Propbank-Br: a Brazilian treebank annotated with semantic role labels

Magali Sanches Duran, Sandra Maria Aluísio
Núcleo Interinstitucional de Linguística Computacional
ICMC – University of São Paulo - São Carlos – SP – Brazil
magali.duran@uol.com.br, sandra@icmc.usp.br

Abstract:

This paper reports the annotation of a Brazilian Portuguese Treebank with semantic role labels following Propbank guidelines. A different language and a different parser output impact the task and require some decisions on how to annotate the corpus. Therefore, a new annotation guide – called Propbank-Br - has been generated to deal with specific language phenomena and parser problems. In this phase of the project, the corpus was annotated by a unique linguist. The annotation task reported here is inserted in a larger projet for the Brazilian Portuguese language. This project aims to build Brazilian verbs frames files and a broader and distributed annotation of semantic role labels in Brazilian Portuguese, allowing inter-annotator agreement measures. The corpus, available in web, is already being used to build a semantic tagger for Portuguese language.

Keywords: Semantic Role Labeling, Brazilian Portuguese, Propbank Guidelines.

1. Introduction

Natural Language Processing (NLP) tackles several tasks using hand-annotated training corpora. The more annotated a corpus is, the more features for statistical learning it offers. A layer of semantic role labels is a desired improvement in a treebank, since it provides input for syntactic parser refinements. However, the major relevance of such kind of annotation is the possibility of building classifiers for automatic identification of role labels, a powerful way to develop information extraction and question answering systems (Shen & Lapata, 2007; Christensen et al., 2011). Moreover, semantic role labels can be used as a sort of “interlingua” to boost machine translation. Nevertheless, only languages that have already developed automatic semantic role labeling (SRL) may benefit from such technological improvements. Portuguese, in spite of being one of the five most spoken languages in the world, has not, until this moment, inspired huge efforts to achieve this goal. Our initiative is a first step to change this situation.

Our aim is to develop resources to enable a large-scale SRL annotation task. The development of the first resource, a layer with SRL annotation in a Treebank of Brazilian Portuguese, is reported here. The resulting corpus will be used as input for the construction of the next resource: a verb lexicon of Portuguese verbal predicates and their predicted argument structure. This lexicon, on its turn, will be used to guide annotators in a broader and distributed SRL annotation task in Portuguese. The corpus may be used, as well, to enable automatic pre-annotation in a larger corpus, requiring only human correction.

The approach we follow is the same of Propbank (Palmer, Gildea & Kingsbury, 2005; Palmer, Gildea & Xue, 2010) and taking this decision we benefit from Propbank´s available resources and open, at the same time, the opportunity for future mappings of Portuguese and English verbal predicates.

The remainder of this paper is organized as follows: in Section 2 we justify the option for Propbank approach instead of Framenet approach to SRL. In Section 3 we explain the methodological decisions we took and in Section 4 we briefly comment the choice of the annotation tool. Section 5 is dedicated to explain preprocessing procedures undertaken in the corpus. In Section 6 we discuss some issues on the assignment of Propbank role labels. Section 7 reports some problems related to the parser output. In Section 8 we report the occurrence of mismatches between syntactic and semantic segments. In Section 9 we discuss some language specific challenges faced during the annotation and in Section 10 we present the additional annotation provided in the corpus. Finally, we envisage further work in Section 11.

2. Propbank and Semantic Role Labeling

When Gildea and Jurafsky (2001) firstly addressed semantic role labeling (SRL) as an NLP task, they used Framenet (Baker, Fillmore & Lowe, 1998) as a training corpus. However, Framenet was not originally conceived to provide a training corpus for machine learning. Its set of semantic role labels, for example, is fine grained and poses some problems of data sparsity for statistical learning methods. Moreover, time, manner, place and other modifiers have different annotation depending on the frame they occur. Propbank initiative (Palmer, Gildea & Kingsbury, 2005), in contrast, took project decisions aiming to facilitate machine learning purposes, like the adoption of a coarse grained set of role labels for arguments, a unique and generic set of role labels for modifiers and, what is essential, annotation over the syntactic tree, as defended by Gildea and Palmer (2002). This project added a layer of semantic role labels in a subcorpus of PennTreebank (the financial subcorpus). Additionally, a verb lexicon with verb senses and rolesets have been built and is available for consultation[1.]

Propbank is a bank of propositions. The underlying idea of the term "proposition" is found in frame semantics proposed by Fillmore (1968). A "proposition" is on the basic structure of a sentence (Fillmore, 1968, p.44), and is a set of relationships between nouns and verbs, without tense, negation, aspect and modal modifiers. Arguments which belong to propositions are annotated by Propbank with numbered role labels (Arg0 to Arg5) and modifiers are annotated with specific ArgM (Argument Modifiers) role labels. Each verb occurrence in the corpus receives also a sense number, which corresponds to a roleset in the frame file of such verb. A frame file may present several rolesets, depending on how many senses the verb may assume. In the roleset, the numbered arguments are "translated" into verb specific role descriptions. Arg0 of the verb "sell", for example, is described as "seller". Thus, human annotators may easily identify the arguments and assign them the appropriate role label.

Recently there have been initiatives to make corpus annotation, following Propbank model, for other languages: Korean (Palmer et al, 2006), Chinese (Xue & Palmer, 2009), Arabic (Palmer et al, 2008) and Basque (Aldezabal et al. 2010). We report here the construction of a Brazilian Portuguese Propbank: Propbank-Br.

3. Methodological decisions

The annotation task, in Propbank, was preceded by the construction of frames files (Kingsbury, Palmer & Marcus, 2002) and by the creation of an Annotator's Guidelines. These two resources enabled them to distribute the task and to reduce inter-annotator disagreements. The Guidelines provide general information about the task and the frames files provide verb-specific examples to be used during the annotation task. To better ensure inter-annotator agreement, Propbank adopted double and blind annotation for each instance and every disagreement, automatically detected, was decided by an adjudicator.

Due to project restrictions, we could not reproduce the same experience of Propbank. This project was one year long and counted with one sole annotator, as it was a post-doctoral research. We were interested in designing Propbank-Br in relatively independent modules to facilitate the collaborative construction of this resource.

Once Propbank Guidelines and Propbank frames files are available for consultation, we decided to adopt a different approach: instead of firstly building frames files and Annotator´s Guidelines, we started Propbank-Br by annotating a corpus using English frames files and guidelines as model. Therefore, unlike Propbank, in this first phase we annotated only semantic role labels and not verb senses.

In this way, we experienced the difficulties of the task, identified language-specific aspects of SRL for Portuguese, and generated a corpus that will be used as base to build frames files. The experience also enabled us to customize Propbank Guidelines for Portuguese. Since we followed a different path, we have now chance to compare approaches and recommend new guidelines for other projects tackling languages with fewer resources.

4. Annotation Tool

When Propbank tools for corpus annotation and frames files edition, respectively Jubilee and Cornestone, (Choi, Bonial & Palmer, 2010a and 2010b) were made available, we had already tested some annotation tools (Duran, Amâncio & Aluísio, 2010) and decided to use SALTO (Burchardt et al., 2006) in our task. We had to customize the use of SALTO, as it was created for the German Framenet project. Jubilee, on its turn, will be very useful in the future, as it allows displaying the frame file of the verb being annotated in a given instance. As in this phase of our project we have not the frames files, such facility would be of no use. SALTO, on the other hand, offers facilities of creating sentence and word flags during the annotation task, which proved to be very useful to add extra annotation that enables clustering similar interesting cases for future analysis.

5. Preparing the corpus for annotation

The underlying motivation of our broader project was to provide a training corpus to build automatic taggers. For this purpose, like in Propbank, it is essential to add a new layer in a corpus syntactically annotated and manually revised. The corpus we chose is the Brazilian portion of Bosque, the manually revised subcorpus of Floresta Sintá(c)tica[2] (Afonso et al., 2002) parsed by PALAVRAS (Bick, 2000). Bosque has 9368 sentences and 4213 of them correspond to the Brazilian Portuguese portion (extracted from the newspaper Folha de São Paulo of 1994). Aiming to shorten manual effort, the corpus was gone through a pre-processing phase as we explain next.

Using SALTO, the first action to annotate a predicate-argument structure is to "invoke" a frame, pressing the right button and choosing a frame. However, Propbank, unlike Framenet, does not define frames and thus there is no frame choice to be made. We defined a unique pseudo frame, called "argumentos", to enable the "invoke frame" action, which is the starting point for the assignment of role labels to the arguments of the verb. Having no choice to be made, we decided to process automatically the step of identification of "argument takers" and subsequent "invoke frame" action. Within this paper, "argument taker" is the verb of the proposition, which has an argument structure and is focus of the annotation. To identify argument takers could have been very simple if we had defined every verb as an argument taker. However, we realized a great opportunity to shorten our annotation effort by electing only main verbs to "invoke" the frame "argumentos". In other words, we decided to disregard,

[1] http://verbs.colorado.edu/propbank/framesets-english/

[2] http://linguateca.pt

for the purpose of assigning semantic role labels, all the verbs that play an auxiliary role, including temporal, modal and aspectual verbs. These verbs are modifiers of the proposition, but do not belong to the argument structure and thus do not integrate the core of the proposition. In Portuguese, these verbs occur at left of the main verb in a verbal chain.

To meet this need, we made a study on auxiliary verbs and built a table which encompasses temporal, aspectual, modal and passive voice auxiliaries. To distinguish the auxiliary use from the full verb use, the table contains the auxiliary verb and the nominal form the auxiliated verb should present (infinitive, gerund or past participle). Using this table, we were able to recognize verbal chains and select only the last verb at right of the chain (which corresponds to the main verb) as argument taker and consequently the focus of the "invoke frame" action. Our previous table has been improved by results from Baptista, Mamede and Gomes (2010), who were working on this subject relating to European Portuguese.

This decision provided a shortcut for our task. Auxiliary verbs are very frequent in the corpus, but their semantic roles as modifiers rarely vary. Being highly predictable, they may be automatically annotated in a next step, that is, they do not require human annotation.

Propbank has an ArgM for auxiliaries (ArgM-AUX), used to annotate auxiliaries of tense and diathesis (interrogative, negative and passive constructions) and an ArgM for modals (ArgM-MOD). However, there is no label defined to annotate aspectual verbs. In Propbank-Br, we plan to adopt four extensions for ArgM-AUX to annotate these verbs:

- ArgM-AUX-Pas, for auxiliaries of passive voice (Portuguese does not require auxiliaries for interrogative and negative constructions)
- ArgM-AUX-Tmp, for temporal auxiliaries
- ArgM-AUX-Mod, for modal verbs
- ArgM-AUX-Asp for aspectual verbs

Electing only the main verbs, we attacked the verbs that present higher levels of polysemy and challenge SRL annotation task.

After that, we repeated the sentences as many times as the number of argument takers they had. In this way, each argument taker of the sentence constitutes a separate instance for annotation. This procedure (of repeating the sentences to create a different instance of annotation for each verb) follows Propbank guidelines. The previous 4213 sentences produced 6142 instances for annotation. with 1068 different argument takers.

6. Assigning role labels

Propbank guidelines are very useful to explain the ArgM´s assignment, because modifier arguments are not verb dependent and, almost always, are very logical. However, for argNs, the guidelines are not sufficient when the argument structure presents several arguments. ArgNs assignment follows some rules for Arg0 and Arg1, but is highly arbitrary for arguments from Arg2 to Arg5. For this, consulting English frames files was essential for our task: following the arbitrary decisions of Propbank we facilitate future mappings between Propbank-Br and Propbank. For example, if we had used only the Propbank Guidelines, we would have been able to infer that the "seller" is the Arg0 of "sell", as Arg0 is related to agent or cause role labels. In the same way, we would have been able to infer that the "good sold" is the Arg1, because Arg1 is related to patient or theme role labels. However, without consulting the frame file of "sell", we would never know that the price is Arg3, because there is no rule to determine ArgNs other than Arg0 and Arg1.

Even in Propbank we may observe the arbitrary character of ArgN's determination. For example, the only roleset of the verb "renegotiate" has the "agreement" as Arg1 and the "other party" as Arg2, whereas the verb "negotiate" has the "agreement" as Arg2 and the "other party" as Arg1. The same inversion may be observed in other not so obvious verbs of a same class, like assist/help and mean/symbolize.

Using English frame files to guide our annotation presented many advantanges and some disadvantages. A single verb in Portuguese may motivate several searches in the English frames files, as each verb sense in Portuguese may be conveyed by different verbs in English. A problem we faced is related to arguments not predicted in the near synonym roleset in English. For example, the verb "relinquish" is the better synonym for "renunciar", but in Portuguese it is frequent to find a beneficiary (somebody in favor of whom somebody relinquishes something) and this role is not defined for "relinquish" in Propbank. In this case, we assigned Arg2 for the beneficiary, an argument number that is not used in the roleset of "relinquish" in Propbank.

Every event of doubt about the role label to be assigned was registered to be used as example in the Guidelines of Propbank-Br. For example, an argument that refers to an event, as "no meu aniversário" (in my birthday), has simultaneously two possibilities of role label in many contexts in Portuguese: the time (the birthday date: ArgM-TMP) and the place (the birthday party: ArgM-LOC) in which the event takes place. In cases like this, the Guidelines have to define clearly which label the annotator must assign; otherwise they will motivate annotation disagreements.

7. Problems arising from parser output

In spite of being a Treebank, the Brazilian portion of corpus Bosque presents some sentence split and spelling errors, as well as parsing inadequacies that affect SRL. Instances that present these problems have been flagged as "Wrongsubcorpus", using SALTO sentence flags facility. Such sentences will be made available separately from the annotated corpus, as they must not to be used for machine learning purposes, but may be valuable for those interested in implementing parser improvements. The main problem found was the lack of a Noun Phrase (NP) at left of the annotated verb, as may be seen in Figure 1, where "um tribunal" is the Arg0, but lacks NP node. This may be due to the fact that our parser is a dependence parser that generates a constituent output.

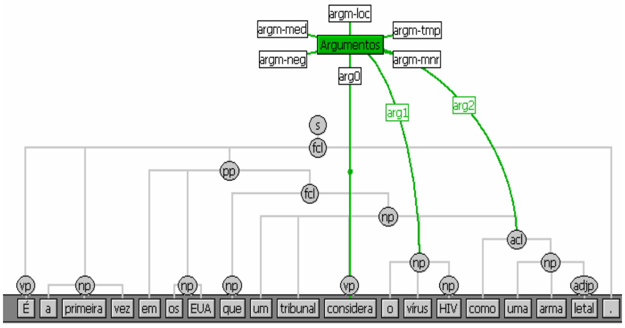

Figure 1. Sentence flagged as Wrongsubcorpus

Another challenge we faced is related to suppressed syntactic elements. The Penn Treebank, used by Propbank, provides "traces" of such elements, which are coindexed with other constituents. This is very important to deal with ellipsis. As our parser output does not have such traces, we have adopted some strategies to circumvent the lack of them. When we identified an empty category in an embedded clause, we assigned the role label to the proper constituent, despite the fact it is in the main clause.

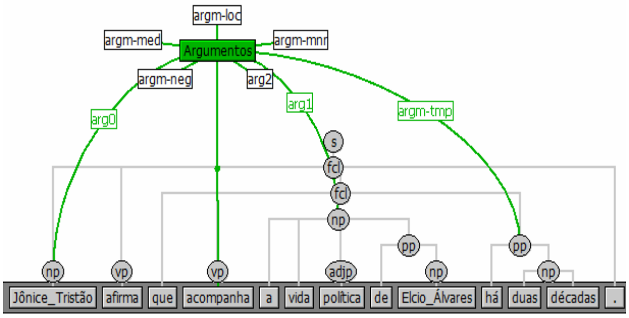

Figure 2. Annotation of an elliptic constituent

In Figure 2, the sentence, without ellipsis, would be:

> *Jônice Tristão afirma que [Jônice Tristão] acompanha a vida política de Élcio Álvares há duas décadas.*

These instances have been flagged as ELIPSE, so that machine learning approaches may identify them and opt for using them or not. They may also be used for improvements in the parser output, creating "traces" like in Penn Treebank.

In the same way, there is no correference resolution in our corpus output and we chose to assign the role label to the correfering constituents. In Figure 3, the Arg0 is assigned to the pronoun "que", which correfers to "Tsunezaemon" and the Arg1 is assigned to the adverbial "onde", which correfers to "Peru". These cases have been flagged as CORREF, keeping track for future work on correference resolution.

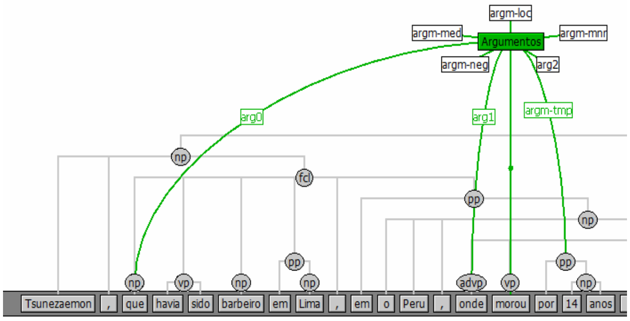

Figure 3. Annotation of correfering constituents

8. Mismatch between syntactic segments and predicate's arguments

Almost always, the syntactic segmented constituents match the required segmentation for predicate's arguments and then the assignment of the semantic role label is one-to-one. However, we found two different occurrences: 1) one syntactic constituent containing more than one predicate's argument and 2) one predicate's argument composed of two or more syntactic constituents.

In the case 1, we assigned a single semantic role label pointing to all syntactic constituents, as seen in Figure 4.

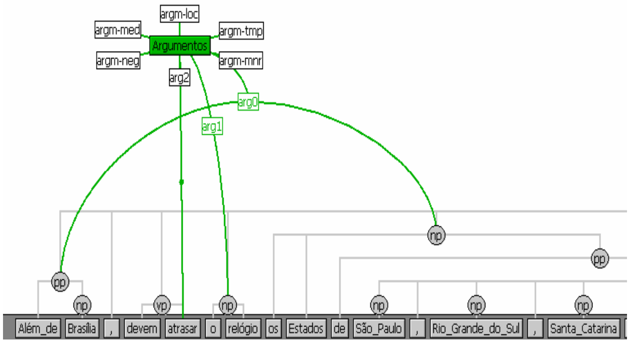

Figure 4. One role label points to two constituents

On the contrary, when a constituent contains more than one argument without possibility of splitting them, we assigned the higher role label (for example, ArgNs prevail over ArgMs and Arg1 prevails over Arg2). In Figure 5 there are two arguments: "do futuro" (Arg3) and "para proteger um garoto" (ArgM-Pnc), a purpose modifier. When there is no rule to decide the higher role label (ArgM of time and manner, for example), we assigned the first semantic role label for the whole constituent.

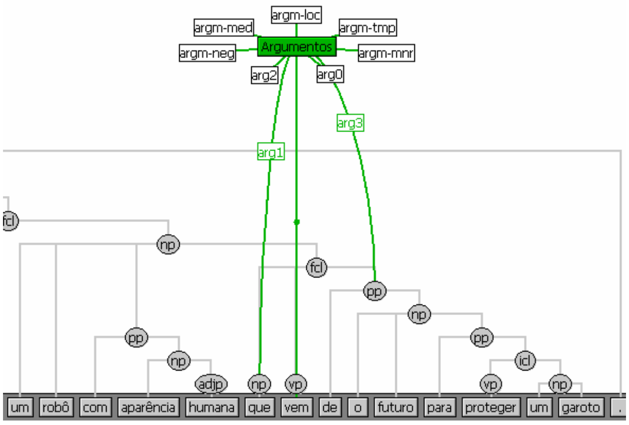

Figure 5. Two arguments in a same node of the syntactic tree

9. Language specific challenges for SRL

We flagged 462 instances with subject omission in our corpus and we think it would be interesting in future work to include a preprocessing phase to automatically infer, from the verb inflection, the personal pronoun that corresponds to the subject, creating a “ghost subject” that could support role label assignment.

Another difficulty is to deal with the particle “se”, a multifunctional word in Portuguese. We provided extra annotation identifying the function of “se” with special sentence flags, which will enable future work on disambiguation.

10. Additional Annotation

The annotation task is a rich opportunity for a linguist to identify relevant occurrences to describe a given language. Besides flagging instances that present subject omission, different functions of particle “se”, correferences and ellipsis, already mentioned, we made other additional annotations. Due to their semantic relevance, we registered 19 multi-word-units not recognized by the parser, 93 complex predicates (including light verb constructions), and 951 embedded infinite clauses with labels describing their respective semantic functions.

11. Future work

The corpus resource is available at PortLex[3] and the next step of Propbank-Br is the creation of frames files for the 1068 predicates annotated in the 6142 instances (from these 1068, 132 have already been done in a pilot study using Cornerstone frame editor). During this phase we will simultaneously add to each instance of the corpus the created roleset identification (sense in Portuguese) and, as often as possible, the equivalent roleset of Propbank (sense in English), thus mapping the Brazilian predicates to the English ones.

As soon as the frames files have been accomplished, we will be prepared to undertake a broader experience of SRL annotation in a larger corpus. For this purpose, our goal for the future is to assemble a group of researchers to expand Propbank-Br through collaborative work.

12. Acknowledgements

Our thanks to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for supporting this research.

13. References

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

Aldezabal, I.; Aranzabe, M. J., Ilarraza, A. D.;Estarrona, A. (2010). Building the Basque PropBank. In: Proceedings of LREC-2010.

Baker, C.F.; Fillmore, C. J.; Lowe. J. B. (1998).The Berkeley FrameNet Project. In: Proceedings of Computational Linguistics 1998 Conference.

Baptista, J.; Mamede, N. J.; Gomes, F. (2010) Auxiliary Verbs and Verbal Chains in European Portuguese. Proceedings of PROPOR 2010, pp. 110-119.

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

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

Choi, J. D. Bonial, C.; Palmer, M. (2010a) Propbank Instance Annotation Guidelines Using a Dedicated Editor, Jubilee. In: Proceedings of LREC-2010.

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

Christensen. J.; Mausam; Soderland, S.; Etzioni, O. (2011) An Analysis of Open Information Extraction based on Semantic Role Labeling. International Conference on Knowledge Capture (KCAP). Banff, Alberta, Canada. June 2011.

Duran, M. S.; Amâncio, M. A.; Aluísio, S. M. (2010) Assigning Wh-Questions to Verbal Arguments: Annotation Tools Evaluation and Corpus. In: Proceedings of LREC 2010.

Fillmore, C.. The Case for Case (1968). In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Gildea, D.; Jurafsky, D. (2001) Identifying Semantic Roles in Text,. In Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), Seattle, Washington.

Gildea, D.; Palmer, M. (2002). The Necessity of Parsing for Predicate Argument Recognition. In: Proceedings of The 40thMeeting of ACL, 2002.

Kingsbury, P.; Palmer, M; Marcus, M. (2002) Adding Predicate Argument Structure to the Penn TreeBank. Proceedings of The Second International Human Language Technology Conference, HLT-02.

Palmer, M.; Gildea, D.; Kingsbury, P. (2005) The Proposition Bank: An Annotated Corpus of Semantic

[3] http://www2.nilc.icmc.usp.br/portlex/index.php/projetos/propbankbr

Roles. Computational Linguistics, 31:1., pp. 71-105, March, 2005.

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

Palmer, M.; Ryu, S.; Choi, J.; Yoon, S.; Jeon, Y. (2006) Korean Propbank. LDC Catalog No.: LDC2006T03 ISBN: 1-58563-374-7

Palmer, M.; Babko-Malaya, O.; Bies, A.; Diab, M.; Maamouri, M.; Mansouri, A.; Zaghouani, W. (2008). A Pilot Arabic Propbank. In: Proceedings of LREC-2008.

Shen, D.; Lapata, M. (2007) Using Semantic Roles to Improve Question Answering. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '07).

Xue, N,; Palmer., M. (2009) Adding semantic roles to the Chinese Treebank. Natural Language Engineering. 15(1) pp. 143-172.