EXTRAÇÃO AUTOMÁTICA DE *SUBCATEGORIZATION FRAMES* A PARTIR DE *CORPORA* EM PORTUGUÊS

1 Introdução

A tarefa de identificar automaticamente *subcategorization frames* (SCFs), que se enquadra como um tipo de aquisição lexical, a partir de *corpora* é um desafio no Processamento de Linguagem Natural (PLN) que tem sido trabalhado em diversas línguas. Tal tarefa, além de servir para a classificação de elementos linguísticos (como, por exemplo, de verbos e substantivos), também pode ser utilizada para várias tarefas de descrição de linguagem, podendo inclusive auxiliar a melhorar o desempenho de um *parser* [1]. Além disso, a aquisição lexical automática pode ser útil para classificação automática de verbos ([2], [3], [4]) e extração de informação [5], entre outros.

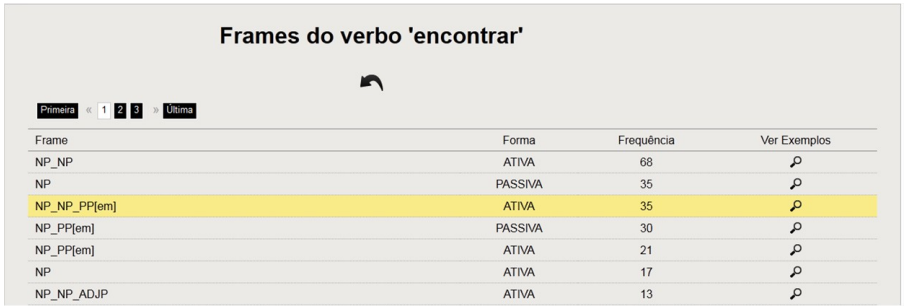
Tendo essa perspectiva em mente, este trabalho apresenta um sistema para a identificação, extração e avaliação de SCFs para o português (omitido devido a revisão cega), o qual iniciou como uma adaptação para o português do sistema de Messiant [6], e mostra a aplicação desse sistema em dois trabalhos de descrição do português baseados em *corpora*. Assim, os objetivos aqui propostos são:

1. Descrever a ferramenta de extração de SCFs
2. Relatar a sua aplicação prática em dois estudos
3. Apresentar brevemente os resultados desses estudos

Na seção a seguir, apresentamos uma definição de *subcategorization frames* e indicamos alguns trabalhos que abordam o tema. Na Seção 3, discorremos brevemente sobre o funcionamento do sistema desenvolvido. A Seção 4 é dedicada à apresentação de dois estudos que utilizaram a ferramenta para diferentes fins, e a Seção 5 é reservada para as considerações finais.

2 Subcategorization Frames

Um *subcategorization frame* (SCF) é uma representação sintática da estrutura de uma sentença ou sintagma que, observada em grandes extensões de texto, permitem a classificação de determinados elementos linguísticos. Neste trabalho, os elementos linguísticos relevantes são os verbos, sendo que os SCFs representam a estrutura argumental dos verbos presentes em sentenças de diferentes *corpora*. Na Figura 1, apresentamos alguns exemplos de frames do verbo "encontrar" extraídos de um corpus de textos de Cardiologia.



Frame	Forma	Frequência	Ver Exemplos
NP_NP	ATIVA	68	
NP	PASSIVA	35	
NP_NP_PP[em]	ATIVA	35	
NP_PP[em]	PASSIVA	30	
NP_PP[em]	ATIVA	21	
NP	ATIVA	17	
NP_NP_ADJP	ATIVA	13	

Figura 1: Visualização de frames do verbo "encontrar" em um corpus de Cardiologia

Os sistemas de SCFs foram desenvolvidos para várias línguas, como inglês ([7] e [1]), alemão ([8], francês ([6]) e italiano ([9]). Tais trabalhos foram desenvolvidos principalmente para o reconhecimento de SCFs verbais. No inglês, os trabalhos de SCFs já foram utilizados, por exemplo, para expandir a VerbNet ([10] e [11]), um repositório de verbos com anotação de papéis semânticos. Para o português, o trabalho com SCFs está ainda em seus primeiros passos. Podemos citar o sistema de Augustini [12], que procura abranger também os frames de preposições, substantivos e adjetivos. A grande restrição desse estudo, porém, é que os SCFs tem de ser determinados antes da extração.

3 Extrator de Subcategorization Frames

O sistema que apresentamos se ocupa exclusivamente de SCFs verbais. A entrada utilizada são *corpora* anotados pelo *parser* PALAVRAS [13], de modo que o sistema extrai uma lista de SCFs para cada verbo, contendo informações de frequência para todas as sentenças no *corpus*, como pôde ser visto na Figura 1. Após a extração, o sistema armazena as informações em um banco de dados.

No momento do armazenamento, o sistema apresenta duas opções: um armazenamento dos dados da forma como foram extraídos, sem qualquer alteração em

sua forma (por exemplo, uma topicalização de objeto direto não seria colocada na forma direta); ou uma avaliação do tipo de estrutura sintática, de modo que as estruturas sejam classificadas em uma ordem direta predeterminada, a qual é apresentada a seguir:

- arg1 - sujeito
- arg2 - objeto direto
- arg3 - objeto indireto
- arg4 - adjetivo (separado de um substantivo)
- arg5 - adjunto adverbial

O processo de funcionamento do sistema é o seguinte:

- Passo 1 - Para cada sentença, o sistema extrai todos os verbos;
- Passo 2 - Para cada verbo na sentença, o sistema busca as suas dependências (ou seja, os elementos que se ligam, de acordo com a anotação do *parser*, ao verbo);
- Passo 3 - As demais informações (morfossintáticas e sintáticas) anotadas pelo parser são avaliadas e os constituintes relevantes são utilizados para construir os SCFs.

O sistema desenvolvido mostrou-se adaptável conforme as necessidades. Esse fato é comprovado pelas diferentes aplicações às quais o sistema se destina, sendo que, na seção a seguir, apresentamos dois trabalhos distintos que fazem uso do sistema.

4 Estudos Desenvolvidos com o Sistema

4.1 Léxico de Verbos

Um dos estudos nos quais o sistema de extração de SCFs vem sendo utilizado envolve a construção de um léxico de verbos para o português brasileiro (omitido devido a revisão cega), baseado na VerbNet [10]. Para essa tarefa, o sistema foi adaptado de modo a apresentar a posição do verbo e do sujeito no frame. Portanto, em comparação com a coluna 1 da Figura 1 os padrões são um pouco diferentes, como por exemplo, SUBJ[NP]_V_NP_PP[em].

O objetivo do trabalho é identificar automaticamente padrões sintáticos para verbos com vistas a agrupá-los em classes que compartilhem os mesmos SCFs.

Os *frames* identificados são usados para dois propósitos. Primeiro, para a escolha dos verbos que farão parte das classes do léxico como parte de um método semi-automático. Segundo, para agrupamento automático de verbos, utilizando aprendizado de máquina. O resultado das duas abordagens serão comparados com um *gold standard* das classes da VerbNet para o português do Brasil, criado com o apoio da doutora Karin Kipper (desenvolvedora da VerbNet).

Inicialmente, utilizou-se o *corpus* Lacio-Ref [14] como entrada para o sistema. Esse *corpus* foi automaticamente anotado pelo *parser* PALAVRAS e possui aproximadamente 9 milhões de palavras. Ele é dividido em 4 gêneros: científico, informativo, judiciário e literário. Usando o extrator de SCFs, identificaram-se 5.131 verbos (considerando apenas verbos com frequência superior a 1) e 4.792 *frames* parametrizados por preposições(considerando apenas *frames* com frequência superior a 1).

Como a quantidade de verbos identificados e a frequência dos mesmos não era suficiente para a tarefa de aprendizado de máquina, foi necessário processar mais textos. Para isso, escolheu-se o *corpus* PLN-BR-FULL [15] do gênero jornalístico, com aproximadamente 26 milhões de palavras, e o *corpus* da Revista FAPESP (`http://revistapesquisa.fapesp.br`) [16] do gênero de divulgação científica, com aproximadamente 6 milhões de palavras. Somando os três *corpora* (Lácio-Ref, PLN-BR-FULL e Revista FAPESP) foram identificados 7.252 verbos (com frequência superior a 1) e 17.448 frames parametrizados por preposições (considerando apenas *frames* com frequência superior a 1).

A próxima etapa do projeto é usar aprendizado de máquina para realizar o agrupamento automático de verbos.

4.2 Anotação de Papéis Semânticos

Outro trabalho que está sendo desenvolvido com o uso da ferramenta de extração de SCFs envolve a anotação manual de papéis semânticos em *corpora* escritos em português brasileiro. O objetivo é gerar uma lista de verbos com seus respectivos SCFs, identificados como a estrutura argumental desses verbos, e papéis semânticos, além de dois *corpora* com orações semanticamente anotadas.

De modo sucinto, os papéis semânticos identificam um significado mais abstrato e esquemático dos argumentos de um elemento linguístico (em nosso caso, um verbo). Um exemplo bastante simples, somente a título de ilustração, seria indicar que na oração:

```
João viu Maria.
```

Existe um verbo que é "ver" (em sua forma canônica) e dois argumentos. A anotação de papéis semânticos indica que o argumento "João" é experienciador e o argumento "Maria" é experienciado.

Esse estudo utiliza dois *corpora* devidamente anotados pelo PALAVRAS, sendo um de linguagem especializada composto por artigos científicos da área de Cardiologia (omitido devido a revisão cega) e outro de linguagem não especializada contendo textos jornalísticos do Diário Gaúcho (compilado pelo projeto PorPopular `http://www6.ufrgs.br/textecc/porlexbras/porpopular/`).

A ferramenta de extração de SCFs que descrevemos neste trabalho auxilia na anotação manual dos papéis semânticos na parte de extração e organização dos dados presentes nos dois *corpora* e também na interface de anotação, que é gerada automaticamente a partir do banco de dados. Essa interface pode ser vista na Figura 2.

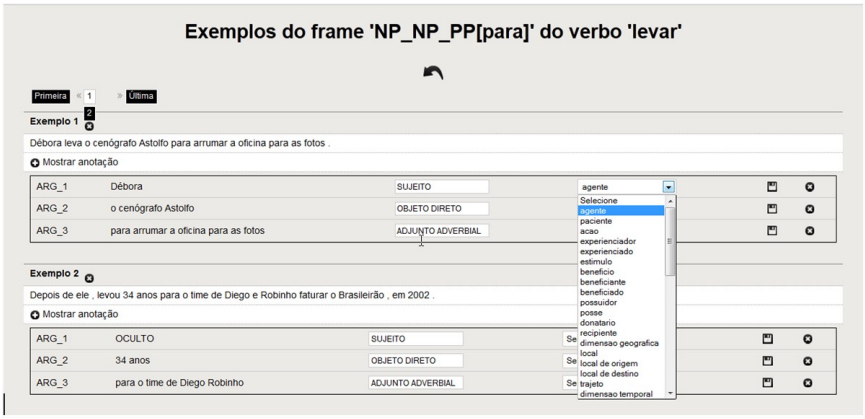


Figura 2: Interface para anotação dos papéis semânticos.

Como podemos ver na Figura 2, cada sentença que contém o verbo "levar" é organizada em exemplos e os argumentos desse verbo em cada um dos exemplos já se apresentam classificados na ordem descrita na Seção 3. Para cada argumento, é apresentada a anotação sintática (que pode ser manualmente alterada, caso o PALAVRAS tenha anotado errado) e uma caixa de rolagem que permite a seleção de um papel semântico a partir de uma lista. Essa lista de papéis semânticos se encontra em um arquivo separado, que pode ser modificado conforme a necessidade do linguista que faz a anotação manual.

Atualmente, foram anotadas sentenças de quatro verbos em cada um dos dois *corpora* como parte de um estudo-piloto que visou a testar se a lista de papéis semânticos selecionada seria aplicável em grande escala. Tal lista contém 46 papéis semânticos e foi proposta por [17] e [18].

Até então, o trabalho resultou, para os quatro verbos, na anotação manual de 482 sentenças, com um total de 138 configurações de papéis semânticos. Tal estudo permitiu observar que alguns papéis semânticos da lista podem ser unificados. Um exemplo disso são os papéis de posse e benefício, que se apresentam

nas mesmas estruturas argumentais e com os mesmos verbos, diferindo apenas na semântica da palavra utilizada.

5 Considerações Finais

A ferramenta aqui apresentada, um extrator e avaliador de *subcategorization frames*, se mostrou bastante útil para a descrição do português brasileiro. Isso é visível a partir da aplicação que demonstramos em dois estudos de cunho diferente, que visam a aumentar a capacidade que temos de processar a linguagem natural. Além disso, a ferramenta pode ser aplicada a outros *corpora* para desenvolver outros tipos de trabalho, sendo bastante flexível, uma possibilidade seria a anotação de *frames* no estilo da FrameNet [19].

Um limitador claro da ferramenta apresentada é o uso do *parser* PALAVRAS como base, pois ele é, infelizmente, um recurso pago ao qual o acesso é bastante restrito. A opção por esse *parser* se deu por ele ser atualmente o mais preciso, porém, seria interessante ter como alternativa a possibilidade de se usar um *parser* gratuito na anotação das dependências. Esse é um dos passos que deve ser levado adiante nos trabalhos futuros relacionados à ferramenta.

Referências

[1] KORHONEN, A.; KRYMOLOWSKI, Y.; BRISCOE T.: **A Large Subcategorization Lexicon for Natural Language Processing Applications**, 2006. In the Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), Genova, Italy.

[2] LI, J.; BREW, C.: **Which are the Best Features for Automatic Verb Classification**, 2008. In Proceedings of the 46th Annual Meeting of the Assoication for Computational Linguistics. Columbus, Ohio.

[3] SUN, L.; Korhonen, A.: **Improving verb clustering with automatically acquired selectional preferences**, 2009. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009). Singapura, pp. 638-647.

[4] SUN, L.; KORHONEN, A.; POIBEAU, T.; MESSIANT, C.: **Investigating the cross-linguistic potential of VerbNet: style classification**, 2010. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, 1056-1064.

[5] SURDEANU, M.; HARABAGIU, S.; WILLIAMS, J.; AARSETH, P.: **Using predicate-argument structures for information extraction**, 2003. In The Proceedings of the 41st Annual Meeting of ACL, Sapporo, Japan, 8-15.

[6] MESSIANT, C.: **A subcategorization acquisition system for French verbs**, 2006. In the proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Techonologies, Columbus, Ohio, 55-60.

[7] BRISCOE, T.; CARROL, J.: **Automatic extracton of subcategorization from corpora**, 1997. Proceedings of the fifth conference on Applied natural language processing, Washington, DC, 356-363.

[8] SCHULTE IM WALDE, S.: **A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG**, 2002. In: Proceedings of the 3rd Conference on Language Resources and Evaluation, v. IV, Las Palmas de Gran Canaria, Espanha, p. 1351-1357. Disponível em: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.8846&rep=rep1&type=pdf`

[9] IENCO, D., VILLATA, S., BOSCO, C.: **Automatic extraction of subcategorization frames for Italian**, 2008. In the Proceedings of the sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marocco, 28-30.

[10] KIPPER-SCHULER, K.: **VerbNet: a broad-coverage, comprehensive verb lexicon**, 2005. University of Pennsylvania. Tese de doutorado orientada por Martha S. Palmer.

[11] KIPPER, K.; KORHONEN, A.; RYANT, N.; PALMER, M.: **Extending VerbNet with Novel Verb Classes**, 2006. In: Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy, June.

[12] AUGUSTINI, A.: **Aquisição Automática de Subcategorização Sintáctico-Semântica e sua Utilização em Sistemas de Processamento da Língua Natural**, 2006. Tese de Doutorado. Orientador: José Gabriel Pereira Lopes. Disponível em: `http://dev-hlt.di.fct.unl.pt/gpl/projects/PATRAS/tese_agustini_10mai.pdf`

[13] BICK, E.: **The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**, 2000. Aarhus: Aarhus University Press.

[14] ALUÍSIO, S.; PINHERO, G. M.; MANFRIM, A. M. P.; OLIVEIRA, L. H. M. de; GENOVES JR., L. C.; TAGNIN, S. E. O.: **The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools**, 2004. In The Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisboa, Portugal, 1779-1782.

[15] BRUCKSCHEN, M., MUNIZ, F., SOUZA, J. G. C., FUCHS, J. T., INFANTE, K., MUNIZ, M., GONÇALVES, P. N., VIEIRA, R. e ALUÍSIO, S. M. **Anotação Lingüística em XML do Corpus PLN-BR**, 2008. Série de Relatórios do NILC. NILC-TR-09-08, 39 p.

[16] AZIZ, W. and SPECIA, L.: **Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation**, 2011. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, Cuiabá, Brasil.

[17] BRUMM, T.: **Erstellung eines Systems thematischer Rollen mit Hilfe einer multiplen Fallstudie**, 2008. Studienarbeit, 103p. Betreuer: Tom Gelhausen. Diponível em: `http://www.ipd.uka.de/Tichy/theses.php?id=135`

[18] GELHAUSEN, T.: **Modellextraktion aus natürlichen Sprachen: Eine Methode zur systematischen Erstellung von Domänenmodellen**, 2010. Karlsruhe: KIT Scientific Publishing. Dissertation, Karlsruher Institut für Technologie. Disponível em: `http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/1437903`

[19] BAKER, C. F., FILLMORE, C. J. e LOWE, J. F.: **The Berkeley FrameNet Project**, 1998. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, University of Montréal, Canadá, pp. 86-90.