Automatic generation of a lexical resource to support semantic role labeling in portuguese

Automatic Generation of a Lexical Resource to support Semantic Role Labeling in Portuguese

Magali Sanches Duran
Center for Computational Linguistics (NILC)
São Paulo University (USP)
São Carlos-SP, Brazil
magali.duran@uol.com.br

Sandra Aluísio
Center for Computational Linguistics (NILC)
São Paulo University (USP)
São Carlos-SP, Brazil
sandra@icmc.usp.br

Abstract

This paper reports an approach to automatically generate a lexical resource to support incremental semantic role labeling annotation in Portuguese. The data come from the corpus Propbank-Br (Propbank of Brazilian Portuguese) and from the lexical resource of English Propbank, as both share the same structure. In order to enable the strategy, we added extra annotation to Propbank-Br. This approach is part of a previous decision to invert the process of implementing a Propbank project, by first annotating a core corpus and only then generating a lexical resource to enable further annotation tasks. The reasoning behind such inversion is to explore the task empirically before distributing the annotation task and to provide simultaneously: 1) a first training corpus for SRL in Brazilian Portuguese and 2) annotated examples to compose a lexical resource to support SRL. The main contribution of this paper is to point out to what extent linguistic effort may be reduced, thereby speeding up the construction of a lexical resource to support SRL for less resourced languages. The corpus Propbank-Br, with the extra annotation described herein, is publicly available.

1 Introduction

The task of semantic role labeling (SRL) consists of identifying a predicate (a verb or a predicate noun) and its arguments, assigning to each argument the semantic roles it play in the argumental structure (Palmer et al. 2010). For example, in the sentence *"Parents complain to education department about schools constantly switching uniforms",* there are two predicates: "complain" and "switching". The argumental structure of "complain" is: "Parents" (agent), "to the education department" (recipient), "about schools constantly switching uniforms" (theme). The argumental structure of "switching" is: "schools" (agent); "constantly" (time/frequency); "uniforms" (theme).

There is no consensus regarding an ideal set of semantic role labels and, for this reason, the first difficult decision in a project of SRL is to choose which set to adopt. No matter which set is used, it is not always easy to decide which label to assign to each argument during the annotation task. In order to facilitate such decision, some projects of SRL developed lexical resources that predict the set of semantic roles required by each predicate. Some of such resources define semantic roles for verb classes, as Verbnet (Kipper et al. 2006); others for semantic frames, as Framenet (Baker et al. 1998); others define semantic roles for verb senses, as Propbank (Palmer et al. 2005) or for predicate nouns, as Nombank (Meyers et al, 2004).

The more detailed and clear is the lexical resource, the easier the decision about which role label to assign during a manual annotation task. This is very important, because when we ease SRL annotation, we increase the likelihood of obtaining a high inter-annotator agreement and, consequently, the likelihood of obtaining a good precision for machine learning classifiers for the task.

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 216–221,
Denver, Colorado, June 4–5, 2015.

Among the lexical resources available for SRL in English, we consider that of Propbank[1] the best one for supporting a distributed task of SRL annotation. From hereafter, we will refer to such lexical resource simply as Propbank, regardless the fact that Propbank encompass both the lexical resource and the annotated corpus.

Propbank does not require any linguistic expertise from the annotators and, instead of using role labels as "agent" and "patient", it uses a small set of numbered arguments, like Arg0 (for agents, causers or experiencers) and Arg1 (for patients and themes), which are described differently for each verb sense. For example, the verb sense "give.01" predicts an Arg0: "giver", an Arg1: "thing given" and an Arg2: "entity given to". This kind of description renders the roles very clear for annotators, regardless their background on semantic role labels.

Propbank has 5649[2] frame files, which are files containing (a) simple and complex predicates associated to a given verb; (b) a coarse distinction of the verb senses; (c) the set of semantic roles of each sense of a verb (rolesets) and (d) several annotated examples to show how the semantic roles may occur in real texts.

In practice, the annotator consults this kind of lexical resource while performing the annotation task. In the frame file of the verb being annotated, he looks for the sense that best suits the instance of annotation in question. Once identified the verb sense, the annotator needs to identify the constituents that play the semantic roles predicted for that verb sense, assigning them the respective role labels.

In short, the lexical resource of verbal frame files works as a repository of knowledge for SRL, accessible during the annotation task, that reduces the learning curve of SRL and facilitates the assignment of annotation tasks to several annotators. Provided that every instance receives a double-blind annotation, the quality of the annotation may be controlled through inter-annotator agreement. Instances with disagreement may be discarded or receive linguists' adjudication. This kind of lexical resource, therefore, is an essential part of the infrastructure to produce large training corpus for SRL classifiers.

It is not a simple task to construct a lexical resource, equivalent to Propbank, to support SRL in another language. Everyone that consults regularly the Unified Verb-Index[3], the system that gives access to Propbank's frame files, may observe that Propbank has been improved over the years, incorporating evidence provided by continuous annotation experience. In a project with limited budget and time, it is natural to think about reusing existing resources in order to maximize the results. In this paper, we report the strategies used to build a lexical resource to support SRL in Portuguese (hereafter referred as Verbo-Brasil), profiting from the English resource developed within the Propbank project and of annotated instances of the corpus Propbank-Br (Duran and Aluísio, 2012).

The remainder of this paper is organized as follows. Section 2 explains the strategies used in minimizing the efforts towards the construction of frame files; Section 3 briefly addresses an extrinsic evaluation of Verbo-Brasil obtained from a particular SRL annotation task. Finally, in Section 4, we present our conclusions and future work.

2 Methodology

Initially, we intended to construct Verbo-Brasil by manually creating frame files for the 1000 most frequent verbs in Portuguese, using the editor of frame files Cornerstone (Choi et. al. 2010), developed within the Propbank project. We envisaged, from the beginning, the possibility of reusing annotated instances of the corpus Propbank-Br, described in the Subsection 2.1, as examples to illustrate verb senses. However, when we started the task, we realized it was possible to automatically construct frame files, reducing the effort required for the task. Automatization entailed the use of two strategies. The first strategy constituted the creation of frame files using the existing data from both the corpus of the earlier version of Propbank-Br and the lexical repository of the English Propbank plus some new data, which was incorporated for this purpose in an updated version of Propbank-Br; this strategy is described in the Subsection 2.2. In the second strategy, described in Subsection 2.3, we duplicated the structure of the framefiles from

[1] http://verbs.colorado.edu/~mpalmer/projects/ace.html

[2] As informed in the site of the Verb-Index, updated on 08/01/2013.

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

the English Propbank to Propbank-Br for every verb which, in English, possessed a single sense.

2.1 The corpus Propbank-Br

The corpus Propbank-Br (Duran and Aluisio, 2012) was annotated by a sole linguist, aiming to provide a training corpus for SRL. During this process, we investigated to which extent the Propbank guidelines were reusable for undertaking an analogous approach to SRL in Portuguese. We ascertained the need of some adjustments in the guidelines in order to deal with differences between the Portuguese and English languages, as well as the differences between the parser outputs of the respective treebanks. As there was no lexical resource to support the annotation task, the sense distinction was made simultaneously to the annotation task, taking as base the guidelines of Propbank [4] [5].

The annotation was over the Brazilian portion of Bosque corpus (Afonso et. al. 2002), containing 4213 sentences. Bosque corpus is a treebank annotated by the parser Palavras (Bick, 2000) and revised by linguists. The sentences produced 6142 instances for annotation. Two SRL classifiers were trained on the resulting corpus. One of them (Alva-Manchego and Rosa, 2012) adopted a semi-supervised approach and obtained an F-Measure of 82.3%; the other (Fonseca and Rosa, 2013) adopted a neural architecture to label semantic arguments, disregarding the syntactic layer of annotation, and obtained an F-Measure of 62.82%.

2.2 Reusing existing data from Propbank-Br and English Propbank

To enable this strategy, it was necessary to add previously some extra data in the corpus Propbank-Br, a manual task that was by far quicker than constructing the frame files from scratch. First, we defined which fields of the frame file could be filled in with information from English Propbank, which ones could be filled in with information from Propbank-Br and which fields would require new information, not available in any one of the existing resources. The idea was to add the extra information required to the corpus Propbank-Br. Aiming this, we created six "word tags" in corpus Propbank-Br, using the same annotation tool used to annotate the original corpus (SALTO – Burchardt et al. 2006), as may be seen in Fig.1.

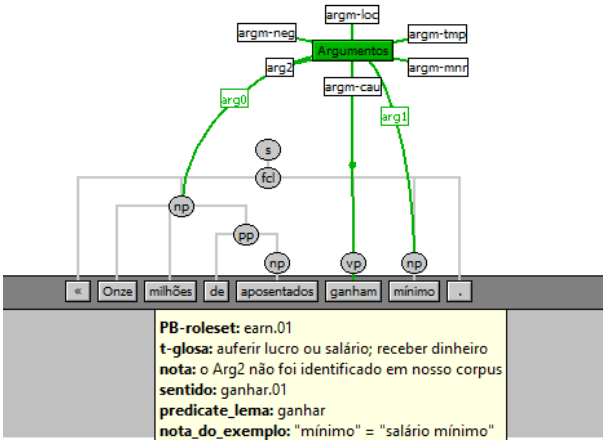

Figure 1. Extra annotation inserted in Propbank-Br.

The word tags are:

(1) *PB-roleset*: an equivalent roleset-id in Propbank, was used as field key to bring, from Propbank, the semantic roles, the semantic roles description, the related Verbnet classes and the Verbnet roles to the framefiles (Fig. 2);

(2) *t-glosa*: field that was filled in only in the first occurrence of a verb sense; it contains a brief description or a synonym of this sense of the verb to distinguish it from the other possible senses.

(3) *Nota* (note): field used for observations regarding a roleset contained within a verb's framefile when further clarification is thought to be helpful to the annotator;

(4) *Predicate_lemma*: field, filled in only in the first occurrence of a verb sense, containing the verb lemma or the name of a complex predicate (phrasal verb) when applicable;

(5) *Sentido* (sense): field that indicates which verb sense is the one being used in the sentence in question, also referred as roleset id, and is filled in for all instances. Once classified, the sentences can be subsequently added as examples of their respec-

[4] http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf

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

tive verb sense within the appropriate frame file;

(6) *Nota_do_exemplo* (example note): field used to convey information about a given example.

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)

Figure 2. Data brought from Propbank using the roleset id as field key.

Once we had created the word tags in the corpus, we undertook the annotation task to fill in them, as showed in the Fig.1. The greater the number of senses of a verb (polysemy), the greater was the difficulty to elect an English equivalent in English Propbank to fill in the word tag "PB_roleset". We realized that highly polysemous verbs would demand special attention in the next phase of the process, that is, during the revision of frame files automatically generated.

The annotation task provided the identification of 1453 verb senses in Portuguese for 1060 verb lemmas (an average of 1.37 senses per lemma). From the 1060 verb lemmas annotated in the corpus, 80% present only one sense, 13% present two senses; 3% present three senses and 4% present four or more senses. Only 109 of the 1453 senses identified in Portuguese did not have an equivalent verb sense in English identified in Propbank. Consequently, as the frame files of such 109 verbs could not obtain the fields brought from Propbank automatically, they required manual edition.

Using the XML frame file structure of Propbank, we defined the automatic generation of frame files, combining data from Propbank-Br and from Propbank, as shown in Fig.3. In the frame file structure, we used the field called "framnet" (aimed to store mappings to Framenet) for the information brought from the word tag "PB-roleset", that is, the equivalent roleset id in the English Propbank. The Propbank roleset id was the field key to access and bring data from the respective English frame file.

```
<predicate lemma="ganhar">

    <roleset    framnet="earn.01" id="ganhar.01"
                name="auferir lucro ou salário; receber dinheiro"
                source=""
                vncls="13.5.1">
        <roles>
            <role descr="earner" f="" n="0">
                <vnrole vncls="13.5.1" vntheta="agent"/>
            </role>
            <role descr="wages" f="" n="1">
                <vnrole vncls="13.5.1" vntheta="theme"/>
            </role>
            <role descr="benefactive" f="" n="2">
                <vnrole vncls="13.5.1" vntheta="beneficiary"/>
            </role>
            <role descr="source" f="" n="3">
                <vnrole vncls="13.5.1" vntheta="source"/>
            </role>

        <note/>
```

Figure 3. Frame file combining data from Propbank-Br and data brought from Propbank

The strategy succeeded, and we achieved 1060 frame files with 1453 verb senses and 6142 annotated examples. After the generation, we began the revision of frame files with the most frequent verbs, translating the description of semantic roles, as may be seen in Fig.4. Currently, 541 frame files are fully revised.

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

Figure 4. Frame file that combines information from corpus Propbank-Br and from Propbank's equivalent roleset.

2.3 Extension of the lexical resource using monosemous verbs

During the task of filling in the word tags in the corpus Propbank-Br, we observed that verbs presenting a unique sense (monosemous verbs) were

the easiest to link to an English verb sense in Propbank and almost always the equivalent verb sense was the unique sense of the respective frame file. This led us to hypothetize that monosemous verbs in Portuguese, would probably correspond to monosemous verbs in English and vice-versa, whenever an equivalent verb exists.

On that basis, we decided to extend our resource taking as the start point the frame files that have a single verb sense in English Propbank. We identified 3737 English frame files that met such condition. We then translated only the verb lemmas of such frame files. Translation was executed automatically using Google translator and revised manually. We chose Google translator because we needed to translate at once 3737 out-of-context verb lemmas in a quickly and uncomplicated manner. It would be ideal if Google translator returned the word class of the results, thus allowing us to filter the verbs (we would only have obtained such result if we had translated the verbs one-by-one).

For several verbs, the automatic translation provided no output. Among the output words in Portuguese, there were several nouns, many of which do not correspond to any verb in Portuguese (eg. "to hangar", "to shark", "to tassle"). We then revised the translation, providing better equivalents when necessary and marking an "N" for those translated lexical items that were not verbs in Portuguese. After eliminating: (1) repetitions of translated verbs (two or more verbs translated into a same verb in Portuguese) and (2) verbs that we already had in our database, we obtained 1538 new verbs to extend our resource.

The next step was to duplicate the respective English frame files, using the name of the verb in Portuguese to substitute the name of the English verb in the fields "roleset id" and "predicate lemma". Subsequently, we replaced the example sentences in English by ones in Portuguese, extracted from corpus PLN-Br (Bruckschen et al., 2008). Lastly, to complete these new frame files, we are now annotating the examples with semantic role labels. Cornerstone frame files editor is being used for this task.

3 Evaluation

The two strategies we reported to automatically generate Portuguese frame files gave us 2598 framefiles. The 541 frame files already revised correspond to the verbs with frequency above 1000 in the corpus PLN-Br, which include the most polysemous verbs in Portuguese. Such verbs were target of a double-blind annotation task of 8345 instances extracted from the same corpus. The annotation task has just been accomplished and will be fully reported in a later date; the Kappa inter annotator agreement (Carletta, 1996) for verb sense identification was 0.93.

This annotation task gave us feedback to evaluate and improve the respective frame files. Among the actions taken during the annotation task we can cite: adding new senses identified in the corpus; merging or splitting senses for verbs that presented low inter-annotator agreement; including new examples to better illustrate a verb sense.

4 Concluding Remarks and Future work

The approach we adopted to build a Propbank-like lexical resource to support SRL in Brazilian Portuguese may be of use for other researchers working on under-represented languages and with a limited budget.

The 541 already revised frame files were used in a double-blind annotation SRL task that obtained a Kappa inter-annotator agreement for sense distinction of 0.93.

In the future, we plan to use Verbnet classes, an information brought from the equivalent verb sense in Propbank, to find in Verbnet-Br (Scarton et al., 2014) verb senses that are not in Verbo-Brasil.

As soon as we accomplish the revision of the frame files, we will make Verbo-Brasil publicly available. The new version of the corpus Propbank-Br, with the extra annotation described in this paper is now available for download at nilc.icmc.usp.br/portlex/index.php/en/downloadsingl.

Acknowledgments

Part of the results presented in this paper were obtained through research activity in the project entitled "Semantic Processing of Texts in Brazilian Portuguese", sponsored by Samsung Eletrônica da Amazônia Ltda, under the terms of Brazilian federal law number 8.248/91. We also thank financial support from FAPESP and CAPES, process number 151/2013 (Portal de Min@s project).

References

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

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. *Proceedings of Computational Linguistics 1998 Conference*.

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

Bruckschen, M., Muniz, F., Souza, J. G. C., Fuchs, J. T., Infante, K., Muniz, M., Gonçalves, P. N., Vieira, R.; Aluísio, S. M. (2008) "Anotação Linguística em XML do Corpus PLN-BR". *NILC-TR-09-08*, 39 p.

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

Carletta, Jean. (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), pp. 249–254.

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

Duran, M. S.; Aluísio, S. M. (2012). Propbank-Br: a Brazilian Treebank annotated with semantic role labels. *Proceedings of LREC 2012*, pp. 1862-1867.

Fonseca, E. R.; Rosa, J. L. G. (2013) A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. Proceedings of IJCNN 2013 International Joint Conference on Neural Networks.

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

Meyers, A.; Reeves, R.; Macleod, C.; Szekely, R.; Zielinska, V.; Young, B.; Grishman, R. (2004), The NomBank Project: An Interim Report. *Proceedings. of HLT-EACL Workshop: Frontiers in Corpus Annotation*.

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

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

Scarton, C. ; Duran, M. S.; Aluísio, S. M. (2014) Using Cross-linguistic Knowledge to Build VerbNet-Style Lexicons: Results for a (Brazilian) Portuguese VerbNet. *Proceedings of the International Conference on Computational Processing of Portuguese* (PROPOR 2014). Heidelberg: Springer Verlag, 2014. v. 1. p. 153-164.