Integrating support verb constructions into a parser

Amanda Rassi1,2, Jorge Baptista2,3, Nuno Mamede3, Oto Vale1,4

1Centro de Ciências Humanas – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – São Carlos – SP – Brazil – 13.565-905

2Faculdade de Ciências Humanas e Sociais – Universidade do Algarve (UAlg)
Campus Gambelas – Faro – Portugal – 8005-139

3Instituto de Engenharia de Sistemas e Computadores (INESC-ID Lisboa/L2F)
Lisboa – Portugal – 1000-029

4Université Catholique de Louvain (UCL)
Louvain-la-Neuve – Bélgica – B-1348

{aprassi,jbaptis}@ualg.pt, nuno.mamede@inesc-id.pt, otovale@ufscar.br

Abstract. This paper describes the process of integrating into a rule-based parser a set of approximately 1,000 nominal predicates forming support verb constructions (SVC) with the verb ‘give’ in Brazilian Portuguese. The system was evaluated on a sample of 580 sentences containing verb-noun combinations candidates to SVC, manually and independently annotated. Best results yield 85% precision, 79% recall, 76% accuracy and 82% F-measure.

1. Introduction

Support verb constructions (SVC) [Gross 1981] pose a challenge to Natural Language Processing (NLP) because they are superficially alike verbal predicates (compare John gave a kiss to Eve vs. John gave a book to Eve), but semantically they present a special configuration since the predicate is actually expressed by the predicative noun (‘kiss’) instead of the verb. This entails a set of SVC-specific properties that distinguish them from ordinary (distributional) verbal constructions (e.g. John gave my *kiss/book to Eve, John’s kiss/*book to Eve). From the perspective of identifying the meaning units of texts, SVC are a combination of verb and noun corresponding to a single semantic unit, although syntactically analysable. In this sense, they do not form a compound word, but a special type of collocation [Mel’cuk 1997], where the verb functions as an auxiliary of the noun, conveying the grammatical values of person-number and tense.

This paper briefly presents the formalization of the linguistic properties of SVC constructions with verb dar ‘give’ in Portuguese, under the Lexicon-Grammar (LG) framework [Gross 1981]; it sketches the integration of the data into a rule-based parser – Xerox Incremental Parser (XIP) [Mokhtar et al. 2002], through an automatic process for generating directly from a Lexicon-Grammar matrix the dependency extraction rules, which are then integrated into a fully-fledged NLP system built for Portuguese – the STRING [Mamede et al. 2012]; and, finally, it evaluates the performance of the system, by comparing it with a golden standard of a manually annotated corpus of SVC candidate sentences [Rassi et al. 2015].

1For more information on XIP and STRING: string.l2f.inesc-id.pt
2. Related work

Most studies on SVC aim at the detection, identification, and extraction from corpora, based only in linguistic information, such as the degree of compositionality of the SVC; or only in statistical information, such as association measures on co-occurrence distribution; or, else, hybrid approaches using both linguistic and statistical information [Calzolari et al. 2002], [Stevenson et al. 2004], [Tan et al. 2006] [Fazly and Stevenson 2007], [de Cruys and Moirón 2007], [Wang and Ikeda 2008], [Diab and Bhutada 2009], [Tu and Roth 2011], [Gurrutxaga and Alegria 2011]. Hybrid methods for identification of MWE are, nowadays, the most commonly used.

In order to parse SVC in texts, two main approaches can be adopted: (i) considering SVC as a whole block, whose constituents are relatively fixed and treated as a subtype of multiword expressions (MWE), such as compound words and many types of idioms (see [Calzolari et al. 2002], [Sag et al. 2002], [Diab and Bhutada 2009], among others); and (ii) an approach that, in spite of some specific syntactic-semantic properties, considers that SVC do have syntactic structure and follow the same constituency rules as the general grammar of the language, systematically admitting several lexically determined syntactic transformations (alternative wordings). To the best of our knowledge, no study reports any attempt to integrated SVC into a parser, under this second perspective.

Portuguese SVC constructions have been intensively studied since the late 80's, and extensive lexicon-grammars of SVC (over 10 thousand predicative nouns) for both the European (EP) and the Brazilian (BP) variety of Portuguese have been produced, including the SVC with support verb dar ‘give’ [Baptista 1997, Rassi et al. 2014b]. For lack of space, see [Rassi et al. 2014a] for a recent overview.

3. Integration of SVC in XIP parser

Firstly, about 1,000 SVC with the verb dar ‘give’ in EP and BP were formalized into a Lexicon-Grammar matrix, where the lines correspond to the lexical entries (predicative nouns) and the columns indicate linguistic properties. In this matrix, the linguistic properties encoded are: (i) formal properties, such as the number of arguments, sub-clausal arguments, prepositions introducing the complements and type of determinant of the predicative noun; (ii) distributional properties, such as the semantic type of arguments (human or non-human nouns, locative complements, etc.) and the arguments’ semantic roles (<agent>, <patient>, etc.); the main support verbs specific to each predicative noun are also explicitly encoded; and (iii) transformational properties, such as Passive, Symmetry, Conversion (see below). Secondly, the original XIP lexicon was enriched with the predicative nouns built with suffix -ada ‘-ed’, which are quite a productive derivational device in Portuguese (particularly in BP), e.g. dar uma cadeirada ‘give an chair-ed’, dar uma mãozada ‘give a hand-ed’, dar uma esquentada ‘give a warm-ed’, etc.

This dataset was integrated into STRING [Mamede et al. 2012], a fully-fledged, hybrid (statistical and rule-based) Natural Language Processing chain for Portuguese. It performs all the basic NLP tasks (tokenization, sentence splitting, part-of-speech (POS) tagging, POS-disambiguation, chunking and deep parsing). The STRING system uses XIP – Xerox Incremental Parser [Mokhtar et al. 2002] for its parsing module, which is rule-based and uses finite-state technology. XIP segments sentences into chunks (NP, PP, VP, etc.) and extracts dependency relations between the chunks’ heads: SUBJECT, CDIR
(Direct Object), MODifier, etc. In this framework, SVC parsing consists in the automatic generation, directly from the LG matrix, of dependency rules in the XIP format, which allow the parser to extract the dependency holding between the support verb and the predicative noun. This dependency is called SUPPORT. A set of programs were built for: (i) validating the linguistic data manually inputed into the LG matrix; and, then, (ii) automatically converting it into XIP dependency extraction rules.

The general strategy towards the implementation of SVC in STRING is sketched as follows: First, all XIP’s normal parsing procedures are applied and the basic syntactic dependencies are extracted, specially SUBJect, CDIR (direct object) and MODifier (for PPs), that is, the dependencies holding between the verb and its arguments, as for any ordinary distributional verb. Then, the special set rules for SVC identification operate upon the parse that has been produced so far in order to extract the SUPPORT dependency. Basically, these rules match, for each support verb and predicative noun combination, the dependencies already extracted (e.g., the CDIR between deu ‘gave’ and abraço ‘hug’ in Rui deu um abraço no João ‘Rui gave a hug to João’). Rules also consider Passive, Relative, and other structures transformationally derived from the base sentence (e.g. O abraço que foi dado pelo Rui no João ‘The hug that was given by Rui to João’). Once this dependency has been extracted, then, the following parsing stages can take it into account, for example, in the assignment of semantic roles.

The dependency rules consider the distinction between two main cases: (i) elementary sentences, whose dependency is called SUPPORT; these include both the standard (or active-like) constructions (Rui deu um beijo na Eva ‘Rui gave a kiss to Eva’) and the converse (or passive-like) constructions [Gross 1989, Baptista 1997, Rassi et al. 2014b] (Eva ganhou um beijo do Rui ‘Ana got a kiss from Rui’); and (ii) causative constructions [Gross 1981, p.23], which involve a causative operator verb (VopC) and a predicative noun, whose dependency is called VOP-CAUSE, and are not considered as elementary sentences (Algo deu raiva na Ana ‘Something gave anger in/on Ana’; cp. Ana tem raiva ‘Ana has hanger’). The SUPPORT dependency receives two features, depending on whether it corresponds the standard and converse constructions. As the causative constructions occurred only 3 times in the 580 sentences of the reference corpus, they ignored in this paper.

4. Evaluation

In order to evaluate the overall performance of the system, a reference corpus containing 2,646 sentences with SVC in (Brazilian) Portuguese was produced [Rassi et al. 2015], constituting a golden standard for SVC processing. These constructions have been manually and independently annotated by 5 annotators, all Portuguese native speakers, professional linguists, and experts in SVC. The average agreement between annotators was 80.8% and Cohen’s Kappa was 0.604, which can be considered in the range between “moderate” and “substantial”. The reference subcorpus for this paper consists of a sample containing 580 sentences with the verb dar ‘give’ and 8 stylistic or aspectual variants.

The evaluation of the new module of the Portuguese grammar for XIP parser in STRING was carried out in two stages: (i) a preliminary evaluation took as reference the

---

2The reference corpus is available at https://sites.google.com/site/amandaprassi/recursos
580 manually annotated sentences, considering the majority agreement among the annotators; (ii) the second evaluation was carried out with the same sample of sentences but after error analysis was performed. This analysis made possible to spot some inconsistencies in the annotation as well as some few errors in the Lexicon-Grammar. For example, some diminutive forms of -ida or -ada ‘-ed’ ending nouns (e.g. fugidinha ‘quick run-ed away’, saidinha ‘quick exit-ed’ and arrumadinha ‘little tidy-ed up’) had not been adequately analyzed by STRING and hence were not associated with their lemmas (fugida ‘run-ed away’, saida ‘exit-ed’ e arrumada ‘tidy-ed up’), which enabled us to improve the lexicon. On the other hand, the inconsistent annotation of some SVC as frozen expressions (or idioms) led us to refine the criteria for a more precise distinction between these two categories. For lack of space a fully detailed error analysis can not be presented here. The new (corrected) reference was then compared with STRING’s new results in a second evaluation run. Results from both runs are compared in Table 1, using standard evaluation metrics (Precision, Recall, Accuracy and F-Measure). In this table, TP=true positives, FP=false-positives, FN=false-negatives and TN=true-negatives.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>First evaluation</td>
<td>350</td>
<td>91</td>
<td>114</td>
<td>25</td>
<td>79%</td>
<td>75%</td>
<td>65%</td>
<td>77%</td>
</tr>
<tr>
<td>Second evaluation</td>
<td>325</td>
<td>56</td>
<td>84</td>
<td>115</td>
<td>85%</td>
<td>79%</td>
<td>76%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Tabela 1. First and second evaluations of STRING’s performance

Comparing the first and second evaluation runs, one can see that the system’s overall performance shows a small improvement. The most important change is the number of true-negative cases (TN), due mostly to a more precise definition and reclassification of frozen sentences (e.g. dar nome ‘give name to’, dar a volta por cima ‘turn things around’). Some errors derive from previous modules of the processing chain, for example errors in POS-tagging and disambiguation, the chunking or in the syntactic parsing. Other errors came from the ambiguity between standard and converse constructions, especially when involving the verb ter ‘have’[Rassi et al. 2014a].

5. Final remarks and future work

This paper reported preliminary experiments in integrating into the STRING NLP system, more precisely in the rule-based parser XIP, a set of about 1,000 SVC, involving the elementary support verb dar ‘give’ and its variants, from European and Brazilian Portuguese. The results are promising and suggest that a rule-based approach is suitable for the analysis of support verb constructions. Furthermore, the methodology presented in this paper proved that it is possible to parse the (sometimes complex) syntactic structure that SVC present, so as to be able to use this for further NLP processing (e.g. semantic role labeling, anaphora resolution).

In the near future, we intend to integrate into STRING the already available Lexicon-Grammar matrices of the remaining predicative nouns, with the support verbs estar Prep ‘be Prep’, ser de ‘be of’, fazer ‘make/do’ and ter “have”, and evaluate the system’s performance, by using the full corpus of 2,646 manually annotated sentences.

Acknowledgments

This work was partially supported by national funds through Portuguese Fundação para a Ciência e a Tecnologia (ref. UID/CEC/50021/2013), and Brazilian CAPES (ref. BEX 12751/13-8).
Referências


