Filling the gap: inserting an artificial constituent where a subject is omitted in Portuguese

Nathan Siegle Hartmann, Magali Sanches Duran and Sandra Maria Aluísio

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo

nathanshartmann@gmail.com, magali.duran@uol.com.br, sandra@icmc.usp.br

Abstract. This paper reports the first efforts to insert null elements to represent omitted subjects in Portuguese. Our aim is to fill some gaps in the syntactic structure in order to facilitate the assignment of semantic role labels and thus provide a better training corpus for SRL classifiers. The main advantage of inserting such null elements is to reduce data sparsity, as all the verbal clauses become similar in what concerns the presence of explicit subjects. The results show a better precision in the insertion of null elements related to subjects of verbs inflected in the first person, both singular and plural.

Keywords: semantic role labeling. null elements. omitted subject.

1 Introduction

In Portuguese language, we can, in most cases, infer the grammatical person from the verb inflection and, for this reason, we can omit the subject without jeopardizing the comprehension. However, subject omission represents an additional difficulty for some language processing tasks, such as semantic role labeling (SRL).

SRL is a Natural Language Processing (NLP) task which aims to detect semantic predicates describing events in a sentence and to assign semantic role labels to each argument of the event structure [1]. Following PropBank's project [2] annotation guidelines [3], there is a ranking of role labels and the higher ranked roles are assigned first. When we annotate SRL over syntactic trees, the subject node generally receives the most important role label of the roleset. If the subject is omitted, the sequence of semantic role labels' assignment is affected. For this reason, in spite of not being a problem for human annotators, subject omission certainly prevents PropBank´s based SRL systems from achieving a better performance.

Despite the fact that English language does not allow subject omission in main clauses, we sought inspiration in Penn Treebank [4] to address the problem of subject omission in Portuguese. The Financial subcorpus of Penn Treebank was annotated with semantic role labels in the PropBank project and the strengths of its syntactic annotation for SRL are widely discussed in PropBank's annotation guidelines [3].

One of them is the “null element”, an artificial constituent used to represent an omitted element.

Marcus et al. [4] argues that “… the easiest mechanism to include information about predicate-argument structure, although indirectly, is by allowing the parse tree to contain explicit null items”. In Penn Treebank, the type of null element that particularly inspired us is *PRO*. Such null element is inserted where there is an underspecified or unrealized subject of a verb. For example, in the sentence “We expected to win the World Cup”, the subject of the verb “to win” is unrealized. In this case, Penn Treebank annotation would represent the unrealized subject of “to win” with *PRO* (1):

(1) We**-1** expected ***PRO*-1** to win the World Cup.

The insertion of null elements contributes to reduce data sparsity, a relevant requirement for machine learning approaches to the task of semantic role labeling.

Penn Treebank would also coindex the *PRO* null element with the realized subject of the higher clause that governs the clause “to win the World Cup”, i.e. “We”. If the reference was not clear, the Penn Treebank would not create the link.

In Portuguese we have the same situation, i.e., unrealized subjects of clausal arguments of higher clauses. But we have also situations in which even the subject of a higher clause is not realized. Nevertheless, no matter whether *PRO* has a referent or not, in SRL annotation it will be assigned the semantic role label corresponding to the subject of the verb.

In Portuguese, the insertion of null elements to represent subjects should cover three phenomena: (i) ellipsed subjects of embedded clauses and coordinated clauses, like in examples (2) and (3), which have referents in the sentence; (ii) omitted subjects, inferable from verb inflection, like in example (4); and (iii) - underspecified subjects, as illustrated by example (5):

(2) Nós queremos *PRO* dar uma festa. (We want to give a party.).

(3) Eles lutaram e *PRO* venceram os adversários. (They fought and *PRO* beat the opponents.).

(4) *PRO* Acho que isso não vai ser possível. (*PRO* think it will not be possible = I think it will not be possible.).

(5) *PRO* Dizem que aqui tem fantasmas. (*PRO* say there are ghosts here = It is said there are ghosts here.).

Brazilian Portuguese has a corpus annotated with semantic role labels, following PropBank’s Guidelines - PropBank-Br [6] - which contains 5942 annotated instances (10% of the English PropBank), for 1025 different verbs. The layer of semantic role labels was added over the Brazilian portion of corpus Bosque, which is part of the Floresta Sintá(c)tica treebank of Portuguese [5]. However, Bosque, unlike Penn Treebank, does not contain null elements, and is one of the reasons that explain the lower precision of classifiers trained on PropBank-Br in comparison to classifiers trained on PropBank.

As we will annotate semantic role labels in another genre of corpus [7], which is not a Treebank, we decided to explore the possibility of inserting a null element to represent omitted subjects as a preprocessing step before the human annotation of semantic role labels. Our aim is to fill some gaps in the syntactic structure in order to facilitate the assignment of semantic role labels and thus provide a better training corpus for SRL classifiers.

For this first study we used a journalistic corpus already annotated with SRL - PropBank-Br - disregarding all the previous syntactic and semantic annotations. Many other corpus would be equally suitable for our purpose, but we chose this in order to leave open the possibility for future comparisons between this new corpus without annotations and PropBank-Br.

We parsed the corpus with the last version of parser PALAVRAS [7], which improved the recognition of auxiliary verbs in verbal chains, a relevant pre-requisite for our task. It is important to stress that we are working with automatically parsed sentences, without human correction and, therefore, parsing errors are expected.

This study is exploratory, as there is no similar effort reported for Portuguese as far as we know. The experience put in evidence the challenges and limitations to be faced by the task. We believe this procedure may be beneficial for other NLP tasks, such as translation systems and question and answering.

2 Rules for insertion of omitted subjects

In the context of SRL, the insertion of artificial subjects in the place of omitted subjects constitutes a preprocessing step once these elements will support latter assignment of semantic role labels. As generally SRL systems are approached via machine learning methods [1] [8] that depend on syntactic features, the insertion of an artificial subject must respect a syntactic tree model. Therefore, we need to parse the corpus before inserting “null elements”.

First of all, we parsed the Brazilian portion of corpus Bosque using the last version of PALAVRAS parser. It generated a non-gold standard set of the corpus. Observing syntactic patterns in the output of the parser PALAVRAS, we defined 16 rules to automatically insert null elements to represent omitted subjects in a scenario of automatic parsing without human correction (non-treebank).

The process of rule creation was exploratory and incremental, since we started with a simple rule: to insert an artificial subject contained in an NP[1] node every time there is no NP at the left side of the target verb (VP). This artificial subject is the missing pronoun “Eu” (*I*) or “Nós” (*We*) for the first person of singular and plural (Fig. 1a) and a generic null element named “SUJ” for the third persons inflected verbs. Third persons are “Ele”, “Ela”, “Isso/Isto” and “Você” (*He*, *She*, *It* and *You*) in the singular

[1] PALAVRAS tagset can be found in http://beta.visl.sdu.dk/visl/pt/symbolset-floresta.html.

and “Eles”, “Elas” and “Vocês” in the plural (*They*, masculine and feminine; and *You*, plural) as may be seen in Fig. 1b.

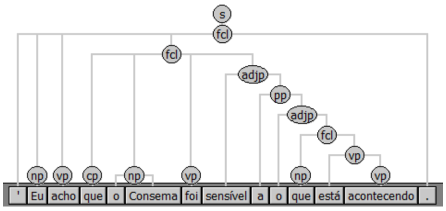

Fig. 1a. Example of insertion of first person ommited subject in the sentence “*Eu acho que o Consema foi sensível ao que está acontecendo*.”.

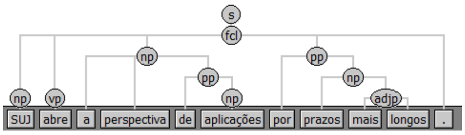

Fig. 1b. Example of insertion of third person omitted subject in the sentence “*Abre a perspectiva de aplicações por prazos mais longos*.”.

The analysis of the results of this first rule showed us the need to insert other restrictions to improve the precision of the insertion. After implementing each restriction, we made a new round of tests to detect unseen problems, and used the results of such tests to refine the rules. We created three sets of rules:

1) to recognize acceptable intervening material between the target verb and the NP at its left side (for example, some adverbs);
2) to recognize target verbs which are impersonal verbs (weather verbs) or verbs used in impersonal constructions, as “ter”, “haver” and “existir” (*to have*, *to have* and *to exist*); and
3) to recognize subject inversion in direct speech constructions, e.g. “Oi, disse ele” (*Hello, said he*).

Specifically, our system does not insert an artificial subject when:

1. A VP contains a unique verb and such verb is an infinitive, gerund or past participle.
2. A VP contains a unique verb and such verb is one of the following impersonal verbs: “haver” (*to have*), “existir” (*to exist*), “chover” (*to rain*), “nevar” (*to snow*) or “ventar” (*to blow*).

Our system inserts artificial subjects when:

1. A VP contains an occurrence of verb “ser” (*to be*) and this verb is starting a sentence or it is an auxiliary verb. We did not succeed in dealing with the verb “ser” when it is used as a copula verb, i.e., it links a subject to a predicative, because there are too many syntactic alternances allowed. The subject corresponds to a topic and the predicative corresponds to a comment on the topic (NP VP NP), but it is difficult to identify which is the topic and which is the comment;
2. There is no syntactic neighbor at the left of a VP (VP starts the sentence). This role does not disaffirm rule 1.

3. There is no NP immediately before a PP between commas, and immediately preceding a VP.
4. A VP immediately follows punctuation; this punctuation does not match with the rule 3 and there is no NP immediately before punctuation.
5. There is no NP or ADVP preceding a VP.
6. There is an ADVP immediately before a VP and there is no NP immediately preceding this ADVP.
7. There is an oblique pronoun[2] "me", "te", "nos", "vos", "lhe" and "lhes" directly preceding a VP and there is no NP immediately before this pronoun.
8. An ACC immediately precedes a VP and there is no NP directly before this direct object. A variation is when there is an ADVP immediately before this ACC but there is no NP directly before this ADVP.

In Portuguese, it is also allowed to change the canonical order Subject + Verb to Verb + Subject. The subject inversion represents an additional difficulty for subject recognition. Whenever there is no NP at the left side of the verb, an NP at the right side may be either a subject or a direct complement. In fact, some sentences are ambiguous due to this possibility:

(6) Antes do nascer do sol, acordou o trabalhador. (Before the sunrise, the worker woke up. OR Before sunrise, [he/she/it] woke up the worker.)

Verbs of reported speech often present subject inversion. These verbs are known as utterance verbs. We created two rules to deal with this case of subject inversion:

1. As a simplification, we considered that an utterance verb should be immediately preceded by punctuation and directly succeeded by an NP or by a sequence of ADVP and NP. The punctuation before the verb delimits the end of a speech and the starting of a description of who has spoken.
2. As a second condition, we consider a verb as an utterance verb if it is in the following list of verbs[3]: acentuar, acrescentar, afirmar, alegar, argumentar, berrar, bradar, clamar, contar, dizer, exclamar, falar, gritar, indagar, informar, perguntar, responder and sustentar.

Finally, there are variations regarding the position in which the null element shall be inserted to represent a subject in a clause. Usually the null element is inserted immediately before the verb. However, depending on the syntactic structure of the sentence, sometimes it is needed to shift it some positions to the left or to insert it at the right of the verb. The following list shows the conditions that must be satisfied to choose the right place to put the subject.

1. We use the tag "ks" (subordinating conjunction) inserted by the parser PALAVRAS to identify which ADVPs must be positioned immediately

[2] We don't deal with the reflexive pronoun *se* because such particle presents high ambiguity in Portuguese.

[3] In English, the utterance verbs list is: to accentuate, to add, to assert, to claim, to argue, to scream, to shout, to acclaim, to count, to say, to exclaim, to speak, to scream, to talk, to shout, to inquire, to inform, to ask, to answer and to maintain.

before a verb and which must be positioned before the subject. However, as the parser sometimes misclassifies some of them, we explicitly set a list of frequent adverbs found in the corpus that must be kept together to the verb: “não” (*no*), “nunca” (*never*), “sempre” (*always*), “quase” (*almost*), “só” (*just*), “já” (*already*) and “também” (*also*).
Whenever we found an ADVP in this list or an ADVP labeled as “ks”, the artificial subject must be shifted one position left, preceding the adverb. Otherwise, the artificial subject must be inserted between the adverb and the verb.

2. In the occurrence of a reflexive pronoun immediately before a VP and not preceded by an NP, the insertion must be done one position left, just before the reflexive pronoun.
3. In the occurrence of a reflexive pronoun before a VP and preceded by one of the adverbs listed in the condition 1 above, the artificial subject must be shifted two positions left. Otherwise, the inserted element must be shifted one position left.
4. In the occurrence of an utterance verb, the artificial subject element must be inserted immediately after the VP.

As most of such rules work with morphosyntactic and syntactic tags, they depend on parsing accuracy to produce the intended results.

To evaluate the accuracy of the developed rules, a gold standard set was created containing 200 sentences randomly selected from the corpus, manually annotated with omitted subjects. The gold standard is composed, therefore, of positive and negative evidence of null elements inserted to represent missing subjects[4].

3 Results

The analysis performed over the automatic insertion of missing subjects on the sample of 200 sentences showed that 157 sentences (78.5%) matched the gold standard after applying our rules and 43 sentences (21.5%) did not. Table 1 shows the precision of insertion of *Eu*/*Nós* and SUJ insertion separately.

Table 1. Results of automatic insertion of omitted subjects.

	Insertions Gold Standard	**Automatic Insertions**	**Hits**	**Wrong insertions**	**Missing Insertions**	**Accuracy**
SUJ	106	120	67	43	10	55.8%
Eu/Nós	26	27	24	2	1	88.8%
Total	132	147	91	45	11	61,9%

[4] The gold standard corpus and the set of sentences with artificial subjects automatically inserted are available at: http://www.nilc.icmc.usp.br/semanticnlp/artificialsubjects

We manually analyzed the causes of errors and realized that part of them was due to parsing errors and part were cases not covered by our rules, as shown in Table 2.

Table 2. Causes of inaccuracy.

	Errors	Due to parsing errors	Due to Insufficient Rules
SUJ	53	29	24
Eu/Nós	3	1	2

Fig. 2 shows an example of error motivated by parsing error. The word "outros" (others) may be an adjective or a noun. In the sentence, "outros" is a noun, but the parser analyzed it as an adjective, assigning the label "adjp" instead of "np".

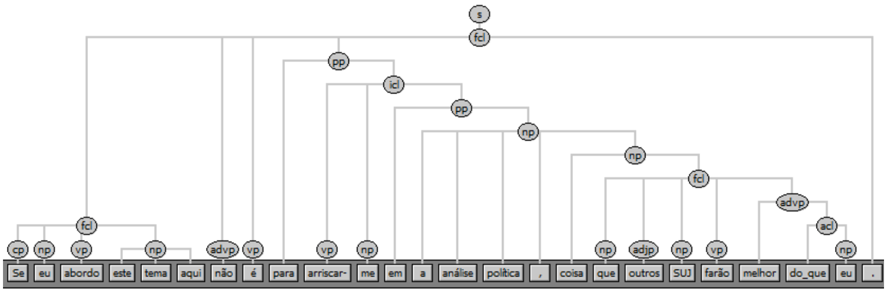

Fig. 2. Example of a POS tag error.

In Fig. 3, all the sequence "qualquer americano medianamente equipado" (any averegily equipped American) should have been analysed as an np.

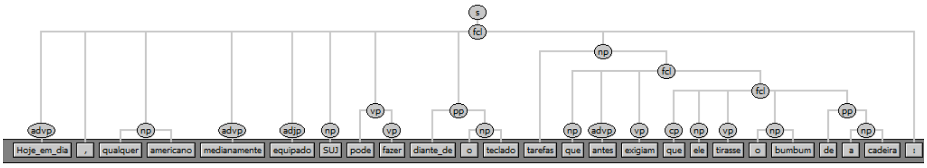

Fig. 3. Example of an NP delimitation error.

4 Conclusions

Aiming to benefit SRL tasks, we developed a set of rules to automatically insert a null element to represent missing subjects in Portuguese. The results showed us this is a task too complex to be accomplished only using rules. The results for first person (*Eu*/*Nós*) is considerably better than for third persons and for this reason we decided to apply the rules only for first persons meanwhile.

A problem that remains is that some verb inflections are ambiguous and in these cases they do not constitute a secure feature to inform the rules. For example, first and third singular persons in Imperfect Past and in Conditional tenses are equal and; for some verbs, first and third singular persons in Future of Subjunctive and Infinitive are equal, e.g. "se eu achar", "se ele achar" (*if I find*, *if he finds*). Such ambiguity is a problem even for taggers and parsers.

We envisage two possibilities for future work. The first of them is to apply the rules as they are and manually correct the resulting corpus, in order to constitute a training corpus for machine learning. The other is to improve our rules in order to recognize patterns besides those that have already been specified.

Whatever the option we take, we can "eliminate" wrong insertions after the manual SRL annotation, as only correctly inserted null elements will be assigned a role label.

Acknowledgements. Part of the results presented in this paper were obtained through research activity in the project titled "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.

References

[1] Palmer, M., Gildea D., and Xue N.: Semantic role labeling. Synthesis Lectures on Human Language Technologies 3 (2010) 1–103.

[2] Palmer, M., Gildea D., and Kingsbury P.: The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31 (2005) 71–106.

[3] Bonial, C., Hwang, J., Bonn, J., Conger, K., Babko-Malaya, O., and Palmer, M.: English PropBank Annotation Guidelines. Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder (2012)

[4] Marcus, M. P., Marcinkiewicz M. A., and Santorini B.: Building a large annotated corpus of English: The Penn Treebank. Computational linguistics 19 (1993) 313–330.

[5] Santos, D., Bick, E., and Afonso, S.: Floresta sintá(c)tica: apresentação e história do projecto. Meeting Um passeio pela Floresta Sintá(c)tica (2007)

[6] Duran, M. S. and Aluísio S. M.: Propbank-Br: a Brazilian treebank annotated with semantic role labels. In LREC, Istanbul, Turkey (2012) 1862–1867.

[7] Bick, E.: THE PARSING SYSTEM "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, Volume 202. Aarhus University Press Aarhus (2000)

[8] Toutanova, K., Haghighi A., and Manning C. D.: A global joint model for semantic role labeling. Computational Linguistics 34 (2008) 161–191