Anaphora – Clause Annotation and Alignment Tool

Borislav Rizov
Department of Computational Linguistics, IBL-BAS
52 Shipchenski Prohod Blvd., bl. 17
1113 Sofia, Bulgaria
boby@dcl.bas.bg

Rositsa Dekova
Department of Computational Linguistics, IBL-BAS
52 Shipchenski Prohod Blvd., bl. 17
1113 Sofia, Bulgaria
rosdek@dcl.bas.bg

Abstract

The paper presents Anaphora – an OS and language independent tool for clause annotation and alignment, developed at the Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences. The tool supports automated sentence splitting and alignment and modes for manual monolingual annotation and multilingual alignment of sentences and clauses. Anaphora has been successfully applied for the annotation and the alignment of the Bulgarian-English Sentence- and Clause-Aligned Corpus (Koeva et al. 2012a) and a number of other languages including French and Spanish.

1 Introduction

For years now corpus annotation has played an essential part in the development of various NLP technologies. Most of the language resources, however, do not include clause annotation and alignment which are considered quite useful in recent research on Machine Translation (MT) and parallel text processing (Piperidis et al., 2000; Sudoh et al., 2010; Ramanathan et al., 2011).

Aiming to facilitate and improve the process of clause annotation and alignment of multilingual texts, we developed Anaphora.

The tool is OS and language independent and supports automated sentence splitting and alignment, manual sentence and clause splitting, validation, correction and alignment, selection and annotation of conjunctions (including compounds (MWE)), and identification of the type of relation between pairs of syntactically connected clauses.

2 User Interface and Functionalities

Anaphora supports two kinds of operating modes: a monolingual and a multilingual one.

The monolingual mode is designed for manual editing and annotation of each part of the parallel corpus. The window consists of three active panes (Fig. 1): **Text view** (top pane), **Sentence view** (bottom left-hand pane) and **Clause view and annotation** (bottom right-hand pane).

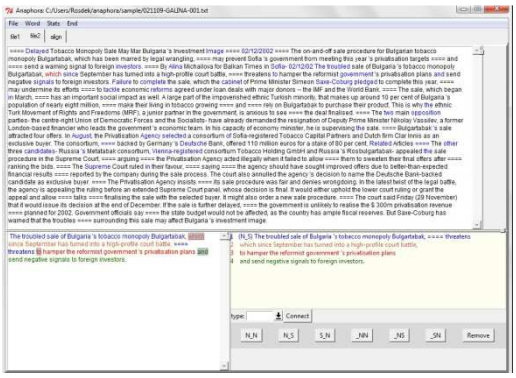

Figure 1. Anaphora – monolingual mode.

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 69–72,
Gothenburg, Sweden, April 26-30 2014.

In this mode the user may chose a file for verification and post-editing of the automatically performed sentence alignment. The individual monolingual texts which are part of an aligned pair are selected by the **file** tabs.

The monolingual mode offers the following functionalities:

- sentence splitting;
- clause splitting;
- correction of wrong splitting (merging of split sentences/clauses);
- annotation of conjunctions;
- selection of compounds;
- identification of the type of relation between pairs of syntactically connected clauses.

The end of a sentence may be changed by choosing the last word of the sentence and marking it using the End button from the top menu. Thus, the selection of the word as a sentence end is toggled and if it was marked as an End word, it is no longer such and the following sentence is automatically merged to the current one. If the word has not been already marked as an end, it is thus marked as one and the sentence is automatically split.

Clicking on any word of a sentence in the Text view pane results in the sentence appearing in the Sentence view pane, where clause segmentation and choice of conjunction are performed. The user defines the boundaries of clauses by selecting the words in them. This is achieved by marking the particular fragment of the text in the Sentence view pane with the mouse and pressing the 'space' key. This operation toggles the selection. Thus, a repeated use causes deselection. Marking a disconnected clause is done by marking the block of text containing it and unmarking the unnecessary words. When a clause is defined, it is listed in the bottom right-hand pane in a new color following the surface order of the sentence. Selection of a clause within another clause is also possible. Then the inner clause is listed directly after the split clause while the order of the split clause in the Clause view pane depends on the position of its first word in the sentence.

Once the clauses are defined, the user may annotate the conjunction of two clauses, also referred to as a marker. The marker may consist of one or more words or an empty word. Empty words (w="====") are artificial elements automatically introduced at the beginning of a potential new clause. An empty word may be selected as a marker when the conjunction is not explicit or the clauses are connected by means of a punctuation mark (for simplicity of annotation punctuation marks are not identified as independent tokens but are attached to the preceding token). When a word or a compound from one clause is selected in the Sentence view pane the user chooses another clause from the Clause view pane to create a pair of syntactically linked clauses. Then the relation for the pair is identified by selecting its type with the grey buttons N_N (coordination), N_S (subordinated clause following the main clause), S_N (subordinated clause preceding the main clause), etc.

The multilingual mode is selected with the **align** tab. In this mode annotators can create, validate and correct the alignment of the parallel units – sentences and/or clauses.

The window (Fig. 2) has two parallel **Text view** panes (on the top) and two parallel **List view** panes (in the bottom). Depending on the chosen menu (Clause or Sentence) the bottom panes show lists of aligned clauses or sentences.

Figure 2. Anaphora – multilingual mode.

The multilingual mode uses the output of the monolingual sentence and clause splitting and supports the following functionalities:

- automated sentence alignment;
- manual sentence alignment;
- manual clause alignment.

Automated sentence alignment is available as a menu command (Auto Sentence Align) in the multilingual mode.

To switch to manual sentence or clause alignment the corresponding menu commands are used – Sentence and Clause.

In the sentence menu the two bottom panes show lists of aligned sentences, each pair in a distinct color. The user may correct the alignment by choosing one or more sentences in each of the bottom panes and pressing the 'space' button to create a new alignment bead.

In the clause menu, when a sentence is selected in one of the two Text panes, its clauses are listed in the respective bottom pane. The corresponding aligned sentence appears in the parallel top pane with its clauses listed in the bottom. Alignment is performed when the user chooses one or more clauses from each of the bottom panes and then presses the 'space' button. Thus a new clause alignment bead is created.

3 Applications

Anaphora was successfully used for the annotation and the alignment of the Bulgarian-English Sentence- and Clause-Aligned Corpus (Koeva et al. 2012a) which was created as a training and evaluation data set for automatic clause alignment in the task of exploring the effect of clause reordering on the performance of SMT (Koeva et al., 2012b).

Since its development the tool is continuously used for annotation and clause alignment of different parts of the Bulgarian-X language Parallel Corpus (Koeva et al. 2012c) covering a number of languages including French and Spanish.

4 Implementation

Anaphora was designed as a self-sufficient module for annotation and clause alignment within the multi-functional platform Chooser (Koeva et al. 2008) which supports various NLP tasks that involve corpora annotation.

The tool is a stand-alone single user application implemented in Python and it uses the standard GUI library *tkinter* (the Tcl/Tk python binding) which makes it highly OS independent.

5 Data Processing and Representation

5.1 Input Data

The used format is a flat xml with root element *text.* The text is a list of word elements with several attributes like ‘w’ – wordform, ‘l’ – lemma, ‘u’ – annotator, ‘t’ – timestamp, ‘e’ – sentence end, etc.

Special attributes are responsible for marking the compounds (MWE) and clauses. The words that are members of a compound share a common value for the attribute ‘p’ (parent). Similarly, the words in a clause share a common value for clause – ‘cl’.

This format is compatible with the other modules of the Chooser platform. Thus, one file can be annotated with several different types of annotation like POS, semantic annotation, etc.

The system provides import scripts for two formats – plain text and the output of the Bulgarian Language Processing Chain (Koeva and Genov, 2011) – a TSV/CSV family format, where the text is tokenized and lemmatized.

Sentence splitting depends on the format of the input text. If it is a plain text, sentence splitting is based on the presence of end of sentence punctuation (full stop, exclamation mark, and question mark) followed by a capital letter. When the file is of the TSV/CSV family format sentence splitting is part of the Language Processing Chain.

5.2 Automated Sentence Alignment

The automated sentence alignment is performed using the Gale-Church aligning algorithm (Gale and Church, 1993).

6 Conclusions and Future Work

We believe that, based on its design and functionalities, Anaphora can be easily used and it will perform well for any given pair of languages, that is, it is to a great extent language independent. The system can also be applied as it is for phrase segmentation and word and phrase alignment. However, if we want to include simultaneous alignment of words, phrases, and clauses the system needs to be adopted.

We work on including additional functionalities to facilitate corpora annotation and parallel text processing such as anaphora annotation.

Our future intentions include also publishing it as an Open Source code so that it can serve the NLP community.

Acknowledgments

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the *Human Resources Development* Operational Programme 2007-2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper and under no conditions can the conclusions made in it be considered an official position of the European Union or the Ministry of Education, Youth and Science of the Republic of Bulgaria.

References

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics* 19(1): 75–102.

Svetla Koeva and Angel Genov. 2011. Bulgarian Language Processing Chain. In: *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*, University of Hamburg.

Svetla Koeva, Borislav Rizov, Svetlozara Leseva. 2008. Chooser - A Multi-Task Annotation Tool. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, ELRA* electronic publication, 728-734.

Svetla Koeva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova, Svetlozara Leseva, Hristina Kukova, Angel Genov. 2012a. Bulgarian-English Sentence- and Clause-Aligned Corpus. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACHR-2)*, Lisboa: Colibri, 51-62.

Svetla Koeva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova, Svetlozara Leseva, Hristina Kukova, and Angel Genov. 2012b. Application of clause alignment for statistical machine translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), Korea, 2012.*

Svetla Koeva, Ivelina Stoyanova, Rositsa Dekova, Borislav Rizov, and Angel Genov. 2012c. Bulgarian X-language parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* N. Calzolari et al. (Eds.) Istanbul: ELRA, 2480-2486.

Stelios Piperidis, Harris Papageorgiou, and Sotiris Boutsis. 2000. From sentences to words and clauses. In *J. Veronis, editor, Parallel Text Processing, Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, 117–138.

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Karthik Visweswariah, Kushal Ladha and Ankur Gandhe. 2011. Clause-based reordering constraints to improve statistical machine translation. In *Proceedings of the 5th International Joint Conference on NLP, Thailand, November 8-13, 2011*, 1351–1355.

Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, Masaaki Nagata. 2010. Divide and translate: improving long distance reordering in statistical machine translation. In *Proceedings of the Joint 5thWorkshop on SMT and Metrics MATR,* 418–427.