

Automatic Semantic-Role Annotation for Portuguese

Eckhard Bick

Institute of Language and Communication, University of Southern Denmark
eckhard.bick@mail.dk, Rugbjergrvej 98, DK-8260 Viby J

Abstract. *This paper describes an automatic annotation system for semantic roles, implemented as an independent add-on module for syntactically parsed text. The formalism uses hand-written Constraint Grammar (CG) rules to map and disambiguate 40 different semantic roles. As a proof-of-concept, an evaluation on treebank data was performed, where the system achieved an average F-score of 88.6.*

1. Introduction

Assigning semantic roles to the arguments of a verb (or proposition in general) is an obvious way of adding deep, semantic structure to the syntactic analysis of a sentence, allowing different surface-syntactic functions, like subject and object, to "code" for one and the same semantic role, e.g. agent or patient depending on the arguments-slots of the governing verb and whether the syntactic structure is active, passive, reflexive or attributive. The idea of semantic roles has a long linguistic history, originally inspired by the concept of case roles (Fillmore 1968), later to be termed thematic or theta roles in Government & Binding theory (e.g. Jackendoff 1982). At the highest level of abstraction, universal semantic functors are postulated (Foley & van Valin 1984, Dowty 1987), obeying general laws of argument precedence based on features like control and animateness.

Within automatic annotation of running Portuguese text, robust results can be achieved for both morphology, syntactic function and even structural relations such as dependency relations. In the semantic area, notable progress has been made in the field of named entity classification (cp. the HAREM competition, Santos & Cardoso 2007), but no automatic systems have yet been published that would assign semantic functions to *all* types of arguments/constituents in a systematic way. In this paper, we will present and evaluate a rule-based method to add semantic roles to Portuguese sentence analyses at the treebank level. Input to the semantic role annotator is provided by the PALAVRAS parser (Bick 2000), using recent extensions regarding dependency and semantic prototype annotation (Bick 2006).

2. Semantic roles

Semantic roles are different from semantic prototypes in that the latter bundle stable lexical features of a given lexeme, while the former are functional and can only be assigned *in context*. Semantic prototype ambiguity does exist, of course, as words may be polysemic, and an automatic parser will use context to resolve such ambiguity, but even one and the same semantic prototype, with one and the same syntactic function, say subject, may fulfill a number of different semantic roles. Thus, the *civitas* prototype (towns, countries etc.) may fulfill the thematic roles of location, origin or destination of movement, but also non-place roles like that of human agent or patient. Here, rather

than hypothesise different senses or lexical types for these cases, a role annotation level can help to build a bridge between syntax and semantics. In our approach, methodology closely matches descriptive issues, by letting the argument slots of predicators (mostly verbs) project a certain semantic interpretation (the roles) onto the slot-fillers (mostly nouns or noun phrases).

We use a set of about 35 semantic roles (e.g. §AG agent, §PAT patient, §TH theme, §LOC location, §ORI origin etc.) covering the major categories of the tectogrammatical annotation layer of the Prague Dependency Treebank (Hajicova et al. 2000), as well as those of the Spanish 3LB-LEX/3LB-SEM project (Taulé et al. 2005). In contrast to the latter, no separate level of argument hierarchy (ARG0, ARG1, ARG2, ARG-M) is used, mainly because the combination of syntactic function and semantic role tags allows the later addition of ARG-attributes on demand, without loss of information. The notational convention is that semantic role tags are assigned to tokens, CG-style, alongside syntactic and dependency tags. Complex (non-terminal) constituents are marked on the semantic head, i.e. the noun in an np, or the argument of a preposition. The semantic role of a subclause is marked on its main verb.

3. The Grammar and lexicon of the system

In order to contextually map semantic roles onto tokens, we have developed a Constraint Grammar of about 500 mapping rules and a small number of disambiguation rules. In the example below, the purpose role (finality, §FIN) onto arguments of the preposition '*para*', if the argument (@P<) in question is a deverbal noun and not ambiguously carrying a human or place prototype tag, in which case the destination or benefactive roles would have been more appropriate.

MAP (§FIN) TARGET @P< (0 N-DEVERBAL)

(NOT 0 N/PROP-LOC OR N/PROP-HUM) (*-1 PRP LINK 0 PRP-PARA) ;

Note that the NOT condition for semantic prototypes here is a safety measure, since PALAVRAS should, in a best case scenario, already have disambiguated semantic prototypes in the input.

A pivotal issue in the assignment of semantic roles is, of course, the lexicon serving the annotation stage as such. On the one hand, role-annotated text will allow to construct an inventory of Portuguese verb-sense argument frames in the style of the PropBank (Palmer et al. 2005), on the other hand exactly this kind of lexical information is necessary to assign semantic roles in the first place. In the absence of a hand-annotated gold resource, we opted for a boot-strapping solution, where we did not manually construct full verb frames, but exploited ordinary (syntactic) valency information - such as <vdt> (ditransitive verb), <ve> (ergative verb), <por^vp> (prepositional valency with *por*), <vk> (copula-verb), <vta> (transobjective adverbial valency) etc. - to construct so-called *CG sets* of verb lexemes that typically *allow* (not *demand*) a given role as subject, object, complement or prepositional argument. All in all, about 80 sets with 1100 verb lexemes were defined in this way - moving part of the lexical information into the grammar.

One such set (VP-EM-TH) contains verbs allowing the theme role (§TH) with the preposition '*em*' - *crer em*, *votar em*, *enganhar-se em*, *gastar ac. em*. The set is then used in rules targeting prepositional objects (@PIV) with a VP-EM-TH class main verb.

4. A new CG formalism for dependency references

It is an obvious efficiency problem for long-distance (verb-argument) rules that head-daughter dependency relations have to be expressed in indirect - in fact topological - terms, using lots of unbounded (*1) and LINKed contexts with safety BARRIERS to locate the head of a dependency arc somewhere far away in the sentence. Furthermore, all head searches have to be conducted both left and right to cover arguments both left and right of their head, basically doubling the number of certain rules. CG's traditional topological coinage, a strength robust disambiguation, becomes, alas, a disadvantage where exact syntactic relations are, in fact, already known and *could* be exploited. Our research group has tried to remedy this problem by designing a new CG rule compiler allowing direct context reference to dependency links. At present, three types are used: (p) parent, (c) child and (s) sibling, all allowing the usual combination with NOT and C (safe) operators. Rule (b) exploits the new formalism to assign the theme role (§TH) to direct objects (@ACC), if it has a dative/indirect object (@DAT) among its siblings, while rule (c) selects an agent role (§AG) for a subject with a speech-verb as parent.

(b) MAP (§TH) TARGET @ACC (s @DAT) ;

(c) SELECT (§AG) (0 @SUBJ) (p V-SPEAK LINK NOT 0 PAS) ;

The current role tagger is hybrid, running the original rules *after* a preprocessing stage using new, dependency-based rules. Currently, this first stage is used to handle valency instantiation, propagation of function labels from first to following conjuncts, and propagation of semantic prototype labels from nouns to pronouns and relative clauses.

5. Evaluation

The semantic role annotator was evaluated on input from the European Portuguese part of the *Floresta Sintá(c)tica* treebank (Afonso et al. 2002). For the test run, a section of 2.500 running words was used. Prior to semantic role annotation, the following steps were performed, using different PALAVRAS modules: (1) dependency conversion, (2) semantic prototype tagging, (3) named entity classification. The advantage of being able to rely on manually revised function labels and phrase structure (as input to 1-2-3) was deemed to outweigh the problem of a certain amount of tokenization incompatibility between the treebank and live PALAVRAS runs.

All in all, the annotator module assigned 884 semantic role labels, of which 84 were wrong. In 38 cases, the label was missing altogether, corresponding to a recall of 86.8 %, a precision of 90.5 % and an F-score of 88.6. In a break-down of categories, the classical case roles (AG, PAT, TH and space roles LOCation, ORIgin) were lumped together in the middle of the field, while adverbial roles (FINality, CAUse, dummy ADVerbial) perform below average, possibly because of their unclear distribution between different syntactic levels (clause types, verb-arguments and np-internal).

<i>label</i>	REFL	PRED	ATR	AG	PAT	LOC	ORI	all	TH	FIN	CAU	BEN	DES	ADV
F	97.3	96.8	94.5	90.0	89.0	88.9	88.9	88.6	84.0	81.7	78.8	76.2	75.9	72.2

Figure 1. Semantic role tagging performance

6. Outlook

Given the tight integration of semantic role annotation with semantic prototype and name tagging, a synergistic trade-off can be expected from improvements in any one of these three areas, and all should be given due consideration in the future. We also hope that hitherto "CG-unexplored" annotation levels, targeted by our new rule compiler formalism, such as anaphora links, co-referent resolution and the integration of statistical data in the form of rule or context thresholds, will ultimately have a positive influence on semantic role annotation.

As a medium term goal, semantic roles could help to resolve the annotation conflict in NER between lexical/stable name classes and functional/contextual classes, by achieving the latter through a combination of the former with semantic roles (e.g. turning a "country" <civ> name into an administrative organisation <admin> by assigning it agent function (<civ> + §AG = <admin>). Long term, we suggest to annotate the Portuguese *Floresta Sintá(c)tica* treebank with semantic roles, in a combination of automatical meta-tagging and human revision. Such a resource would ultimately support the construction of a Portuguese PropBank or FrameNet.

References

- Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002). Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC'2002*. pp. 1698-1703, Paris: ELRA
- Bick, Eckhard (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press
- Bick, Eckhard (2005), Turning Constraint Grammar Data into Running Dependency Treebanks, In: Civit, M. & Kübler, S. & Martí, M. A. (red.), *Proceedings of TLT05, Barcelona*, pp.19-27
- Bick, Eckhard (2006), "Noun Sense Tagging: Semantic Prototype Annotation of a Portuguese Treebank". In: Hajic, J. & Nivre, J. (red.), *Proceedings of TLT 2006, Prague*, pp.127-138
- Dowty, D. (1987). Thematic proto roles and argument selection, *Language* **67**: 547-619.
- Fillmore, C. (1968). The case for case, in E. Bach & R. Harms (eds), *Universals in linguistic theory*, Holt, Rinehart and Winston, New York.
- Foley, W. & van Valin, R. (1984). *Functional syntax and Universal Grammar*, CUP, Cambridge.
- Hajicova, E. & J. Panevova & P. Sgall (2000). A Manual for Tectogrammatic Tagging of the PDT. UFAL/CKL Technical Report *TR-2000-09*, Charles University, Czech Republic
- Jackendoff, R. (1972). *Semantic interpretation in generative grammar*, The MIT Press, Cambridge, Ma.
- Karlsson, F., Voutilainen, A., Heikkilä, J. & Antilla, A. (1995). *Constraint Grammar: A language-independent system for parsing unrestricted text*, Mouton de Gruyter, Berlin.
- Palmer M, Kingsbury P, Gildea D. (2005) "The Proposition Bank: An Annotated Corpus of Semantic Roles". *Computational Linguistics* **31** (1): 71-106
- Santos, D. & Cardoso, N. (2007). HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro. *Linguatca*
- Taulé, M. et al. (2005). "Mapping Syntactic Functions into Semantic Roles". In: *Proceedings of TLT2005*: 185-194