Um repositório de verbos para a anotação de papéis semânticos disponível na web

Magali Sanches Duran, Jhonata Pereira Martins, Sandra Maria Aluísio
Núcleo Interinstitucional de Linguística Computacional
ICMC – University of São Paulo - São Carlos – SP – Brazil
magali.duran@uol.com.br, jhonata.p.martins@gmail.com; sandra@icmc.usp.br

Abstract. *This paper describes the semi-automatic construction of a large repository of verbs in Portuguese, using as source of examples a corpus annotated with semantic role labels, created by the PropBank-Br project. To enable the task, additional annotation has been included in the PropBank-Br corpus: the identification, in Propbank's lexical resource for English, of a sense equivalent to the annotated sense in Portuguese; an identification of the verb sense in Portuguese; a gloss of the verb sense; predicate lemma and sense and example notes. The resulting lexical resource will be used in an annotation task to evaluate whether its sense granularity is adequate to reach good inter-annotator agreement measures.*

Resumo. *Este artigo descreve a construção semi-automatica de um grande repositório de verbos em português, usando como fonte de exemplos um corpus anotado com rótulos de papéis semânticos, criado pelo projeto PropBank-Br. Para viabilizar a tarefa, novos campos de anotação foram incluídos no corpus PropBank-Br: a identificação, no repositório do Propbank inglês, de um sentido equivalente ao sentido anotado em português; uma identificação do sentido do verbo em português; uma glosa para cada sentido do verbo; o lema do predicado e notas do sentido e do exemplo. O recurso lexical resultante será usado em uma tarefa de anotação para julgar se sua granularidade de sentidos é adequada para se atingir um bom índice de concordância entre anotadores.*

1. Introdução

A tarefa de anotação de papéis semânticos consiste em identificar/delimitar o predicador (normalmente um verbo) e seus argumentos e atribuir a cada argumento um rótulo de papel semântico (Palmer et al. 2010). Por exemplo, na oração "O homem reclamou ao patrão sobre as péssimas condições de trabalho", o predicador é "reclamou" e seus argumentos são: "o homem" (Agente), "ao patrão" (Receptor) e "sobre as péssimas condições de trabalho" (Tema ou Tópico).

Como existem várias propostas de conjuntos de rótulos de papéis semânticos e como nem sempre é simples decidir qual o rótulo mais adequado para anotar um argumento, grandes projetos de anotação de papéis semânticos (SRL: Semantic Role Labeling) desenvolveram repositórios lexicais onde estão definidos os papéis semânticos previstos pelos sentidos dos predicadores. Alguns desses repositórios definem papéis semânticos para as classes verbais, como é o caso da Verbnet (Kipper et al. 2006), outros para os frames semânticos, como é o caso da Framenet (Baker et al. 1998) e outros ainda para os sentidos de cada verbo, como o Propbank (Palmer et al. 2005). Quanto mais completo for o repositório lexical e quanto mais clara for a

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 168–172,
Fortaleza, CE, Brazil, October 21–23, 2013.

distinção que ele faz dos sentidos do predicador, mais simples a tarefa se tornará para os anotadores, aumentando a probabilidade de se atingir bons índices de concordância entre anotadores e de se obter boa precisão no aprendizado de máquina da tarefa (Duffield et al. 2007; Palmer et al. 2007; Hovy et al. 2006). No repositório do Propbank, por exemplo, o anotador pode consultar o arquivo do predicador (*frame file*), ver qual o sentido que coincide com o sentido que está sendo anotado e ver quais os papéis semânticos previstos para esse sentido (*roleset*).

A alternativa para a anotação de SRL, quando não se tem um repositório lexical para guiar a tarefa, é utilizar apenas um guia contendo regras e alguns exemplos de anotação. Essa alternativa foi adotada na anotação do PropBank-Br (Duran e Aluísio, 2010), anotado por um único linguista. Os classificadores que usaram esse corpus para treinamento (Alva-Manchego e Rosa, 2012; Fonseca e Rosa, 2013), no entanto, não atingiram o estado da arte, em parte porque o corpus é pequeno (possui 1068 verbos e 6142 instâncias anotadas, o que é menos de 10% do tamanho do Propbank do inglês), e em parte porque ele não contém distinção de sentidos dos verbos.

A fim de superar esses pontos fracos, é preciso empreender um projeto de anotação de SRL em larga escala. Uma das condições para isso é a existência de um repositório lexical para guiar um processo de anotação que envolva diversos anotadores. A construção semi-automática desse repositório, ainda em curso, é o tema deste artigo.

2. Metodologia

O modelo de banco de dados adotado foi o do repositório do projeto Propbank, composto por um arquivo xml para cada verbo, que possui um editor dedicado - Cornerstone (Choi et al. 2010). As telas do Cornerstone possuem campos para inserir verbos, sentidos de verbos, definição de cada sentido dos verbos, os papéis semânticos previstos pelos sentidos dos verbos e exemplos anotados de cada um desses sentidos.

Tínhamos algumas alternativas para construir o repositório:

1) Traduzir o recurso inglês. Descartamos essa alternativa pois ela excluiria todos os verbos e sentidos de verbos que não tivessem equivalentes em inglês. Além disso, a tradução dos exemplos seria trabalhosa;

2) Construir a partir do zero, usando informações de dicionários. Descartamos essa alternativa porque os dicionários trazem muitos sentidos que não se verificam em corpus e fazem uma divisão de sentidos muito detalhada, o que não era nosso interesse;

3) Construir a partir do zero utilizando evidências de corpus. Essa alternativa foi adotada por ser relativamente mais rápida e mais automatizável, já que poderíamos aproveitar as sentenças do Propbank-Br para suprir a necessidade de exemplos anotados com papéis semânticos.

A criação manual dos arquivos do repositório no Cornerstone envolvia, de um lado, tarefas muito rudimentares, como copiar e colar os exemplos anotados nos campos reservados e, de outro lado, tarefas muito complexas, como definir e criar identificações para cada sentido dos verbos. Visando acelerar o processo, decidimos separar as tarefas automatizáveis das que exigiam trabalho linguístico. Percebemos que, se o corpus contivesse todas as informações necessárias, os arquivos do repositório poderiam ser construídos automaticamente. Por isso, complementamos a anotação do corpus com as informações exigidas pelos arquivos do repositório. Para isso, criamos seis campos de

anotação ou *wordtags*, um recurso da ferramenta de anotação de corpus adotada (SALTO – Burchardt et al. 2006): (1) *PB-roleset*: sentido no repositório do Propbank inglês, equivalente ao sentido anotado em português, que permitirá herdarmos as definições dos papéis semânticos e seus mapeamentos para os papéis semânticos e classes da Verbnet; (2) Nota: campo utilizado sempre que for necessário fazer alguma observação sobre o sentido do verbo para os anotadores; (3) Nota do exemplo: esse campo é utilizado para chamar a atenção dos anotadores para algum aspecto do exemplo; (4) *Predicate lemma*: campo obrigatório na primeira ocorrência de um sentido; é onde se coloca o nome do predicado, incluindo predicados complexos (idiomáticos ou não), como por exemplo "abrir_mão"; (5) Sentido: identificação do sentido do verbo; (6) t-glosa: campo obrigatório na primeira ocorrência de um sentido; é onde se coloca uma definição clara do sentido do verbo que permita ao anotador distinguir um sentido de outro. A Figura 1 apresenta uma instância do corpus com quatro dessas *wordtags* preenchidas.

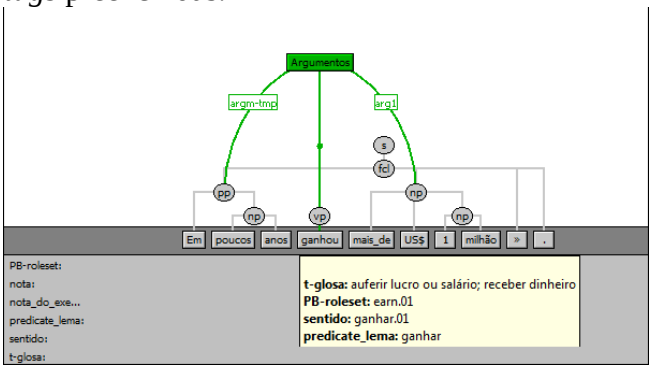


Figura 1 - Instância do Propbank-Br e novos campos preenchidos

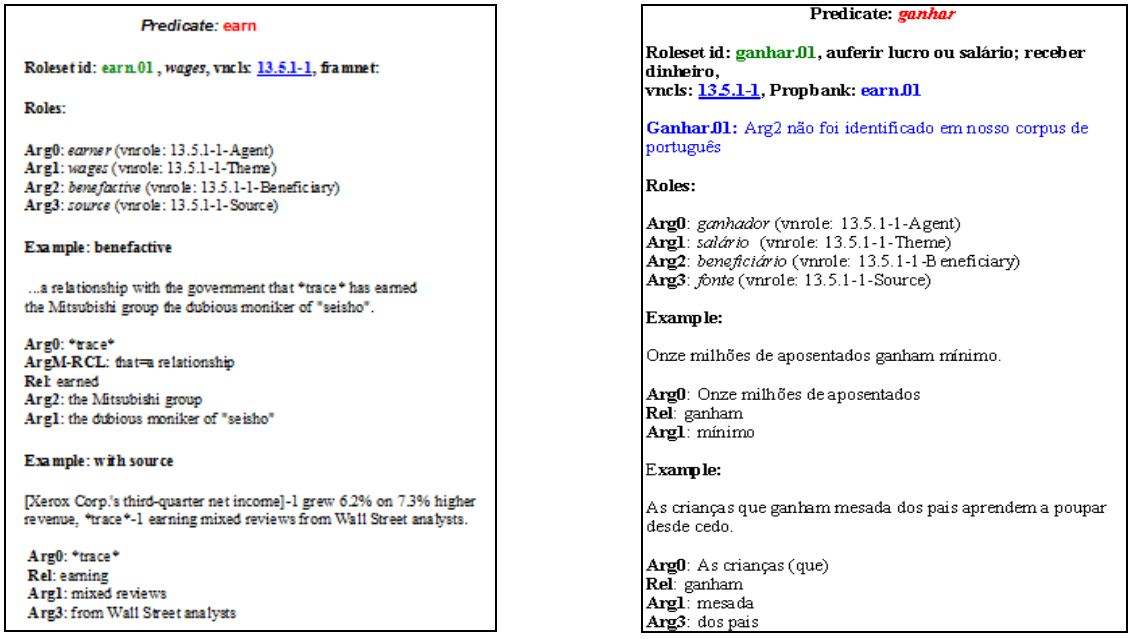
Das seis *wordtags*, as duas mais desafiadoras são "PB-roleset" e "Sentido". "PB-roleset" só deve ser deixada em branco quando não for identificado nenhum sentido no repositório inglês que corresponda ao sentido em português, como é o caso, por exemplo, do sentido caber.01 (*caber a* alguém fazer alguma coisa), que em inglês é expresso por uma expressão "*to be up to* someone to do something". Decidir qual o sentido equivalente mais adequado em inglês implica consultar o repositório do inglês e identificar um sentido que apresente estrutura argumental mais similar à estrutura argumental do sentido em português. Se este campo estiver em branco, o arquivo terá que ser complementado manualmente, no editor Cornerstone, com todas as informações que não puderem ser herdadas automaticamente.

A *wordtag* "Sentido", por sua vez, é a única que deve ser preenchida obrigatoriamente em todas as instâncias anotadas do corpus. É constituída pelo nome do verbo seguido de um número sequencial de dois dígitos (ganhar.01, por exemplo). Todas as instâncias com o mesmo sentido devem ter o mesmo número de identificação, porém os sentidos de um verbo nem sempre são facilmente delimitáveis e muitas vezes o contexto da sentença não é suficiente para inferi-los, o que torna a tarefa complexa.

A decisão quanto à granularidade de sentidos (mais genéricos ou mais específicos) impacta tanto a anotação (se os anotadores não forem capazes de distingui-los, a concordância entre anotadores diminui) quanto o futuro aprendizado de máquina que usa o corpus anotado (se não houver nenhuma "pista" explícita da diferença de sentido, a precisão do classificador automático ficará prejudicada). Por exemplo, o verbo "esperar" pode ter o sentido de "aguardar" e o de "ter esperança ou expectativa".

O primeiro prevê um NP como argumento (Arg1) após o verbo e o segundo prevê uma oração subordinada (introduzida por "que" ou reduzida de infinitivo).

Nas 4554 instâncias anotadas até o momento, foram contemplados 794 verbos e identificados 1092 sentidos, o que dá uma média de 1,37 sentido por verbo. Desses verbos, 81,2% apresentam um único sentido, 12,1% dois sentidos, 2,9 três sentidos, 1,8% quatro sentidos e 2% cinco ou mais sentidos. Tal distribuição assemelha-se à relatada para os verbos do inglês no Propbank. O próximo passo é fazer um programa que mapeie alguns campos dos arquivos do repositório inglês para os respectivos campos dos arquivos do repositório português, usando como chave o identificador de sentido informado no campo "PB-roleset" do corpus. Os campos a serem aproveitados do repositório inglês são: vnclass (classe da Verbnet), vnrole (papel semântico correspondente na Verbnet), definições dos papéis semânticos para futura tradução (*earner*, *wages*, *benefective* e *source* na Figura 2a).



Predicate: earn

Roleset id: earn.01 , *wages*, **vncls:** 13.5.1-1, **framnet:**

Roles:

Arg0: *earner* (vnrole: 13.5.1-1-Agent)
Arg1: *wages* (vnrole: 13.5.1-1-Theme)
Arg2: *benefactive* (vnrole: 13.5.1-1-Beneficiary)
Arg3: *source* (vnrole: 13.5.1-1-Source)

Example: benefactive

...a relationship with the government that *trace* has earned the Mitsubishi group the dubious moniker of "seisho".

Arg0: *trace*
ArgM-RCL: that=a relationship
Rel: earned
Arg2: the Mitsubishi group
Arg1: the dubious moniker of "seisho"

Example: with source

[Xerox Corp.'s third-quarter net income]-1 grew 6.2% on 7.3% higher revenue, *trace*-1 earning mixed reviews from Wall Street analysts.

Arg0: *trace*
Rel: earning
Arg1: mixed reviews
Arg3: from Wall Street analysts

Predicate: *ganhar*

Roleset id: ganhar.01, **auferir lucro ou salário; receber dinheiro,**
vncls: 13.5.1-1, **Propbank:** earn.01

Ganhar.01: Arg2 não foi identificado em nosso corpus de português

Roles:

Arg0: *ganhador* (vnrole: 13.5.1-1-Agent)
Arg1: *salário* (vnrole: 13.5.1-1-Theme)
Arg2: *beneficiário* (vnrole: 13.5.1-1-Beneficiary)
Arg3: *fonte* (vnrole: 13.5.1-1-Source)

Example:

Onze milhões de aposentados ganham mínimo.

Arg0: Onze milhões de aposentados
Rel: ganham
Arg1: mínimo

Example:

As crianças que ganham mesada dos pais aprendem a poupar desde cedo.

Arg0: As crianças (que)
Rel: ganham
Arg1: mesada
Arg3: dos pais

Figura 2 – (a) Roleset do repositório do Propbank. (b) Repositório do Português

3. Trabalhos futuros e conclusão

Quando o repositório estiver pronto (Figura 2b), deverá ser validado em uma tarefa de anotação, a fim de julgarmos se a divisão de sentidos é apropriada para a tarefa de anotação de papéis semânticos. Os sentidos que não obtiverem um bom índice de concordância entre anotadores deverão ser reavaliados linguisticamente e provavelmente mesclados a outros sentidos. Além disso, futuramente as informações de classes da Verbnet poderão ser utilizadas a fim de promover a inclusão dos verbos da Verbnet.Br (Scarton e Aluísio, 2012) que ainda não estejam no repositório, da mesma forma que foi feito para estender o Propbank inglês com os verbos da VerbNet. Nossa estratégia de complementar a anotação do corpus Propbank-Br com sentidos de verbos e outras informações sobre o predicador anotado gerará um corpus mais rico para o aprendizado automático. Esse corpus será disponibilizado na Web, no endereço http://www.nilc.icmc.usp.br/portlex, assim como os arquivos do repositório.

Agradecimentos

Agradecemos à FAPESP e ao CNPq pelo apoio aos pesquisadores.

Referências bibliográficas

Alva Manchego, F. E.; Rosa, J. L. G. (2012). Semantic Role Labeling for Brazilian Portuguese: A Benchmark. In IBERAMIA 2012, *Lecture Notes in Artificial Intelligence*, v. 7637 p. 481–490. Springer.

Baker, C.F.; Fillmore, C. J.; Lowe. J. B. (1998).The Berkeley FrameNet Project. In: *Proceedings of Computational Linguistics 1998 Conference*.

Burchardt, A.; Erk, K.; Frank, A.; Kowalski, A.; Pado, S. (2006) SALTO - A Versatile Multi-Level Annotation Tool. In: *Proceedings of LREC 2006*.

Choi, J. D.; Bonial, C.; Palmer, M. (2010) Propbank Frameset Annotation Guidelines Using a Dedicated Editor, Cornerstone. In: *Proceedings of LREC-2010*.

Duffield, C. J.; Hwang, J. D.; Brown, S. W.; Dligach, D.; Vieweg, S. E.; Davis, J.; Palmer, M. (2007). Criteria for the Manual Grouping of Verb Senses. In: *Proceedings of the Linguistic Annotation Workshop*, p. 49–52, Prague, June 2007. Association for Computational Linguistics.

Fonseca, E. R.; Rosa, J. L. G. (2013) A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. In: *Proceedings of IJCNN 2013 International Joint Conference on Neural Networks* (no prelo).

Gildea, D.; Jurafsky, D. (2001) Identifying Semantic Roles in Text. In: *Seventeenth International Joint Conference on Artificial Intelligence* (IJCAI-01), Seattle, Washington.

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; Weischedel, R. (2006). OntoNotes: The 90% Solution. In: *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, p. 57-60.

Kipper,K.; Korhonen, Anna; Ryant, N.; Palmer, M. (2006). Extensive Classifications of English verbs. *Proceedings of the 12th EURALEX International Congress*. Turin, Italy.

Palmer, M.; Dang, H.; Fellbaum, C. (2007). Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*,13:2, 137-163.

Palmer, M.; Gildea, D.; Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31:1., pp. 71-105, March, 2005.

Palmer, M.; Gildea, D.; Xue, N. (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technology Series, ed. Graeme Hirst, Mogan & Claypoole.

Scarton, C. And Aluisio, S. (2012).Towards a cross-linguistic VerbNet-style lexicon to Brazilian Portuguese. In: *LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, 2012, Istambul, Turkey.