Abordagens para Estimar Relevância de Relações Não-Taxonômicas Extraídas de Corpus de Domínio

Lucelene Lopes[1], **Maria José Finatto**[2], **Alena Ciulla**[2], **Renata Vieira**[1]

Faculdade de Informática – PUCRS
Porto Alegre – RS – Brazil

[2]Faculdade de Letras – UFRGS
Porto Alegre – RS – Brazil

{lucelene.lopes,renata.vieira}@pucrs.br {alena.ciulla,mjose.finatto}@ufrgs.br

Abstract. *This paper performs a comparison between two approaches to weight the relevance of extracted non-taxonomic relations found in domain corpora. The first approach computes the relevance according to the verb absolute frequency. The second approach computes the relevance according to the verb frequency and uniqueness in each corpus using tf-dcf relevance index, an index that takes into account the high frequency of verbs in the target corpus, and the low frequency in contrasting corpora. The preliminary results are evaluated for three domain corpora and the top relevant relations are evaluated by expert terminologists.*

Resumo. *Este artigo apresenta uma comparação entre duas abordagens de ponderação de relevância de relações não-taxonômicas extraídas de corpora de domínio. A primeira abordagem calcula a relevância de acordo com a frequência absoluta dos verbos. A segunda abordagem calcula a relevância de acordo com a frequência do verbo e sua especificidade em cada corpus utilizando o índice de relevância tf-dcf, um índice que leva em consideração a alta frequência no corpus alvo, e a baixa frequência em corpora contrastantes. Os resultados preliminares foram avaliados para três corpora de domínio e as relações mais relevantes foram avaliadas por terminologistas.*

1. Introdução

A descoberta de relações não-taxonômicas é uma tarefa difícil da aprendizagem de ontologias [Kavalec and Átek 2005]. Esta tarefa se divide em dois momentos: detectar que conceitos estão relacionados; e etiquetar cada relação detectada (em geral pela definição de um verbo que descreve a relação) [Sánchez and Moreno 2008]. Diversos trabalhos científicos seguem esse processo, por exemplo [Weichselbraun et al. 2009, Serra and Girardi 2011, Ferreira et al. 2013]. Esses trabalhos baseiam-se na detecção de verbos transitivos que relacionam dois sintagmas nominais, usualmente sujeito e objeto. Caso esses sintagmas nominais representem conceitos, ou instâncias de conceitos, esta tripla (sujeito,verbo,objeto) representará uma instância de uma relação da ontologia.

O estudo e a descrição dos verbos do português do Brasil (PB) são elementos importantes no âmbito dos Estudos da Linguagem, visto que, entre outros elementos da linguagem escrita, são elementos vitais para subsidiar uma série de recursos de representação

e de recuperação de informação com apoio computacional a partir de acervos documentais. Afinal, os verbos, além do papel fundamental para o funcionamento gramatical de qualquer língua, oferecem via predicação, elementos para a representação de diferentes relações não-taxonômicas, que extrapolam relações hierárquicas do tipo parte-todo. Os elementos relacionados assim pelos verbos podem corresponder a termos ou a conceitos de um domínio.

Um exemplo dessa importância do verbo, para a depreensão de relações entre itens de uma ontologia ou mapa conceitual é a expressão da relação não-taxonômica do tipo "X causa Y" que se depreende, por exemplo, nas seguintes frases: (1) Tabagismo causa câncer./ (2) Tabagismo pode causar câncer de pulmão./ (3) Tabagismo favorece o aparecimento de câncer de pulmão. Todavia, muitos dos trabalhos de que dispomos em PLN e mesmo em Linguística, tem se centrado na descrição de estruturas sintáticas de verbos, como a anotação de papéis semânticos ou de argumentos, que visa reconhecer padrões de associação entre determinados sujeitos e complementos nominais e um dado tipo de verbo [Scarton 2013, Zilio 2015], sem preocupação específica de caracterizar padrões de verbos em diferentes domínios.

Assumindo a existência de uma forma eficiente e eficaz de recuperar automaticamente este tipo de relação de um corpus de domínio [Lopes 2012], o desafio é filtrar dentre as relações extraídas quais são particularmente relevantes para o domínio. Na verdade, esse tipo de detecção das relações frequentemente mostra um número grande de relações e poucos estudos se dedicam a estabelecer uma ordem de relevância entre as relações detectadas.

Este artigo apresenta um trabalho inicial que propõe uma forma alternativa de estimar a relevância de relações não-taxonômicas de um domínio baseado no contraste com outros domínios. Dessa forma, a próxima seção apresenta a abordagem elementar que considera a relevância dos verbos extraídos pela sua frequência absoluta e a abordagem proposta. Em seguida, a seção 3 apresenta a comparação destas duas abordagens sobre três corpora de domínio.

2. Abordagens de Ponderação

Nesta seção apresenta-se a abordagem elementar que assume a frequência absoluta e a abordagem proposta que utiliza a frequência contrastada com outros corpora como indicador de relevância.

2.1. Abordagem por Frequência Absoluta

A primeira abordagem considerada neste trabalho toma os verbos mais frequentes como sendo as relações não-taxonômicas mais relevantes. Dessa forma, esse processo analisa o corpus alvo e identifica os verbos que relacionam dois termos, ou seja, os textos anotados são percorridos e todo verbo que conecta dois sintagmas nominais (um sujeito e um objeto) é considerado uma instância de relação. As instâncias são contabilizadas, considerando-se os verbos em sua forma canônica (infinitivo), ou seja, a frequência absoluta dos verbos é contabilizada, ignorando-se tempos verbais e flexões de pessoa e número.

A vantagem dessa abordagem é que os verbos com maior frequência absoluta serão naturalmente mais produtivos na quantidade de relações geradas, pois quanto maior

o número de instâncias, maior o número de relações a considerar. Segundo o levantamento de um grande corpus do PB [Biderman 1998], que gerou um dicionário de frequências, temos a situação sobre verbos mais frequentemente empregados, independentemente de domínio, conforme apresentado na tabela 1. No entanto, essa abordagem por frequência absoluta tenderá a privilegiar os verbos usuais semelhante aos encontrados por [Biderman 1998].

Tabela 1. Os 20 verbos lematizados no infinitivo mais frequentes no corpus de Biderman - fonte: [Finatto 2012].

ranking	verbo	ranking	verbo	ranking	verbo	ranking	verbo	ranking	verbo
1	ser	5	poder	9	dar	13	ficar	17	chegar
2	ter	6	dizer	10	ver	14	achar	18	precisar
3	ir	7	haver	11	saber	15	dever	19	começar
4	estar	8	fazer	12	querer	16	falar	20	olhar

2.2. Abordagem por Frequência Contrastada (*tf-dcf*)

Buscando ter mais especificidade nas relações a considerar, a abordagem proposta neste trabalho baseia-se na aplicação do índice *tf-dcf* (term frequency, disjoint corpora frequency) [Lopes et al. 2012]. Este índice é originalmente empregado para calcular a relevância de um termo em um corpus alvo, diretamente proporcional à frequência absoluta do termo no corpus alvo e inversamente proporcional a sua frequência em corpora contrastantes. Dessa forma, essa abordagem inicia contabilizando as ocorrências dos verbos no corpus alvo e em todos os corpora contrastantes. Em seguida, os valores de frequência absoluta dos verbos são utilizados como entrada para a fórmula do índice *tf-dcf* aplicada a verbos, em vez de termos[1] (Eq. 1).

$$\textit{tf-dcf}_v^{(c)} = \frac{\textit{tf}_v^{(c)}}{\prod_{g \in \mathcal{G}} 1 + \log\left(1 + \textit{tf}_v^{(g)}\right)} \tag{1}$$

Onde $\textit{tf}_v^{(c)}$ representa a frequência absoluta do verbo v no corpus c; e $\mathcal{G}$ representa o conjunto de corpora contrastantes.

3. Experimentos

Para ilustrar as diferenças das duas abordagens apresentadas escolhemos três corpora de domínio, um sobre Geologia (Geo), um sobre Pneumopatias (Pneumo), e o Curso de Linguística Geral (CLG), de Ferdinand de Saussure, um texto fundamental para a área de Línguistica. Adicionalmente, como a abordagem baseada no índice *tf-dcf* requer o uso de corpora contrastantes, foram utilizados três outros corpora sobre Modelagem estocástica (SM), Mineração de dados (DM) e Processamento paralelo (PP) como contrastantes. Assim, para calcular os índices *tf-dcf* de cada corpus são usados como contrastantes os dois outros corpora, além dos três corpora adicionais (SM, DM, PP). A tabela 2 apresenta as características desses corpora e indica, para os três corpora alvos o número de relações extraídas. A tabela 3 apresenta as dez relações consideradas mais relevantes para cada um dos corpora segundo a frequência absoluta (*tf*) e o índice *tf-dcf*.

Tabela 2. Características dos corpora utilizados.

corpus	Número de Textos	Número de Sentenças	Número de Tokens	Relações Extraídas
Geo	139	39,648	1,165,220	1,395
Pneumo	71	9,239	241,806	433
CLG	25	3,486	34,295	192
SM	88	44,222	1,173,401	
DM	53	42,932	1,127,816	
PP	62	40,928	1,086,771	

Tabela 3. Relações mais relevantes de cada corpus segundo ambas abordagens.

	Geo		Pneumo		CLG	
#	*tf*	*tf-dcf*	*tf*	*tf-dcf*	*tf*	*tf-dcf*
1	ser	recobrir	ser	acometer	ser	obscurecer
2	apresentar	cortar	apresentar	inalar	ter	acentuar
3	ter	aflorar	ter	contaminar	constituir	consagrar
4	mostrar	erodir	estar	contraindicar	estar	pode equiparar
5	estar	condicionar	mostrar	dever intimidar	apresentar	falsear
6	representar	retrabalhar	poder ser	poder agravar	tornar	suscitar
7	constituir	cristalizar	demonstrar	poder contaminar	fazer	unificar
8	possuir	ser depositar	revelar	poder justificar	formar	pode exprimir
9	indicar	postular	fazer	recomendar	produzir	transtornar
10	permitir	drenar	ser considerar	infectar	dar	apagar

Conforme [Biderman 1998], na sua lista dos verbos mais frequentes do PB, encabeçando-a temos os auxiliares "ser", "estar", "ter". Até o verbo "ir" registrou um elevado número de valores modais e aspectuais, razão para estar também nos primeiros lugares da hierarquia dos verbos usuais. Constam dessa lista ainda verbos modalizadores como "poder", ou vicários, e/ou suportes como "fazer", "dar"; entre os de significação plena, apenas "dizer", "falar", "olhar" e "ver" [Biderman 1998] (p. 174). Se excluirmos os verbos que integram uma locução ou que são auxiliares do levantamento por domínio com *tf-dcf* , temos que ("recobrir", "cortar" e "aflorar"); ("acometer", "inalar" e "contaminar"); e ("obscurecer", "acentuar" e "consagrar") seriam, respectivamente, os verbos de maior especificidades nos domínios de Geologia, Pneumologia e Linguística, considerando-se os corpora sob exame e os tipos de textos envolvidos.

Tabela 4. Exemplos de relações mais relevantes para o corpus Geo.

#	Frequência Absoluta	Índice *tf-dcf*
1	superfície → ser → molhável	cascalho → recobrir → formação ferruginosa
	É mostrado que a ausência de ácidos não garante que a superfície **será** molhável por a fase aquosa.	Horizonte cascalhento ferruginoso friável de superfície corresponde ao solo ou os cascalhos que eventualmente **recobrem** as formações ferruginosas.
2	footwall → apresentar → soerguimento	corpo de granito → cortar → foliação gnáissica
	Desta forma, o footwall **apresenta** sempre um soerguimento, enquanto o hangingwall é o domínio subsidente.	Os corpos de granito e pegmatito são usualmente subconcordantes, mas com freqüência **cortam** a foliação gnáissica.
3	empregado → ter → gerente	tonalito → aflorar → belt de Crixás
	Utilizando-se este paradigma, pode-se induzir que cada empregado **tem** um gerente, o que é uma generalização a partir dos dados existentes naquelas relações.	No extremo sudoeste da área o tonalito **aflora** como um corpo triangular, limitado a nordeste pelos Gnaisses Crixás Açu e a oeste pelo greenstone belt de Crixás.

[1]A única adaptação da formulação do índice *tf-dcf* para termos ao considerar verbos consiste em considerar frequência absoluta de verbos (tf_v) ao invés de frequência de termos (tf_t).

Tabela 5. Exemplos de relações mais relevantes para o corpus Pneumo.

#	Frequência Absoluta	Índice *tf-dcf*
1	efeito → ser → fator importante	espondilite tuberculosa → acometer → disco intervertebral
	O efeito idade **é** um fator importante na chance de abandono do hábito de fumar.	A espondilite tuberculosa **acomete** o disco intervertebral mais tardiamente no curso da doença.
2	moxifloxacina → apresentar → metabolização hepática	nadador → inalar → grande quantidade de ar
	A moxifloxacina, entretanto, **apresenta** metabolização hepática, e a principal via de excreção é a biliar.	Durante a prática do esporte, os nadadores **inalam** grandes quantidades de ar logo acima de a superfície da água.
3	stress → ter → papel relevante	balangeroíta → contaminar → corpos minerais
	Como exemplos, podemos citar as doenças coronarianas, em as quais o stress **tem** um papel relevante.	A balangeroíta **contamina** os corpos minerais da Itália, e assim por diante.

Tabela 6. Exemplos de relações mais relevantes para o corpus CLG.

#	Frequência Absoluta	Índice *tf-dcf*
1	língua → ser → sistema	escrita → obscurecer → visão da língua
	Visto **ser** a língua um sistema em que todos os termos são solidários e o valor de um resulta tão somente da presença simultânea de outros, segundo o esquema:	O resultado evidente de tudo isso é que a escrita **obscurece** a visão da língua.
2	língua → ter → caráter de fixidez	evolução de som → acentuar → diferença existente
	Se a língua **tem** um caráter de fixidez, não é somente porque está ligada ao peso da coletividade, mas também porque está situada no tempo.	A evolução dos sons não faz mais que **acentuar** as diferenças existentes antes de ela.
3	língua → constituir → sistema	uso → consagrar → dupla grafia
	Uma língua **constitui** um sistema.	Vimos na que, contrariamente ao que se verifica para outros sons, o uso **consagrou** para aqueles uma dupla grafia.

As tabelas 4, 5 e 6 apresentam exemplos (sentenças do corpus) das três relações mais relevantes para cada um dos corpora, respectivamente, segundo cada uma das abordagens. Observando estes exemplos, percebe-se que as relações mais relevantes segundo abordagem baseada no índice *tf-dcf* apresentam características claras de relações não-taxonômicas. Por exemplo, observa-se as triplas geradas por *tf-dcf* "cascalho **recobre** formação ferruginosa", "espondite tuberculosa **acomete** disco intervertebral", e "escrita **obscurece** visão da língua".

Já os exemplos das relações mais relevantes segundo a frequência absoluta tem um caracter que se assemelha mais a definição de propriedades/atributos, como é o caso de "superfície **é** molhável", ou ainda de "stress **tem** papel relevante". Ainda encontra-se casos que podem ser vistos como uma relação taxonômica, como por exemplo: "língua **é** sistema", ou seja, uma língua é um tipo de sistema.

4. Considerações Finais e Trabalhos Futuros

Neste estudo, mostramos dois tipos de abordagens no que diz respeito ao tratamento automático dos verbos em corpora de domínio com o propósito de identificar relações não-taxonômicas mais relevantes. Enquanto que a primeira abordagem, que considera a frequência em termos absolutos, aponta para aqueles verbos que são mais gerais da língua, a segunda abordagem, que se vale do índice *tf-dcf*, fornece uma lista de verbos que são mais específicos do domínio a que pertencem os textos.

Acreditamos, portanto, que atingimos nosso objetivo de identificar as relações mais relevantes para o domínio, contribuição do estudo através do índice *tf-dcf* que consiste no auxílio à construção de ontologias e na recuperação automática de informações, visto que acrescenta dados importantes sobre o verbo, um elemento vital - e pouco explorado, do ponto de vista do processamento automático - para o funcionamento da língua.

Além disso, temos também uma importante contribuição para os Estudos da Linguagem, ressaltando o papel dos verbos em diferentes domínios.

Cabe observar, contudo, que, quanto aos corpora em exame neste estudo, o CLG destaca-se dos outros corpora analisados, por vários motivos. Em primeiro lugar, ainda que se trate de um texto importante dentro do domínio da Linguística, não é uma compilação de textos científicos, como os corpora de Geologia e de Pneumopatias e, além disso, é uma tradução de um texto escrito originalmente em francês, em 1916. Outro aspecto é o de que é o único representante de um domínio de áreas humanas, enquanto que todos os outros são das áreas Exatas, da Saúde ou das Ciências Naturais, incluindo-se os corpora contrastantes. Por isso, fica como sugestão para trabalhos futuros, a contraposição dos verbos do CLG com os verbos de um corpus de textos de jornais, por exemplo, em que a linguagem ordinária desse gênero pode, em contraste, oferecer um panorama mais específico do domínio da Linguística.

Referências

Biderman, M. T. C. (1998). *A face quantitativa da linguagem: um dicionário de freqüências do português*. Alfa, São Paulo, Brasil.

Ferreira, V. H., Lopes, L., Vieira, R., and Finatto, M. J. B. (2013). Automatic extraction of domain specific non-taxonomic relations from portuguese corpora. In *Knowledge Discovery in Ontologies - KDO 2013*, Proc. of WI-IAT 2013, pages 161–165.

Finatto, M. J. B. (2012). Projeto porpopular, frequência de verbos em português e no jornal popular popular brasileiro. In *As Ciências do Léxico: lexicologia, lexicografia, terminologia, volume VI*, pages 277–244. Edit. da UFMS/Lab. de Edição FALE-UFMG.

Kavalec, M. and Átek, V. S. (2005). A study on automated relation labelling in ontology learning. In *Ontology Learning from Text: Methods, Evaluation and Applications*, pages 44–58. IOS Press.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Sánchez, D. and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.*, 64(3):600–623.

Scarton, C. E. (2013). Verbnet.br: construção semiautomática de um léxico verbal online e independente de domínio para o português do brasil. Master's thesis, ICMC - USP.

Serra, I. and Girardi, R. (2011). A process for extracting non-taxonomic relationships of ontologies from text. *Intelligent Information Management*, 3:119–124.

Weichselbraun, A., Wohlgenannt, G., Scharl, A., Granitzer, M., Neidhart, T., and Juffinger, A. (2009). Discovery and evaluation of non-taxonomic relations in domain ontologies. *International Journal of Metadata, Semantics and Ontologies*, 4(3):212–222.

Zilio, L. (2015). *Um Recurso Léxico com Anotação de Papéis Semânticos para o Português*. PhD thesis, PPG Letras - UFRGS.