TextLink: Structuring Discourse in Multilingual Europe

COST Action Number IS1312

STSM Report

COST STSM Reference Number: COST-STSM-IS1312-26899
Period: from 2015-04-25 to 2015-05-10
STSM Applicant: Teresa Gonçalves, University of Évora (PT), tcg@uevora.pt
STSM Topic: Linguistic cues for automatic identification of attribution
Host: Prof. Deniz Zeyrek, Middle East Technical University, Ankara(TR), dezeyrek@metu.edu.tr

1 Introduction

The current modality tagger for Portuguese only detects and classifies the trigger [10]. PALAVRAS syntactic parser [1] is used to detect the modal triggers and a machine learning approach is used to label each modal trigger with the appropriate modal value in context (building a different classifier model for each verb).

The preliminary work is now extended to incorporate more linguistic cues allowing a better model construction [9]. These cues come from the **target** (attributes about the target word and the three nodes in the syntactic tree above the target – father, grandfather and great grandfather), the **context** (attributes of the words in the context around the target word) and the **path** (attributes of the tree path between the target word and the root and the attributes of the right and left brothers).

Besides the trigger (which is the lexical element conveying the modal value), the modality corpus [2] annotates the target, the source of the event mention (speaker or writer) and the source of the modality (agent or experiencer).

The **source of the event mention** is also annotated in the Penn Discourse TreeBank (PDTB) [8] as the **source** property of attribution feature. In PDTB each discourse relation and its arguments are labelled with **attribution** capturing four salient properties [7]: (a) source (distinguishing between different types of agents), (b) type (reflecting the degree of factuality), (c) scopal polarity (indicating polarity reversals due to surface negated attributions), and (d) determinacy (indicating the presence of contexts cancelling the entailment of attribution).

Purpose of the STSM

Considering the commonality between these two schemes, this two week mission aimed at studying the issues and exploring the similarities and differences cross-linguistically and to investigate the applicability of the used approach for tagging modality in the Portuguese corpus to tagging source attribution in the PDTB and the TDB [11] corpora.

After examining the modality issue in Portuguese and attribution in English it would be possible to plan a joint work on (automatically) tagging source in modality and/or attribution problems, not just in Portuguese and English but also in Turkish.

Work description

In the first week the PDTB and TDB corpora were studied by:

- reading papers and reports describing their annotation schemes;
- installing the PDTB and TDB annotators and browsing over the corpora;
- searching for papers aiming at automatically identifying some discourse features (e.g. connectives, arguments spans, etc).

A comparative study between the tagging of the source of the event mention in the Portuguese modality corpus and source attribution in PDTB was also conducted looking at descriptions, but most importantly at corpora examples (the modality corpus use MMAX2 [3] annotation software tool for tagging, that was also installed).

In the second week the Attribution problem was addressed by studying Silvia Pareti's work on attribution, namely:

- the development of the PARC corpus [5], an Attribution Relations corpus based on PDTB;
- the automatic detection of attribution features, namely the quote status [4, 6].

Finally, a shallow assessment of the PDTB and PARC attribution schemas was made aiming at their inclusion into/adaptation to the TDB.

Main results

One of the main results of this STSM was the awareness of the different stages on the annotation of Discourse Relations in three different languages: English, Turkish and Portuguese. For english, the PDTB annotates the connectives, arguments and their supplementary information text spans and also the connective sense and attribution information of the relation (for connective and arguments). On the other hand, TDB, besides annotating the same text spans adds a shared (between both arguments) one; it uses an expanded PDTB style annotation for the discourse sense (three-level hierarchy), adds the class feature but misses attribution information. Finally, currently there are no Discourse Relation corpus for Portuguese; the creation of such corpus is just starting.

Another result was that, in spite of the commonality between the concepts of source of the event mention in modality and the source of attribution in discourse relations, since the corpora address different problems, their sentences are different regarding the writing style, the linguistic features and structures difficulting the use of the same linguistic cues on the development of an automatic source annotator (for modality and discourse relations).

The PARC corpus was built from the PDTB corpus attribution information, enriching it with more information. For example, while in PDTB attribution is classified through its type, source type, factuality and scopal polarity, in PARC, besides these ones, there is also room

for the authorial stance, source attitude and quote status. Even if not yet fully defined the feature set to include in the TBD attribution annotation, authorial stance seems to be of high importance because of specific Turkish sentence elements that reveals the commitment the sentence's author expresses towards the given statement.

Finally, in what concerns the realization of a set of linguistic cues to be extracted from texts aiming at the construction of robust machine learning models for the identification of attribution features, even being possible as presented in [6] for the identification of the quote status, it needs further study; one needs to focus first on a specific feature of attribution and after devise the set of cues. At the present time this is only possible for the english language since the is no attribution annotation on Turkish nor Portuguese.

Future collaboration

Since Discourse Relation corpora for English, Turkish and Portuguese are in three different stages, there is plenty of room for collaboration contributing mainly to the goals of WG1 and WG2:

- the Portuguese team (led by Prof. Amália Mendes) can use the TDB team rich experience on discourse annotation and definition of guidelines; this will accelerate the process of building an annotated corpus;
- for attribution information, it is possible for both teams to work together and try to define an interoperable annotation guideline taking into account the PDTB and PARC corpora annotations.

While working on the attribution annotation special attention will be given on unveiling linguistic cues that could help build, using a machine learning approach, automatic annotation tools for each/some of the attribution features. This will contribute mainly to the goals of WG4 and also WG1 since its use would allow a more rapidly annotation of new texts aiming at extending these corpora.

References

[1] E. Bick. *The parsing system PALAVRAS*. Aarhus University Press, 1999.

[2] I. Hendrickx, A. Mendes, , and S. Mencarelli. Modality in text: a proposal for corpus annotation. In *LREC 2012: Eighth International Conference on Language Resources and Evaluation*, 2012.

[3] C. Müller and M. Strube. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

[4] T. O'Keefe, S. Pareti, J. R. Curran, I. Koprinska, and M. Honnibal. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, pages 790–799, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[5] S. Pareti. A database of attribution relations. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[6] S. Pareti, T. O'Keefe, I. Konstas, J. R. Curran, and I. Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[7] R. Prasad, N. Dinesh, A. Lee, A. Joshi, and B. Webber. Annotating attribution in the penn discourse treebank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 31–38, Sydney, Australia, July 2006. Association for Computational Linguistics.

[8] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *LREC*, 2008.

[9] P. Quaresma, A. Mendes, T. Gonçalves, and J. Sequeira. Modality in portuguese language: Influence of attributes in an svm approach. *to be published*.

[10] P. Quaresma, A. Mendes, I. Hendrickx, and T. Gonçalves. Tagging and Labelling Portuguese Modal Verbs. In J. Baptista, N. J. Mamede, S. Candeias, I. Paraboni, T. A. S. Pardo, and M. das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014. Proceedings*, volume 8775 of *Lecture Notes in Computer Science*, pages 70–81. Springer, 2014.

[11] D. Zeyrek, I. Demirşahin, A. Sevdik-Çallı, and R. Çakıcı. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2):174–184, 2013.