Development of a Lexical Resource Annotated with Semantic Roles for Portuguese

Leonardo Zilio

Universidade Federal do Rio Grande do Sul (UFRGS)
lzilio@ig.com.br

1 Introduction

Many applications in Natural Language Processing (NLP) may benefit from the use of semantic information, for example, text summarization (Yoshikawa et al. 2012), resolution of anaphora (Kong and Zhou 2012) and machine translation (Jones et al. 2012). For Portuguese, there are some important studies on semantic role labeling (Salomão 2009; Scarton 2013; Duran and Aluisio 2011, 2012), which will be presented in the following sections, but there is still much to do in this field.

This paper aims at presenting ongoing work in the development of a lexical resource for Portuguese which contains semantic role information. Semantic roles linked to verb arguments can be considered as a way to comprehend the basic semantic of a sentence without having to rely directly on lexical semantics. As an example we can use the following simple sentence:

Thomas saw the dog.

In this sentence, the syntactic arguments are the subject *Thomas* and the direct object *the dog*. With only this information, it is still not possible to comprehend whether the subject is doing something to the object, is being affected by it, or is simply experiencing it without intervention. When we describe those arguments with the roles of EXPERIENCER and THEME, respectively, then we can have a better (although not yet full) understanding of the event that is being described.

As we will see further on in this paper, semantic roles can be descriptive or numbered, and the choices made by the designer of the resource regarding the extent and grain size of the roles are of extreme importance for the final product, as well as for the development process.

The resource we present in this paper is based on VerbNet (Kipper-Schuler 2005) and PropBank (Palmer, Gildea and Kingsbury 2005), and contains sentences annotated with descriptive semantic roles.

The structure of this paper goes as follows: Section 2 will present related work, debating some of the methods used for semantic role labeling, specially the semantic role lists and the resulting resource; Section 3 will show the method used in this work, along with the materials we used; Section 4 discusses our partial results; and Section

J. Baptista et al. (Eds.): PROPOR 2014, LNAI 8775, pp. 195–200, 2014.
© Springer International Publishing Switzerland 2014

5 is reserved for our final remarks, including information about the availability of the resource being developed.

2 Related Work

In this section we present two semantic resources for English along with their Brazilian counterparts. We will give an overview of their methodology and type of semantic roles, and of the resulting resource.

2.1 Propositional Bank

PropBank (Palmer, Gildea and Kingsbury 2005) is a resource which presents sentences annotated with numbered semantic roles. That is, instead of using a descriptive name like AGENT or THEME, the PropBank uses semantic roles like ARG0 or ARG1. The numbered roles vary from ARG0 to ARG5 and there are also roles reserved for modifiers, which vary in name according to the type of adjunct being annotated (ARGM-TMP for temporal adjunct; ARGM-LOC for local adjuncts, etc.). These semantic roles were annotated as a semantic level for the sentences present in the Penn Tree-Bank (Marcus, Marcinkiewicz and Santorini 1993), so that the resource would have a semantic role level on top of the already existent syntactic and morphological levels. The PropBank has a total of 3,185 annotated verbs and can be consulted on-line[1].

For Portuguese, we have the PropBank.Br (Duran and Aluisio 2011, 2012) which consists of a set of annotated sentences extracted from the Bosque corpus, a part of Floresta Sintá(c)tica corpus[2]. The resource has 1,025 distinct verbs annotated with numbered semantic roles.

2.2 VerbNet

Unlike PropBank, which used corpus sentences to apply a new level of semantic information, the VerbNet (Kipper-Schuler 2005) relied on the verb classes, as organized by Levin (1993). For each class the possible meaning associated with it is provided along with an example. VerbNet uses descriptive semantic roles (such as AGENT, THEME, PACIENT, etc.) and has 272 classes, which comprise 3,769 verbs. The semantic role list consists of 36 roles, but only 30 are actually being used in the latest update (version 3.2).

VerbNet has also a Portuguese version, VerbNet.Br (Scarton 2013), which was semi-automatically imported from the English VerbNet through the existing links between VerbNet and WordNet, and between WordNet and WordNet.Br. Since the resource was semi-automatically constructed, the results may contain some noise.

---

[1] `http://verbs.colorado.edu/verb-index/`.

[2] `http://www.linguateca.pt/Floresta/principal.html`.

3 Method

For the development of our resource, we used corpora and a subcategorization frame extractor that offers a user interface for the annotation of semantic roles. We also compiled a semantic role list from the VerbNet (Kipper-Schuler 2005) and extended it with other semantic roles, as we explain below.

3.1 Corpora

As a base for the extraction of sentences, we used two corpora: one with Brazilian Cardiology papers compiled by Zilio (2009) and the other with Brazilian newspaper articles compiled by the project PorPopular[3]. The different domains of the corpora are intended for a comparison between the behavior of semantic roles in specialized and non-specialized texts. Both corpora were analyzed with the parser PALAVRAS (Bick 2000) and present dependency trees for all sentences.

3.2 Subcategorization Frames Extractor

Subcategorization Frames can be seen as simpler syntactic representations of sentences (or phrases). They show what types of phrases are present in the sentence, without regard to whether it is a subject, complement or adjunct. As such, the representation of our early example "Thomas saw the dog" would be something like NP_NP or NP_V_NP (i.e. with or without the verb position).

This type of representation shows the basic arguments of a sentence. For this reason it can be used for the purposes of semantic role labeling. In our study we used a subcategorization frame extractor developed originally by Zanette (2010) and further modified by Zilio, Zanette and Scarton (2014).

The extractor uses corpora analyzed with the parser PALAVRAS (Bick 2000) as input, and processes all the sentences, organizing them according to verb, subcategorization frame and frequency in a MySQL database.

3.3 Semantic Roles

For the annotation we selected the roles used in VerbNet and adapted them for Portuguese. The basic elements from VerbNet were preserved, but some roles for adverbial adjuncts (from PropBank) were also included, so as to make the list more robust vis-à-vis the sentences from real written texts that are present in our corpora. The list itself contains 47 semantic roles covering arguments and adjuncts. This is much more fine-grained than PropBank's list and a bit more than the VerbNet's one.

Some of the roles we add to the list, other than those for adjuncts taken from PropBank, are roles for the pronoun "se", which can have many functions in a sentence, all

[3] http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php.

of which are well explained in the tutorial of PropBank.Br[4]. Another role was created specifically for agents that are, in fact, places or organizations, normally used as metonymical references to the people within it, such as "New York" in the sentence: "New York does the job." All the created semantic roles either were necessary because of the language or can be joined back together with the original role from VerbNet without much effort.

3.4 Annotation Process

The annotation is made by one linguist annotator and follows these criteria:

- The most frequent verbs from the newspaper corpus are annotated first, and then follows the annotation of the same verbs in the Cardiology corpus.
- Ten sentences per subcategorization frame are annotated.
- The verbs "ser" (to be), "estar" (to be), "haver" (to exist), and "ter" (to have) were not annotated, because they present too many subcategorization frames in both corpora and would require too much work for the sake of only a few verbs.

Following this method, samples of sentences are annotated for each verb, ensuring that we annotate many verbs and many types of subcategorization frames for each of them, maximizing the reward regarding the effort.

4 Results

Until now we have annotated 85 verbs in the newspaper corpus, and 42 of them were also annotated in the Cardiology corpus (the other 43 verbs were not available or did not present ten sentences in a subcategorization frame). This represents 4,640 annotated sentences (3,470 in the newspaper corpus and 1,170 in the Cardiology corpus). From the 47 semantic roles present in our list, 45 were used so far.

The most frequent semantic roles were AGENT and THEME, but, if we look at the individual corpora, only THEME is highly frequent in both of them. The sentences in the Cardiology corpus showed a tendency to suppress the agents, highlighting the instruments, results and the objects of the domain.

A list of the most frequent semantic roles in both corpora can be seen in Table 1. In that list, one can see that semantic roles like EXPERIENCER and TOPIC, which are highly ranked in the newspaper corpus, have very few correspondents in the Cardiology corpus. We have not yet made any statistical computation of these results, since we have not yet finished the annotation process, but the results so far seem to corroborate the ones found in a study by Zilio, Ramisch and Finatto (2013).

[4] http://www.nilc.icmc.usp.br/portlex/images/arquivos/propbank-br/propbank.br%20tutorial.pdf.

5 Final Remarks

The partial results of the annotation were already made available for online consultation on the Jibiki platform[5] and can also be downloaded in XML format from the CAMELEON project website[6]. There is still much work to be done, though, and we will annotate many more verbs before releasing the final version of our database.

For now, the results already show some important information, like the difference in the semantic roles annotated in both corpora. If this tendency remains after more verbs are annotated, than it leads to the conclusion that semantic roles are not dependent of domain, and it seems to be possible to achieve a domain independent list. However, the frequency of each semantic role oscillates greatly from corpus to corpus, showing that their ranking may be domain specific.

Table 1. Semantic roles annotated so far in both corpora, organized by total frequency

Semantic Roles	Newspaper	Cardiology	Total
THEME	1922	841	2763
AGENT	1673	158	1831
PLACE	430	117	547
EXPERIENCER	458	46	504
RESULT	245	211	456
VERB	316	133	449
PIVOT	197	172	369
MOMENT	248	73	321
PATIENT	266	49	315
TOPIC	282	19	301
GOAL	189	96	285
INSTRUMENT	107	153	260
SITUATION	144	107	251
ATTRIBUTE	129	51	180
CAUSE	102	76	178
DESTINATION	150	7	157
BENEFICIARY	100	46	146
MANNER	71	68	139
RECIPIENT	119	3	122

Acknowledgements. We would like to thank CNPq and CAPES for funding and project CAMELEON (CAPES/Cofecub 707/11) for all the support.

[5] http://jibiki.univ-savoie.fr/jibiki/Home.po.

[6] http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic%20role%20labels%20corpus%20-%20Brazilian%20Portuguese.

References

1. Bick, E.: The Parsing System "Palavras" : Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, vol. 202. Aarhus University Press Aarhus (2000)
2. Duran, M.S., Aluísio, S.M.: Propbank-Br: A Brazilian Portuguese corpus annotated with semantic role labels. In: Proceedings of the 8th Symposium in Information and Human Language Technology, Cuiabá/MT, Brazil, October 24-26 (2011)
3. Duran, M.S., Aluísio, S.M.: Propbank-Br: A Brazilian treebank annotated with semantic role labels. In: Proceedings of the LREC 2012, Istanbul, Turquia, May 21-27 (2012)
4. Jones, B., Andreas, J., Bauer, D., Hermann, K.M., Knight, K.: Semantics-based machine translation with hyperedge replacement grammars. In: Proceedings of COLING 2012, Mumbai, India, pp. 1359–1376. The COLING 2012 Organizing Committee (2012)
5. Kipper-Schuler, K.: VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania. Tese de doutorado orientada por Martha S. Palmer (2005)
6. Kong, F., Zhou, G.: Exploring local and global semantic information for event pronoun resolution. In: Proceedings of COLING 2012, Mumbai, India, pp. 1475–1488. The COLING 2012 Organizing Committee (December 2012)
7. Levin, B.: English verb classes and alternations: A preliminary investigation, vol. 348. University of Chicago press, Chicago (1993)
8. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational linguistics 19(2), 313–330 (1993)
9. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics Journal 31(1) (2005)
10. Salomão, M.: FrameNet Brasil: Um trabalho em progresso. Calidoscópio 7(3), 171–182 (2009)
11. Scarton, C.: VerbNet.Br: Construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. NILC/USP. Dissertação de mestrado orientada por Sandra Maria Aluísio (2013)
12. Yoshikawa, K., Iida, R., Hirao, T., Okumura, M.: Sentence compression with semantic role constraints. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea. Short Papers, vol. 2, pp. 349–353. Association for Computational Linguistics (2012)
13. Zanette, A.: Aquisição de Subcategorization Frames para Verbos da Língua Portuguesa. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio (2010)
14. Leonardo, Z., Ramischand, C., BocornyFinatto, M.J.: Desenvolvimento de um recurso léxico com papéis semânticos para o português. Linguamática 2, 23–41 (2013)
15. Leonardo, Z., Zanetteand, A., Scarton, C.: Automatic extraction of subcategorization frames from portuguese corpora. In: Aluisio, S.M., Tagnin, S.E.O. (eds.) New Languages Technologies and Linguistic Research: A Two-Way Road, pp. 78–96. Cambridge Scholars Publishing (2014)