Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels

Magali Sanches Duran, Sandra Maria Aluísio

Núcleo Interinstitucional de Linguística Computacional
ICMC – Universidade de São Paulo - São Carlos – SP – Brasil
magali.duran@uol.com.br, sandra@icmc.usp.br

Abstract. *Semantic Role Labeling is a task in Natural Language Processing often carried out through annotated corpus. So far, there is no available corpus of Portuguese annotated with semantic role labels. This paper reports the annotation of a Brazilian Portuguese corpus following Propbank guidelines. This is the first step of a larger annotation effort and aims to pave the way for a distributed annotation task. Annotation decisions are discussed to stress language specific aspects involved in the Project.*

Resumo. *A Anotação de Papéis Semânticos é uma tarefa de Processamento de Línguas Naturais frequentemente realizada por meio de corpus anotado. Até o momento não há um corpus de português disponível que esteja anotado com rótulos de papéis semânticos. Este artigo relata a anotação de um corpus de português do Brasil seguindo as instruções do Propbank. Este é o primeiro passo de um esforço mais amplo de anotação e tem por objetivo abrir caminho para uma tarefa de anotação distribuída. As decisões de anotação são discutidas a fim de salientar os aspectos específicos de língua envolvidos no projeto.*

1. Introduction

Semantic role labeling (SRL), as an NLP task based on annotated corpus, was first addressed by Gildea e Jurafsky (2002), employing Framenet corpus (Baker et al. 1998). Since then several projects dealt with SRL (Gildea & Palmer, 2002, Surdeanu et. al, 2003, Gildea & Hockenmaier, 2003, Yi, Loper and Palmer, 2007, Palmer et al. 2010, among others).

There are at least two ways for improving SRL classifiers based on annotated corpus: one of them is trying out different machine learning methods; the other way is providing a large and properly annotated training corpus. Framenet was not originally conceived to provide a training corpus for machine learning. Its set of semantic role labels, for example, is fine grained and poses a problem of data sparsity for statistical learning methods. Propbank initiative, in contrast, focus specifically on this purpose and presents project decisions that contribute for machine learning, like a coarse grained set of role labels and annotation over the syntactic tree.

The annotation of a corpus with semantic role labels consists of three subtasks: 1) identification of the "argument taker", which may be a single verb or a complex predicate

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 164–168,
Cuiabá, MT, Brazil, October 24–26, 2011. ©2011 Sociedade Brasileira de Computação

(light verb constructions or phrasal verbs, for example); 2) identification and delimitation of arguments associated with the "argument taker", and 3) assignment of a semantic role label to each of these arguments. Annotation over a syntactic tree eliminates the step of arguments delimitation, as the syntactic constituents delimitated by the parser are kept for arguments annotation. Hence, the quality of SRL annotation is dependant of syntactic parsing quality.

Recently there are initiatives to make corpus annotation, following Propbank model, for other languages besides English: Korean (Palmer et al, 2006), Chinese (Xue, 2009), Arabic (Palmer et al, 2008) and Basque (Aldezabal et al. 2010). However, as far as we know, there is not until this date such a corpus of Brazilian Portuguese. To fulfill this gap, we report here the construction of a Brazilian Portuguese Propbank: Propbank-Br. This first step of the research aims to pave the way for a broader and distributed annotation task. Language specific challenges became evident during the annotation task and several decisions have been taken to deal with them. This experience enabled us to customize Propbank guidelines and build frames files for Portuguese verbs, essential resources to guide annotators and ensure inter-annotator agreement.

2. A brief outline of Propbank

Propbank (Palmer et al 2005) produced a new layer of annotation, adding semantic role labels in a subcorpus of PennTreebank (the financial subcorpus). Additionally, a verb lexicon with verb senses and rolesets have been built and is available for consultation[1].

Propbank is a bank of propositions. The underlying idea of the term "proposition" is found in frame semantics proposed by Fillmore (1968). A "proposition" is on the basic structure of a sentence (Fillmore, 1968, p.44), and is a set of relationships between nouns and verbs, without tense, negation, aspect and modality modifiers. Arguments which belong to propositions are annotated by Propbank with numbered role labels (Arg0 to Arg5) and modifiers are annotated with specific ArgMs (Argument Modifiers) role labels. Each verb occurrence in the corpus receives also a sense number, which corresponds to a roleset in the frame file of such verb. A frame file may present several rolesets, depending on how many senses the verb may assume. In the roleset, the numbered arguments are "translated" into verb specific role descriptions. Arg0 of the verb "give", for example, is described as "giver".

3. Methods and Tools

In the same way as Propbank, our aim is to provide a training corpus to build automatic taggers. For this purpose, it was interesting to annotate a corpus syntactically annotated and manually revised. We decided to annotate the Brazilian portion of Bosque, the manually revised subcorpus of Floresta Sintá(c)tica[2] (Affonso et al, 2002), parsed by Palavras (Bick, 2000). Bosque has 9368 sentences and 4213 of them correspond to the Brazilian portion (extracted from the journal Folha de São Paulo of 1994).

analisador
sintático?

The annotation tool we have chosen was SALTO (Burchardt et al, 2008) due to a previous successful experience we had on assigning wh-questions to verbal arguments, a

[1] http://verbs.colorado.edu/propbank/framesets-english/
[2] http://linguateca.pt

related task. After annotation started, we have notice of Jubilee (Choi et al. 2010), a dedicated annotation tool developed by the Propbank team. SALTO has been developed for annotation of German Framenet, but its resources were adequate for our annotation purposes not requiring tool customization (we customized only the use). A facility of SALTO that we have extensively used is the sentence flag. For example, we have flagged as Wrongsubcorpus all sentences that present some error, and for this we created three parameters: EP corresponds to parsing errors or inadequacies, EC corresponds to corpus errors, like spelling or punctuation errors, and EV corresponds to invocation errors, like past participles used as adjectives. The other SALTO sentence flags (Reexamine, Interesting and Later) have been similarly used to annotate sentences into types that offer further study possibilities.

Aiming to shorten manual effort, we decided to process automatically the step of identification of argument takers. The problem was to distinguish modifier verbs (auxiliaries) from main verbs. To meet this need, we made a study on auxiliary verbs and built a table which encompasses temporal, aspectual, modal and passive voice auxiliaries, followed by the infinite form imposed to the auxiliated verb (infinitive, past participle or gerund). This table has been improved by results from Baptista and Mamede (2010) and enabled us to identify verbal chains and select, as argument taker, only the last verb at right of the chain (which corresponds to the main verb). Following Propbank guidelines, we repeated the sentences as many times as the number of argument takers they had. In this way, each argument taker of the sentence constitutes a separate instance for annotation. The previous 4213 sentences produced 7107 instances for annotation.

c/ Jubilee?

Due to time and resources restrictions of the project, there is a significant difference between Propbank´s methodology and ours. Propbank involved several annotators, and each instance was double annotated to control annotator's agreement. Propbank-Br, on the other hand, has been annotated only by the main researcher of the project in order to produce customized guidelines and Portuguese frames files to guide a future distributed annotation task. Besides that, Propbank annotated simultaneously role labels and verb senses, because verb frames files had been built previously to guide annotators. We, on the other hand, decided to assign role labels using English rolesets and to annotate senses in the future. When there is no equivalent of a Portuguese verb in English, we use Propbank framing guidelines[3] to determine the roleset.

4. Discussion

Propbank-Br faced several challenges. In spite of following Propbank guidelines as often as possible, there are differences and the major of them are due to the parser output. The Penn Treebank, used by Propbank, has "traces" of suppressed syntactic elements. This is very important to deal with ellipsis and co-references. We have adopted some strategies to circumvent the lack of such traces in our corpus.

Sentences without expressed subject have been flagged with parameter OCULTO, if the subject is inferable from verb inflection or INDETERMINADO, if the verb is in third person of plural (mark of subject indeterminacy in Portuguese). Embedded clauses that have one argument represented by a pronoun have been flagged with the parameter

[3] http://verbs.colorado.edu/~mpalmer/projects/ace/FramingGuidelines.pdf

CORREF since the co-referred element is in the main clause. In these cases, the role label has been assigned to the pronoun, but we kept track for future work on anaphora resolution. Embedded clauses or coordinate clauses which present a suppressed element which is in the main clause have been flagged with the parameter ELIPSE. In these cases, the role label has been assigned to the corresponding element of the main clause. We kept track for future work on anaphora resolution with zero-related elements.

Besides that, parser Palavras is a dependency parser that provides a constituent parsing output. However, this output is not as good as that provided by a true constituent parser and many internal NPs on the left of the verb are not annotated, affecting role label assignment. The parse tree in Figure 1 does not have a constituent corresponding to the subject of the complex predicate “dizer respeito” (relate to) and consequently we can not assign the respective role label. In these cases, we did not edit parse trees, we simply flagged the instance as “Wrongsubcorpus”.

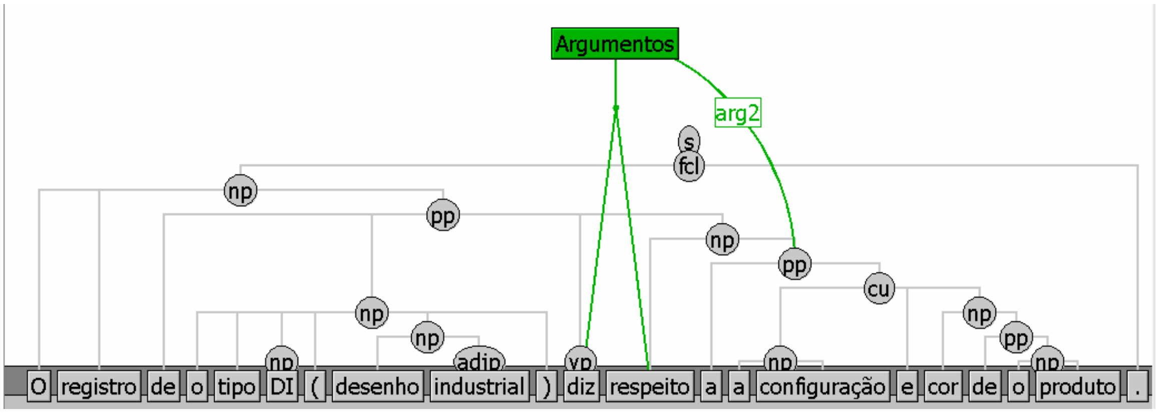

Figure 1. Parse tree flagged as “Wrongsubcorpus”

However, we take the decision to maintain in our corpus the instances where two or three arguments are concatenated in a unique syntactic constituent. Following Propbank guidelines, we assign them the role the more important (for example, numbered Args prevail on ArgMs and Arg 1 prevails on Arg2).

5. Future work

não dá para acessar

We are almost finishing the annotation of the 7107 instances and the corpus will soon be available at PortLex (http://www2.nilc.icmc.usp.br/portlex/). We followed Propbank guidelines and registered our decisions related to Portugues in order to elaborate Propbank-Br Guidelines. Such a guide will enable us to extend semantic role label annotation to a larger corpus with several annotators. The complement of the corpus resource is the construction of frames files with verb senses and respective rolesets, which will make it possible to add verb senses annotation to Propbank-Br. We have already built 132 frames files using Cornerstone (Choi et al. 2010), a dedicated editor developed by Propbank team, but this effort will be reported in future work.

6. Acknowledgements

Our thanks to São Paulo Research Foundation (FAPESP) for supporting this work.

References

Afonso S. ; Bick, E. ; Haber, E. ; Santos, D. (2002) Floresta sintá(c)tica: a treebank for Portuguese. In: Proceedings of LREC-2002.

Aldezabal, I.; Aranzabe, M. J., Ilarraza, A. D.;Estarrona, A. (2010). Building the Basque PropBank. In: Proceedings of LREC-2010.

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

Baptista, J., Mamede, N.J., and Gomes, F. (2010) Auxiliary Verbs and Verbal Chains in European Portuguese. In: Proceedings of PROPOR 2010.

Bick, E. (2000). The Parsing System Palavras Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus, Denmark, Aarhus University Press.

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

Fillmore, C.. The Case for Case (1968). In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

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

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

Gildea, D. ;Hockenmaier, J. (2003). Identifying Semantic Roles Using Combinatory Categorial Grammar. In: Proceedings of 2003 Conference on Empirical Methods in NLP.

Gildea, D.; Palmer, M. (2002). The Necessity of Parsing for Predicate Argument Recognition. In: Proceedings of The 40thMeeting of ACL, 2002.

Palmer, M.; Ryu, S.; Choi, J.; Yoon, S.; Jeon, Y. (2006) Korean Propbank. LDC Catalog No.: LDC2006T03 ISBN: 1-58563-374-7

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.; Babko-Malaya, O.; Bies, A.; Diab, M.; Maamouri, M.; Mansouri, A.; Zaghouani, W. (2008). A Pilot Arabic Propbank. In: Proceedings of LREC-2008.

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

Surdeanu, M.; Harabagiu, S.; Williams, J.; Aarseth, P. (2003) Using Predicate-Argument structures for information extraction. In: Proceedings of ACL 2003.

Xue, N,; Palmer., M. (2009) Adding semantic roles to the Chinese Treebank. Natural Language Engineering. 15(1) pp. 143-172.

Yi, S.; Loper, E.; Palmer, M. (2007) Can Semantic Roles Generalize Across Genres? In: Proceedings of HLT/NAACL-2007.