Semantic Role Labeling for Portuguese – A Preliminary Approach –

João Sequeira, Teresa Gonçalves, and Paulo Quaresma

Universidade de Évora
m5071@alunos.uevora.pt, {tcg,pq}@uevora.pt

Abstract. Currently there are increasingly more private and academic publications in the form of digital content on the Internet making extremely difficult to extract and maintain the content information manually. Normally, these tasks follow approximations based on natural language processing. This paper presents a preliminary approach for obtaining a semantic role labeler for Portuguese, a little explored aspect of natural language processing for this language. The approach was evaluated for the 3 most frequent semantic roles (relation, subject and object) with a subset of Bosque 8.0 corpus. The same approach was applied to an English corpus – the CONLL'2004 one and its results were compared to the ones obtained on the CONLL'2004 shared task. At the same time it presents BosqueUE, a Portuguese corpus for semantic role labeling that can be the basis material for future research in the area. This corpus has the same format as the CONLL'2004 one, facilitating multi-language evaluations.

1 Introduction

Currently there is a large amount of digital content (academic, personal,news and other) available on the internet. The task of extracting information content from these different kind of sources became practically impossible [22]. With the increase of digital content published there was also an increase in research applications able to automatically analyze and extract information from them [7].

The semantic role labeling, portraying the semantic relationships between the different constituents of the sentence, has been an area of increasing interest due to their importance in applications of information extraction, question-answering, document summarization and others that require semantic information [6]. This aspect of natural language processing already has several available resources for the English language, product of several projects under international conferences [6], but there is still much material to be explored in other languages, such as the Portuguese one.

This paper describes the construction of a Portuguese corpus for the Semantic Role Labeling task and the use of the MinorThird tool [9] as a preliminary work for this NLP task. To have a means of comparison, MinorThird is also used with

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 193–203, 2012.
© Springer-Verlag Berlin Heidelberg 2012

the English corpus built for the CONLL'2004 Conference[1] and its results are compared with the ones obtained there [6].

It is organized as follows: Section 2 introduces the Semantic Role Labeling task and some systems built to perform it, Section 3 describes the construction of a Portuguese corpus for the SRL task and Section 4 presents a preliminary semantic role labeler built with MinorThird. Finally Section 5 discusses the obtained results and enumerates future work.

2 Semantic Role Labeling

This section introduces the semantic role labeling task and presents some work done in the field.

2.1 The Task

Semantic Role Labeling is currently one of the most active subgroups in the area of natural language processing. It intends to identify the verbs in sentences and its syntactic arguments [7], such as the subject of the action and the object of the action among others.

Figure 1 shows the semantic roles for a Portuguese sentence present in Bosque 8.0 [1], a Portuguese corpus parsed by the Palavras tool [3] and manually revised by linguists. The sentence has a subject ("`Vera`"), a verb that makes up the predicate ("`apagou`") and an object ("`a luz`"). The predicate is tagged as the *Relation*, the subject as *Arg0* and the object as *Arg1* (*Arg0* and *Arg1* are numbered arguments of the predicate used by the Propbank annotation).

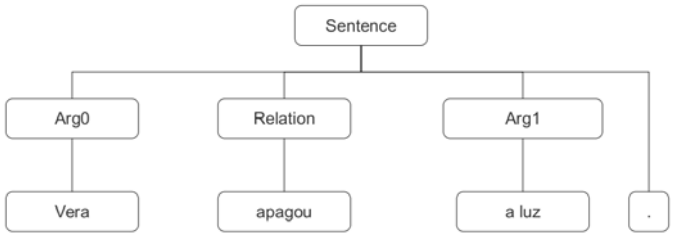

Fig. 1. A Portuguese sentence and the semantic roles present in it

Gildea and Jurafsky [13], pioneers in the semantic role labeling task, listed two prominent methods to perform the analysis of texts: the grammar based systems and the data-driven ones. The process of creating grammars is very time consuming as they are created by hand and need to include a description for each existing case of the language. On the other hand, data-driven systems

[1] Conference on Computational Natural Language Learning.

need a classified corpus, and an application to create a model from them. This model is then used to classify new texts. Examples of these applications are the participants of CONLL shared tasks, specifically for the years of 2004 and 2005 [6,7].

Palmer *el. al* [24] introduces SRL and discusses the most important and discriminating features used by this task. Next sub-section presents some semantic role labeling systems.

2.2 Related Work

Bick [5] describes a grammar based semantic-role annotator for the Portuguese language: it uses a constraint grammar to map and disambiguate 40 different semantic roles. The grammar has 500 mapping rules and a small number of disambiguation rules. It is reported to have an average f-score of 88.6%.

Amancio *et. al.* [2] describe a similar task to SRL, which is the assigning of Wh-Questions to Verbal Arguments via machine learning, using the same features used in SRL.

Gildea and Hockenmaier [14] present a systems that uses a combinatory categorial grammar to identify the semantic roles for the English language and is reported to have a f-score of 80.4 on PropBank [16,23].

CONLL'2004 and CONLL'2005 shared tasks aimed at developing data-driven systems for semantic role labeling. They used the Penn Treebank [21,26] with predicate-argument annotations from the PropBank. On CONLL'2005 the Brown corpus [12] was also used to realize tests.

On CONLL'2004 the systems with better performance were the ones developed by Hacioglu et al. [15] with an f-score of 69.49 for the test set and Punyakanok et al. [28] with an f-score of 66.39.

The Hacioglu system was developed using TinySVM [17], a Support Vector Machine implementation, with a polynomial kernel of degree 2 using features from three categories [15]:

- **Base** Features. These can be inferred directly from the simplest data: words, verb lemmas, part-of-speech, BP positions using IOB, clause labels present in the sentence and named entities (using person, local, organization and others);
- **Token** Features. These are those that can be inferred at the level above the basic features (the level of BP) such as: token position with respect to the predicate, path between the token and the predicate, clause patterns, distances between the token and the predicate and the number of words in a the token
- **Sentence** level features such as part-of-speech of the predicate and the words that precede and follow it, frequent/rare predicate, context window of the predicate, semantic frames of the predicates present in PropBank and number of predicates in the sentence.

The Punyakanok et al. system [28] is composed by a set of classifiers and an inference procedure used both to clean the classification results and to ensure structural integrity of the final role labeling. The learning algorithm used is a

variation of the Winnow update rule incorporated in SNoW [31,32], a multi-class classifier that is specifically tailored for large scale learning tasks.

The system consists of three phases [28]:

1. Find **Argument Candidates.** The system tries to filter out unlikely candidates with two classifiers: one to detect beginning phrase locations and other to detect end phrase locations. Both use the following features: word, POS tag, IOB tags for chunks, lemma and POS tag of the active predicate, active/passive voice of the current predicate, word position with respect to the predicate, the boundary of clauses, the sequence of chunks from the current word to the predicate, the path formed from a semi-parsed tree containing clauses and chunk and the position of the target word relative to the predicate;
2. **Phrase Classification.** This phase is accomplished with a multi-class classifier used to supply confidence scores of how likely individual phrases have specific argument types and The most likely solution is chosen using the matrix of confidences and linguistic information. It uses a multitude of linguistic, position and features.
3. **Filter Function.** This phase applies global constraints derived from linguistic information and structural considerations.

In CONLL'2005, the systems that obtained the best results were [27] and [25] with f-scores of 77.92 and 77.30 respectively for test set combining the Penn Treebank and the Brown corpus.

The Punyakanok et al. system [27] uses the same learning algorithm as the system presented in CONLL'2004. It has four stages:

1. **Pruning.** Very unlikely constituents are filtered by means of an heuristic presented in [36];
2. **Argument Identification.** Uses binary classification to identify whether a candidate is an argument or not;
3. **Argument Classification.** Uses a multi-class classifier trained to differentiate the types of the arguments supplied by the previous stage;
4. **Inference.** It incorporates linguistic and structural knowledge to resolve any inconsistencies of argument classification. The process is formulated as an integer linear programming (ILP) problem that takes as inputs the confidences over each type of the arguments supplied by the argument classifier using several constraints.

The Pradhan et al. system [25] uses TinySVM to train one-vs-all classifiers with Support Vector Machines developing a binary classifier to each semantic class plus a "NULL" class. It uses two systems: one chunk based that are very efficient and robust and other based on full syntactic parses that normally are more accurate; the goal was to preserve the robustness and flexibility of the segmentation of the phrase-based chunker, and take advantage of features from full syntactic parses. For an input sentence, syntactic constituent structure parses are generated by a Charniak [8] parser and a Collins [10] parser. Semantic role labels are assigned to the constituents of each parse using Support Vector Machine classifiers. The

resulting semantic role labels are converted to an IOB representation. These IOB representations are used as additional features, along with flat syntactic chunks, by a chunking SVM classifier that produces the final SRL output [25].

3 BosqueUE – A Portuguese Corpus for SRL

As already mention, to build a data-driven SRL system for the Portuguese language it is necessary to have a classified corpus. Besides including the semantic roles of each sentence's chunk, it helps to have the morphologic, syntactic and other kind of features.

This enriched corpus is primally based on Bosque 8.0[2], a corpus incorporated in Forest Sintá(c)tica project [1]. This project consists of plain text with syntactically analyzed sentences (tree structures) by the PALAVRAS parser [4]. Figure 2 shows the information retained in Bosque 8.0 for the sentence '`Vera apagou a luz.`'.

```
'source' => 'CP429-7 Vera apagou a luz.',
'number' => 1,
'cod' => 'CETEMPúblico n=429 sec=clt sem=96a',
't' => [
  'fcl||STA',
  [
    'np||SUBJ',
    'prop(\'Vera\' F S)||H::Vera'
  ],
  [
    'vp||P',
    'v-fin(\'apagar\' PS 3S IND)||MV::apagou'
  ],
  [
    'np||ACC',
    'art(\'o\' <artd> F S)||>N::a',
    'n(\'luz\' <np-def> F S)||H::luz'
  ],
  'jjpunct(-.-)'
]
```

Fig. 2. Bosque 8.0 representation for the sentence '`Vera apagou a luz.`'

Bosque 8.0 consists of 9368 sentences from the first 1000 extracts from the CETEMPúblico and CETEMFolha, prioritizing quality over quantity [20]. CETEMPúblico uses news taken from the Público newspaper and CETEMFolha uses news excerpts taken from the Folha de S. Paulo newspaper.

A subset of 4416 sentences from Bosque 8.0 was used to build BosqueUE; this subset was obtained by selecting the CETEMPublico sentences that ended with a punctuation mark.

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

BosqueUE corpus was built using the same format as the corpus used in CONLL'2004; Figure 3 presents the extract for same sentence.

```
Vera      Vera     PROP    B-np    B-SUBJ     B-PER (S*
apagou    apagar   V       B-vp    B-p        O     *
a         o        DET     B-np    B-ACC      O     *
luz       luz      N       I-np    I-ACC      O     *
.         .        PU      O       O          O     *S)
```

Fig. 3. BosqueUE representation for the sentence '`Vera apagou a luz.`'

This format has a word per line and contains the following 7 features: word, lemma, part-of-speech tag, chunks with IOB, semantic roles with IOB, named entities with IOB and clauses. Words, lemmas, chunks and clauses were extracted from the Bosque8.0 corpus; the part-of-speech column uses LABEL-LEX [18] tags having performed a manual review in ambiguous situations; the named entities were obtained with the application presented in [22].

A similar annotated corpus for SRL for Portuguese language, is the Propbank-Br [11], a Brazilian treebank annotated with semantic role labels. This corpus consists of 6142 instances for SRL annotation, with 1068 different predicates. This corpus follows the Propbank guidelines and uses the syntactic trees of the Brazilian portion of Bosque.

4 A Preliminary Semantic Role Labeler

This section presents a preliminary approach of using a Portuguese corpus for the SRL task: it starts by introducing MinorThird tool and then presents corpora and the experimental setup. It ends by displaying the results obtained.

4.1 MinorThird

MinorThird is an open source set of Java classes to perform tasks over texts, such as text classification and named entity extraction. It was created by Professor William W. Cohen of the Carnegie Mellon University and is currently maintained Frank Lin [9].

It uses a collection of documents to create a database called *TextBase* and logical statements over text chunks are stored in *TextLabels* objects. Since the annotations on *TextLabels* are independent of the contents of the documents it is possible to have different annotations on same set of documents. The annotations in *TextLabels* describe syntactic or semantic properties for words, documents or spans and can be created manually or automatically through an application [9].

It implements sequential learning methods such as Conditional Random Fields [35,19] and semi, conditional, maximum entropy and hidden Markov Models [33].

4.2 Corpora

The BosqueUE corpus was transformed into XML documents with syntactic annotations. For example, the sentence '`Vera apagou a luz`' is represented as showed in Figure 4 (tag`<P>` stands for *Predicate*, `<Arg0>` for *Subject* and `<Arg1>` for *Object*).

```
<Arg0>Vera</Arg0> <P>apagou</P> <Arg1>a luz</Arg1>.
```

Fig. 4. SRL tags for the sentence '`Vera apagou a luz.`'

In order to compare the outcome of the BosqueUE Portuguese corpus with an English one, a similar process was carried out with the corpus used for the CONLL'2004 shared task [6][3]. This corpus consists of six sections of the Wall Street Journal part of the Penn Treebank [21] adding information from predicate-argument syntactic structures [16,23].

4.3 Experimental Setting

In order to obtain the best results several MinorThird algorithms with the default values were tried with a context window of size three. The best results were obtained for:

- SVMCMM: Conditional Markov Models [30,29] trained with Support Vector Machines [34];
- CRF: Conditional Random Fields [35,19]

As already said, BosqueUE is composed of 4416 sentences; from all its syntactic tags, models were built for the ones with statistic validity, namely, *Predicate*, *Subject* and *Object*.

CONLL'2004 corpus is divided into train, development and test sets composed respectively of 8936 (sections 15–18 of Penn TreeBank), 1671 (section 20) and 2012 sentences (section 21). Only train and test sets were used, since MinorThird algorithms were tried with default values and models were built for the same tags.

Table 1 shows the syntactic tags with the number of times they appear in each corpus.

Table 1. Main tags, semantic roles and count for BosqueUE and CONLL'2004 corpora

		BosqueUE	CONLL'2004	
tag	semantic role	#	# train	#test
`P`	Predicate	7268	19098	3627
`Arg0`	Subject	4673	12709	1671
`Arg1`	Object	3802	18046	3429

[3] Retrieved from: `http://www.lsi.upc.edu/srlconll/st04/st04.html`

A 10-fold cross-validation procedure was applied over the BosqueUE corpus and a train/test procedure was applied to CONLL'2004 one. Model's performance was analyzed through precision (π), recall (ρ) and F_1 (f_1) measures.

4.4 Results

Table 2 shows the results obtained with SVMCMM and CRF algorithms for the BosqueUE corpus. In it, it's possible to observe that CRF algorithm consistently presents better precision values while SVMCMM presents better recall ones.

For both algorithms precision values are at least 0.1 above recall ones for all tags (for CRF algorithm and `Arg0` and `Arg1` tags precision is 0.2 higher than recall). As expected the *Predicate* tag presents the best results with f_1 values above 52%, while the *Object* tag has f_1 values below 20%.

Table 2. BosqueUE: Precision, recall and F_1 for SVMCMM and CRF algorithms

	SVMCMM			CRF		
tag	π	ρ	f_1	π	ρ	f_1
`P`	.588	.477	.527	.648	.477	.549
`Arg0`	.388	.259	.311	.434	.237	.306
`Arg1`	.269	.147	.190	.317	.117	.171

Table 3 shows the results for the CONLL'2004 corpus obtained with SVMCMM and CRF algorithms.

As opposed to BosqueUE, in CONLL'2004 corpus both algorithms present similar precision and recall values (except for `Arg0` tag where CRF has a 0.1 higher precision value). Once again, the *Predicate* tag presents the best results with f_1 values above 82%, while the *Object* tag has f_1 values below 24%.

Table 3. CONLL'2004: Precision, recall and F_1 for SVMCMM and CRF algorithms

	SVMCMM			CRF		
tag	π	ρ	f_1	π	ρ	f_1
`P`	.850	.823	.836	.842	.805	.823
`Arg0`	.599	.464	.523	.699	.463	.557
`Arg1`	.372	.170	.234	.414	.151	.221

Comparing Table 2 and Table 3, one can conclude that the ones obtained for the Portuguese corpus are below the corresponding ones for English corpus. This gap could be explained by the different sizes of the datasets: CONLL'2004 is around 3 times bigger than BosqueUE.

Table 4 compares F_1 values for `Arg0` and `Arg1` tags obtained by SVMCMM Minorthird algorithm with the best and worst ones from CONLL'2004 shared task as reported on [6] (*Predicate* values are not shown since they were not reported on the shared task).

Table 4. F_1 values obtained by SVMCMM and the best algorithm from CONLL'2004 shared task

		CONLL'2004	
tag	SVMCMM	best	worst
`Arg0`	.523	.814	.562
`Arg1`	.234	.716	.490

From the table one can see that the use of linguistic information such as words, part-of-speech and chunk labels, clauses and named entities are useful for the semantic role labeling problem and the sequential learning methods alone are not enough. While these features improve both labelers (`Arg0` and `Arg1`) the increase is greater for *Object* role than for the *Subject* one.

5 Conclusions and Future Work

This work attempts to apply a line of research still little explored for the Portuguese language. It was found that the preliminary results obtained with a Portuguese corpus are below those obtained with an English corpus. As already mentioned this difference could be explained by the different corpus size. Another possible explanation is the use of more complex syntactic structures and many word flexions that exists in the Portuguese language when compared with the English one.

On the other hand it is possible to conclude that the use of linguistic information such as words, part-of-speech and chunk labels, clauses and named entities are useful for the semantic role labeling problem and the sequential learning methods alone does not produce good results.

As future work we intend to increase the size of the Portuguese corpus and develop a classifier that makes use of all the linguistic information that the built BosqueUE corpus provides. Only then a comparison between both languages will be fair.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Florestra sintá(c)tica: A treebank for portuguese. In: LREC 2002, the Third International Conference on Language Resources and Evaluation, pp. 1698–1703 (2002)
2. Amancio, M.A., Duran, M.S., Aluisio, S.M.: Automatic question categorization: a new approach for text elaboration. Procesamiento del Lenguaje Natural (46), 43–50 (March 2011)
3. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, Aarhus University, Aarhus, Denmark (November 2000)
4. Bick, E.: The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)

5. Bick, E.: Automatic semantic-role annotation for portuguese. In: Anais do XXVII Congresso de SBC (2007)
6. Carreras, X., Màrquez, L.: Introduction to the conll-2004 shared task: Semantic role labeling. In: Proceedings of CoNLL 2004 (2004)
7. Carreras, X., Màrquez, L.: Introduction to the conll-2005 shared task: Semantic role labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005 (2005)
8. Charniak, E.: A maximum-entropy inspired parser. In: Proceedings of NAACL 2000 (2000)
9. Cohen, W.: Minorthird: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data (2004), http://minorthird.sourceforge.net
10. Collins, M.: Head-driven statistical models for natural language parsing. Computational Linguistics 29(4), 589–637 (2003)
11. Duran, M.S., Aluisio, S.M.: Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In: STIL 2011 – 8th Symposium in Information and Human Language Technology (October 2011)
12. Francis, W., Kucera, H.: Brown corpus manual (1997), http://icame.uib.no/brown/bcm.html
13. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28, 245–288 (2002)
14. Gildea, D., Hockenmaier, J.: Identifying semantic roles using combinatory categorial grammar. In: Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, EMNLP 2003, pp. 57–64. Association for Computational Linguistics, Stroudsburg (2003)
15. Hacioglu, K., Pradhan, S., Ward, W., Martin, J., Jurafsky, D.: Semantic role labeling by tagging syntactic chunks. In: Proceedings of CoNLL 2004 Shared Task, pp. 110–113 (2004)
16. Kingsbury, P., Palmer, M.: From treebank to propbank (2002), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.7566
17. Kudo, T.: Tinysvm: Support vector machines (2002), http://chasen.org/~taku/software/TinySVM
18. Laboratório de Engenharia da Linguagem: Label-lex (1995), http://label.ist.utl.pt/pt/apresentacao.php
19. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of 18th International Conference on Machine Learning, pp. 282–289 (2001)
20. Linguateca: Florestra sintá(c)tica (2009), http://www.linguateca.pt/floresta/corpus.html
21. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of english: The penn treebank. Computational Linguistics 19(2), 313–330 (1993)
22. Miranda, N., Raminhos, R., Seabra, P., Sequeira, J., Gonçalves, T., Quaresma, P.: Named entity recognition using machine learning techniques. In: EPIA 2011, 15th Portuguese Conference on Artificial Intelligence, Lisbon, PT (October 2011)
23. Palmer, M., Gildea, D., Kingsbury, P.: The preposition bank: An annotated corpus of semantic roles. Computational Linguistics 31 (2005)
24. Palmer, M., Gildea, D., Xue, N.: Semantic Role Labeling. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
25. Pradhan, S., Hacioglu, K., Ward, W., Martin, J., Jurafsky, D.: Semantic role chunking combining complementary syntactic views. In: Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005 (2005)

26. Project, T.P.T.: The penn treebank project (1999), http://www.cis.upenn.edu/~treebank/
27. Punyakanok, V., Koomen, P., Roth, D., Yih, W.: Generalized inference with multiple semantic role labeling systems. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005), pp. 181–184 (2005)
28. Punyakanok, V., Roth, D., Yih, W., Zimak, D., Tu, Y.: Semantic role labeling via generalized inference over classifiers. In: Proceedings of CoNLL 2004 Shared Task (2004)
29. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
30. Rabiner, L., Juang, B.: An introduction to hidden markov models. IEEE ASSP Magazine (Janeiro 1986)
31. Roth, D.: Learning to resolve natural language ambiguities: A unified approach. In: Proc. of AAAI, pp. 806–813 (1998)
32. Roth, D., Yih, W.: Probabilistic reasoning for entity & relation recognition. In: The 19th International Conference on Computational Linguistics, COLING 2002, pp. 835–841 (2002)
33. Stamp, M.: A revealing introduction to hidden markov models (2004), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.136.137&rank=1
34. Vapnik, V.: Statistical Learning Theory. Wiley-Interscience (Setembro 1998)
35. Wallach, H.: Conditional random fields: An introduction (2004), http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.124.6711
36. Xue, N., Palmer, M.: Calibrating features for semantic role labeling. In: Proc. of the EMNLP 2004, pp. 88–94 (2004)