) na fase "Extração de relacionamentos" sejam maiores que os apresentados por AERA para todo o intervalo de recomendações observado, considerando o mesmo corpus e conceitos como entrada. Isso ocorre pois os relacionamentos de referência tendem a estar mais dispersos em AERA que em TARNT. TARNT obteve o ápice de precisão no intervalo das 10 primeiras recomendações tendo sido igual a 0,6. Já o ápice de precisão de AERA correspondeu a 0,4, nos grupos de 20 a 30 primeiros relacionamentos recomendados. A partir da faixa 40, tanto a precisão de TARNT quanto a de AERA apresentam uma trajetória descendente. Entretanto, AERA decresce de forma um pouco mais acentuada. A diferença entre os valores de precisão de TARNT e AERA é de 0,0333 na faixa dos 30; 0,0572 na faixa dos 70 e de 0,0666 na faixa dos 90. Essa observação sugere que os relacionamentos de referência de TARNT estão mais concentrados nos primeiros conjuntos de recomendações (de 5 a 30). Além disso, a perda de precisão de TARNT, com o crescente aumento do número de recomendações, tende a ser menor, uma vez que os relacionamentos de referência restantes estão distribuídos por um número menor de relacionamentos candidatos. Com relação a Medida-F, para o mesmo conjunto das primeiras 100 recomendações, TARNT também obteve valores iguais ou superiores aos obtidos por AERA, tendo sido de 0,0678 (aproximadamente 6,7%) a diferença média.

Por fim, considerando o procedimento de avaliação adotado (RAPR) e os valores obtidos para as medidas de avaliação *recall* e precisão, considera-se que TARNT foi mais efetiva que AERA na aprendizagem de relacionamentos não-taxonômicos a partir do corpus Genia [53] nas condições descritas nesse experimento. Além disso, conclui-se que *Bag of labels* [61] foi

uma solução mais adequada que a "Extração de regras de associação" [1] [2] para a fase de "Refinamento", fato que sugere a confirmação da hipótese postulada nesse trabalho (seção 1.3).

5.7.Avaliação *Bag of labels* versus "Extração de regras de associação" utilizando RMMA e o corpus Genia

No experimento realizado na seção 5.6 as medidas de avaliação (*recall*, precisão e medida-F) foram calculadas para grupos de cinco recomendações consideradas cumulativamente para as primeiras cem recomendações das técnicas avaliadas. O presente experimento utiliza o procedimento RMMA para avaliar comparativamente a efetividade de TARNT e AERA na extração de relacionamentos na situação na qual deseja-se maximizar uma medida de avaliação via o ajuste dos parâmetros das respectivas técnicas.

Considerou-se que para AERA um casamento entre um relacionamento recomendado e um de referência ocorreu sempre que os três elementos "<c_1, fv, c_2>" ("c_1" e "c_2" são conceitos da ontologia e "fv" é uma frase verbal) de um relacionamento de referência coincidiram com os correspondentes elementos de um relacionamento recomendado pela ferramenta na forma de uma regra de associação "<c_1, c_2> → fv". Já para TARNT um casamento ocorreu sempre que um par de conceitos e sua frase verbal correspondente, no caso de um relacionamento de referência, coincidiram respectivamente com um par de conceitos recomendado por TARNT e uma frase verbal em seu *bag of labels*. Para a maximização das medidas de avaliação *recall*, precisão e medida-F respectivamente, TARNT e AERA foram configuradas conforme especificado nas tabelas 40 a 45.

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0	0

Tabela 40: Configuração de AERA para a maximização do *recall* utilizando o corpus *Genia* [53]

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0

Tabela 41: Configuração de TARNT para a maximização do *recall* utilizando o corpus *Genia* [53]

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0,0019	0,6667

Tabela 42: Configuração de AERA para a maximização da precisão utilizando o corpus *Genia* [53]

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0,0147

Tabela 43: Configuração de TARNT para a maximização da precisão utilizando o corpus *Genia* [53]

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0,0038	0,5000

Tabela 44: Configuração de AERA para a maximização da medida-F utilizando o corpus *Genia* [53]

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0,0134

Tabela 45: Configuração de TARNT para a maximização da medida-F utilizando o corpus *Genia* [53]

A tabela 46 apresenta os valores máximos de *recall,* precisão e medida-F e as correspondentes quantidades de relacionamentos recomendados e válidos obtidos por TARNT e AERA. Tanto TARNT quanto AERA obtiveram o mesmo valor máximo para o *recall*, fato esperado, uma vez que apesar de estarem distribuídos por um número maior de recomendações no caso de AERA, os mesmos relacionamentos de referência estão todos presentes nos grupos de todas as recomendações feitas por ambas as abordagens. Generalizando, quando TARNT, na fase "Extração de relacionamentos" for configurada com a "Regra de sentença", obterá o mesmo *recall* que o de AERA; quando configurada com a "Regra de sentença com frase verbal", TARNT apresentará *recall* igual ou maior; já quando configurada com a "Regra de apóstrofo", TARNT apresentará *recall* menor ou igual.

Com relação à precisão máxima, TARNT obteve aproximadamente 0,6153, um valor 0,1923 superior ao obtido por AERA que foi de aproximadamente 0,4230. Isso se deve ao fato da melhor separação entre os verdadeiros e falsos relacionamentos feita por TARNT, que tende a concentrá-los no inicio da lista de recomendações. Com relação ao valor máximo para a medida-F, TARNT obteve 0,5274 e AERA 0,4438; sendo de 8,3% sua diferença. Por fim, considerando o procedimento de avaliação adotado (RMMA) e os valores obtidos para as medidas de avaliação *recall*, precisão e medida-F, considera-se que TARNT foi mais efetiva na aprendizagem de relacionamentos não-taxonômicos a partir do corpus Genia [53]. Além disso, conclui-se que *Bag of labels* [61] foi uma solução mais adequada que a "Extração de regras de associação" [1] [2] para a fase de "Refinamento", fato que sugere a confirmação da hipótese postulada nesse trabalho (seção 1.3).

	TARNT			AERA			Diferença entre as medidas de avaliação (A - B)%
	Qtd. de rel. recomen-dados	Qtd. de rel. válidos	Valor da medida de avaliação (A)	Qtd. de rel. recomen-dados	Qtd. de rel. válidos	Valor da medida de avaliação (B)	
Recall	134	37	0,9736	524	37	0,9736	0%
Precisão	13	8	0,6153	26	11	0,4230	19,2%
Medida-F	17	9	0,5274	29	13	0,4438	8,3%

Tabela 46: Valores máximos de *recall* e precisão para TARNT e AERA utilizando o corpus Genia [53]

5.8. Avaliação *Bag of labels* versus "Extração de regras de associação" utilizando RAPR e o corpus *Family law doctrine*

Nesse experimento o procedimento de avaliação RAPR (seção 5.5.1) foi aplicado a AERA e TARNT com as configurações apresentadas nas tabelas 47 e 48 respectivamente. O corpus e ontologia *Family law doctrine* [62] (seção 5.4) foram utilizados como fonte de extração dos relacionamentos e ontologia de referência respectivamente.

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0	0

Tabela 47: Configuração de AERA para avaliação utilizando RAPR e o corpus *Family law doctrine*

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0

Tabela 48: Configuração de TARNT para avaliação utilizando RAPR e o corpus *Family law doctrine*

Os relacionamentos recomendados por cada abordagem ordenados pelos parâmetros, frequência de co-ocorrência (TARNT) e confiança (AERA) estão nos anexos G e H respectivamente. No anexo F estão os relacionamentos de referência obtidos da ontologia *Family law doctrine*. Da mesma forma que para os experimentos das seções 5.6 e 5.7, considerou-se que para AERA um casamento entre um relacionamento recomendado e um de referência ocorreu sempre que os três elementos "$<c_1$, fv, $c_2>$" de um relacionamento de referência coincidiram com os correspondentes elementos de um relacionamento recomendado pela técnica na forma de uma regra de associação "$<c_1, c_2> \rightarrow$ fv". Para TARNT, um casamento ocorreu sempre que um par de conceitos e sua frase verbal correspondente, no caso de um relacionamento de referência, coincidiram respectivamente com um par de conceitos recomendado por TARNT e uma frase verbal em seu *bag of labels*. Por exemplo, o relacionamento "<COURT, DIVORCE> → grant" (anexo H)

recomendado por AERA casou com o relacionamento de referência “<COURT, grant, DIVORCE>” (anexo F). Houve também o casamento desse relacionamento de referência com uma recomendação de TARNT (anexo G), já que os conceitos “Court” e “Divorce” coincidiram e a frase verbal “grant” está presente em seu respectivo *bag of labels*. As tabelas 49, 50 e 51 apresentam as quantidades de relacionamentos válidos para cada grupo de cinco recomendações e também os valores de *recall*, precisão e medida-F para esses grupos considerados cumulativamente a partir do primeiro. As figuras 26, 27 e 28 apresentam os gráfico de *recall*, precisão e medida-F correspondentes as tabelas 49, 50 e 51 para ambas as abordagens.

	TARNT		AERA		A – B
Grupos de relacionamentos	**Qtd. rel. válidos**	***Recall* (A)**	***Recall* (B)**	**Qtd. rel. válidos**	
5	2	0,0476	0,0238	1	0,0238
10	3	0,1190	0,0714	2	0,0476
15	3	0,1904	0,0952	1	0,0952
20	2	0,2380	0,1666	3	0,0714
25	3	0,3095	0,2142	2	0,0952
30	3	0,3809	0,2619	2	0,1190
35	2	0,4285	0,2857	1	0,1429
40	2	0,4761	0,3095	1	0,1667
45	1	0,5000	0,3095	0	0,1905
50	2	0,5476	0,3333	1	0,2143
55	1	0,5714	0,3333	0	0,2381
60	0	0,5714	0,3571	1	0,2143
65	1	0,5952	0,3571	0	0,2381
70	2	0,6428	0,3571	0	0,2857
75	0	0,6428	0,3571	0	0,2857
80	1	0,6666	0,3809	1	0,2857
85	1	0,6904	0,3809	0	0,3095
90	1	0,7142	0,4047	1	0,3095
95	2	0,7619	0,4047	0	0,3571
100	1	0,7857	0,4047	0	0,3810

Tabela 49: *Recall* de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

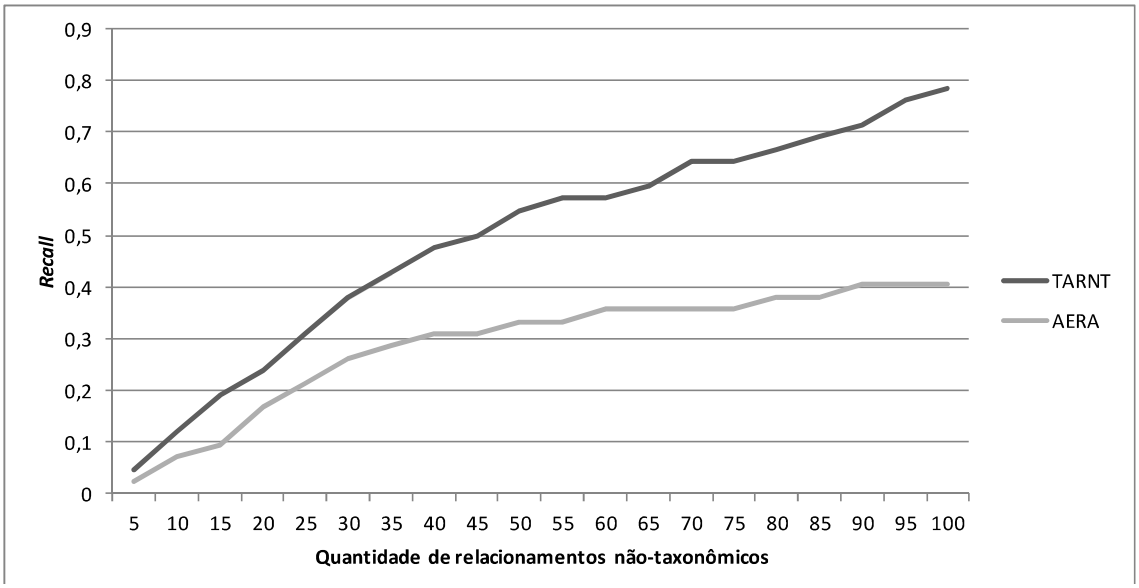


Figura 26: *Recall* de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

	TARNT		AERA		
Grupos de relacionamentos	**Qtd. rel. válidos**	**Precisão (A)**	**Precisão (B)**	**Qtd. rel. válidos**	**A – B**
5	2	0,4000	0,2000	1	0,2000
10	3	0,5000	0,3000	2	0,2000
15	3	0,5333	0,2666	1	0,2667
20	2	0,5000	0,3500	3	0,1500
25	3	0,5200	0,3600	2	0,1600
30	3	0,5333	0,3666	2	0,1667
35	2	0,5142	0,3428	1	0,1714
40	2	0,5000	0,3250	1	0,1750
45	1	0,4666	0,2888	0	0,1778
50	2	0,4600	0,2800	1	0,1800
55	1	0,4363	0,2545	0	0,1818
60	0	0,4000	0,2500	1	0,1500
65	1	0,3846	0,2307	0	0,1539
70	2	0,3857	0,2142	0	0,1715
75	0	0,3600	0,2000	0	0,1600
80	1	0,3500	0,2000	1	0,1500
85	1	0,3411	0,1882	0	0,1529
90	1	0,3333	0,1888	1	0,1445
95	2	0,3368	0,1789	0	0,1579
100	1	0,3300	0,1700	0	0,1600

Tabela 50: Precisão de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

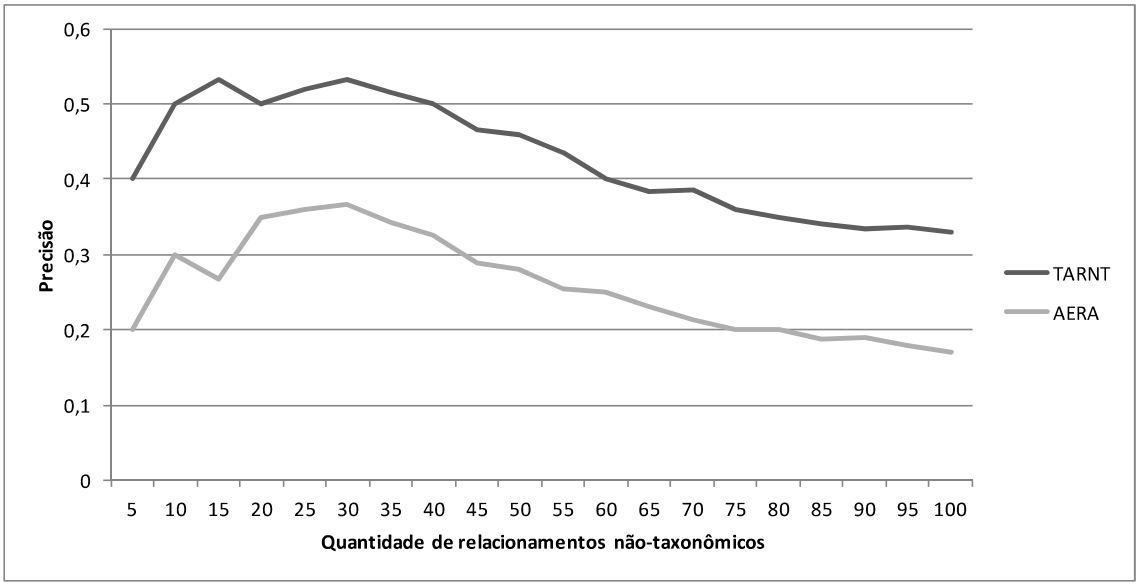


Figura 27: Precisão de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

	TARNT		AERA		A – B
Grupos de relacionamentos	**Qtd. rel. válidos**	**Medida-F (A)**	**Medida-F (B)**	**Qtd. rel. válidos**	
5	2	0,0851	0,0426	1	0,0425
10	3	0,1923	0,1154	2	0,0769
15	3	0,2807	0,1404	1	0,1403
20	2	0,3226	0,2258	3	0,0968
25	3	0,3881	0,2687	2	0,1194
30	3	0,4444	0,3056	2	0,1388
35	2	0,4675	0,3117	1	0,1558
40	2	0,4878	0,3171	1	0,1707
45	1	0,4828	0,2989	0	0,1839
50	2	0,5000	0,3043	1	0,1957
55	1	0,4948	0,2887	0	0,2061
60	0	0,4706	0,2941	1	0,1765
65	1	0,4673	0,2804	0	0,1869
70	2	0,4821	0,2679	0	0,2142
75	0	0,4615	0,2564	0	0,2051
80	1	0,4590	0,2623	1	0,1967
85	1	0,4567	0,2520	0	0,2047
90	1	0,4545	0,2576	1	0,1969
95	2	0,4672	0,2482	0	0,2190
100	1	0,4648	0,2394	0	0,2254

Tabela 51: Medida-F de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

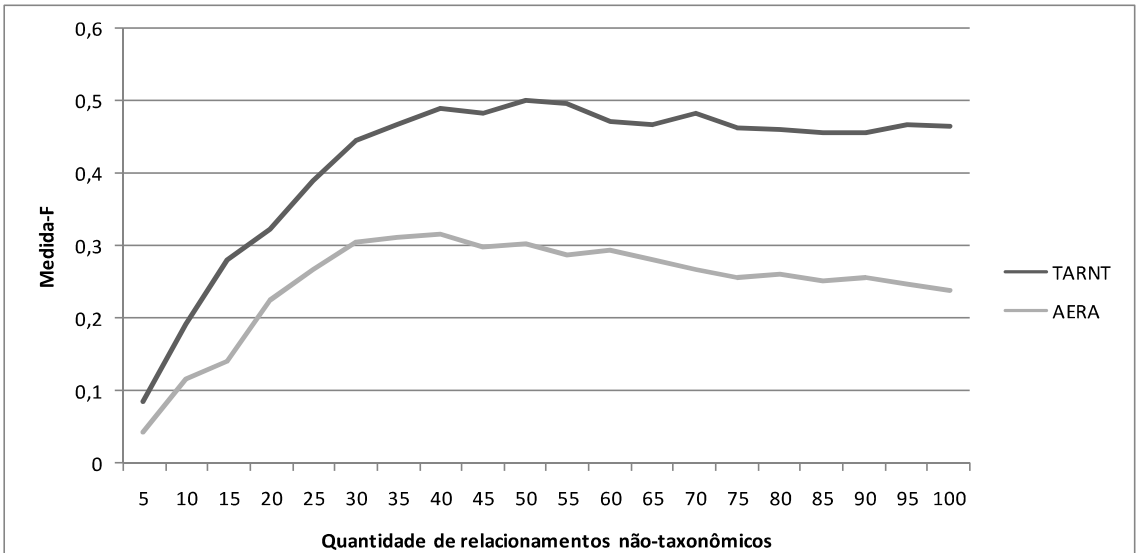


Figura 28: Medida-F de TARNT e AERA para as cem primeiras recomendações a partir do corpus *Family law doctrine*

Em todo o intervalo de recomendações observado (da primeira a centésima), TARNT obteve valores de *recall* superiores aos obtidos por AERA, tendo sido de 0,2035 (aproximadamente 20%) sua diferença média. Essa observação pode ser explicada pelo fato de a mesma quantidade de relacionamentos de referência identificados por ambas as abordagens (34) estarem distribuídos por um número maior de relacionamentos recomendados por AERA (551) que pelos 108 recomendados no caso de TARNT. O número de recomendações feita por AERA foi 5,1 vezes superior ao de TARNT.

De forma geral a tendência é que os valores de *recall* de TARNT, quando configurada com a "Regra de sentença com frase verbal" [61] na fase "Extração de relacionamentos" sejam iguais ou maiores que os apresentados por AERA considerando o mesmo corpus e conceitos como entrada, uma vez que o conjunto de tuplas extraídas por AERA (relacionamentos candidatos) será maior ou igual ao de TARNT. A situação de igualdade entre o número de relacionamentos candidatos de TARNT e AERA é possível apenas na situação hipotética na qual todas as frases verbais de cada par de conceitos extraídos do corpus coincidam.

A diferença entre os valores de *recall* obtidos por TARNT e AERA aumentou a medida que aumentou a quantidade de relacionamentos recomendados. A diferença no grupo das 5 primeiras recomendações foi de

aproximadamente 0,0238, no grupo das 10 primeiras foi de 0,0476 e nos grupos das 80 e 100 primeiras foi de 0,2857 e 0,3810, respectivamente. Isso significa que AERA foi gradativamente identificando menos relacionamentos válidos que TARNT com o crescimento do número de recomendações. Esse fato sugere que o algoritmo *Bag of labels* [61] e a medida "frequência de co-ocorrencia" foram mais eficazes na realização da separação entre verdadeiros e falsos relacionamentos não-taxonômicos e que *Bag of labels* é capaz de identificar com o menor número de recomendações um maior número de relacionamentos em relação a "Extração de regras de associação" [1] [2].

Com relação à precisão (figura 26), da mesma forma que para o *recall*, TARNT, em todo o intervalo de recomendações observado (da primeira a centésima), obteve valores superiores aos obtidos por AERA, tendo sido de 0,1715 (aproximadamente 17%) sua diferença média. Essa observação pode ser explicada pelo fato de a mesma quantidade de relacionamentos de referência identificados por ambas as abordagens (34) estarem distribuídos por um número maior de relacionamentos recomendados por AERA (551) que pelos 108 recomendados no caso de TARNT.

De forma geral, a tendência é que os valores de precisão de TARNT quando configurada com quaisquer das regras de extração ("Regra de apóstrofo" [61] ou "Regra de sentença com frase verbal" [61]) na fase "Extração de relacionamentos" sejam maiores que os apresentados por AERA para todo o intervalo de recomendações observado, considerando o mesmo corpus e conceitos como entrada. Isso ocorre, pois os relacionamentos de referência tendem a estar mais dispersos em AERA que em TARNT. Com relação a Medida-F, para o mesmo conjunto das primeiras 100 recomendações, TARNT também obteve valores superiores aos obtidos por AERA, tendo sido de 0,1676 (aproximadamente 16%) a diferença média.

Essas observações sugerem que a medida "Frequência de co-ocorrencia" adotada por *Bag of labels* [61] é mais eficaz em concentrar maior quantidade de relacionamentos de referência nas primeiras recomendações que a medida de confiança adotada pela "Extração de Regras de associação" [1] [2].

Por fim, considerando o procedimento de avaliação adotado (RAPR) e os valores obtidos para as medidas de avaliação *recall*, precisão e medida-F, considera-se que TARNT [60] [61] foi mais efetiva que AERA na aprendizagem de relacionamentos não-taxonômicos a partir do corpus *Family law doctrine* (seção 5.4) nas condições descritas nesse experimento. Além disso, conclui-se que *Bag of labels* [61] foi uma solução mais adequada que a "Extração de Regras de associação" [1] [2] para a fase de "Refinamento", fato que sugere a confirmação da hipótese postulada nesse trabalho (seção 1.3).

5.9. Avaliação *Bag of labels* versus "Extração de regras de associação" utilizando RMMA e o corpus *Family law doctrine*

O presente experimento utiliza o procedimento RMMA para avaliar comparativamente a efetividade de TARNT e AERA na extração de relacionamentos na situação na qual deseja-se maximizar uma medida de avaliação via o ajuste dos parâmetros das respectivas técnicas. Considerou-se que para AERA e TARNT um casamento entre um relacionamento recomendado e um de referência ocorreu conforme descrito na seção 5.8.

Para a maximização das medidas de avaliação *recall*, precisão e medida-F respectivamente, TARNT e AERA foram configuradas conforme especificado nas tabelas 52 a 57.

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0	0

Tabela 52: Configuração de AERA para a maximização do *recall* utilizando o corpus *Family law doctrine*

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0

Tabela 53: Configuração de TARNT para a maximização do *recall* utilizando o corpus *Family law doctrine*

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0,0054	0,5000

Tabela 54: Configuração de AERA para a maximização da precisão utilizando o corpus *Family law doctrine*

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0,0203

Tabela 55: Configuração de TARNT para a maximização da precisão utilizando o corpus *Family law doctrine*

AERA			
Extração de Relacionamentos Candidatos	**Refinamento**		
Regra de extração	Solução de refinamento	Sup. min.	Conf. min.
Regra de sentença com frase verbal	Extração de regras de associação	0,0036	0,4286

Tabela 56: Configuração de AERA para a maximização da medida-F utilizando o corpus *Family law doctrine*

TARNT		
Extração de Relacionamentos Candidatos	**Refinamento**	
Regra de extração	Solução de refinamento	Freq. min.
Regra de sentença com frase verbal	*Bag of labels*	0,0122

Tabela 57: Configuração de TARNT para a maximização da medida-F utilizando o corpus *Family law doctrine*

A tabela 58 apresenta os valores máximos de *recall*, precisão e medida-F e as correspondentes quantidades de relacionamentos recomendados e válidos obtidos por TARNT e AERA. Verifica-se que tanto TARNT quanto AERA obtiveram o mesmo valor máximo para o *recall*, fato esperado, uma vez que apesar de estarem distribuídos por um número maior de recomendações no caso de AERA, os mesmos relacionamentos de referência estão todos presentes nos grupos de todas as recomendações feitas por ambas as técnicas. Generalizando, quando TARNT, na fase "Extração de relacionamentos" for configurada com a "Regra de sentença" [61], obterá o mesmo *recall* que o de AERA; quando configurada com a "Regra de sentença com frase verbal" [61], TARNT apresentará *recall* igual ou maior; já quando configurada com a "Regra de apóstrofo" [61], TARNT apresentará *recall* menor ou igual.

Com relação à precisão máxima, TARNT obteve aproximadamente 0,5555; um valor aproximadamente 0,17 superior ao obtido por AERA que foi de 0,3846. Isso se deve ao fato da melhor separação entre os verdadeiros e falsos relacionamentos feita por TARNT, que tende a concentrá-los no inicio da lista de recomendações. Com relação ao valor máximo para a medida-F, TARNT obteve 0,5384 e AERA 0,3333; sendo de 20,5% sua diferença. Por fim, considerando o procedimento de avaliação adotado (RMMA) e os valores obtidos para as medidas de avaliação, considera-se que TARNT foi mais efetiva na aprendizagem de relacionamentos não-taxonômicos a partir do corpus *Family law doctrine* [62].

	TARNT			AERA			Diferença entre as medidas de avaliação (A - B)%
	Qtd. de rel. recomen-dados	Qtd. de rel. válidos	Valor da medida de avaliação (A)	Qtd. de rel. recomen-dados	Qtd. de rel. válidos	Valor da medida de avaliação (B)	
Recall	108	34	0,8095	551	34	0,8095	0%
Precisão	9	5	0,5555	13	5	0,3846	17%
Medida-F	23	12	0,5724	28	13	0,4602	11,2%

Tabela 58: Valores máximos de *recall,* precisão e medida-F para TARNT e AERA utilizando o corpus *Family law doctrine*

5.10. Conclusões dos resultados das avaliações

Os experimentos realizados nas seções 5.6 a 5.9 demonstraram experimentalmente a maior efetividade de TARNT [60] [61] em relação a AERA na tarefa de ARNT a partir dos corpora Genia [14] [52] (seção 5.3) e *Family law doctrine* [62] (seção 5.4). Nessa seção iremos justificar formalmente o motivo de tal superioridade e generalizar esses resultados com o intuito de verificar a hipótese de pesquisa postulada nesse trabalho (seção 1.3).

Para a execução dos experimentos, TARNT foi configurada com a mesma solução adotada por AERA para a fase "Extração de relacionamentos", a "regra de sentença com frase verbal" (seção 4.2.3). Já para a fase de refinamento foi utilizada uma solução (*Bag of labels*) que, diferentemente do algoritmo de "Extração de regras de associação" [1] [2], que utiliza relacionamentos "<c_1, fv, c_2>", trabalha com relacionamentos do tipo "<c_1, c_2>".

Começamos definindo o conceito de taxa de casamento de relacionamentos (TCR) para uma solução de refinamento (r_i) de uma técnica de ARNT como sendo a razão entre o número de relacionamentos de referência que coincidiram com os da lista de recomendados ("Qtd_rel_válidos") e o número total de relacionamentos recomendados ao especialista ("Qtd_rel_recomendados"). Formalmente, TCR é definido pela equação 27.

$$TCR(r_i) = \frac{\text{Qtd\_rel\_válidos}}{\text{Qtd\_rel\_recomendados}} \tag{27}$$

Considera-se que para a "Extração de regras de associação" [1] [2] um casamento entre um relacionamento recomendado e um de referência ocorre sempre que os três elementos "<c_1, fv, c_2>" de um relacionamento de referência coincidiram com os correspondentes elementos de um relacionamento recomendado na forma de uma regra de associação "<c_1, c_2> → fv" ("c_1" e "c_2" são conceitos da ontologia e "fv" uma frase verbal). Já para *Bag of labels* [61] um casamento ocorreu sempre que um par de conceitos e sua frase verbal correspondente, no caso de um relacionamento de referência, coincidiram respectivamente com um par de conceitos recomendado ("<c_1, c_2>") e uma frase verbal em seu *bag of labels*.

Nesse ponto, para *Bag of labels* [61] e "Extração de regras de associação" afirmamos que, para um mesmo conjunto de relacionamentos candidatos, produto da fase "Extração de relacionamentos" e mesmo conjunto de relacionamentos de referência tem-se a equação 28:

$$TCR(Bag\ of\ labels) >= TCR(Extração\ de\ regras\ de\ associação) \quad (28)$$

Disso vimos que o valor de TCR fica em função do número de relacionamentos recomendados ("Qtd_rel_recomendados"). É então necessário saber como calcular essa medida para cada uma das duas técnicas.

Para *Bag of labels* [61], o número de relacionamentos recomendados ("Qtd_rel_recomendados") é igual ao número de pares de conceitos diferentes presentes nos relacionamentos candidatos ("Qtd_pc"), *input* para o algoritmo de refinamento. Formalmente, o número de relacionamentos recomendados ("Qtd_rel_recomendados") por *Bag of labels* [61] é dado pela equação 29:

$$Qtd\_rel\_recomendados(Bag\ of\ labels) = Qtd\_pc \quad (29)$$

Já para a "Extração de regras de associação" [1] [2] o número de relacionamentos recomendados é dado pela fórmula 30.

$$\text{Qtd\_rel\_ recomendados (Extração de regras de asociação)} = SQFP(\{pc_1 \dots pc_n\}) = \sum_{i=1}^{n} QFP(pc_i) \quad (30)$$

"QFP" corresponde à quantidade de frases verbais diferentes associadas a um par de conceitos obtido a partir dos relacionamentos candidatos. Já o "SQFP" de um conjunto de pares de conceitos corresponde à soma dos "QFP" de todos os pares de conceitos diferentes presentes no conjunto de relacionamentos candidatos. Dessa forma, "Qtd_rel_recomendados(Extração de regras de associação)" está em função do "SQFP", ou seja, quanto maior/menor o número de frases verbais diferentes associadas aos pares de conceitos obtidos com as regras de extração, maior/menor será o "Qtd_rel_recomendados(Extração de regras de

associação)" e consequentemente maior/menor será o TCR(Extração de regras de associação).

Já a quantidade de relacionamentos recomendados por *Bag of labels* [61] é igual ao número de pares de conceitos diferentes obtidos a partir de uma regra de extração, independentemente do número de frases verbais diferentes associadas a cada um deles (ver equação 29). Então TCR(*Bag of labels*), por não estar em função do "SQFP", não tem alterado o número de relacionamentos recomendados conforme o número de frases verbais diferentes associadas a cada par de conceitos.

Sedo assim, considerando a situação hipotética na qual todas as frases verbais associadas a cada par de conceitos sejam iguais têm-se: TCR(*Bag of labels*) = TCR(Extração de regras de associação). Para qualquer outra situação, ou seja, que haja pelo menos uma frase verbal diferente das demais associadas para qualquer par de conceitos, TCR(*Bag of labels*) > TC(Extração de regras de associação). Em outros termos, a "Extração de regras de associação" [1] [2] tem seu resultado, em termos da medida TCR, pior quanto maior o número de frases verbais diferentes presentes entre os conceitos que casaram. Já *Bag of labels* mantém seu resultado independentemente do número de frases verbais diferentes.

A relevância de TCR na discussão sobre *Bag of labels* [61] e "Extração de regras de associação" [1] [2] reside no fato de que por possuir maior TCR que a "Extração de regras de associação", para um mesmo corpus e ontologia de referência, *Bag of labels* tende a apresentar maiores valores para as medidas de avaliação. Isso se deve ao fato de que, conforme já explicado nos parágrafos anteriores, os relacionamentos válidos estão distribuídos por um maior ou igual número de relacionamentos recomendados no caso de "Extração de regras de associação" que *Bag of labels* (equação 30). Outro fator que corrobora para que a "Extração de regras de associação" tenda a apresentar menores valores para as medias de avaliação é o fato de essa explicitar a informação de quais frases verbais normalmente ocorrem com cada par de conceitos, o que a princípio é adequado para uma solução de refinamento que se propõem a recomendar rótulos aos pares de conceitos. Entretanto, regras de associação que possuem os maiores valores de

confiança (seção 2.6.2.1) podem conter pares de conceitos que apresentam baixa ocorrência no corpus. Em suma, a “Extração de regras de associação” [1] [2] valoriza uma informação menos relevante que é a de qual rótulo está associado a determinado par de conceitos e em contrapartida perde efetividade na realização da tarefa mais importante que é a da identificação de quais pares de conceitos estão relacionados.

Os argumentos apresentados no parágrafo anterior são ainda confirmados com evidencias obtidas a partir de resultados experimentais apresentados nas seções 5.6 a 5.9. Para os experimentos das duas primeiras seções (5.6 e 5.7), que consistiram na aplicação dos procedimentos de avaliação RAPR e RMMA às extrações a partir do corpus Genia [14] [52], os valores para TCR(*Bag of labels*) e TCR(Extração de regras de associação) são calculados conforme as equações 31 e 32.

$$TCR(Bag\ of\ labels) = \frac{37}{134} = 0{,}2761 \tag{31}$$

$$TCR(Extração\ de\ regras\ de\ associação) = \frac{37}{524} = 0{,}0706 \tag{32}$$

A diferença entre TCR(*Bag of labels*) e TCR(Extração de regras de associação) foi de 0,2055 o que corresponde a aproximadamente 20,5%. Essa diferença se refletiu nos valores das medidas de avaliação. Na avaliação realizada com o procedimento RAPR (seção 5.6), os valores médios de diferença, em termos das medidas *recall*, precisão e medida-F, para as 100 primeiras recomendações feitas por *Bag of labels* [61] e “Extração de regras de associação” [1] [2] foram de 0,0842; 0,0852 e 0,0678 respectivamente. Na avaliação realizada com o procedimento RMMA (seção 5.7) a diferença entre os valores máximos de *recall*, precisão e medida-F obtidos pelas duas técnicas foram 0; 0,1923 e 0,0836 respectivamente.

Nos experimentos das seções 5.8 e 5.9, que consistiram na aplicação dos procedimentos de avaliação RAPR e RMMA às extrações a partir do corpus *Family law doctrine*, os valores para TCR(*Bag of labels*) e

TCR(Extração de regras de associação) foram calculados conforme as equações 33 e 34.

$$TCR(Bag\ of\ labels) = \frac{34}{108} = 0{,}3148 \quad (33)$$

$$TCR(Extração\ de\ regras\ de\ associação) = \frac{34}{551} = 0{,}0617 \quad (34)$$

A diferença entre TCR(*Bag of labels*) e TCR(Extração de regras de associação) foi de 0,2531 o que corresponde a aproximadamente 25%. Essa diferença se refletiu nos valores das medidas de avaliação. Na avaliação realizada com o procedimento RAPR (seção 5.8), os valores médios de diferença, em termos das medidas *recall*, precisão e medida-F, para as 100 primeiras recomendações feitas por *Bag of labels* [61] e "Extração de regras de associação" [1] [2] foram de 0,2035; 0,1715 e 0,1676 respectivamente. Na avaliação realizada com o procedimento RMMA (seção 5.9), a diferença entre os valores máximos de *recall*, precisão e medida-F obtidos pelas duas técnicas foram 0; 0,1709 e 0,1102 respectivamente.

Pelo exposto afirma-se então que *Bag of Labels* [61], por apresentar um menor ou igual número de relacionamentos recomendados que a "Extração de regras de associação" [1] [2] e consequentemente a tendência em apresentar maiores valores para as medidas de avaliação é uma solução mais efetiva para a fase de refinamento de ARNT.

Por fim, podemos generalizar as observações feitas sobre *Bag of labels* [61] e "Extração de regras de associação" [1] [2] para quaisquer soluções que usem relacionamentos dos tipos "<c_1, c_2>" e "<c_1, fv, c_2>" para a fase de "Refinamento" baseado em duas considerações. Em primeiro lugar, qualquer algoritmo que utilize relacionamentos "<c_1, fv, c_2>" no processo de refinamento terá que verificar a coincidência de três elementos para poder aglutiná-los da forma particular utilizada por cada algoritmo. O algoritmo "Extração de regras de associação" [1] [2], por exemplo, utiliza regras na forma "(c_1, c_2) → fv". Por outro lado, algoritmos que utilizem relacionamentos do tipo "<c_1, c_2>" no processo de refinamento terão que verificar a coincidência de apenas dois

elementos, o que aumenta significativamente a probabilidade de aglutinação de cada tupla. Esse fato reduz significativamente o número total de relacionamentos gerados pelo algoritmo (elemento "Qtd_rel_recomendados" da fórmula 27) e consequentemente tende a aumentar os valores das medidas de avaliação, como já discutido nessa seção. Em segundo lugar, em se tratando de avaliações, caso o critério de casamento adotado para técnicas que utilizem relacionamentos "$<c_1, c_2>$" seja o de coincidencia desses dois elementos com os correspondentes de um relacionamento de referência; relacionamentos que não casariam com um recomendado por uma técnica do tipo "$<c_1, fv, c_2>$", pela não coicidencia apenas do rótulo, casariam com a recomendação feita pela técnica do tipo "$<c_1, c_2>$". A consequencia seria um aumento do TCR da técnica do tipo "$<c_1, c_2>$" e um provável aumento de sua efetividade em termos de medidas de avaliação.

Concluímos então, baseado em todo o exposto, que soluções para a fase de "Refinamento" que utilizam relacionamentos do tipo "$<c_1, c_2>$" são mais efetivas que as que utilizam "$<c_1, fv, c_2>$", uma vez que para um mesmo conjunto de relacionamentos candidatos e ontologia de referência tendem a apresentar melhores resultados em termos das medidas de avaliação.

5.11. Considerações finais

Nesse capítulo foram definidos dois procedimentos de avaliação de técnicas de ARNT, ambos baseados no princípio da comparação dos resultados obtidos com uma ontologia de referência. O procedimento RAPR (seção 5.5.1) tem o objetivo de avaliar comparativamente as técnicas em termos de uma medida de avaliação cujo valor é calculado para grupos de relacionamentos por elas sugeridos. Os relacionamentos sugeridos devem estar ordenados por algum dos parâmetros da solução de refinamento e os grupos devem ser considerados em igual aridade. Já o procedimento RMMA (seção 5.5.2) tem o objetivo de avaliar comparativamente técnicas de ARNT em termos de uma medida de avaliação a ser maximizada. Dessa forma, cada técnica deve ser executada com a configuração que lhe permita obter o maior valor para a medida de avaliação considerada.

Para verificar a hipótese de pesquisa postulada nesse trabalho as soluções para a fase de refinamento de ARNT, *Bag of labels* [61] e "Extração de regras de associação" [1] [2] foram avaliadas na aprendizagem de relacionamentos a partir dos corpora Genia [14] [53] e *Family law doctrine* [62]. Os resultados obtidos foram discutidos e então generalizados para quaisquer soluções que recomendem relacionamentos dos tipos "$<c_1, c_2>$" e "$<c_1, fv, c_2>$". O algoritmo de "Extração de regras de associação" [1] [2] foi o adotado para efeito da verificação da hipótese de pesquisa, uma vez que dentre as soluções de refinamento utilizadas pelas técnicas de ARNT avaliadas nesse trabalho [29] [42] [48] [55] [56] [69] (capítulo 3) é a que emprega relacionamentos do tipo "$<c_1, fv, c_2>$".

O próximo capítulo apresenta uma discussão sobre os resultados científicos e tecnológicos alcançados e limitações referentes a esse trabalho. São ainda discutidos trabalhos que podem vir a ser desenvolvidos a partir da presente pesquisa.

6. Conclusão

Algumas técnicas tem sido propostas para Aprendizagem de Relacionamentos Não-Taxonômicos de ontologias (ARNT) [29] [42] [48] [55] [56] [69]. Todas elas utilizam técnicas de Processamento de Linguagem Natural (PLN) [4] [8] [19] [22] [23] [68] para anotar o corpus com informações necessárias ao processamento subsequente e Aprendizagem de Máquina (AM) [9] [46] [49] [50] para refinar os relacionamentos provenientes da fase “Extração de relacionamentos” do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3).

As técnicas de ARNT [29] [42] [48] [55] [56] [69] da mesma forma que outras na área de Aprendizagem de Ontologias (AO) [10] [11] [15] [16] [17] [33] [37] estão sujeitas a uma grande quantidade de ruído uma vez que a fonte de informação da qual extraem os relacionamentos ser não estruturada. Portanto, soluções customizáveis são necessárias para que essas técnicas sejam aplicáveis a maior variedade possível de situações.

O presente trabalho apresentou TARNT [60] [61] (capítulo 4), uma técnica semi-automática para a aprendizagem de relacionamentos não-taxonômicos de ontologias a partir de textos na língua inglesa que utiliza técnicas de PLN [4] [8] [19] [22] [23] [68] e estatísticas [61]. O controle sobre a execução de suas regras de extração e consequentemente sobre o *recall* e precisão na fase "Extração de relacionamentos candidatos"; a “regra de apóstrofo” [61] (seção 4.2.3), que confere tratamento particular às extrações que tem maior probabilidade de serem relacionamentos válidos e *Bag of labels* [61] (seção 4.2.4.2), solução para a fase de refinamento que prioriza a informação de quais pares de conceitos estão relacionados, mas ao mesmo tempo recomenda rótulos ao relacionamento, assim como a técnica de “ARNT baseada na Extração de regras de associação” [69] estão entre seus aspectos positivos.

Avaliações experimentais de TARNT [60] [61] foram realizadas conforme os procedimentos RAPR (Recomendação de relacionamentos com anulação dos paramentros de refinamento) e RMMA (Recomendação de relacionamentos com maximização da medida de avaliação), baseados no

princípio de comparação dos relacionamentos aprendidos com os de referência. Esses experimentos consistiram em mensurar com as medidas de avaliação *recall*, precisão e medida-F, definidas na seção 2.3, a efetividade da técnica na aprendizagem de relacionamentos não-taxonômicos a partir de dois corpora nos domínios da biologia [14] [53] (seção 5.3) e do direito da família [62] (seção 5.4). Os resultados obtidos foram ainda comparados aos de outra abordagem (seção 5.2) que utiliza o algoritmo de extração de regras de associação [1] [2] na fase de "Refinamento". Uma discussão detalhada sobre os resultados dessas avaliações comparativas foi apresentado nas seções 5.6 a 5.9.

Esse trabalho demonstrou ainda a hipótese de pesquisa de que: soluções para a fase de refinamento de ARNT que recomendam relacionamentos compostos por dois conceitos de uma ontologia e um rótulo são menos efetivas que as que recomendam relacionamentos compostos apenas por dois conceitos. Formalmente, diz-se que soluções de refinamento que recomendam relacionamentos do tipo "<c_1, fv, c_2>" ("c_1", "c_2" e "fv" são respectivamente conceitos de uma ontologia e uma frase verbal obtida do corpus, ver seção 2.4) são menos efetivas que os que recomendam relacionamentos do tipo "<c_1, c_2>", uma vez que tendem a apresentar menores valores para as medidas de avaliação quando considerados os mesmos corpus e ontologia de referência.

A demonstração foi realizada por meio de uma exposição teórica (seção 5.10) que consistiu na generalização das observações realizadas sobre os resultados obtidos pelas técnicas de refinamento *Bag of labels* [61] e "Extração de regras de associação" [1] [2] que recomendam respectivamente relacionamentos dos tipos "<c_1, c_2>" e "<c_1, fv, c_2>".

6.1. Resultados Científicos e Tecnológicos

Na busca para alcançar o objetivo dessa proposta de Tese, que é o de propor uma técnica para a Aprendizagem de relacionamentos não-taxonômicos de ontologias e a verificação da hipótese de pesquisa de que soluções para a fase de refinamento de ARNT que recomendam

relacionamentos compostos por dois conceitos de uma ontologia e um rótulo são menos efetivas que as que recomendam relacionamentos compostos apenas por dois conceitos, contribuições para esse campo da AO [10] [11] [15] [16] [17] [33] [37] foram feitas e algumas delas são a seguir comentadas:

1. Realização de uma avaliação qualitativa das técnicas do estado da arte em ARNT [29] [42] [48] [55] [56] [69] (seção 3.6). Nela foram evidenciadas as soluções por elas adotadas para cada uma das fases do processo genérico de ARNT [57] [58] [59] [60] e discutidos seus aspectos positivos e limitações. Essa avaliação serviu como subsidio para propôr o processo genérico para ARNT [57] [58] [59] [60] e também TARNT [60] [61].
2. Definição de um processo genérico para ARNT [57] [58] [59] [60] (seção 2.3) baseado em técnicas do estado da arte [29] [42] [48] [55] [56] [69] (capítulo 3). Esse processo serve como *framework* para guiar o desenvolvimento de novas propostas para ARNT, além de permitir a avaliação comparativa de diferentes propostas em termos das soluções que adotam para cada uma de suas fases.
3. Desenvolvimento de TARNT [60] [61] (capítulo 4), uma técnica para a aprendizagem do conjunto “R” da definição formal de ontologias (seção 2.1). Esta técnica realiza a extração de forma semiautomática de relacionamentos não-taxonômicos de classes de ontologias utilizando técnicas de PLN [61] e estatísticas [61].

   TARNT possui seis fases: na primeira, "construção do corpus", um conjunto de documentos em língua inglesa no domínio de interesse é construído; na fase anotação do corpus, técnicas de PLN são utilizadas para marcar o corpus com anotações necessárias à extração dos relacionamentos candidatos; na fase “Extração de relacionamentos”, regras de extração são utilizadas para obter do corpus previamente anotado os possíveis relacionamentos não-taxonômicos; na fase de “Refinamento” os relacionamentos candidatos são filtrados com o objetivo de sugerir ao engenheiro de conhecimento os mais prováveis; na

fase de “Avaliação do especialista”, esse seleciona e possivelmente edita os relacionamentos a serem acrescentados a ontologia; por fim, na fase “Atualização da ontologia”, um procedimento de atualização do arquivo OWL da ontologia com os relacionamentos não-taxonômicos aprendidos é executado.

4. Desenvolvimento de TARNTool (seção 4.4), uma ferramenta de software que provê suporte automatizado a aplicação de TARNT. TARNTool utiliza a API Gate [20] [21] na fase “Anotação do corpus” e implementa todas as regras de extração da fase “Extração de relacionamentos”: regra de sentença [61], regra de sentença com frase verbal [61] e regra de apóstrofo [61]. São ainda implementadas as soluções de refinamento, *Bag of labels* [61] e “Frequência de co-ocorrência” [61].
5. Desenvolvimento dos algoritmos *Bag of labels* [61] e “Frequência de co-ocorrência” [61] (seções 4.2.4.1 e 4.2.4.2, respectivamente) para a fase de “Refinamento” de TARNT. O algoritmo “Frequência de co-ocorrência” é a solução adotada por TARNT para o refinamento dos relacionamentos candidatos (produto da fase “Extração de relacionamentos candidatos”) extraídos pela regra de sentença [61] e pela regra de apóstrofo [61] (seção 4.2.3). O algoritmo *Bag of labels* é a solução adotada por TARNT para o refinamento dos relacionamentos candidatos extraídos pela regra de sentença com frase verbal (seção 4.2.3).
6. Especificação formal das regras de extração de TARNT, regra de sentença [61], regra de sentença com frase verbal [61] e regra de apóstrofo [61] (seção 4.2.3), com expressões regulares no padrão PCRE (“Perl Compatible Regular Expressions”) [39].
7. Separação dos relacionamentos candidatos extraídos pela regra de apóstrofo [61] (seção 4.2.3.3), dos relacionamentos extraídos pela regra de sentença [61] ou pela regra de sentença com frase verbal [61]. Essa característica é uma particularidade de TARNT em relação às demais técnicas avaliadas [29] [42] [48] [55] [56] [69] e permite que os relacionamentos candidatos extraídos pela

regra de apóstrofo, que tem maior probabilidade de serem relacionamentos válidos, sejam tratados de forma exclusiva nas fases subsequentes.

8. Desenvolvimento da ferramenta de software AERA-Tool (seção 5.2) que utiliza a técnica de "Extração de regras de Associação" [1] [2] na fase de "Refinamento". Sua construção envolveu a implementação da regra de sentença com frase verbal para a fase "Extração de relacionamentos" e utilização do algoritmo Apriori [46], para a extração de regras de associação. Essa ferramenta foi necessária à realização dos experimentos cujos resultados foram utilizados na demonstração da hipótese de pesquisa dessa Tese.
9. Definição do procedimento de avaliação de técnicas de ARNT RAPR, apresentado na seção 5.5.1. Esse procedimento tem o objetivo de avaliar comparativamente técnicas em termos de uma medida de avaliação cujo valor é calculado para grupos de relacionamentos por elas recomendados. Os relacionamentos recomendados devem estar ordenados por algum dos parâmetros da solução de refinamento e os grupos devem ser considerados em igual aridade. No contexto do presente trabalho, o procedimento RAPR foi utilizado na avaliação comparativa dos resultados obtidos pelas abordagens TARNT e AERA com o objetivo de demonstrar a hipótese de pesquisa.
10. Definição do procedimento de avaliação de técnicas de ARNT RMMA, apresentado na seção 5.5.2. Esse procedimento tem como objetivo avaliar comparativamente técnicas de ARNT em termos de uma medida de avaliação a ser maximizada. Dessa forma, cada técnica deve ser executada com a configuração que lhe permita obter o maior valor para a medida de avaliação considerada. No contexto do presente trabalho, esse procedimento também foi utilizado com o objetivo de demonstrar a hipótese de pesquisa.
11. Demonstração da hipótese de pesquisa definida nessa Tese (seção 1.3), por meio de uma exposição teórica que consistiu na

generalização de conclusões feitas sobre as técnicas de refinamento *Bag of labels* [61] e “Extração de regras de associação” [1] [2] (seção 5.10). A argumentação foi ainda confirmada experimentalmente pelos resultados obtidos na aprendizagem de relacionamentos a partir de dois corpora nos domínios da biologia [14] [53] e do direito da família [62] (seções 5.6 a 5.9).

6.2.Limitações

Algumas limitações de TARNT [60] [61] e também de TARNTool que merecem destaque, assim como suas justificativas e consequências são a seguir mencionadas:

1. Em sua atual versão, TARNT não aborda a tarefa custosa de construção do corpus, a primeira fase do processo genérico de ARNT [57] [58] [59] [60] (seção 2.3). Da mesma forma, TARNTool não provê suporte automatizado a construção do corpus, sendo o processo de aprendizagem dos relacionamentos iniciado na fase "Anotação do corpus" quando o especialista informa ao sistema o caminho dos arquivos no formato *txt* que constituem o corpus. Das técnicas do estado arte, “ARNT baseada na extração de regras de associação generalizadas” [42], “ARNT baseada na extração de regras de associação” [69] e “ARNT baseada na regressão logística” [29] semelhantemente a TARNT também não proveem soluções para a fase de construção do corpus. Por outro lado, “ARNT baseada em consultas a Web” [55] [56] automatiza essa tarefa por meio de consultas a mecanismos de busca Web conforme descrito na seção 3.3. Por entender que o resultado da ARNT é diretamente dependente da adequação do corpus a referida tarefa e que a inspeção visual e julgamento humano tendem sempre a conduzir a melhores resultados, em TARNT optou-se por atribuir ao engenheiro de conhecimento a responsabilidade sobre como proceder na fase de "Construção do corpus".

2. TARNT não realiza a aprendizagem de relacionamentos não-taxonômicos a partir de textos com instancias. A técnica “ARNT baseada na extração de regras de associação” [69], de forma semelhante à TARNT, também extrai somente de textos com classes. Já a “ARNT baseada na extração de regras de associação generalizadas” [42] e a “ARNT baseada na classificação de relacionamentos” [48] extraem de textos com instancias. A desvantagem de trabalhar com textos com instancias é que é necessário utilizar a técnica de PLN, Reconhecimento de entidades nomeadas [4] [8]. Essa técnica utiliza listas predefinidas de instancias (lista *gazzeter*) associadas a cada classe. Dessa forma o processo de reconhecimento das classes por meio de suas instancias fica dependente da abrangência dessas listas. Dependendo do domínio, as listas *gazzeters* disponibilizadas por ferramentas automatizadas de PLN, como o GATE [20] [21], podem ser insatisfatórias.
3. Em sua atual versão, TARNT não dispõem, nem TARNTool implementa, a funcionalidade de sugestão de nível hierárquico presente na técnica “ARNT baseada na extração de regras de associação generalizadas” [42] (seção 3.2). Entretanto, a disponibilização dessa funcionalidade em TARNT [60] [61] consiste basicamente em incluir o algoritmo de extração de regras de associação generalizadas [67] na fase de “Refinamento”; substituir o conjunto “C” dado como entrada para a fase “Extração de relacionamentos” pelo conjunto “H”; e também do desenvolvimento de um procedimento para retornar os ancestrais e os descendentes de determinado conceito da hierarquia da ontologia.
4. A aplicação de TARNT tem como pré-requisito a disponibilização do conjunto de conceitos da ontologia (conjunto C, seção 2.1) para o qual deseja-se realizar a ARNT. Entretanto, essa é uma exigência aceitável e realista também compartilhada pelas técnicas de “ARNT baseada na classificação de

relacionamentos" [48] e "ARNT baseada na extração de regras de associação" [69].

5. TARNT não possui a funcionalidade de busca por sinônimos dos conceitos da ontologia. Essa funcionalidade pode ser incluída na fase "Extração de relacionamentos" com o objetivo de possivelmente obter um acrescimo nos valores das medidas de avaliação como consequência da possível identificação de um maior número de conceitos no corpus. Para implementar essa proposta na atual versão de TARNTool pode-se recuperar os sinônimos dos conceitos da ontologia a partir de uma base léxica, como o difundido *Wordnet* [13] [30].
6. Este trabalho não apresenta estudos que mensurem a influência da busca por sinônimos dos conceitos da ontologia na fase "Extração de relacionamentos candidatos" de TARNT e que consequentemente recomendem a inclusão dessa funcionalidade a técnica. Essa verificação pode ser realizada experimentalmente realizando diferentes execuções de TARNT sobre um mesmo corpus e com a mesma configuração alterando-se apenas a busca ou não por sinônimos. Em seguida, os valores obtidos para as medidas de avaliação devem ser avaliados comparativamente. Tanto o procedimento RAPR, definido na seção 5.5.1 quanto o procedimento RMMA definido na seção 5.5.2, para avaliação de técnicas de ARNT podem ser utilizados para essa finalidade.
7. A funcionalidade de atualização do arquivo da ontologia com os relacionamentos não-taxonômicos obtidos pelo processo de aprendizagem, conforme previsto em TARNT na fase "Atualização da ontologia" (seção 4.2.6), não está implementada na atual versão de TARNTool. Entretanto, essa funcinalidade não era impressindível à realização dos experimentos de ARNT das seções 5.6 a 5.9 e tão pouco à verificação da hipótese de pesquisa.

6.3.Trabalhos Futuros

Alguns trabalhos que podem vir a ser desenvolvidos com relação a TARNT [60] [61] e TARTool são a seguir comentados:

1. Incluir a funcionalidade de resolução de co-referências [47] em TARNT e TARNTool. A resolução de coreferência é normalmente dividida conceitualmente e também implementada como duas técnicas de PLN: a resolução de coreferência nominal e pronominal. Essas técnicas de PLN seriam incluídas na fase "Anotação do corpus" com o objetivo de possivelmente obter um acrescimo nos valores das medidas de avaliação como consequência da possível identificação de um maior número de conceitos. Para operacionalizar essa proposta, na atual versão de TARNTool é necessário incluir os correspondentes *plugins* no *pipeline* de PLN na fase de "Anotação do corpus" e atualizar as regras de extração de relacionamentos para reconhecer as anotações feitas por essas técnicas.
2. Verificar experimentalmente a real influencia da inclusão da técnica de resolução de co-referência ao *pipeline* de PLN da fase de "Anotação do corpus" de TARNT. Para operacionalizar essa proposta, TARNT deve ser executada duas vezes sobre um mesmo corpus e com a mesma configuração sendo alteranda apenas a execução ou não da resolução de co-referência. Em seguida os valores para a medida de avaliação devem ser avaliados comparativamente. Tanto o procedimento RAPR, definido na seção 5.5.1 quanto o procedimento RMMA definido na seção 5.5.2 para avaliação de técnicas de ARNT podem ser utilizados para essa finalidade.
3. Disponibilização de uma nova versão do algoritmo *Bag of labels* [61], na qual as frases verbais em cada *bag* (conjunto de todas as frases verbais associadas a cada par de conceitos) tenham suas frequências calculadas. Com essa medida será disponibilizada uma informação adicional ao engenheiro de conhecimento para a escolha do rótulo do relacionamento.

4. Desenvolvimento de uma interface usuário para a ferramenta de software TARNTool. O objetivo é torná-la mais usual e intuitiva ao engenheiro de conhecimento.
5. Desenvolvimento de ferramentas de software que automatizem totalmente ou parcialmente os procedimentos de avaliação de técnicas de ARNT, RMMA e RAPR. RMMA tem como objetivo avaliar técnicas de ARNT em termos de uma medida de avaliação a ser maximizada. Dessa forma, cada técnica deve ser executada com a configuração que lhe permita obter o maior valor para a medida de avaliação considerada. RAPR tem como objetivo avaliar técnicas de ARNT em termos de uma medida de avaliação cujo valor é calculado para grupos de relacionamentos por elas sugeridos. Os relacionamentos sugeridos devem estar ordenados por algum dos parâmetros da solução de refinamento e os grupos devem ser considerados em igual aridade.
6. Nesta Tese, TARNT foi utilizada para a realização de Aprendizagem de relacionamentos não-taxonômicos a partir de corpora na lingua Inglesa. Entretanto, sugere-se como trabalho futuro sua adaptação para utilização com corpora na língua Portuguesa. Para tanto, é necessário que na fase de "Anotação do corpus" sejam utilizadas ferramentas de anotação para esta língua, como o NLTK [8] e o Palavras [7] e não o Gate [20] [21], como na atual versão. Na fase de "Extração de relacionamentos", as regras de sentença e de sentença com frase verbal podem ser mantidas, entretanto a regra de apótrofo deve ser excluída. Na fase de "Refinamento", ambas as soluções, *Bag of labels* e Frequêcia de co-ocorrência podem ser utilizadas. Além disso, para permitir a busca por sinônimos dos conceitos da ontologia para a qual os relacionamentos não-taxonômicos estejam sendo aprendidos, em uma possível extenção de TARNT, a base léxica para lingua portuguesa *WordNet Br* [26] pode ser utilizada.
7. Neste trabalho foi realizada a demonstração da hipótese de pesquisa postulada na seção 1.3 de que soluções de refinamento de ARNT que recomendam relacionamentos não-taxonômicos do

tipo “<c_1, fv, c_2>” são menos efetivas que as que recomendam relacionamentos do tipo “<c_1, c_2>”, uma vez que tendem a apresentar menores valores para as medidas de avaliação quando executados no mesmo contexto. Entretanto, assim como esta, outras hipóteses que levam em consideração a influência do tipo de representação de relacionamentos adotado na efetividade das técnicas de refinamento e de ARNT podem ser definidas e demonstradas. Representações não contempladas nesta Tese foram “<fn_1, fv, fn_2>” e “<fn_1, fn_2>”. Iniciativas nesse sentido representam relevante contribuição teórica à área de ARNT e são sugeridas como trabalho futuro.

8. Neste trabalho, TARNTool foi utilizada para automatizar a aplicação de TARNT, com a finalidade de mensurar sua efetividade em termos de medidas de avaliação (seção 2.3). Entretanto, nenhuma avaliação qualitativa da ferramenta TARNTool foi realizada por especialistas de domínio para mensurar sua adequação a tarefa e conformidade com TARNT, além de outros aspectos como usabilidade e desempenho. Estas questões são sugeridas como trabalhos futuros.
9. Nesta Tese foram apresentadas e discutiudas avaliações quantitativas de duas abordagens para ARNT: TARNT e AERA. Estes experimentos consitiram em avaliar a efetividade dessas técnicas na Aprendizagem de relacionamentos não-taxonômicos com o uso de métricas bem definidas (*recall*, precisão e medida-F) e conforme os procedimentos de avaliação RAPR e RMMA. Entretanto, avalições incluindo outras técnicas de ARNT podem ser executadas com o objetivo de mensurar comparativamente a eficácia dessas abordagens e de identificar as que, considerando o mesmo contexto, conduzem a melhores resultados em termos de medidas de avaliação.

7. Publicações

As três seguintes publicações se referem ao estado da arte de ARNT e ao processo genérico de ARNT proposto no presente trabalho e correspondem principalmente aos conteúdos dos capítulos 2 e 3 desse manuscrito:

- Serra, I., Girardi, R. Extracting Non-taxonomic Relationships of Ontologies from Texts, Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications - SOCO 2011, Ed. Springer, vol. 87, pp. 329-338. Salamanca, Spain. April 6th to 8th, 2011. Qualis Capes B4.

- Serra, I., Girardi, R., Novais, P. The Problem of Learning Non-Taxonomic Relationships of Ontologies from Text, Proceedings of the 9th International Symposium on Distributed Computing and Artificial Intelligence, Ed. Springer, vol. 151, pp. 485-492. Salamanca, Spain. 28-30 March de 2012. Qualis Capes B4.

- Serra, I., Girardi, R., Novais, P. Reviewing the Problem of Learning Non-Taxonomic Relationships of Ontologies from Text, International Journal of Semantic Computing. Vol. 6-4, December 2012. Qualis Capes B4 Ciência da computação e B5 em Engenharias IV.

As cinco seguintes publicações se referem à abordagem para ARNT proposta nesse trabalho e sua avaliação e correspondem principalmente aos conteúdos dos capítulos 4 e 5 desse manuscrito:

- Serra, I., Girardi, R. A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text, Intelligent Information Management, Vol. 3 No. 4, pp. 119-124, 2011. Qualis Capes B5 em Engenharias IV.

- Serra, I., Girardi, R., Novais, P. PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text. In: 2013 Tenth International Conference on Information Technology: New Generations (ITNG), 2013, Las Vegas. 2013 10th International Conference on Information Technology: New Generations. p. 561-566. Qualis Capes B1.

- Serra I. Girardi R. A Comparative Evaluation of Refinement Solutions for Learning Non-taxonomic Relationships of Ontologies from Text, artigo submetido em 4.12.2013 para Transactions on Knowledge and Data Engineering. Qualis Capes A1 em Ciência da Computação e A1 Interdisciplinar.

- Serra I. Girardi R. An Effective Solution for Learning Non-taxonomic Relationships of Ontologies from Text, artigo submetido em 19.12.2013 para Knowledge-Based Systems. Qualis Capes B1 em Ciência da Computação e A2 em Engenharias IV.

- Serra, I., Girardi, R., Novais, P. Evaluating Techniques for Learning Non-Taxonomic Relationships of Ontologies from Text, artigo submetido em 28.12.2013 para Expert Systems with Applications. Qualis Capes A1 em Ciência da Computação e A2 em Engenharias IV.

Referências

[1] Agrawal, R., Imielinski, T. and Swami, A. N. Mining association rules between sets of items in large databases. Proceedings of the International Conference on Management of Data, ACM SIGMOD'93, Washington, D.C., USA, page 207-216, May 1993.

[2] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules, In J. B. Bocca, M. Jarke and C. Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, p.p. 478-499, June 1994.

[3] Alexiev, V., Breu, M., de Bruijn, J., Fensel, D., Lara, R. and Lausen, H., Information Integration with Ontologies: Experiences from an Industrial Showcase, Wiley, 2005.

[4] Allen, J. Natural Language Understanding. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc., 1995.

[5] Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval. New York: ACM Press, 1999.

[6] Benz, D. Collaborative Ontology Learning. Dissertação. 2007. 95 f. Dissertação (Mestrado) - Universität Freiburg, Freiburg, 2007.

[7] Bick, Eckhard. The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework, PhD thesis, Arhus University, Arhus, Danemark, 2000.

[8] Bird, S., Klein, E., Loper, E. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. Editora O'Really Media, 2009.

[9] Bishop, C. M. Pattern Recognition and Machine Learning, Springer, 2006.

[10] Buitelaar, P., Cimiano, P. and Magnini, B. Ontology Learning from Text: An Overview. DFKI, Language Technology Lab. AIFB, University of Karlsruhe. ITC-irst. 2003.

[11] Buitelaar, P., Cimiano, P. and Magnini, P. Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2006.

[12] Bunescu, R. and Mooney, R. Learning to extract relations from the web using minimal supervision. In proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 576-583, 2007.

[13] Buscaldi, D., Rosso, P., Arnal, E. S. A WordNet-based Query Expansion method for Geographical Information Retrieval. Dpto. de Sistemas Informáticos y Computación (DSIC). Universidad Politécnica de Valencia. Spain: 2005.

[14] Buyko, E. et al. Automatically mapping an NLP core engine to the biology domain. In Proceedings of the ISMB 2006 joint BioLINK/Bio-Ontologies meeting, 2006.

[15] Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.

[16] Cimiano, P. et al. Ontology learning from text: Methods, applications, evaluation. In: IOS Press, cap. Learning taxonomic relations from heterogenous sources of evidence, 2005.

[17] Cimiano, P., Volker, J. and Studer, R. Ontologies on Demand? – A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. In Information, Wissenschaft und Praxis 57 (6-7): 315-320. October 2006.

[18] Cowie, J. and Lenhnert, W. Information Extraction. Communications of the ACM 39(1), 80-91, 1996.

[19] Cowie, J. and Wilks, Y. Information Extraction, Handbook of Natural Language Processing, Robert Dale, Hermann Moisl and Harold Somers, p. 241–260, 2000.

[20] Cunningham, H. Developing Language Processing Components with GATE - Version 5 (a User Guide). University of Shelfield, 2009.

[21] Cunningham, H. et al. GATE: A Framework and Graphical Development Environment for Robust PLN Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, jul. 2002.

[22] Cunningham, H. Information Extraction, Encyclopedia of Language and Linguistics, 2nd Edition, 2005.

[23] Dale, R., Moisl, H., Somers, H. L. Handbook of natural language processing. CRC Press, 2000.

[24] Dellschaft, K. Measuring the Similarity of Concept Hierarchies and its Influence on the Evaluation of Learning Procedures. 2005. 83 f. Dissertação (Mestrado) - Universität Koblenz Landau, Koblenz, 2005.

[25] Dellschaft, K., Staab, S. On how to perform a gold standard based evaluation of ontology learning. In: International Semantic Web Conference, 5., 2006, Athens. Proceedings. Springer, 2006. p. 228 - 241.

[26] Dias-da-Silva, B.C., Rocha, M.A.E, Nunes, M.G.V. Projeto Montagem da Base Wordnet para o Português do Brasil-Processo CNPq 552057/01-0. Relatório Técnico. Araraquara: FCL-Unesp, 52p, 2004.

[27] Drake, Miriam A. Encyclopedia of Library and Information Science: Lib-Pub. CRC Press, 2003.

[28] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., Soderland, S., Weld, D., and Yates, A. Web-scale information extraction in KnowItAU (preliminary results). In Proceedings of the 13th World Wide Web Conference (WWW), pages 100-109, 2004.

[29] Fader, A., Soderland, S., and Etzioni, O. Identifying Relations for Open Information Extraction. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July, 2011.

[30] Fellbaum, C. WordNet: An Electronic Lexical Database.Cambridge: MIT Press. 23-24 p. 1998

[31] Finin, T., Fritzson, R., McKay, D. and McEntire, R. KQML as an Agent Communication Language. In: Proceedings of the 3rd International Conference on Information and Knowledge management (CIKM'94), p. 456-463, 1994.

[32] Freitag, D. Information extraction from HTML: Application of a general machine learning approach. In Proceedings of the 15th Conference on Artificial Intelligence (AAAI'98). pp. 517-523, 1998.

[33] Girardi, R. Guiding Ontology Learning and Population by Knowledge System Goals, In: Proceedings of International Conference on

Knowledge Engineering and Ontology Development, Ed. INSTIIC, Valence, October, 2010.

[34] Gonzalez, M., Lima, V. L. S. Recuperação de Informação e Processamento da Linguagem Natural. XXIII Congresso da Sociedade Brasileira de Computação, Campinas. Anais do III Jornada de Mini-Cursos de Inteligência Artificial, Volume III, p.347-395, 2003.

[35] Gruber, T. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, p. 907-928, 1995.

[36] Guarino, N., Masolo, C. and Vetere, C. Ontoseek: Content-based Access to the web, IEEE Intelligent Systems, v. 14(3), p. 70-80, 1999.

[37] Haase, P., Volker, J. Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency. In: Workshop on uncertainty reasoning for the semantic web, Galway. p. 45 – 55, 2005.

[38] Hearst, M. Automatic acquisition of hyponyms from large text corpora. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14., 1992, Nantes. Morristown: Association For Computational Linguistics, 1992. p. 539 - 545.

[39] Jeffrey, E. F. Mastering Regular Expressions. O'Reilly Media, 3rd Edition, 2006.

[40] Koyuncu, M. and Yazici, A. "A fuzzy knowledge-based system for intelligent retrieval", IEEE Trans. Fuzzy Syst., vol. 13, no. 3, pp.317 - 330, 2005

[41] Lehmann, J., Hitzler, P. A refinement operator based learning algorithm for the alc description logic. In: International Conference on Inductive Logic Programming, 17, Corvallis. Berlin: Springer-verlag. p. 147 – 160, 2007.

[42] Maedche, A. and Staab, S. Mining non-taxonomic conceptual relations from text". In Knowledge Engineering and Knowledge Management. Methods, models and tools: 12th International Conference, EKAW 2000. Proceedings. 189-202. Berlin: Springer.

[43] Marcus, M., Santorini, B., Marcinkiewicz, M. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics:

Special Issue on Using Large Corpora, [S. l.], v. 19, n. 2, p. 313-330, 1993.

[44] Marinho, L., Buza, K., Schmidt-Thieme, L. Folksonomy-based Collabulary Learning. In Proceedings of the 7th International Semantic web conference. Berlin: Springer-verlag, 2008. p. 261 - 276.

[45] Marneffe, M., Manning, C. The Stanford typed dependencies representation. In: Workshop on cross-framework and cross-domain parser evaluation, 2008, Manchester. p. 1 - 8.

[46] Mitchell, T. Machine Learning, McGraw Hill, 1997.

[47] Mitkov, R., Orasan, C., Evans, R. The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting, In Proceedings of Corpora and NLP: Reflecting on Methodology Workshop, TALN'99, 1999.

[48] Mohamed, T., Junior, E. R. H. and Mitchell, T. Discovering Relations between Noun Categories. In Proceedings of the conference on empirical methods in natural language processing (EMNLP 2011), Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1447-1455, 2011.

[49] Monard, M. C. and Baranauskas, J. A. Conceitos sobre Aprendizado de Máquina. In: REZENDE, Solange Oliveira (coord.). Sistemas Inteligentes. Barueri: Manole, 2005.

[50] Monard, M. C. and Baranauskas, J. A. Indução de Regras e Árvores de Decisão. In: REZENDE, Solange Oliveira (coord.). Sistemas Inteligentes. Barueri: Manole, 2005.

[51] Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T. and Swartout, W. R. Enabling technology for knowledge sharing. AI Magazine, 12(3):16--36, 1991.

[52] Nunberg, G. The linguistics of punctuation. C.S.L.I. Lecture Notes, Number 18. Center of the study of language and Information, Stanford, CA, 1990.

[53] Rinaldi, F., Schneider, G., Kaljurand, K., Dowdal, J., Andronis, C., Persidis A. and Konstanti, O. Mining relations in the GENIA corpus. Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, pp. 61 - 68, 2004.

[54] Russell, S., Norvig, P. Inteligência artificial. Rio de Janeiro: Editora Campus, 2004.

[55] Sanchez, D. and Moreno, A. A methodology for knowledge acquisition from the web. International Journal of Knowledge-Based and Intelligent Enginnering Systems. 10(6), 453-475, 2006.

[56] Sanchez, D. and Moreno, A. Learning non-taxonomic relationships from web documents for domain ontology construction. Data and Knowledge Engineering, 64(3), 600-623, 2008

[57] Serra, I., Girardi, R. Extracting Non-taxonomic Relationships of Ontologies from Texts, Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications - SOCO 2011, Ed. Springer, vol. 87, pp. 329-338. Salamanca, Spain. April 6th to 8th, 2011.

[58] Serra, I., Girardi, R. A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text, Intelligent Information Management, Vol. 3 No. 4, pp. 119-124, 2011.

[59] Serra, I., Girardi, R., Novais, P. The Problem of Learning Non-Taxonomic Relationships of Ontologies from Text, Proceedings of the 9th International Symposium on Distributed Computing and Artificial Intelligence, Ed. Springer, vol. 151, pp. 485-492. Salamanca, Spain. 28-30 March de 2012.

[60] Serra, I., Girardi, R., Novais, P. Reviewing the Problem of Learning Non-Taxonomic Relationships of Ontologies from Text, International Journal of Semantic Computing. Vol. 6-4, December 2012.

[61] Serra, I., Girardi, R., Novais, P. PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text. In: Tenth International Conference on Information Technology: New Generations (ITNG), p. 561-566, 2013.

[62] Serra I. Girardi R. A Comparative Evaluation of Refinement Solutions for Learning Non-taxonomic Relationships of Ontologies from Text, Transactions on Knowledge and Data Engineering. Artigo submetido em 4.12.2013.

[63] Shamsfard, M., Barforoush, A. An introduction to HASTI: An ontology learning system. In Proceedings of the 6th Conference on artificial intelligence and soft computing, 2002, Banff. ACTA Press, 2002.

[64] Shamsfard, M. and Abdollahzadeh Barforoush, A., The state of the art in ontology learning: a framework for comparison, The Knowledge Engineering Review, v. 18(4), p. 293-316, 2003.

[65] Sirin, E., Hendler, J. and Parsia, B. Semi-automatic composition of web services using semantic descriptions. In: Proceedings of the ICEIS Workshop on Web Services: Modeling, Architecture and Infrastructure, 2002.

[66] Smith, B. Ontology. In: FLORIDI, L. (Ed.). The Blackwell guide to philosophy of computing and information. Malden: Blackwell, p. 155-166, 2003.

[67] Srikant, R. and Agrawal, R. Mining generalized association rules. In Proceedings of VLDB' 95, pages 407-419, 1995.

[68] Vieira, R., Lima, V. L. S. Lingüística Computacional: princípios e aplicações. Anais do Congresso da Sociedade Brasileira de Computação, v. 2, p. 47-88, Fortaleza: SBC, 2001.

[69] Villaverde, J., Persson, A., Godoy, D., Amandi, A. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Syst. Appl. 36(7): 10288-10294, 2009.

[70] Voorhees, E. and Tice, D. Building a question-answering test collection. In Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-2000) Athens, Greece. pp. 200–207, 2000.

[71] Welty, C. A. and Ide, N. Using the right tolls: enhancing retrieval from marked-up documents, Computers and Humanities, v. 33(10), p. 59-84, 1999.

[72] Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead M., Soderland, S. TextRunner: open information extraction on the web. In Proceedings of HLT: NAACL. 2007.

Anexo A – Conceitos da ontologia Genia

1	PROTEIN
2	CELL
3	PEPTIDE
4	AMINO ACID
5	DNA
6	NUCLEOTIDE
7	PROTEIN COMPLEX
8	RNA
9	CELLULAR PROCESS
10	TRANSLATION
11	GENE REGULATION
12	GENE EXPRESSION
13	MORPHOGENESIS
14	TRANSCRIPTION
15	CELL DIFFERENTIATION
16	BINDING
17	POSITIVE REGULATION
18	NEGATIVE REGULATION
19	VIRAL LIFE CICLE
20	CELL ADHESION
21	METABOLISM
22	CELL CULTURE
23	VIRUS
24	CELL COMPONENT
25	TISSUE
26	GENE
27	NUCLEIC ACID

Anexo B – Relacionamentos não-taxonômicos de referência (lematizados) da ontologia Genia

Nº	Conceito 1	Rótulo	Conceito 2
1	PROTEIN	MAKE OF	AMINO ACID
2	PROTEIN	SYNTHESIZE	TRANSLATION
3	DNA	CONTAIN	NUCLEOTIDE
4	CELL	HAVE	GENE
5	TRANSCRIPTION	CONTROL	CELL DIFFERENTIATION
6	NUCLEOTIDE	CONTAIN	RNA
7	CELL	HOST	VIRUS
8	GENE	MAKE OF	NUCLEIC ACID
9	CELL	CONTAIN	CELL COMPONENT
10	VIRUS	CONTAIN	NUCLEOTIDE
11	GENE	CONSIST OF	DNA
12	DNA	GENERATE	GENE EXPRESSION
13	PROTEIN	CREATE	CELLULAR PROCESS
14	PROTEIN	BE NEED	CELL ADHESION
15	CELL	MAKE OF	NUCLEOTIDE
16	VIRUS	HAVE	GENE
17	POSITIVE REGULATION	ENABLE	TRANSCRIPTION
18	GENE REGULATION	GENERATE	TRANSCRIPTION
19	TISSUE	NEED	CELL ADHESION
20	CELL	PRODUCE	CELL CULTURE
21	VIRUS	EXECUTE	BINDING
22	PROTEIN	MAKE OF	PROTEIN COMPLEX
23	GENE EXPRESION	INCLUDE	TRANSLATION
24	AMINO ACID	CONNECT	PEPTIDE
25	TRANSCRIPTION	READ	DNA
26	VIRUS	HAVE	RNA
27	PROTEIN COMPLEX	GENERATE	BINDING

Nº	Conceito 1	Rótulo	Conceito 2
28	CELL	HAVE	METABOLISM
29	PROTEIN	REGULATE	BINDING
30	TRANSCRIPTION	DERIVE	PEPTIDE
31	GENE REGULATION	LEAD	CELL DIFFERENTIATION
32	VIRUS	HAVE	DNA
33	TRANSCRIPTION	CREATE	RNA
34	NEGATIVE REGULATION	CANCEL	TRANSCRIPTION
35	GENE EXPRESSION	GENERATE	RNA
36	GENE REGULATION	REGULATE	MORPHOGENESIS
37	TRANSCRIPTION	ENABLE	BINDING
38	TISSUE	HAVE	CELL

Anexo C – Relacionamentos não-taxonômicos recomendados por TARNT a partir do corpus Genia

Lista de relacionamentos não-taxonômicos recomendados por TARNT quando executada com a configuração da tabela 36 (seção 5.6) e o corpus Genia [53]. Os relacionamentos de referência estão destacados.

Freq. de co-ocorrência	tupla	rótulos	Nº
0,0281	CELL,CELL	need, might play, demonstrate, deplete, do not show, should provide, provide, express, affect, represent, replicate, combine, may also play	1
0,0256	CELL,NUCLEOTIDE	govern, should provide, normally regulate, prepare, express, interfere, require, carry, make	2
0,0256	CELL,TRANSCRIPTION	conclude, do not affect, achieve, prepare, can induce, affect, participate, locate, reflect, produce	3
0,0232	GENE REGULATION,TRANSCRIPTION	generate, allow, can serve, compare, indicate, express, interfere, analyze, describe, be, produce, follow	4
0,0220	PROTEIN,AMINO ACID	be not, introduce, can provide, enter, prepare, may explain, inhibit, make	5
0,0220	PROTEIN,CELL	only induce, enter, express, fold, replicate, transform, require, characterize, go	6
0,0195	TRANSLATION,PROTEIN	serve, identify, permit, can induce, activate, should always be, participate, synthesize, go	7
0,0195	TRANSCRIPTION,CELL DIFFERENTIATION	evaluate, control, deplete, provide, have, select, induce, include, operate, correlate	8
0,0195	PROTEIN,TRANSCRIPTION	do not show, mediate, take, may also play, correlate, characterize, find, go	9
0,0183	DNA,NUCLEOTIDE	be not, connect, contribute, oppose, reduce, replicate, mediate, observe, contain	10
0,0171	NUCLEOTIDE,PROTEIN	consist, give, originate, play, can induce, alter, require, specify, synthesize	11
0,0147	CELL,VIRUS	trigger, host, generate, consist, involve, prepare, replicate	12
0,0147	GENE,NUCLEIC ACID	differentiate, determine, recognize, may also play, require, make, be, synthesize	13
0,0134	NUCLEOTIDE,PEPTIDE	induce, describe, establish, determine, translate, may not only play, follow	14
0,0134	AMINO ACID,TRANSCRIPTION	evaluate, do not show, give, construct, derive, do not give, interfere	15

0,0134	CELL,GENE	mediate, consist, may develop, have, derive, may differ	16
0,0134	PROTEIN,GENE	need, reveal, establish, express, carry	17
0,0134	CELL,PEPTIDE	give, have, play, highly induce, may not only play, regulate, be	18
0,0122	AMINO ACID,PROTEIN COMPLEX	should provide, may explain, may be, modify, specify, restrict, represent	19
0,0122	PROTEIN,DNA	obtain, appear, modify, reflect, alter, be	20
0,0122	VIRUS,NUCLEOTIDE	provide, remain, may develop, contain, may also play, highly induce, follow	21
0,0122	AMINO ACID,RNA	examine, normally regulate, contribute, prepare, match, receive	22
0,0122	TRANSCRIPTION,NUCLEOTIDE	do not contain, govern, fail, recognize, enter	23
0,0122	NUCLEOTIDE,PROTEIN COMPLEX	do not contain, govern, can serve, enable, encompass, reduce, treat	24
0,0110	NUCLEOTIDE,GENE EXPRESSION	confer, turn, implicate, differentiate, inhibit	25
0,0110	GENE,DNA	serve, consist, grow, require, be	26
0,0110	PROTEIN,PEPTIDE	do not express, consist, differentiate, express, know, treat	27
0,0110	DNA,GENE EXPRESSION	generate, normally regulate, lead, enter, reduce, regulate	28
0,0110	GENE,GENE EXPRESSION	relate, evaluate, permit, can provide, require, do not give, be	29
0,0110	PROTEIN,CELLULAR PROCESS	do not express, generate, combine, create, know	30
0,0110	GENE,CELL COMPONENT	should provide, depend, describe, involve, produce	31
0,0110	PROTEIN,CELL ADHESION	activate, be need, characterize, highly induce, suppress	32
0,0110	GENE,PROTEIN COMPLEX	evaluate, analyze, should always be, introduce	33
0,0098	TRANSLATION,CELL DIFFERENTIATION	deplete, allow, determine, go	34
0,0098	TRANSLATION,RNA	allow, observe, become, may also play, inhibit	35
0,0085	VIRUS,CELL DIFFERENTIATION	generate, require, correlate, find	36
0,0085	NUCLEOTIDE,NUCLEOTIDE	play, suggest, restrict, follow	37
0,0085	CELL,CELL COMPONENT	contain, enable, establish, indicate	38
0,0085	AMINO ACID,TISSUE	submit, permit, characterize, go, be	39

0,0085	PROTEIN,CELL DIFFERENTIATION	detect, consist, coordinate, reduce	40
0,0085	PROTEIN,RNA	encompass, express, highly induce, reduce, suppress	41
0,0085	VIRUS,GENE	trigger, fail, have, inhibit, interfere	42
0,0085	AMINO ACID,GENE	remain, have, require, specify, react, include	43
0,0085	TRANSLATION,PEPTIDE	might be, can be, originate, develop, oppose	44
0,0085	TRANSCRIPTION,PROTEIN COMPLEX	demonstrate, cause, can serve, permit, translate	45
0,0085	DNA,AMINO ACID	may develop, express, suppress, regulate	46
0,0073	PROTEIN,MORPHOGENESIS	reveal, translate, specify, employ	47
0,0073	CELL,RNA	reveal, normally regulate, become, may also play, know	48
0,0073	CELL DIFFERENTIATION,CELL	consist, may explain, indicate, require	49
0,0073	POSITIVE REGULATION,TRANSCRIPTION	turn, enable, derive	50
0,0073	NUCLEIC ACID,PEPTIDE	be not, may also play, occur, be	51
0,0073	GENE,PEPTIDE	grow, may be, combine, can provide	52
0,0073	PROTEIN,PROTEIN COMPLEX	need, express, make	53
0,0073	AMINO ACID,GENE EXPRESSION	do not express, consist, normally regulate	54
0,0073	PROTEIN,VIRUS	demonstrate, should provide, can be, fold	55
0,0073	CELL,GENE EXPRESSION	involve, create, synthesize	56
0,0073	CELL,DNA	turn, generate, may explain, enter, translate	57
0,0073	VIRUS,VIRUS	take, introduce, require, do not give	58
0,0073	TRANSLATION,CELL	should provide, originate, indicate, translate	59
0,0061	PROTEIN,GENE EXPRESSION	do not express, induce, permit, may also play	60
0,0061	AMINO ACID,AMINO ACID	conclude, appear, describe, may differ	61
0,0061	TISSUE,CELL ADHESION	need, describe, can provide, do not give	62
0,0061	TRANSCRIPTION,GENE EXPRESSION	cause, induce, contain	63
0,0061	NUCLEIC ACID,GENE REGULATION	need, do not contain, find, occur	64
0,0061	CELL,CELL CULTURE	do not involve, block, produce	65
0,0061	NUCLEOTIDE,MORPHOGENESIS	observe, enable, lead	66
0,0061	TRANSLATION,DNA	normally regulate, grow, be	67
0,0061	VIRUS,PROTEIN COMPLEX	operate, introduce, express	68

0,0061	VIRUS,BINDING	execute, encompass, may differ, represent	69
0,0061	GENE REGULATION,CELL	establish, ascribe, represent	70
0,0061	TRANSCRIPTION,GENE	detect, develop, affect	71
0,0061	CELL ADHESION,CELL	take, ascribe, produce	72
0,0061	AMINO ACID,TRANSLATION	express, alter, employ, be	73
0,0061	PROTEIN,GENE REGULATION	demonstrate, observe, establish	74
0,0049	CELL,BINDING	demonstrate, highly induce, include	75
0,0049	AMINO ACID,CELL DIFFERENTIATION	select, describe, lead, correlate	76
0,0049	NUCLEIC ACID,GENE EXPRESSION	may also play, block, react	77
0,0049	NUCLEOTIDE,TRANSLATION	transform, contain, occur	78
0,0049	NUCLEOTIDE,RNA	have, contain	79
0,0049	GENE EXPRESSION,TRANSLATION	specify, include, replicate	80
0,0049	TRANSLATION,GENE REGULATION	achieve, locate, reduce	81
0,0049	GENE,MORPHOGENESIS	take, identify	82
0,0049	NEGATIVE	do not show, cancel	83
0,0049	CELL,PROTEIN COMPLEX	allow, indicate, express	84
0,0049	NUCLEIC ACID,METABOLISM	serve, occur	85
0,0037	VIRUS,PEPTIDE	select, reflect, restrict	86
0,0037	TRANSLATION,PROTEIN COMPLEX	coordinate, create	87
0,0037	TRANSCRIPTION,BINDING	enable, be	88
0,0037	NUCLEOTIDE,GENE REGULATION	compare, become	89
0,0037	NUCLEOTIDE,CELL DIFFERENTIATION	become, correlate	90
0,0037	AMINO ACID,VIRUS	cause, may explain, require	91
0,0037	VIRUS,RNA	have, can induce	92
0,0037	PROTEIN COMPLEX,BINDING	generate, originate	93
0,0037	GENE,TISSUE	mediate, derive, decrease	94
0,0037	VIRAL LIFE CICLE,CELL COMPONENT	confer, do not have, develop	95
0,0037	CELL CULTURE,GENE	introduce, may differ, produce	96
0,0037	CELL COMPONENT,VIRUS	grow, possess	97
0,0037	TISSUE,CELL	have, select	98
0,0037	NUCLEIC ACID,VIRUS	induce, require	99
0,0037	PROTEIN,BINDING	transform, regulate	100
0,0037	TRANSCRIPTION,PEPTIDE	derive, carry	101
0,0037	NUCLEIC ACID,CELLULAR PROCESS	need, generate, may not only play	102
0,0037	CELL,POSITIVE REGULATION	confer, turn, do not have	103
0,0037	CELL CULTURE,TISSUE	confer, fail	104
0,0037	DNA,RNA	turn, originate, make	105
0,0024	GENE REGULATION,CELL DIFFERENTIATION	be not, lead	106
0,0024	CELL COMPONENT,RNA	fail, decrease	107

0,0024	CELL COMPONENT,TISSUE	have, make	108
0,0024	VIRUS,DNA	consist, have	109
0,0024	CELL ADHESION,DNA	implicate, produce	110
0,0024	VIRUS,TRANSLATION	differentiate, reduce	111
0,0024	NUCLEIC ACID,TISSUE	can induce, be	112
0,0024	TRANSCRIPTION,RNA	create	113
0,0024	TRANSLATION,CELL ADHESION	conclude, interfere	114
0,0024	VIRUS,GENE REGULATION	transform, require	115
0,0024	NUCLEIC ACID,BINDING	activate, permit	116
0,0024	AMINO ACID,PEPTIDE	connect	117
0,0024	TRANSCRIPTION,CELL ADHESION	recognize, permit	118
0,0024	AMINO ACID,GENE REGULATION	might play, obtain	119
0,0024	TRANSCRIPTION,DNA	select, read	120
0,0024	AMINO ACID,CELL ADHESION	coordinate, operate	121
0,0024	CELL COMPONENT,AMINO ACID	might be, interfere	122
0,0024	TRANSCRIPTION,VIRUS	participate	123
0,0024	TRANSCRIPTION,MORPHOGENESIS	induce, suppress	124
0,0024	VIRUS,GENE EXPRESSION	do not affect	125
0,0024	NUCLEOTIDE,BINDING	follow	126
0,0024	TRANSCRIPTION,TISSUE	mediate, derive	127
0,0012	GENE REGULATION,MORPHOGENESIS	regulate	128
0,0012	VIRUS,MORPHOGENESIS	carry	129
0,0012	AMINO ACID,MORPHOGENESIS	activate	130
0,0012	CELL,MORPHOGENESIS	do not have	131
0,0012	CELL COMPONENT,DNA	permit	132
0,0012	GENE EXPRESSION,RNA	generate	133
0,0012	PROTEIN,TISSUE	differentiate	134

Anexo D – Relacionamentos não-taxonômicos recomendados por AERA a partir do corpus Genia

Lista de relacionamentos não-taxonômicos recomendados por AERA quando executada com a configuração da tabela 35 (seção 5.6) e o corpus Genia [53]. Os relacionamentos de referência estão destacados.

Conf.	Sup.	Regra de associação	Nº
1,0000	0,0019	CELL COMPONENT,DNA => permit	1
1,0000	0,0019	VIRUS,MORPHOGENESIS => carry	2
1,0000	0,0038	TRANSCRIPTION,VIRUS => participate	3
1,0000	0,0038	VIRUS,GENE EXPRESSION => do not affect	4
1,0000	0,0019	GENE EXPRESSION,RNA => generate	5
1,0000	0,0038	NUCLEOTIDE,BINDING => follow	6
1,0000	0,0038	AMINO ACID,PEPTIDE => connect	7
1,0000	0,0019	AMINO ACID,MORPHOGENESIS => activate	8
1,0000	0,0019	CELL,MORPHOGENESIS => do not have	9
1,0000	0,0019	GENE REGULATION,MORPHOGENESIS => regulate	10
1,0000	0,0038	TRANSCRIPTION,RNA => create	11
1,0000	0,0019	PROTEIN,TISSUE => differentiate	12
0,7500	0,0057	NUCLEOTIDE,RNA => contain	13
0,7500	0,0057	NEGATIVE REGULATION,TRANSCRIPTION => do not show	14
0,6667	0,0038	TRANSLATION,PROTEIN COMPLEX => coodinate	15
0,6667	0,0038	TRANSCRIPTION,PEPTIDE => derive	16
0,6667	0,0038	CELL COMPONENT,VIRUS => grow	17
0,6667	0,0038	TRANSCRIPTION,BINDING => be	18
0,6667	0,0038	PROTEIN,BINDING => regulate	19
0,6667	0,0038	TRANSCRIPTION,BINDING => enable	20
0,6667	0,0038	NUCLEOTIDE,CELL DIFFERENTIATION => correlate	21
0,6667	0,0038	TISSUE,CELL => have	22
0,6667	0,0038	NUCLEIC ACID,VIRUS => require	23
0,6667	0,0038	PROTEIN COMPLEX,BINDING => generate	24
0,6667	0,0038	CELL CULTURE,TISSUE => serve	25
0,6667	0,0038	VIRUS,RNA => have	26
0,5000	0,0057	AMINO ACID,GENE EXPRESSION => regulate	27
0,5000	0,0019	NUCLEIC ACID,TISSUE => be	28
0,5000	0,0038	GENE EXPRESION,TRANSLATION => include	29
0,5000	0,0019	NUCLEIC ACID,TISSUE => can induce	30
0,5000	0,0038	GENE,MORPHOGENESIS => identify	31
0,5000	0,0019	TRANSCRIPTION,DNA => read	32
0,5000	0,0038	NUCLEIC ACID,METABOLISM => serve	33
0,5000	0,0019	GENE REGULATION,CELL DIFFERENTIATION => lead	34

0,5000	0,0019	CELL COMPONENT,RNA => develop	35
0,5000	0,0019	CELL COMPONENT,TISSUE => determine	36
0,5000	0,0019	VIRUS,DNA => have	37
0,5000	0,0019	CELL ADHESION,TISSUE => construct	38
0,5000	0,0019	VIRUS,TRANSLATION => have show	39
0,5000	0,0038	CELL,BINDING => coordinate	40
0,5000	0,0019	TRANSLATION,CELL ADHESION => synthesize	41
0,5000	0,0019	VIRUS,GENE REGULATION => use	42
0,5000	0,0019	NUCLEIC ACID,BINDING => be observe	43
0,5000	0,0019	TRANSCRIPTION,CELL ADHESION => form	44
0,5000	0,0019	AMINO ACID,GENE REGULATION => use	45
0,5000	0,0038	PROTEIN,PROTEIN COMPLEX => make	46
0,5000	0,0019	AMINO ACID,CELL ADHESION => include	47
0,5000	0,0019	CELL COMPONENT,AMINO ACID => represent	48
0,5000	0,0019	TRANSCRIPTION,MORPHOGENESIS => be induce	49
0,5000	0,0019	TRANSCRIPTION,TISSUE => form	50
0,5000	0,0019	GENE REGULATION,CELL DIFFERENTIATION => lead	51
0,5000	0,0019	TRANSCRIPTION,MORPHOGENESIS => show	52
0,5000	0,0038	NUCLEIC ACID,GENE EXPRESSION => depend	53
0,5000	0,0038	TRANSLATION,GENE REGULATION => trigger	54
0,5000	0,0038	GENE,MORPHOGENESIS => use	55
0,5000	0,0038	NUCLEIC ACID,METABOLISM => serve	56
0,5000	0,0019	GENE REGULATION,CELL DIFFERENTIATION => transform	57
0,5000	0,0019	CELL COMPONENT,RNA => implicate	58
0,5000	0,0057	POSITIVE REGULATION,TRANSCRIPTION => enable	59
0,5000	0,0019	VIRUS,DNA => be	60
0,5000	0,0019	CELL ADHESION,TISSUE => function	61
0,5000	0,0019	VIRUS,TRANSLATION => indicate	62
0,5000	0,0038	NUCLEOTIDE,TRANSLATION => activate	63
0,5000	0,0019	TRANSLATION,CELL ADHESION => link	64
0,5000	0,0019	VIRUS,GENE REGULATION => be also require	65
0,5000	0,0019	NUCLEIC ACID,BINDING => modify	66
0,5000	0,0019	TRANSCRIPTION,CELL ADHESION => connect	67
0,5000	0,0019	AMINO ACID,GENE REGULATION => be obtain	68
0,5000	0,0019	TRANSCRIPTION,DNA => be take out	69
0,5000	0,0019	AMINO ACID,CELL ADHESION => obtain	70
0,5000	0,0019	CELL COMPONENT,AMINO ACID => employ	71
0,5000	0,0057	CELL,GENE EXPRESSION => form	72
0,5000	0,0019	TRANSCRIPTION,TISSUE => be associate	73
0,5000	0,0019	CELL COMPONENT,TISSUE => can provide	74

0,4286	0,0057	CELL,CELL COMPONENT => can be find	75
0,4000	0,0038	TRANSLATION,DNA => be activate	76
0,4000	0,0038	TRANSCRIPTION,GENE => contain	77
0,4000	0,0038	NUCLEOTIDE,PEPTIDE => identify	78
0,4000	0,0038	TISSUE,CELL ADHESION => need	79
0,4000	0,0038	CELL ADHESION,CELL => activate	80
0,4000	0,0038	TRANSCRIPTION,GENE => produce	81
0,4000	0,0038	TRANSCRIPTION,GENE EXPRESSION => associate	82
0,4000	0,0038	NUCLEOTIDE,PEPTIDE => be induce	83
0,4000	0,0038	TRANSLATION,DNA => cause	84
0,4000	0,0038	VIRUS,PROTEIN COMPLEX => exhibit	85
0,4000	0,0038	GENE REGULATION,CELL => result	86
0,4000	0,0038	CELL,CELL CULTURE => produce	87
0,4000	0,0038	CELL ADHESION,CELL => be require	88
0,4000	0,0038	AMINO ACID,TRANSLATION => consist	89
0,4000	0,0038	PROTEIN,GENE REGULATION => differentiate	90
0,4000	0,0038	PROTEIN,GENE EXPRESSION => be know	91
0,4000	0,0038	TRANSCRIPTION,GENE EXPRESSION => be characterize	92
0,4000	0,0038	NUCLEIC ACID,GENE REGULATION => sustain	93
0,4000	0,0038	CELL,CELL CULTURE => be observe	94
0,4000	0,0038	VIRUS,PROTEIN COMPLEX => serve	95
0,4000	0,0038	GENE REGULATION,CELL => serve	96
0,4000	0,0038	VIRUS,BINDING => execute	97
0,4000	0,0038	PROTEIN,GENE REGULATION => produce	98
0,4000	0,0038	AMINO ACID,AMINO ACID => be develop	99
0,3750	0,0057	TRANSLATION,CELL DIFFERENTIATION => oppose	100
0,3333	0,0038	NUCLEIC ACID,PEPTIDE => serve	101
0,3333	0,0057	GENE,PROTEIN COMPLEX => need	102
0,3333	0,0038	PROTEIN,MORPHOGENESIS => treat	103
0,3333	0,0038	POSITIVE REGULATION,TRANSCRIPTION => involve	104
0,3333	0,0057	GENE,DNA => consist of	105
0,3333	0,0038	PROTEIN,VIRUS => carry	106
0,3333	0,0038	CELL,DNA => derive	107
0,3333	0,0038	VIRUS,VIRUS => confer	108
0,3333	0,0019	VIRUS,PEPTIDE => be characterize	109
0,3333	0,0019	TRANSCRIPTION,BINDING => be identify	110
0,3333	0,0019	NUCLEOTIDE,GENE REGULATION => regulate	111
0,3333	0,0019	NUCLEOTIDE,CELL DIFFERENTIATION => be functionally link	112
0,3333	0,0019	AMINO ACID,VIRUS => be absolutely require	113
0,3333	0,0019	PROTEIN COMPLEX,BINDING => regulate	114

0,3333	0,0057	PROTEIN,CELLULAR PROCESS => create	115
0,3333	0,0019	VIRAL LIFE CICLE,CELL COMPONENT => use	116
0,3333	0,0019	CELL CULTURE,GENE => encode	117
0,3333	0,0019	NUCLEIC ACID,VIRUS => be regulate	118
0,3333	0,0019	PROTEIN,BINDING => be associate	119
0,3333	0,0019	TRANSCRIPTION,PEPTIDE => provide	120
0,3333	0,0019	NUCLEIC ACID,CELLULAR PROCESS => be involve	121
0,3333	0,0019	CELL,POSITIVE REGULATION => be closely relate	122
0,3333	0,0019	CELL CULTURE,TISSUE => form	123
0,3333	0,0019	DNA,RNA => find	124
0,3333	0,0019	GENE,TISSUE => support	125
0,3333	0,0057	GENE,CELL COMPONENT => can lead	126
0,3333	0,0038	GENE,PEPTIDE => define	127
0,3333	0,0038	PROTEIN,VIRUS => require	128
0,3333	0,0019	VIRUS,PEPTIDE => might be	129
0,3333	0,0019	TRANSLATION,PROTEIN COMPLEX => relate	130
0,3333	0,0019	AMINO ACID,VIRUS => operate	131
0,3333	0,0019	VIRUS,RNA => determine	132
0,3333	0,0019	GENE,TISSUE => be compare	133
0,3333	0,0019	VIRAL LIFE CICLE,CELL COMPONENT => obtain	134
0,3333	0,0019	CELL CULTURE,GENE => test	135
0,3333	0,0019	CELL COMPONENT,VIRUS => be require	136
0,3333	0,0019	TISSUE,CELL => be locate	137
0,3333	0,0019	NUCLEIC ACID,CELLULAR PROCESS => cause	138
0,3333	0,0057	PROTEIN,CELL ADHESION => be need	139
0,3333	0,0019	DNA,RNA => do not contain	140
0,3333	0,0057	GENE,PROTEIN COMPLEX => code	141
0,3333	0,0038	CELL,RNA => rest	142
0,3333	0,0038	CELL DIFFERENTIATION,CELL => be not specify	143
0,3333	0,0038	GENE,PEPTIDE => result	144
0,3333	0,0038	PROTEIN,PROTEIN COMPLEX => be	145
0,3333	0,0038	AMINO ACID,GENE EXPRESSION => control	146
0,3333	0,0038	CELL,GENE EXPRESSION => prepare	147
0,3333	0,0038	TRANSLATION,CELL => produce	148
0,3333	0,0019	VIRUS,PEPTIDE => be associate	149
0,3333	0,0019	AMINO ACID,VIRUS => obtain	150
0,3333	0,0019	GENE,TISSUE => be not	151
0,3333	0,0019	VIRAL LIFE CICLE,CELL COMPONENT => support	152
0,3333	0,0019	CELL CULTURE,GENE => have not be identify	153
0,3333	0,0019	NUCLEIC ACID,CELLULAR PROCESS => find	154

0,3333	0,0019	CELL,POSITIVE REGULATION => determine	155
0,3333	0,0019	DNA,RNA => provide	156
0,3333	0,0057	NUCLEOTIDE,GENE EXPRESSION => do not express	157
0,3333	0,0019	CELL,POSITIVE REGULATION => do not have	158
0,3333	0,0038	PROTEIN,MORPHOGENESIS => analyze	159
0,3333	0,0038	CELL DIFFERENTIATION,CELL => coordinate	160
0,3333	0,0038	NUCLEIC ACID,PEPTIDE => be	161
0,3333	0,0038	VIRUS,VIRUS => be not activate	162
0,3333	0,0038	TRANSLATION,CELL => recognize	163
0,3000	0,0057	TRANSCRIPTION,NUCLEOTIDE => be also observe	164
0,2857	0,0038	VIRUS,CELL DIFFERENTIATION => contain	165
0,2857	0,0038	AMINO ACID,TISSUE => can be	166
0,2857	0,0038	TRANSCRIPTION,PROTEIN COMPLEX => detect	167
0,2857	0,0038	DNA,AMINO ACID => compose	168
0,2857	0,0038	CELL,CELL COMPONENT => contain	169
0,2857	0,0038	NUCLEOTIDE,NUCLEOTIDE => release	170
0,2857	0,0038	PROTEIN,CELL DIFFERENTIATION => have be show	171
0,2857	0,0038	VIRUS,CELL DIFFERENTIATION => clone	172
0,2857	0,0038	TRANSLATION,PEPTIDE => be submit	173
0,2857	0,0038	DNA,AMINO ACID => possess	174
0,2857	0,0038	NUCLEOTIDE,NUCLEOTIDE => may also play	175
0,2857	0,0038	PROTEIN,CELL DIFFERENTIATION => have	176
0,2857	0,0038	TRANSLATION,PEPTIDE => belong	177
0,2857	0,0038	TRANSCRIPTION,PROTEIN COMPLEX => treat	178
0,2857	0,0038	VIRUS,CELL DIFFERENTIATION => should provide	179
0,2857	0,0038	NUCLEOTIDE,NUCLEOTIDE => do not show	180
0,2857	0,0038	VIRUS,GENE => have	181
0,2857	0,0038	AMINO ACID,TISSUE => occur	182
0,2857	0,0038	PROTEIN,CELL DIFFERENTIATION => represent	183
0,2857	0,0038	PROTEIN,RNA => compare	184
0,2857	0,0038	DNA,AMINO ACID => might play	185
0,2857	0,0038	PROTEIN,RNA => contain	186
0,2857	0,0038	VIRUS,GENE => be	187
0,2857	0,0038	AMINO ACID,GENE => carry	188
0,2727	0,0057	PROTEIN,GENE => can be find	189
0,2727	0,0057	PROTEIN,GENE => differ	190
0,2500	0,0038	TRANSLATION,RNA => be induce	191
0,2500	0,0019	AMINO ACID,CELL DIFFERENTIATION => have develop	192
0,2500	0,0019	NUCLEIC ACID,GENE EXPRESSION => do not contain	193
0,2500	0,0019	NUCLEOTIDE,RNA => be employ	194

0,2500	0,0019	TRANSLATION,GENE REGULATION => lead	195
0,2500	0,0019	NUCLEOTIDE,TRANSLATION => transform	196
0,2500	0,0019	CELL,BINDING => construct	197
0,2500	0,0019	AMINO ACID,CELL DIFFERENTIATION => recognize	198
0,2500	0,0019	NUCLEOTIDE,TRANSLATION => regulate	199
0,2500	0,0019	GENE EXPRESION,TRANSLATION => establish	200
0,2500	0,0019	CELL,PROTEIN COMPLEX => produce	201
0,2500	0,0038	TRANSLATION,CELL DIFFERENTIATION => process	202
0,2500	0,0019	CELL,BINDING => be not specify	203
0,2500	0,0019	AMINO ACID,CELL DIFFERENTIATION => find	204
0,2500	0,0019	NUCLEIC ACID,GENE EXPRESSION => follow	205
0,2500	0,0019	NEGATIVE REGULATION,TRANSCRIPTION => cancel	206
0,2500	0,0019	GENE EXPRESION,TRANSLATION => be	207
0,2500	0,0019	TRANSLATION,GENE REGULATION => have be demonstrate	208
0,2500	0,0019	CELL,PROTEIN COMPLEX => obtain	209
0,2500	0,0038	TRANSLATION,CELL DIFFERENTIATION => combine	210
0,2500	0,0038	TRANSLATION,RNA => interfere	211
0,2500	0,0019	AMINO ACID,CELL DIFFERENTIATION => be need	212
0,2500	0,0038	TRANSLATION,RNA => play	213
0,2222	0,0038	PROTEIN,PEPTIDE => compose	214
0,2222	0,0038	PROTEIN,CELLULAR PROCESS => be activate	215
0,2222	0,0038	GENE,CELL COMPONENT => activate	216
0,2222	0,0038	PROTEIN,CELL ADHESION => be block	217
0,2222	0,0038	PROTEIN,CELL ADHESION => be achieve	218
0,2222	0,0038	DNA,GENE EXPRESSION => generate	219
0,2222	0,0038	GENE,PROTEIN COMPLEX => be induce	220
0,2222	0,0038	GENE,DNA => be direct	221
0,2222	0,0038	NUCLEOTIDE,GENE EXPRESSION => generate	222
0,2222	0,0076	PROTEIN,AMINO ACID => make	223
0,2222	0,0038	PROTEIN,PEPTIDE => play	224
0,2222	0,0038	DNA,GENE EXPRESSION => activate	225
0,2222	0,0038	GENE,GENE EXPRESSION => emerge	226
0,2222	0,0038	PROTEIN,CELLULAR PROCESS => result	227
0,2222	0,0038	GENE,CELL COMPONENT => extend	228
0,2222	0,0038	NUCLEOTIDE,GENE EXPRESSION => reveal	229
0,2222	0,0038	GENE,DNA => turn	230
0,2222	0,0038	DNA,GENE EXPRESSION => play	231
0,2222	0,0038	PROTEIN,PEPTIDE => effect	232
0,2222	0,0038	GENE,GENE EXPRESSION => persist	233
0,2000	0,0038	PROTEIN,DNA => develop	234

0,2000	0,0038	AMINO ACID,RNA => be associate	235
0,2000	0,0038	TRANSCRIPTION,NUCLEOTIDE => have generate	236
0,2000	0,0019	PROTEIN,GENE EXPRESSION => find	237
0,2000	0,0019	AMINO ACID,AMINO ACID => activate	238
0,2000	0,0019	TRANSCRIPTION,GENE EXPRESSION => contain	239
0,2000	0,0038	VIRUS,NUCLEOTIDE => contain	240
0,2000	0,0019	VIRUS,PROTEIN COMPLEX => activate	241
0,2000	0,0019	GENE REGULATION,CELL => express	242
0,2000	0,0019	CELL ADHESION,CELL => do not give	243
0,2000	0,0019	AMINO ACID,TRANSLATION => be detect	244
0,2000	0,0019	PROTEIN,GENE REGULATION => induce	245
0,2000	0,0038	NUCLEOTIDE,PROTEIN COMPLEX => depend	246
0,2000	0,0019	PROTEIN,GENE EXPRESSION => be use	247
0,2000	0,0019	AMINO ACID,AMINO ACID => process	248
0,2000	0,0019	TISSUE,CELL ADHESION => have be identify	249
0,2000	0,0019	NUCLEIC ACID,GENE REGULATION => do not express	250
0,2000	0,0019	CELL,CELL CULTURE => control	251
0,2000	0,0019	VIRUS,VIRAL LIFE CYCLE => require	252
0,2000	0,0038	AMINO ACID,PROTEIN COMPLEX => modify	253
0,2000	0,0038	PROTEIN,DNA => be detect	254
0,2000	0,0019	TRANSLATION,DNA => obtain	255
0,2000	0,0038	AMINO ACID,RNA => may explain	256
0,2000	0,0038	TRANSCRIPTION,NUCLEOTIDE => support	257
0,2000	0,0038	NUCLEOTIDE,PROTEIN COMPLEX => evidence	258
0,2000	0,0019	AMINO ACID,AMINO ACID => have be transform	259
0,2000	0,0019	TISSUE,CELL ADHESION => cause	260
0,2000	0,0019	NUCLEOTIDE,PEPTIDE => drive	261
0,2000	0,0038	VIRUS,NUCLEOTIDE => be not	262
0,2000	0,0019	VIRUS,VIRAL LIFE CYCLE => be	263
0,2000	0,0019	TRANSCRIPTION,GENE => be not restrict	264
0,2000	0,0019	AMINO ACID,TRANSLATION => treat	265
0,2000	0,0019	NUCLEIC ACID,GENE REGULATION => induce	266
0,2000	0,0019	PROTEIN,GENE EXPRESSION => serve	267
0,2000	0,0019	TISSUE,CELL ADHESION => act	268
0,2000	0,0038	VIRUS,NUCLEOTIDE => determine	269
0,2000	0,0019	VIRUS,VIRAL LIFE CYCLE => exhibit	270
0,2000	0,0019	AMINO ACID,TRANSLATION => evidence	271
0,2000	0,0038	AMINO ACID,PROTEIN COMPLEX => should always be employ	272
0,2000	0,0038	PROTEIN,DNA => be increase	273
0,2000	0,0019	NUCLEIC ACID,GENE REGULATION => implicate	274

0,2000	0,0038	AMINO ACID,RNA => perform	275
0,2000	0,0038	TRANSCRIPTION,NUCLEOTIDE => generate	276
0,2000	0,0038	AMINO ACID,PROTEIN COMPLEX => permit	277
0,2000	0,0038	PROTEIN,DNA => cluster	278
0,2000	0,0038	VIRUS,NUCLEOTIDE => play	279
0,2000	0,0038	AMINO ACID,RNA => regulate	280
0,2000	0,0038	NUCLEOTIDE,PROTEIN COMPLEX => derive	281
0,1905	0,0076	CELL,TRANSCRIPTION => originate	282
0,1875	0,0057	PROTEIN, TRANSCRIPTION => activate	283
0,1875	0,0057	TRANSCRIPTION,CELL DIFFERENTIATION => lead	284
0,1875	0,0057	PROTEIN, TRANSCRIPTION => suppress	285
0,1875	0,0057	PROTEIN,TRANSLATION => be determine	286
0,1875	0,0057	TRANSCRIPTION,CELL DIFFERENTIATION => have	287
0,1818	0,0038	NUCLEOTIDE,CELL => display	288
0,1818	0,0038	CELL,GENE => use	289
0,1818	0,0038	CELL,GENE => have	290
0,1818	0,0038	CELL,PEPTIDE => be initially express	291
0,1818	0,0038	NUCLEOTIDE,CELL => implicate	292
0,1818	0,0038	CELL,GENE => appear	293
0,1818	0,0038	AMINO ACID,TRANSCRIPTION => be know	294
0,1818	0,0038	PROTEIN,GENE => may develop	295
0,1818	0,0038	CELL,PEPTIDE => differentiate	296
0,1818	0,0038	CELL,GENE => be involve	297
0,1818	0,0038	PROTEIN,GENE => know	298
0,1818	0,0038	CELL,PEPTIDE => might be relate	299
0,1818	0,0038	NUCLEOTIDE,CELL => mediate	300
0,1818	0,0038	AMINO ACID,TRANSCRIPTION => may develop	301
0,1818	0,0038	NUCLEOTIDE,CELL => measure	302
0,1818	0,0038	AMINO ACID,TRANSCRIPTION => can be find	303
0,1818	0,0038	CELL,GENE => find	304
0,1818	0,0038	CELL,PEPTIDE => fail	305
0,1818	0,0038	AMINO ACID,TRANSCRIPTION => increase	306
0,1667	0,0057	PROTEIN,CELL => be involve	307
0,1667	0,0038	GENE,NUCLEIC ACID => be observe	308
0,1667	0,0038	CELL,VIRUS => host	309
0,1667	0,0019	CELL,RNA => remain	310
0,1667	0,0019	CELL DIFFERENTIATION,CELL => contribute	311
0,1667	0,0019	GENE,PEPTIDE => have not be	312
0,1667	0,0019	PROTEIN,PROTEIN COMPLEX => regulate	313
0,1667	0,0019	CELL,GENE EXPRESSION => have be characterize	314

0,1667	0,0019	TRANSLATION,CELL => establish	315
0,1667	0,0038	CELL,VIRUS => affect	316
0,1667	0,0038	GENE,NUCLEIC ACID => appear	317
0,1667	0,0019	PROTEIN,MORPHOGENESIS => have prepare	318
0,1667	0,0019	CELL,RNA => can induce	319
0,1667	0,0019	CELL DIFFERENTIATION,CELL => control	320
0,1667	0,0019	POSITIVE REGULATION,TRANSCRIPTION => correlate	321
0,1667	0,0019	NUCLEIC ACID,PEPTIDE => involve	322
0,1667	0,0019	AMINO ACID,GENE EXPRESSION => be associate	323
0,1667	0,0019	CELL,DNA => be determine	324
0,1667	0,0019	VIRUS,VIRUS => propagate	325
0,1667	0,0019	TRANSLATION,CELL => be activate	326
0,1667	0,0057	PROTEIN,CELL => may differ	327
0,1667	0,0019	PROTEIN,MORPHOGENESIS => indicate	328
0,1667	0,0019	NUCLEIC ACID,PEPTIDE => appear	329
0,1667	0,0019	PROTEIN,VIRUS => function	330
0,1667	0,0019	CELL,DNA => indicate	331
0,1667	0,0019	VIRUS,VIRUS => contribute	332
0,1667	0,0038	CELL,VIRUS => infect	333
0,1667	0,0038	GENE,NUCLEIC ACID => be	334
0,1667	0,0019	CELL,RNA => evidence	335
0,1667	0,0019	GENE,PEPTIDE => use	336
0,1667	0,0019	PROTEIN,VIRUS => be regulate	337
0,1667	0,0019	CELL,DNA => be highly induce	338
0,1667	0,0057	PROTEIN,AMINO ACID => be	339
0,1667	0,0019	CELL,RNA => produce	340
0,1667	0,0019	CELL,DNA => identify	341
0,1667	0,0038	CELL,VIRUS => normally regulate	342
0,1667	0,0038	CELL,VIRUS => be know	343
0,1667	0,0038	GENE,NUCLEIC ACID => express	344
0,1500	0,0057	GENE REGULATION,TRANSCRIPTION => be govern	345
0,1429	0,0019	NUCLEOTIDE,NUCLEOTIDE => affect	346
0,1429	0,0019	CELL,CELL COMPONENT => have also be	347
0,1429	0,0019	PROTEIN,CELL DIFFERENTIATION => reduce	348
0,1429	0,0019	PROTEIN,RNA => be involve	349
0,1429	0,0019	VIRUS,GENE => be require	350
0,1429	0,0019	AMINO ACID,GENE => construct	351
0,1429	0,0019	TRANSLATION,PEPTIDE => connect	352
0,1429	0,0019	AMINO ACID,TISSUE => introduce	353
0,1429	0,0057	CELL,NUCLEOTIDE => make	354

0,1429	0,0019	AMINO ACID,GENE => have be find	355
0,1429	0,0019	TRANSCRIPTION,PROTEIN COMPLEX => derive	356
0,1429	0,0019	PROTEIN,RNA => be express	357
0,1429	0,0038	NUCLEOTIDE,PROTEIN => replicate	358
0,1429	0,0019	VIRUS,CELL DIFFERENTIATION => become	359
0,1429	0,0019	CELL,CELL COMPONENT => sustain	360
0,1429	0,0019	AMINO ACID,TISSUE => be identify	361
0,1429	0,0019	PROTEIN,RNA => be compose	362
0,1429	0,0019	VIRUS,GENE => associate	363
0,1429	0,0019	AMINO ACID,GENE => encode	364
0,1429	0,0019	DNA,AMINO ACID => derive	365
0,1429	0,0038	NUCLEOTIDE,PROTEIN => be not influence	366
0,1429	0,0019	VIRUS,GENE => activate	367
0,1429	0,0019	AMINO ACID,GENE => prepare	368
0,1429	0,0019	TRANSLATION,PEPTIDE => associate	369
0,1429	0,0019	TRANSCRIPTION,PROTEIN COMPLEX => indicate	370
0,1429	0,0038	NUCLEOTIDE,PROTEIN => associate	371
0,1429	0,0019	AMINO ACID,TISSUE => appear	372
0,1429	0,0019	TRANSLATION,PEPTIDE => turn	373
0,1429	0,0019	TRANSCRIPTION,PROTEIN COMPLEX => analyze	374
0,1429	0,0057	CELL,NUCLEOTIDE => retain	375
0,1429	0,0057	CELL,TRANSCRIPTION => be not	376
0,1429	0,0019	AMINO ACID,GENE => use	377
0,1429	0,0038	NUCLEOTIDE,PROTEIN => have be	378
0,1429	0,0057	CELL,NUCLEOTIDE => suggest	379
0,1429	0,0038	NUCLEOTIDE,PROTEIN => remain	380
0,1333	0,0038	DNA,NUCLEOTIDE => have identify	381
0,1333	0,0038	DNA,NUCLEOTIDE => be locate	382
0,1333	0,0038	DNA,NUCLEOTIDE => be observe	383
0,1333	0,0038	DNA,NUCLEOTIDE => be	384
0,1333	0,0038	DNA,NUCLEOTIDE => contain	385
0,1333	0,0038	DNA,NUCLEOTIDE => may be	386
0,1304	0,0057	CELL,CELL => produce	387
0,1304	0,0057	CELL,CELL => be functionally link	388
0,1304	0,0057	CELL,CELL => transform	389
0,1250	0,0038	PROTEIN, TRANSCRIPTION => be find	390
0,1250	0,0019	TRANSLATION,CELL DIFFERENTIATION => obtain	391
0,1250	0,0038	PROTEIN,TRANSLATION => give	392
0,1250	0,0019	TRANSLATION,RNA => be use	393
0,1250	0,0038	PROTEIN,TRANSLATION => express	394

0,1250	0,0038	TRANSCRIPTION,CELL DIFFERENTIATION => may not only play	395
0,1250	0,0038	PROTEIN, TRANSCRIPTION => indicate	396
0,1250	0,0019	TRANSLATION,RNA => react	397
0,1250	0,0038	PROTEIN,TRANSLATION => cause	398
0,1250	0,0038	PROTEIN,TRANSLATION => translate	399
0,1250	0,0038	PROTEIN, TRANSCRIPTION => be only express	400
0,1250	0,0038	PROTEIN,TRANSLATION => bind	401
0,1250	0,0038	TRANSCRIPTION,CELL DIFFERENTIATION => control	402
0,1250	0,0038	PROTEIN, TRANSCRIPTION => decrease	403
0,1111	0,0038	PROTEIN,AMINO ACID => reveal	404
0,1111	0,0019	NUCLEOTIDE,GENE EXPRESSION => turn	405
0,1111	0,0019	DNA,GENE EXPRESSION => be examine	406
0,1111	0,0019	GENE,GENE EXPRESSION => need	407
0,1111	0,0038	PROTEIN,AMINO ACID => have	408
0,1111	0,0038	PROTEIN,CELL => require	409
0,1111	0,0019	NUCLEOTIDE,GENE EXPRESSION => participate	410
0,1111	0,0019	GENE,DNA => be use	411
0,1111	0,0019	PROTEIN,PEPTIDE => reflect	412
0,1111	0,0019	GENE,GENE EXPRESSION => lead	413
0,1111	0,0019	GENE,DNA => involve	414
0,1111	0,0019	PROTEIN,PEPTIDE => can provide	415
0,1111	0,0019	DNA,GENE EXPRESSION => be use	416
0,1111	0,0019	GENE,GENE EXPRESSION => attract	417
0,1111	0,0019	PROTEIN,CELLULAR PROCESS => be express	418
0,1111	0,0019	GENE,CELL COMPONENT => process	419
0,1111	0,0019	PROTEIN,CELL ADHESION => implicate	420
0,1111	0,0038	PROTEIN,AMINO ACID => crossreact	421
0,1111	0,0038	PROTEIN,CELL => have conclude	422
0,1111	0,0019	GENE,PROTEIN COMPLEX => extend	423
0,1111	0,0038	PROTEIN,CELL => use	424
0,1111	0,0019	PROTEIN,PEPTIDE => enable	425
0,1111	0,0019	GENE,GENE EXPRESSION => initiate	426
0,1111	0,0019	PROTEIN,CELLULAR PROCESS => induce	427
0,1111	0,0019	GENE,CELL COMPONENT => combine	428
0,1111	0,0019	PROTEIN,CELL ADHESION => have recently be	429
0,1111	0,0038	PROTEIN,AMINO ACID => have sustain	430
0,1111	0,0038	PROTEIN,CELL => have be establish	431
0,1111	0,0019	DNA,GENE EXPRESSION => match	432
0,1111	0,0019	GENE,GENE EXPRESSION => obtain	433
0,1111	0,0038	PROTEIN,AMINO ACID => include	434

0,1111	0,0038	PROTEIN,CELL => do not affect	435
0,1000	0,0019	AMINO ACID,PROTEIN COMPLEX => be only induce	436
0,1000	0,0019	VIRUS,NUCLEOTIDE => affect	437
0,1000	0,0019	NUCLEOTIDE,PROTEIN COMPLEX => be demonstrate	438
0,1000	0,0038	GENE REGULATION,TRANSCRIPTION => suggest	439
0,1000	0,0019	AMINO ACID,PROTEIN COMPLEX => confer	440
0,1000	0,0019	PROTEIN,DNA => be alter	441
0,1000	0,0019	VIRUS,NUCLEOTIDE => encompass	442
0,1000	0,0019	AMINO ACID,RNA => deplete	443
0,1000	0,0019	AMINO ACID,PROTEIN COMPLEX => make of	444
0,1000	0,0019	PROTEIN,DNA => allow	445
0,1000	0,0019	AMINO ACID,RNA => control	446
0,1000	0,0019	TRANSCRIPTION,NUCLEOTIDE => analyze	447
0,1000	0,0019	NUCLEOTIDE,PROTEIN COMPLEX => be require	448
0,1000	0,0038	GENE REGULATION,TRANSCRIPTION => describe	449
0,1000	0,0019	NUCLEOTIDE,PROTEIN COMPLEX => evaluate	450
0,1000	0,0038	GENE REGULATION,TRANSCRIPTION => be provide	451
0,1000	0,0019	AMINO ACID,PROTEIN COMPLEX => characterize	452
0,1000	0,0019	NUCLEOTIDE,PROTEIN COMPLEX => enter	453
0,1000	0,0038	GENE REGULATION,TRANSCRIPTION => act	454
0,1000	0,0038	GENE REGULATION,TRANSCRIPTION => selectively inhibit	455
0,0952	0,0038	CELL,NUCLEOTIDE => lead	456
0,0952	0,0038	CELL,TRANSCRIPTION => employ	457
0,0952	0,0038	CELL,NUCLEOTIDE => increase	458
0,0952	0,0038	CELL,TRANSCRIPTION => activate	459
0,0952	0,0038	CELL,NUCLEOTIDE => activate	460
0,0952	0,0038	CELL,NUCLEOTIDE => be require	461
0,0952	0,0038	CELL,TRANSCRIPTION => can be	462
0,0952	0,0038	CELL,NUCLEOTIDE => fold	463
0,0952	0,0038	CELL,NUCLEOTIDE => have be	464
0,0952	0,0038	CELL,TRANSCRIPTION => be examine	465
0,0952	0,0038	CELL,TRANSCRIPTION => do not involve	466
0,0952	0,0038	CELL,TRANSCRIPTION => treat	467
0,0909	0,0019	AMINO ACID,TRANSCRIPTION => express	468
0,0909	0,0019	AMINO ACID,TRANSCRIPTION => involve	469
0,0909	0,0019	CELL,PEPTIDE => identify	470
0,0909	0,0019	NUCLEOTIDE,CELL => be not express	471
0,0909	0,0019	NUCLEOTIDE,CELL => have clone	472
0,0909	0,0019	AMINO ACID,TRANSCRIPTION => form	473
0,0909	0,0019	CELL,GENE => be constitutively activate	474

0,0909	0,0019	PROTEIN,GENE => identify	475
0,0909	0,0019	CELL,PEPTIDE => be express	476
0,0909	0,0019	NUCLEOTIDE,CELL => be know	477
0,0909	0,0019	CELL,PEPTIDE => trigger	478
0,0870	0,0038	CELL,CELL => derive	479
0,0870	0,0038	CELL,CELL => play	480
0,0870	0,0038	CELL,CELL => be require	481
0,0870	0,0038	CELL,CELL => bind	482
0,0833	0,0019	CELL,VIRUS => follow	483
0,0833	0,0019	GENE,NUCLEIC ACID => present	484
0,0833	0,0019	CELL,VIRUS => can serve	485
0,0833	0,0019	GENE,NUCLEIC ACID => be involve	486
0,0833	0,0019	GENE,NUCLEIC ACID => make	487
0,0833	0,0019	GENE,NUCLEIC ACID => be use	488
0,0714	0,0019	NUCLEOTIDE,PROTEIN => exhibit	489
0,0714	0,0019	NUCLEOTIDE,PROTEIN => alter	490
0,0714	0,0019	NUCLEOTIDE,PROTEIN => have examine	491
0,0714	0,0019	NUCLEOTIDE,PROTEIN => receive	492
0,0667	0,0019	DNA,NUCLEOTIDE => require	493
0,0667	0,0019	DNA,NUCLEOTIDE => clone	494
0,0667	0,0019	DNA,NUCLEOTIDE => construct	495
0,0625	0,0019	TRANSLATION,PROTEIN => may be	496
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => be express	497
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => be find	498
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => have be ascribe	499
0,0625	0,0019	PROTEIN, TRANSCRIPTION => restrict	500
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => involve	501
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => play	502
0,0625	0,0019	PROTEIN, TRANSCRIPTION => grow	503
0,0625	0,0019	TRANSCRIPTION,CELL DIFFERENTIATION => activate	504
0,0625	0,0019	PROTEIN,TRANSLATION => produce	505
0,0625	0,0019	PROTEIN,TRANSLATION => synthesize	506
0,0556	0,0019	PROTEIN,AMINO ACID => show	507
0,0556	0,0019	PROTEIN,CELL => construct	508
0,0556	0,0019	PROTEIN,CELL => synthesize	509
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => turn	510
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => be not	511
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => involve	512
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => generate	513
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => be know	514

0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => affect	515
0,0500	0,0019	GENE REGULATION,TRANSCRIPTION => synthesize	516
0,0476	0,0019	CELL,TRANSCRIPTION => can induce	517
0,0476	0,0019	CELL,TRANSCRIPTION => produce	518
0,0435	0,0019	CELL,CELL => replicate	519
0,0435	0,0019	CELL,CELL => do not show	520
0,0435	0,0019	CELL,CELL => represent	521
0,0435	0,0019	CELL,CELL => might play	522
0,0435	0,0019	CELL,CELL => express	523
0,0435	0,0019	CELL,CELL => combine	524

Anexo E – Conceitos da ontologia *Family law doctrine*

1	JUDGE
2	COURT
3	DIVORCE
4	PARTY
5	MARRIAGE
6	SPOUSE
7	MARITAL AGREEMENT
8	REGIME
9	PROPERTY
10	LAWYER
11	VISITATION
12	CHILD
13	EMANCIPATION
14	LEGAL SEPARATION
15	CHILD CUSTODY
16	DECREE
17	ALIMONY
18	CHILD SUPPORT
19	COUPLE
20	ADOPTION
21	ANNULMENT
22	INHERITANCE
23	FEE
24	WILL
25	SPOUSAL SUPPORT
26	TRIAL
27	WITNESS

Anexo F – Relacionamentos não-taxonômicos de referência (lematizados) da ontologia *Family law doctrine*

Nº	Conceito 1	Rótulo	Conceito 2
1	COURT	GRANT	DIVORCE
2	PARTY	ASK FOR	LEGAL SEPARATION
3	PARTY	ASK FOR	DIVORCE
4	DIVORCE	DISSOLVE	MARRIAGE
5	COURT	GRANT	SPOUSAL SUPPORT
6	MARRIAGE	HAVE	SPOUSE
7	MARRIAGE	HAVE	MARITAL AGREEMENT
8	MARRIAGE	HAVE	REGIME
9	PARTY	OWN	PROPERTY
10	LAWYER	REPRESENT	PARTY
11	PARTY	ASK FOR	VISITATION
12	COURT	GRANT	VISITATION
13	WITNESS	PARTICIPATE	TRIAL
14	SPOUSE	INHERIT	PROPERTY
15	COURT	GIVE OUT	DECREE
16	COURT	GRANT	LEGAL SEPARATION
17	PARTY	ASK FOR	CHILD CUSTODY
18	COUPLE	SOLICIT	ADOPTION
19	COURT	GRANT	CHILD CUSTODY
20	COURT	GRANT	ALIMONY
21	COURT	GRANT	CHILD SUPPORT
22	CHILD	INHERIT	PROPERTY
23	CHILD	ASK FOR	EMANCIPATION
24	JUDGE	GRANT	ADOPTION
25	PARTY	SIGN	MARITAL AGREEMENT
26	COUPLE	ASK FOR	ANNULMENT
27	COUPLE	HAVE	CHILD

Nº	Conceito 1	Rótulo	Conceito 2
28	SPOUSE	RECEIVE	INHERITANCE
29	CHILD	RECEIVE	INHERITANCE
30	LAYWER	TAKE	FEE
31	PARTY	PAY	FEE
32	SPOUSE	HAVE	WILL
33	COURT	GRANT	EMANCIPATION
34	PARTY	ASK FOR	SPOUSAL SUPPORT
35	COURT	CANCEL	MARRIAGE
36	PARTY	ASK FOR	CHILD SUPPORT
37	COURT	GRANT	ANNULMENT
38	PARTY	ASK FOR	ALIMONY
39	PARTY	PARTICIPATE	TRIAL
40	JUDGE	PRESIDE	COURT
41	LAWYER	PARTICIPATE	TRIAL
42	JUDGE	CONDUCT	TRIAL

Anexo G – Relacionamentos não-taxonômicos recomendados por TARNT a partir do corpus *Family law doctrine*

Lista de relacionamentos não-taxonômicos recomendados por TARNT quando executada com a configuração da tabela 48 (seção 5.8) e o corpus *Family law doctrine* [62]. Os relacionamentos de referência estão destacados.

Freq. de co-ocorrência	tupla	rótulos	Nº
0,0271	MARRIAGE,SPOUSE	can agree, reject, should have, allow, should contract, accept, determine, establish, should not be, should wait, must ask, do not preserve, have to accept, should consult, may determine	1
0,0271	PARTY,DIVORCE	decide, apply, feel, terminate, resolve, give, have, become, have to prove, should wait, must justify, may consent, can use, must know, ask, can assist	2
0,0271	DIVORCE,LAWYER	be not, look, can agree, may award, justify, can require, indicate, involve, recommend, should wait, do not guarantee, draw, should consider, may need, want	3
0,0244	COURT,DIVORCE	can not refuse, need, conclude, do not show, grant, may reject, give, may consult, may seem, assist, involve, may consent, must know, may determine	4
0,0244	PARTY,PARTY	need, apply, may seem, know, do not guarantee, represent, register, may represent, remain, may agree, assess,	5
0,0244	PARTY,CHILD SUPPORT	can consult, may consult, meet, do not share, represent, do not modify, must present, pay, may accept, participate, ask, make, find, agree	6
0,0230	COURT,CHILD CUSTODY	should have, grant, become, can require, assist, to justify, have to share, can find, must represent, require, can assist, live	7
0,0230	PARTY,COURT	can consider, have, accept, expect, be set, do not modify, may consent, might present, may bring, recognize, can	8
0,0203	LAWYER,PARTY	do not justify, resolve, do not show, assist, recommend, may revoke, may deny, represent, obtain, must consider, require,	9
0,0189	COUPLE,DIVORCE	relate, evaluate, accept, solicit, may revoke, have to accept, leave, may bring, can use, receive	10
0,0189	LAWYER,COURT	have to examine, may help, leave, may bring, deny, pay, may accept, defend, assume, go	11
0,0162	PARTY,CHILD CUSTODY	resolve, have, draw, leave, must decide, may agree, ask	12

0,0162	COUPLE,MARRIAGE	can order, do not justify, should have, grant, get, disagree, may deny, should be, must agree	13
0,0149	COURT,CHILD SUPPORT	need, grant, do not accept, can allege, participate, make, preserve, find	14
0,0149	PARTY,ALIMONY	have to examine, save, have to deal, ask, meet, may revoke	15
0,0149	COUPLE,CHILD SUPPORT	define, hear, should have, may use, deny, may consult, must represent, do not share	16
0,0149	PARTY,PROPERTY	must ask, can consult, may justify, resolve, own	17
0,0149	MARRIAGE,MARITAL AGREEMENT	can not refuse, evaluate, can consider, obtain, may help, submit, can solicit, may agree	18
0,0135	COUPLE,CHILD	have to share, allege, terminate, may require, have, proceed, require	19
0,0135	DIVORCE,PROPERTY	allege, feel, get, bring, may determine, preserve, do not share	20
0,0135	PARTY,SPOUSAL SUPPORT	define, should consider, may enter, require, ask, may result, should wait, be	21
0,0135	SPOUSE,INHERITANCE	can order, should contract, determine, can assist, receive, may result, do not guarantee	22
0,0135	MARRIAGE,MARRIAGE	may help, consider, solicit, file, proceed, ask, preserve, declare	23
0,0135	DIVORCE,MARITAL AGREEMENT	may claim, must consider, must provide, may require, can find, may allege	24
0,0122	DIVORCE,MARRIAGE	can use, consider, indicate, disagree, dissolve	25
0,0122	COURT,MARRIAGE	must be, help, may award, do not accept, may consult, cancel	26
0,0108	DIVORCE,DIVORCE	expect, conclude, help, may allege, may result	27
0,0108	SPOUSE,PROPERTY	can order, have to evaluate, can solicit, assess, inherit	28
0,0108	PROPERTY,PROPERTY	must consider, should have, get, give, may obtain	29
0,0108	CHILD,INHERITANCE	do not make, must agree, receive, may approve, do not share	30
0,0108	COUPLE,PROPERTY	approve, should have, allow, may modify, involve	31
0,0108	DIVORCE,CHILD CUSTODY	should have, have to deal, receive, include, declare	32
0,0108	SPOUSE,WILL	replace, do not modify, have, must decide, should not be	33
0,0108	LAWYER,MARITAL AGREEMENT	have to examine, may modify, defend, consult, agree	34
0,0108	CHILD,PROPERTY	look, have to accept, should consider, inherit, prepare, do not share	35
0,0094	DIVORCE,CHILD SUPPORT	be set, break, may claim, must provide, determine	36
0,0094	MARRIAGE,LAWYER	look, break, recognize, intend, may deny	37
0,0094	COURT,SPOUSAL SUPPORT	replace, need, may award, grant, to justify	38

0,0094	DIVORCE,ALIMONY	can not refuse, can agree, give, have, agree	39
0,0094	PARTY,VISITATION	hear, grant, bring, ask, choose	40
0,0094	MARRIAGE,CHILD CUSTODY	can include, pay, defend, may approve, do not share	41
0,0094	PARTY,MARITAL AGREEMENT	sign, evaluate, be set, do not show, do not make	42
0,0094	LAWYER,CHILD CUSTODY	expect, draw, may justify, terminate, leave, mean	43
0,0094	MARRIAGE,ADOPTION	apply, must present, may use, assist, ask, know	44
0,0081	LAWYER,INHERITANCE	draw, may justify, may need, pay, may allege	45
0,0081	COURT,PROPERTY	look, assume, may determine, go	46
0,0081	COURT,ALIMONY	grant, establish, indicate, do not guarantee	47
0,0081	LAWYER,SPOUSAL SUPPORT	register, file, involve, inherit, may revoke	48
0,0081	INHERITANCE,WILL	can determine, decide, may consent, must present	49
0,0081	COUPLE,ADOPTION	be set, feel, do not show, solicit	50
0,0081	LAWYER,LAWYER	obtain, may grant, deny, do not make, should ask, can assist	51
0,0081	VISITATION,CHILD	may include, must consider, must provide, may enter, mean, involve	52
0,0081	DIVORCE,INHERITANCE	stay, belong, do not have, should wait	53
0,0081	PARTY,LEGAL SEPARATION	register, may represent, ask	54
0,0081	LAWYER,WILL	have to examine, can change, may consent, agree	55
0,0081	JUDGE,INHERITANCE	may not accept, give, accept	56
0,0081	LAWYER,ADOPTION	may represent, conclude, must consider, indicate, should wait	57
0,0081	LEGAL SEPARATION,LAWYER	be not, mean, intend, may revoke	58
0,0067	VISITATION,LAWYER	be set, terminate, leave, accept	59
0,0067	DIVORCE,SPOUSAL SUPPORT	obtain, do not modify, indicate, abandon	60
0,0067	LAWYER,FEE	can change, take, bring, may approve	61
0,0067	MARRIAGE,ALIMONY	define, resolve, have to evaluate, may consult, choose	62
0,0067	LAWYER,CHILD SUPPORT	must be, can consult, terminate, may award, go	63
0,0067	DIVORCE,DECREE	should be, must know, justify, assist	64
0,0054	COURT,WILL	relate, can solicit, assist	65
0,0054	LAWYER,PROPERTY	expect, may consent, file, must represent	66
0,0054	MARRIAGE,DECREE	do not make, may allege, choose, do not share	67
0,0054	COURT,VISITATION	grant, may deny, agree	68

0,0054	WILL,DECREE	have to accept, may include, provide, make	69
0,0054	CHILD,EMANCIPATION	can change, stay, ask	70
0,0054	DIVORCE,WILL	allow, deny, receive	71
0,0054	CHILD,WILL	allege, may require, may reject	72
0,0054	DIVORCE,TRIAL	have to examine, may claim, abandon, do not share	73
0,0054	WILL,MARITAL AGREEMENT	can change, may reject, give, prepare	74
0,0054	COUPLE,LAWYER	do not accept, must represent, preserve	75
0,0054	PARTY,FEE	can order, take, pay	76
0,0054	COURT,TRIAL	have to accept, may enter, may determine	77
0,0054	PARTY,TRIAL	replace, allow, find	78
0,0054	VISITATION,COUPLE	can not refuse, define, allege, refrain	79
0,0054	CHILD,CHILD	draw, must ask, should consult, accept	80
0,0054	PROPERTY,WILL	can not refuse, can consider, live	81
0,0054	COURT,LEGAL SEPARATION	must consider, grant	82
0,0040	COUPLE,TRIAL	do not make, must justify	83
0,0040	SPOUSE,COURT	must ask, must represent, must justify	84
0,0040	ADOPTION,TRIAL	may require, recommend, go	85
0,0040	WILL,TRIAL	allow, do not make, ask	86
0,0040	PROPERTY,LEGAL SEPARATION	give, mean, consult	87
0,0040	MARRIAGE,REGIME	can consider, leave, submit	88
0,0040	JUDGE,ADOPTION	grant, notify	89
0,0040	CHILD,TRIAL	have to share, may consult, defend	90
0,0040	MARITAL AGREEMENT,DECREE	must present, can assist, should not be	91
0,0040	COUPLE,ANNULMENT	should ask, ask	92
0,0040	DIVORCE,REGIME	become, must justify	93
0,0040	COURT,EMANCIPATION	grant, do not make, involve	94
0,0040	ANNULMENT,LAWYER	belong, must justify, represent	95
0,0040	MARITAL AGREEMENT,TRIAL	relate, violate, should wait	96
0,0040	INHERITANCE,REGIME	provide, notify, may result	97
0,0040	JUDGE,SPOUSE	save, should wait	98

0,0027	COURT,ANNULMENT	be set, grant	99
0,0027	LAWYER,CHILD	solicit, ask	100
0,0027	EMANCIPATION,PROPERTY	can order, can assist	101
0,0027	LAWYER,WITNESS	feel	102
0,0027	MARITAL AGREEMENT,WITNESS	might present, refrain	103
0,0027	JUDGE,WILL	may include, may accept	104
0,0027	ADOPTION,INHERITANCE	inherit	105
0,0027	JUDGE,COURT	may use, preside	106
0,0027	TRIAL,FEE	have to evaluate, become	107
0,0027	PARTY,WITNESS	get, pay	108

Anexo H – Relacionamentos não-taxonômicos recomendados por AERA a partir do corpus *Family law doctrine*

Lista de relacionamentos não-taxonômicos recomendados por AERA quando executada com a configuração da tabela 47 (seção 5.8) e o corpus *Family law doctrine* [62] Os relacionamentos de referência estão destacados.

Conf.	Sup.	Regra de associação	Nº
1,0000	0,0036	LAWYER,WITNESS => feel	1
1,0000	0,0036	ADOPTION,INHERITANCE => inherit	2
0,6667	0,0036	JUDGE,SPOUSE => save	3
0,6667	0,0036	JUDGE,ADOPTION => grant	4
0,6667	0,0036	COUPLE,TRIAL => do not make	5
0,6667	0,0036	DIVORCE,REGIME => become	6
0,6667	0,0036	COUPLE,ANNULMENT => ask	7
0,5000	0,0073	LAWYER,MARITAL AGREEMENT => may modify	8
0,5000	0,0054	PARTY,LEGAL SEPARATION => ask	9
0,5000	0,0054	LEGAL SEPARATION,LAWYER => intend	10
0,5000	0,0018	DIVORCE,WILL => allow	11
0,5000	0,0054	COUPLE,LAWYER => must represent	12
0,5000	0,0073	CHILD,INHERITANCE => receive	13
0,5000	0,0018	EMANCIPATION,PROPERTY => can order	14
0,5000	0,0018	COURT,ANNULMENT => be set	15
0,5000	0,0073	SPOUSE,WILL => have	16
0,5000	0,0036	COURT,LEGAL SEPARATION => grant	17
0,5000	0,0018	MARITAL AGREEMENT,WITNESS => might present	18
0,5000	0,0036	JUDGE,COURT => preside	19
0,5000	0,0018	JUDGE,WILL => may accept	20
0,5000	0,0018	TRIAL,FEE => have to evaluate	21
0,5000	0,0018	PARTY,WITNESS => get	22
0,5000	0,0036	PARTY,FEE => pay	23
0,5000	0,0018	LAWYER,CHILD => solicit	24
0,5000	0,0036	COURT,ALIMONY => grant	25
0,5000	0,0036	CHILD,WILL => have	26
0,5000	0,0073	SPOUSE,PROPERTY => inherit	27
0,5000	0,0054	COURT,TRIAL => approve	28
0,5000	0,0036	PROPERTY,WILL => resolve	29
0,5000	0,0054	COUPLE,ADOPTION => solicit	30
0,5000	0,0018	TRIAL,FEE => become	31
0,5000	0,0018	LAWYER,CHILD => ask	32
0,5000	0,0036	CHILD,EMANCIPATION => ask	33
0,5000	0,0018	MARITAL AGREEMENT,WITNESS => refrain	34

0,5000	0,0018	JUDGE,WILL => may include	35
0,5000	0,0018	JUDGE,COURT => may use	36
0,5000	0,0018	COURT,ANNULMENT => grant	37
0,5000	0,0018	PARTY,WITNESS => pay	38
0,5000	0,0073	DIVORCE,CHILD CUSTODY => may obtain	39
0,5000	0,0018	COURT,LEGAL SEPARATION => be	40
0,5000	0,0054	LAWYER,WILL => should ask	41
0,5000	0,0018	COURT,WILL => do not make	42
0,5000	0,0036	COURT,VISITATION => may revoke	43
0,5000	0,0018	EMANCIPATION,PROPERTY => can assist	44
0,5000	0,0054	DIVORCE,DIVORCE => can find	45
0,5000	0,0054	JUDGE,INHERITANCE => may award	46
0,5000	0,0018	PROPERTY,PROPERTY => belong	47
0,5000	0,0018	PARTY,TRIAL => obtain	48
0,5000	0,0018	COUPLE,PROPERTY => inherit	49
0,4286	0,0054	COURT,SPOUSAL SUPPORT => grant	50
0,4286	0,0054	MARRIAGE,LAWYER => assist	51
0,4286	0,0054	DIVORCE,CHILD SUPPORT => indicate	52
0,4286	0,0054	DIVORCE,ALIMONY => can consult	53
0,4286	0,0054	MARRIAGE,CHILD CUSTODY => resolve	54
0,4000	0,0036	VISITATION,LAWYER => need to justify	55
0,4000	0,0036	DIVORCE,SPOUSAL SUPPORT => may obtain	56
0,4000	0,0036	DIVORCE,DECREE => can consider	57
0,4000	0,0036	LAWYER,FEE => be	58
0,4000	0,0036	COURT,MARRIAGE => cancel	59
0,4000	0,0036	PROPERTY,LEGAL SEPARATION => do not preserve	60
0,3333	0,0018	COUPLE,TRIAL => should wait	61
0,3333	0,0018	SPOUSE,COURT => can allege	62
0,3333	0,0018	ADOPTION,TRIAL => obtain	63
0,3333	0,0018	WILL,TRIAL => can determine	64
0,3333	0,0036	LAWYER,ADOPTION => be recognize	65
0,3333	0,0018	MARRIAGE,REGIME => represent	66
0,3333	0,0018	JUDGE,ADOPTION => must consider	67
0,3333	0,0018	CHILD,TRIAL => obtain	68
0,3333	0,0018	MARITAL AGREEMENT,DECREE => remain	69
0,3333	0,0018	DIVORCE,REGIME => become	70
0,3333	0,0018	COURT,EMANCIPATION => do not make	71
0,3333	0,0018	ANNULMENT,LAWYER => belong	72
0,3333	0,0054	COURT,PROPERTY => may award	73
0,3333	0,0018	INHERITANCE,REGIME => provide	74

0,3333	0,0018	MARITAL AGREEMENT,TRIAL => should wait	75
0,3333	0,0054	DIVORCE,MARRIAGE => dissolve	76
0,3333	0,0018	SPOUSE,COURT => judge	77
0,3333	0,0018	ADOPTION,TRIAL => have to prove	78
0,3333	0,0018	WILL,TRIAL => do not have	79
0,3333	0,0018	PROPERTY,LEGAL SEPARATION => remain	80
0,3333	0,0018	MARRIAGE,REGIME => must agree	81
0,3333	0,0018	CHILD,TRIAL => be not	82
0,3333	0,0018	MARITAL AGREEMENT,DECREE => resolve	83
0,3333	0,0036	INHERITANCE,WILL => determine	84
0,3333	0,0018	JUDGE,SPOUSE => should wait	85
0,3333	0,0018	ANNULMENT,LAWYER => must justify	86
0,3333	0,0018	MARITAL AGREEMENT,TRIAL => violate	87
0,3333	0,0018	INHERITANCE,REGIME => notify	88
0,3333	0,0018	COURT,EMANCIPATION => grant	89
0,3333	0,0036	LAWYER,SPOUSAL SUPPORT => ask for	90
0,3333	0,0018	PROPERTY,LEGAL SEPARATION => establish	91
0,3333	0,0054	PARTY,LEGAL SEPARATION => can require	92
0,3333	0,0036	JUDGE,INHERITANCE => consider	93
0,3333	0,0018	SPOUSE,COURT => approve	94
0,3333	0,0018	ADOPTION,TRIAL => can consider	95
0,3333	0,0018	WILL,TRIAL => justify	96
0,3333	0,0036	INHERITANCE,WILL => do not show	97
0,3333	0,0018	MARRIAGE,REGIME => determine	98
0,3333	0,0018	CHILD,TRIAL => be judge	99
0,3333	0,0018	MARITAL AGREEMENT,DECREE => be	100
0,3333	0,0018	COURT,EMANCIPATION => involve	101
0,3333	0,0018	ANNULMENT,LAWYER => represent	102
0,3333	0,0036	COURT,PROPERTY => be not require	103
0,3333	0,0018	INHERITANCE,REGIME => may consult	104
0,3333	0,0036	LAWYER,INHERITANCE => indicate	105
0,3333	0,0036	DIVORCE,INHERITANCE => gain	106
0,3333	0,0018	MARITAL AGREEMENT,TRIAL => relate	107
0,2857	0,0036	COUPLE,ANNULMENT => should ask	108
0,2857	0,0036	DIVORCE,INHERITANCE => become	109
0,2857	0,0036	PARTY,VISITATION => ask	110
0,2857	0,0036	PARTY,MARITAL AGREEMENT => be	111
0,2857	0,0036	MARRIAGE,ADOPTION => obtain	112
0,2857	0,0036	PARTY,VISITATION => have	113
0,2857	0,0036	PARTY,MARITAL AGREEMENT => may include	114

0,2857	0,0036	LAWYER,CHILD CUSTODY => should consider	115
0,2727	0,0054	PARTY,ALIMONY => receive	116
0,2727	0,0054	PARTY,PROPERTY => have to share	117
0,2500	0,0018	COURT,WILL => decide	118
0,2500	0,0018	LAWYER,PROPERTY => leave	119
0,2500	0,0018	MARRIAGE,DECREE => represent	120
0,2500	0,0036	CHILD,PROPERTY => inherit	121
0,2500	0,0018	WILL,DECREE => should not be	122
0,2500	0,0018	CHILD,EMANCIPATION => obtain	123
0,2500	0,0018	CHILD,WILL => preserve	124
0,2500	0,0018	DIVORCE,TRIAL => may enter	125
0,2500	0,0018	WILL,MARITAL AGREEMENT => indicate	126
0,2500	0,0036	PARTY,FEE => be	127
0,2500	0,0018	COURT,TRIAL => may need	128
0,2500	0,0018	VISITATION,COUPLE => disagree	129
0,2500	0,0036	CHILD,CHILD => assist	130
0,2500	0,0018	PROPERTY,WILL => be not require	131
0,2500	0,0018	COURT,WILL => should consider	132
0,2500	0,0018	LAWYER,PROPERTY => must be	133
0,2500	0,0018	MARRIAGE,DECREE => should wait	134
0,2500	0,0018	COURT,VISITATION => be	135
0,2500	0,0036	WILL,DECREE => be subject	136
0,2500	0,0018	CHILD,EMANCIPATION => can allege	137
0,2500	0,0018	DIVORCE,WILL => may need	138
0,2500	0,0018	DIVORCE,TRIAL => be require	139
0,2500	0,0018	WILL,MARITAL AGREEMENT => may award	140
0,2500	0,0018	COUPLE,LAWYER => indicate	141
0,2500	0,0018	PARTY,TRIAL => must be	142
0,2500	0,0018	VISITATION,COUPLE => have to accept	143
0,2500	0,0018	CHILD,CHILD => do not accept	144
0,2500	0,0054	LAWYER,PROPERTY => represent	145
0,2500	0,0018	MARRIAGE,DECREE => mean	146
0,2500	0,0018	WILL,DECREE => take	147
0,2500	0,0018	DIVORCE,WILL => can consider	148
0,2500	0,0018	CHILD,WILL => obtain	149
0,2500	0,0054	DIVORCE,TRIAL => do not justify	150
0,2500	0,0018	WILL,MARITAL AGREEMENT => determine	151
0,2500	0,0018	COUPLE,LAWYER => resolve	152
0,2500	0,0018	PARTY,FEE => may allege	153
0,2500	0,0018	COURT,TRIAL => be require	154

0,2500	0,0018	PARTY,TRIAL => resolve	155
0,2500	0,0036	VISITATION,COUPLE => remain	156
0,2500	0,0018	CHILD,CHILD => gain	157
0,2500	0,0018	PROPERTY,WILL => determine	158
0,2500	0,0036	COURT,VISITATION => grant	159
0,2500	0,0018	LAWYER,PROPERTY => assist	160
0,2500	0,0018	MARRIAGE,DECREE => approve	161
0,2500	0,0018	WILL,DECREE => remain	162
0,2500	0,0018	DIVORCE,TRIAL => decide	163
0,2500	0,0036	CHILD,PROPERTY => may inquire	164
0,2500	0,0018	VISITATION,COUPLE => resolve	165
0,2500	0,0018	CHILD,CHILD => participate	166
0,2500	0,0018	WILL,MARITAL AGREEMENT => be write	167
0,2222	0,0036	COURT,MARRIAGE => grant	168
0,2222	0,0036	DIVORCE,MARRIAGE => need	169
0,2222	0,0036	DIVORCE,MARRIAGE => may obtain	170
0,2000	0,0036	COUPLE,CHILD => want	171
0,2000	0,0018	MARRIAGE,ALIMONY => should consider	172
0,2000	0,0036	LAWYER,CHILD SUPPORT => need	173
0,2000	0,0018	VISITATION,LAWYER => be	174
0,2000	0,0018	DIVORCE,SPOUSAL SUPPORT => ask for	175
0,2000	0,0018	LAWYER,FEE => include	176
0,2000	0,0073	DIVORCE,PROPERTY => share	177
0,2000	0,0036	SPOUSE,INHERITANCE => may obtain	178
0,2000	0,0018	DIVORCE,DECREE => justify	179
0,2000	0,0054	PARTY,SPOUSAL SUPPORT => may need	180
0,2000	0,0018	LAWYER,FEE => can consider	181
0,2000	0,0036	DIVORCE,MARITAL AGREEMENT => define	182
0,2000	0,0018	VISITATION,LAWYER => defend	183
0,2000	0,0018	DIVORCE,SPOUSAL SUPPORT => have	184
0,2000	0,0036	MARRIAGE,MARRIAGE => defend	185
0,2000	0,0018	MARRIAGE,ALIMONY => take	186
0,2000	0,0036	LAWYER,CHILD SUPPORT => represent	187
0,2000	0,0018	DIVORCE,DECREE => determine	188
0,2000	0,0036	COUPLE,CHILD => have	189
0,2000	0,0036	DIVORCE,MARITAL AGREEMENT => consider	190
0,2000	0,0018	LAWYER,FEE => take	191
0,2000	0,0018	MARRIAGE,ALIMONY => receive	192
0,2000	0,0073	LAWYER,CHILD SUPPORT => ask for	193
0,2000	0,0036	DIVORCE,MARITAL AGREEMENT => be subject	194

0,2000	0,0018	VISITATION,LAWYER => might present	195
0,2000	0,0018	DIVORCE,SPOUSAL SUPPORT => should wait	196
0,2000	0,0036	DIVORCE,MARITAL AGREEMENT => respect	197
0,2000	0,0018	LAWYER,CHILD SUPPORT => be	198
0,2000	0,0018	DIVORCE,DECREE => may approve	199
0,2000	0,0054	DIVORCE,PROPERTY => ask for	200
0,2000	0,0036	PARTY,SPOUSAL SUPPORT => ask	201
0,2000	0,0036	SPOUSE,INHERITANCE => may include	202
0,2000	0,0018	MARRIAGE,MARRIAGE => be available	203
0,2000	0,0018	MARRIAGE,ALIMONY => may allege	204
0,2000	0,0018	LAWYER,CHILD SUPPORT => expect	205
0,2000	0,0018	DIVORCE,PROPERTY => be subject	206
0,2000	0,0036	SPOUSE,INHERITANCE => should consult	207
0,2000	0,0018	MARRIAGE,ALIMONY => must prove	208
0,1818	0,0036	COUPLE,DIVORCE => concern	209
0,1818	0,0073	PARTY,PROPERTY => inherit	210
0,1818	0,0036	MARRIAGE,MARITAL AGREEMENT => should consider	211
0,1818	0,0036	PARTY,ALIMONY => pay	212
0,1818	0,0036	PARTY,PROPERTY => leave	213
0,1818	0,0036	MARRIAGE,MARITAL AGREEMENT => include	214
0,1818	0,0036	COUPLE,DIVORCE => be not allow	215
0,1818	0,0036	PARTY,PROPERTY => own	216
0,1818	0,0036	COURT,CHILD SUPPORT => must justify	217
0,1818	0,0036	PARTY,ALIMONY => must prove	218
0,1818	0,0036	PARTY,PROPERTY => can use	219
0,1818	0,0036	PARTY,ALIMONY => ask	220
0,1818	0,0036	COUPLE,DIVORCE => desire	221
0,1818	0,0036	MARRIAGE,MARITAL AGREEMENT => be	222
0,1818	0,0036	COURT,CHILD SUPPORT => will consider	223
0,1667	0,0018	COURT,ALIMONY => will define	224
0,1667	0,0018	LAWYER,SPOUSAL SUPPORT => must prove	225
0,1667	0,0018	INHERITANCE,WILL => may bring	226
0,1667	0,0073	PARTY,CHILD CUSTODY => have	227
0,1667	0,0018	COURT,PROPERTY => decide	228
0,1667	0,0036	COUPLE,MARRIAGE => mean	229
0,1667	0,0018	COUPLE,ADOPTION => look for	230
0,1667	0,0018	LAWYER,LAWYER => should ask	231
0,1667	0,0036	VISITATION,CHILD => grant	232
0,1667	0,0018	DIVORCE,INHERITANCE => will continue	233
0,1667	0,0018	LAWYER,WILL => be not require	234

0,1667	0,0054	JUDGE,INHERITANCE => determine	235
0,1667	0,0073	PARTY,CHILD CUSTODY => ask	236
0,1667	0,0018	LAWYER,INHERITANCE => may justify	237
0,1667	0,0018	COURT,ALIMONY => hear	238
0,1667	0,0018	LAWYER,SPOUSAL SUPPORT => decide	239
0,1667	0,0018	INHERITANCE,WILL => indicate	240
0,1667	0,0018	COUPLE,ADOPTION => have	241
0,1667	0,0018	LAWYER,LAWYER => may agree	242
0,1667	0,0036	VISITATION,CHILD => allege	243
0,1667	0,0018	PARTY,LEGAL SEPARATION => receive	244
0,1667	0,0036	LAWYER,WILL => assist	245
0,1667	0,0018	LAWYER,ADOPTION => justify	246
0,1667	0,0036	LEGAL SEPARATION,LAWYER => may need	247
0,1667	0,0036	LAWYER,INHERITANCE => need	248
0,1667	0,0036	COURT,PROPERTY => be	249
0,1667	0,0018	COURT,ALIMONY => be subject	250
0,1667	0,0018	LAWYER,LAWYER => talk	251
0,1667	0,0018	VISITATION,CHILD => expect	252
0,1667	0,0018	DIVORCE,INHERITANCE => do not modify	253
0,1667	0,0036	LAWYER,ADOPTION => be	254
0,1667	0,0018	LEGAL SEPARATION,LAWYER => advocate	255
0,1667	0,0073	PARTY,CHILD CUSTODY => gain	256
0,1667	0,0036	COUPLE,MARRIAGE => intend	257
0,1667	0,0018	LAWYER,INHERITANCE => help	258
0,1667	0,0018	LAWYER,SPOUSAL SUPPORT => may represent	259
0,1667	0,0018	COUPLE,ADOPTION => may allege	260
0,1667	0,0018	LAWYER,LAWYER => expect	261
0,1667	0,0018	VISITATION,CHILD => can consult	262
0,1667	0,0036	PARTY,CHILD CUSTODY => may consent	263
0,1667	0,0018	LAWYER,ADOPTION => should consult	264
0,1667	0,0018	LEGAL SEPARATION,LAWYER => assist	265
0,1667	0,0018	LAWYER,INHERITANCE => should consider	266
0,1667	0,0018	LAWYER,SPOUSAL SUPPORT => do not guarantee	267
0,1667	0,0054	PARTY,CHILD CUSTODY => be	268
0,1667	0,0018	VISITATION,CHILD => consider	269
0,1667	0,0018	LAWYER,ADOPTION => may need	270
0,1667	0,0018	LAWYER,LAWYER => call	271
0,1667	0,0036	COUPLE,MARRIAGE => become	272
0,1667	0,0018	LAWYER,LAWYER => be	273
0,1667	0,0018	VISITATION,CHILD => have to evaluate	274

0,1667	0,0018	LAWYER,WILL => consider	275
0,1429	0,0018	DIVORCE,CHILD SUPPORT => solicit	276
0,1429	0,0036	MARRIAGE,LAWYER => may consult	277
0,1429	0,0036	COURT,SPOUSAL SUPPORT => judge	278
0,1429	0,0018	DIVORCE,ALIMONY => may award	279
0,1429	0,0036	COUPLE,DIVORCE => can change	280
0,1429	0,0018	MARRIAGE,CHILD CUSTODY => ask for	281
0,1429	0,0018	PARTY,MARITAL AGREEMENT => violate	282
0,1429	0,0018	LAWYER,CHILD CUSTODY => represent	283
0,1429	0,0018	MARRIAGE,ADOPTION => can consider	284
0,1429	0,0054	DIVORCE,CHILD SUPPORT => obtain	285
0,1429	0,0018	MARRIAGE,LAWYER => be not require	286
0,1429	0,0018	COURT,SPOUSAL SUPPORT => do not accept	287
0,1429	0,0018	DIVORCE,ALIMONY => have	288
0,1429	0,0018	PARTY,VISITATION => be	289
0,1429	0,0018	MARRIAGE,CHILD CUSTODY => gain	290
0,1429	0,0036	PARTY,MARITAL AGREEMENT => sign	291
0,1429	0,0018	LAWYER,CHILD CUSTODY => can assist	292
0,1429	0,0018	MARRIAGE,ADOPTION => resolve	293
0,1429	0,0054	COUPLE,DIVORCE => solicit	294
0,1429	0,0018	DIVORCE,CHILD SUPPORT => may award	295
0,1429	0,0036	LAWYER,COURT => allege	296
0,1429	0,0018	MARRIAGE,CHILD CUSTODY => judge	297
0,1429	0,0018	LAWYER,CHILD CUSTODY => charge	298
0,1429	0,0036	COUPLE,DIVORCE => be not allow	299
0,1429	0,0036	LAWYER,COURT => must prove	300
0,1429	0,0018	MARRIAGE,LAWYER => be	301
0,1429	0,0018	COURT,SPOUSAL SUPPORT => determine	302
0,1429	0,0018	MARRIAGE,CHILD CUSTODY => look for	303
0,1429	0,0018	PARTY,MARITAL AGREEMENT => be not require	304
0,1429	0,0018	LAWYER,CHILD CUSTODY => may require	305
0,1429	0,0018	MARRIAGE,ADOPTION => ask for	306
0,1429	0,0036	LAWYER,COURT => deny	307
0,1429	0,0018	DIVORCE,CHILD SUPPORT => be charge	308
0,1429	0,0018	MARRIAGE,LAWYER => represent	309
0,1429	0,0018	COURT,SPOUSAL SUPPORT => be subject	310
0,1429	0,0018	DIVORCE,ALIMONY => evaluate	311
0,1429	0,0018	PARTY,VISITATION => can require	312
0,1429	0,0018	MARRIAGE,ADOPTION => should consult	313
0,1429	0,0054	LAWYER,CHILD CUSTODY => ask for	314

0,1429	0,0018	MARRIAGE,ADOPTION => justify	315
0,1429	0,0036	COUPLE,DIVORCE => ask for	316
0,1429	0,0018	DIVORCE,ALIMONY => be	317
0,1429	0,0036	LAWYER,COURT => represent	318
0,1429	0,0018	PARTY,VISITATION => claim	319
0,1333	0,0036	LAWYER,PARTY => ask	320
0,1333	0,0073	LAWYER,PARTY => have	321
0,1333	0,0036	LAWYER,PARTY => should contract	322
0,1250	0,0018	DIVORCE,DIVORCE => proceed	323
0,1250	0,0073	SPOUSE,PROPERTY => have	324
0,1250	0,0018	PROPERTY,PROPERTY => be relate	325
0,1250	0,0036	CHILD,INHERITANCE => may obtain	326
0,1250	0,0073	COUPLE,PROPERTY => share	327
0,1250	0,0018	DIVORCE,CHILD CUSTODY => do not justify	328
0,1250	0,0036	SPOUSE,WILL => must decide	329
0,1250	0,0018	CHILD,PROPERTY => have	330
0,1250	0,0036	DIVORCE,DIVORCE => may use	331
0,1250	0,0018	SPOUSE,PROPERTY => can determine	332
0,1250	0,0018	PROPERTY,PROPERTY => represent	333
0,1250	0,0036	COUPLE,PROPERTY => become	334
0,1250	0,0018	DIVORCE,CHILD CUSTODY => resolve	335
0,1250	0,0036	LAWYER,MARITAL AGREEMENT => be not require	336
0,1250	0,0018	CHILD,PROPERTY => gain	337
0,1250	0,0018	DIVORCE,DIVORCE => know	338
0,1250	0,0054	SPOUSE,PROPERTY => own	339
0,1250	0,0018	PROPERTY,PROPERTY => may award	340
0,1250	0,0036	CHILD,INHERITANCE => have	341
0,1250	0,0036	COUPLE,PROPERTY => own	342
0,1250	0,0018	SPOUSE,WILL => replace	343
0,1250	0,0018	LAWYER,MARITAL AGREEMENT => may help	344
0,1250	0,0054	CHILD,PROPERTY => may claim	345
0,1250	0,0018	DIVORCE,DIVORCE => may seem	346
0,1250	0,0018	CHILD,INHERITANCE => may claim	347
0,1250	0,0018	COUPLE,PROPERTY => be subject	348
0,1250	0,0018	DIVORCE,CHILD CUSTODY => gain	349
0,1250	0,0018	SPOUSE,WILL => may need	350
0,1250	0,0018	LAWYER,MARITAL AGREEMENT => draw up	351
0,1250	0,0036	SPOUSE,PROPERTY => decide	352
0,1250	0,0018	PROPERTY,PROPERTY => claim	353
0,1250	0,0018	CHILD,INHERITANCE => be not allow	354

0,1250	0,0018	DIVORCE,CHILD CUSTODY => may include	355
0,1250	0,0018	SPOUSE,WILL => can include	356
0,1250	0,0018	LAWYER,MARITAL AGREEMENT => hear	357
0,1250	0,0036	CHILD,PROPERTY => become	358
0,1176	0,0036	COURT,CHILD CUSTODY => reject	359
0,1176	0,0018	PARTY,COURT => have hold	360
0,1176	0,0054	COURT,CHILD CUSTODY => grant	361
0,1176	0,0036	COURT,CHILD CUSTODY => must decide	362
0,1176	0,0018	PARTY,COURT => have rule	363
0,1176	0,0036	COURT,CHILD CUSTODY => have	364
0,1176	0,0018	PARTY,COURT => notify	365
0,1176	0,0036	COURT,CHILD CUSTODY => do not accept	366
0,1176	0,0036	PARTY,COURT => declare	367
0,1111	0,0036	PARTY,PARTY => violate	368
0,1111	0,0018	DIVORCE,MARRIAGE => leave	369
0,1111	0,0036	COURT,DIVORCE => help	370
0,1111	0,0036	PARTY,PARTY => ask for	371
0,1111	0,0036	PARTY,CHILD SUPPORT => may award	372
0,1111	0,0018	COURT,MARRIAGE => be	373
0,1111	0,0018	COURT,MARRIAGE => ask	374
0,1111	0,0036	PARTY,PARTY => give up	375
0,1111	0,0054	PARTY,CHILD SUPPORT => be oblige	376
0,1111	0,0036	COURT,DIVORCE => grant	377
0,1111	0,0018	COURT,MARRIAGE => may not accept	378
0,1111	0,0036	DIVORCE,MARRIAGE => grant	379
0,1111	0,0018	COURT,MARRIAGE => should be	380
0,1111	0,0018	PARTY,PARTY => deny	381
0,1111	0,0073	PARTY,CHILD SUPPORT => ask	382
0,1111	0,0018	COURT,DIVORCE => hear	383
0,1111	0,0036	PARTY,PARTY => disagree	384
0,1111	0,0036	PARTY,CHILD SUPPORT => be issue	385
0,1111	0,0036	COURT,DIVORCE => be	386
0,1000	0,0054	DIVORCE,PROPERTY => have	387
0,1000	0,0018	PARTY,SPOUSAL SUPPORT => include	388
0,1000	0,0054	PARTY,SPOUSAL SUPPORT => claim	389
0,1000	0,0018	MARRIAGE,MARRIAGE => help	390
0,1000	0,0036	MARRIAGE,SPOUSE => do not share	391
0,1000	0,0018	COUPLE,CHILD => leave	392
0,1000	0,0036	PARTY,DIVORCE => want	393
0,1000	0,0018	SPOUSE,INHERITANCE => have	394

0,1000	0,0018	DIVORCE,PROPERTY => may determine	395
0,1000	0,0018	PARTY,SPOUSAL SUPPORT => be	396
0,1000	0,0036	SPOUSE,INHERITANCE => receive	397
0,1000	0,0018	MARRIAGE,MARRIAGE => move toward	398
0,1000	0,0036	PARTY,DIVORCE => will depend	399
0,1000	0,0018	DIVORCE,PROPERTY => lose	400
0,1000	0,0054	DIVORCE,LAWYER => represent	401
0,1000	0,0018	SPOUSE,INHERITANCE => leave	402
0,1000	0,0018	MARRIAGE,MARRIAGE => advocate	403
0,1000	0,0036	MARRIAGE,SPOUSE => can make	404
0,1000	0,0073	DIVORCE,LAWYER => be	405
0,1000	0,0018	COUPLE,CHILD => abandon	406
0,1000	0,0018	DIVORCE,MARITAL AGREEMENT => do not guarantee	407
0,1000	0,0018	COUPLE,CHILD => live apart	408
0,1000	0,0054	PARTY,SPOUSAL SUPPORT => pay	409
0,1000	0,0018	MARRIAGE,MARRIAGE => become	410
0,1000	0,0036	PARTY,DIVORCE => establish	411
0,1000	0,0018	COUPLE,CHILD => be	412
0,1000	0,0018	DIVORCE,PROPERTY => decide	413
0,1000	0,0018	PARTY,SPOUSAL SUPPORT => may award	414
0,1000	0,0036	PARTY,DIVORCE => file for	415
0,1000	0,0018	MARRIAGE,MARRIAGE => refrain	416
0,1000	0,0036	MARRIAGE,SPOUSE => want	417
0,1000	0,0036	DIVORCE,LAWYER => must prove	418
0,1000	0,0018	PARTY,SPOUSAL SUPPORT => can agree	419
0,1000	0,0018	MARRIAGE,MARRIAGE => sign	420
0,1000	0,0036	MARRIAGE,SPOUSE => assume	421
0,1000	0,0018	DIVORCE,MARITAL AGREEMENT => be use	422
0,1000	0,0054	DIVORCE,LAWYER => should consult	423
0,1000	0,0036	MARRIAGE,SPOUSE => be involve	424
0,1000	0,0018	SPOUSE,INHERITANCE => claim	425
0,1000	0,0018	COUPLE,CHILD => bring up	426
0,0909	0,0036	COURT,CHILD SUPPORT => decide	427
0,0909	0,0018	PARTY,ALIMONY => look for	428
0,0909	0,0018	COURT,CHILD SUPPORT => give out	429
0,0909	0,0036	COUPLE,DIVORCE => file for	430
0,0909	0,0018	MARRIAGE,MARITAL AGREEMENT => determine	431
0,0909	0,0018	COURT,CHILD SUPPORT => may not accept	432
0,0909	0,0036	COUPLE,DIVORCE => solicit	433
0,0909	0,0054	COURT,CHILD SUPPORT => grant	434

0,0909	0,0018	PARTY,ALIMONY => need	435
0,0909	0,0018	MARRIAGE,MARITAL AGREEMENT => must represent	436
0,0909	0,0018	COUPLE,DIVORCE => sign	437
0,0909	0,0018	MARRIAGE,MARITAL AGREEMENT => formalise	438
0,0909	0,0018	COURT,CHILD SUPPORT => resolve	439
0,0909	0,0054	COURT,CHILD SUPPORT => order	440
0,0909	0,0018	COUPLE,DIVORCE => ask for	441
0,0909	0,0018	MARRIAGE,MARITAL AGREEMENT => use	442
0,0909	0,0018	COUPLE,DIVORCE => allege	443
0,0909	0,0018	MARRIAGE,MARITAL AGREEMENT => be set	444
0,0909	0,0036	COURT,CHILD SUPPORT => can change	445
0,0833	0,0018	COUPLE,MARRIAGE => desire	446
0,0833	0,0018	PARTY,CHILD CUSTODY => may award	447
0,0833	0,0018	COUPLE,MARRIAGE => break up	448
0,0833	0,0054	PARTY,CHILD CUSTODY => claim	449
0,0833	0,0018	COUPLE,MARRIAGE => terminate	450
0,0833	0,0036	COUPLE,MARRIAGE => should consider	451
0,0833	0,0018	COUPLE,MARRIAGE => must know	452
0,0833	0,0018	COUPLE,MARRIAGE => assume	453
0,0714	0,0018	COUPLE,DIVORCE => concern	454
0,0714	0,0018	LAWYER,COURT => may accept	455
0,0714	0,0018	COUPLE,DIVORCE => file for	456
0,0714	0,0036	LAWYER,COURT => solicit	457
0,0714	0,0018	LAWYER,COURT => conduct	458
0,0714	0,0018	COUPLE,DIVORCE => sign	459
0,0714	0,0018	COUPLE,DIVORCE => desire	460
0,0714	0,0018	LAWYER,COURT => ask	461
0,0714	0,0054	COUPLE,DIVORCE => obtain	462
0,0714	0,0018	LAWYER,COURT => may grant	463
0,0714	0,0018	COUPLE,DIVORCE => may allege	464
0,0714	0,0018	LAWYER,COURT => defend	465
0,0667	0,0018	LAWYER,PARTY => may need	466
0,0667	0,0018	LAWYER,PARTY => assist	467
0,0667	0,0018	LAWYER,PARTY => represent	468
0,0667	0,0018	LAWYER,PARTY => do not guarantee	469
0,0667	0,0036	LAWYER,PARTY => consult	470
0,0667	0,0018	LAWYER,PARTY => recommend	471
0,0667	0,0036	LAWYER,PARTY => pay	472
0,0667	0,0018	LAWYER,PARTY => may help	473
0,0667	0,0018	LAWYER,PARTY => advice	474

0,0588	0,0036	COURT,CHILD CUSTODY => recieve	475
0,0588	0,0018	PARTY,COURT => ask for	476
0,0588	0,0018	COURT,CHILD CUSTODY => may deny	477
0,0588	0,0018	PARTY,COURT => consider	478
0,0588	0,0018	PARTY,COURT => get	479
0,0588	0,0018	COURT,CHILD CUSTODY => assess	480
0,0588	0,0018	PARTY,COURT => be file	481
0,0588	0,0018	PARTY,COURT => register	482
0,0588	0,0018	COURT,CHILD CUSTODY => reach	483
0,0588	0,0018	PARTY,COURT => hear	484
0,0588	0,0036	COURT,CHILD CUSTODY => be base	485
0,0588	0,0036	COURT,CHILD CUSTODY => provide	486
0,0588	0,0018	PARTY,COURT => meet	487
0,0588	0,0036	COURT,CHILD CUSTODY => give out	488
0,0588	0,0018	PARTY,COURT => review	489
0,0588	0,0018	PARTY,COURT => have go	490
0,0556	0,0036	COURT,DIVORCE => be approve	491
0,0556	0,0018	PARTY,CHILD SUPPORT => obtain	492
0,0556	0,0036	COURT,DIVORCE => prepare	493
0,0556	0,0018	PARTY,PARTY => leave	494
0,0556	0,0054	PARTY,CHILD SUPPORT => claim	495
0,0556	0,0018	COURT,DIVORCE => may result	496
0,0556	0,0018	PARTY,PARTY => should consult	497
0,0556	0,0018	COURT,DIVORCE => provide	498
0,0556	0,0036	PARTY,CHILD SUPPORT => may enter	499
0,0556	0,0018	PARTY,PARTY => do not have	500
0,0556	0,0054	PARTY,CHILD SUPPORT => recieve	501
0,0556	0,0036	COURT,DIVORCE => decide	502
0,0556	0,0018	COURT,DIVORCE => be available	503
0,0556	0,0018	PARTY,PARTY => choose	504
0,0556	0,0018	PARTY,CHILD SUPPORT => can change	505
0,0556	0,0018	COURT,DIVORCE => look after	506
0,0556	0,0018	PARTY,PARTY => inherit	507
0,0556	0,0018	PARTY,CHILD SUPPORT => contest	508
0,0556	0,0018	PARTY,PARTY => be	509
0,0556	0,0018	PARTY,CHILD SUPPORT => may modify	510
0,0556	0,0036	COURT,DIVORCE => recommend	511
0,0556	0,0018	PARTY,PARTY => claim	512
0,0556	0,0018	PARTY,CHILD SUPPORT => can not refuse	513
0,0556	0,0018	COURT,DIVORCE => apply for	514

0,0556	0,0036	PARTY,PARTY => have	515
0,0556	0,0036	PARTY,CHILD SUPPORT => allege	516
0,0556	0,0018	COURT,DIVORCE => issue	517
0,0556	0,0036	PARTY,CHILD SUPPORT => gain	518
0,0500	0,0018	MARRIAGE,SPOUSE => stay	519
0,0500	0,0036	MARRIAGE,SPOUSE => be	520
0,0500	0,0018	PARTY,DIVORCE => mean	521
0,0500	0,0054	DIVORCE,LAWYER => be not require	522
0,0500	0,0018	MARRIAGE,SPOUSE => save	523
0,0500	0,0054	PARTY,DIVORCE => ask	524
0,0500	0,0018	DIVORCE,LAWYER => look for	525
0,0500	0,0018	MARRIAGE,SPOUSE => take place	526
0,0500	0,0073	PARTY,DIVORCE => require	527
0,0500	0,0018	DIVORCE,LAWYER => accept	528
0,0500	0,0018	PARTY,DIVORCE => should be	529
0,0500	0,0018	MARRIAGE,SPOUSE => feel	530
0,0500	0,0036	PARTY,DIVORCE => agree	531
0,0500	0,0018	DIVORCE,LAWYER => support	532
0,0500	0,0018	MARRIAGE,SPOUSE => must provide	533
0,0500	0,0018	DIVORCE,LAWYER => know	534
0,0500	0,0036	PARTY,DIVORCE => must justify	535
0,0500	0,0036	PARTY,DIVORCE => consent	536
0,0500	0,0018	DIVORCE,LAWYER => need	537
0,0500	0,0018	PARTY,DIVORCE => apply	538
0,0500	0,0018	DIVORCE,LAWYER => solicit	539
0,0500	0,0018	MARRIAGE,SPOUSE => reject	540
0,0500	0,0054	PARTY,DIVORCE => resolve	541
0,0500	0,0018	DIVORCE,LAWYER => should wait	542
0,0500	0,0018	MARRIAGE,SPOUSE => must ask	543
0,0500	0,0018	PARTY,DIVORCE => submit	544
0,0500	0,0036	DIVORCE,LAWYER => indicate	545
0,0500	0,0018	MARRIAGE,SPOUSE => determine	546
0,0500	0,0018	DIVORCE,LAWYER => want	547
0,0500	0,0018	MARRIAGE,SPOUSE => establish	548
0,0500	0,0018	PARTY,DIVORCE => decide	549
0,0500	0,0018	DIVORCE,LAWYER => involve	550
0,0500	0,0036	PARTY,DIVORCE => must know	551