MultiDPS – A multilingual Discourse Processing System

Daniel Alexandru Anechitei
"Al. I. Cuza" University of Iasi
Faculty of Computer Science
16, General Berthelot St., 700483, Iasi, Romania
daniel.anechitei@info.uaic.ro

Abstract

[1]This paper presents an adaptable online Multilingual Discourse Processing System (MultiDPS), composed of four natural language processing tools: named entity recognizer, anaphora resolver, clause splitter and a discourse parser. This NLP Meta System allows any user to run it on the web or via web services and, if necessary, to build its own processing chain, by incorporating knowledge or resources for each tool for the desired language. In this paper is presented a brief description for each independent module, and a case study in which the system is adapted to five different languages for creating a multilingual summarization system.

1 Introduction

This paper describes a multilingual discourse processing system (MultiDPS) consisting in four different modules: Named Entity Recognizer (NER), Anaphora Resolver (AR), Clause Splitter (CS), Discourse Parser (DP), and for the summarization scope, the proper summarizer (SUM). This system can run online via web services such that it can be accessed from any programming environment and the architecture allows each tool to be individually trained. Each task, except for discourse parsing, MultiDPS's component tools combines machine learning techniques with heuristics to learn from a manually created corpus (a gold corpus of discourse trees is very difficult to obtain due to the complexity of the task). The complexity of the processing tasks (reaching to discourse analysis) and the multilingual capabilities, make MultiDPS an important system in the field of natural language processing.

2 System Design

The MultiDPS architecture includes two main parts as it can be seen in Figure **1**.

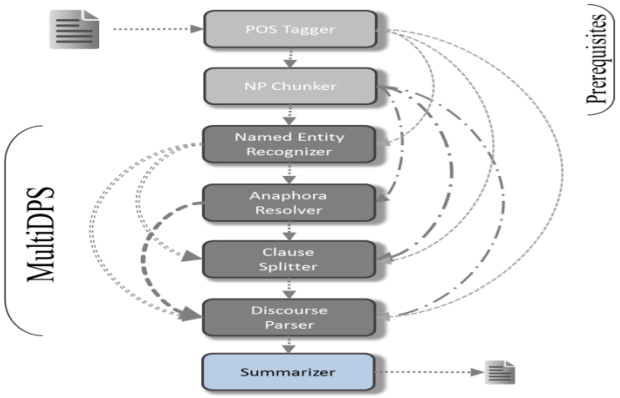

Figure 1: The MultiDPS's component modules and supported workflows

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 44–47, Dublin, Ireland, August 23-29 2014.

The Prerequisite part includes usually known basic NLP tools and it is a primary step for obtaining the input for MultiDPS. The system consists of four different modules which will be discussed in detail in the next sections. All modules implement a language independent vision in which the algorithm is separated from linguistic details. Each phase and the output of each module is an input for a next phase, not necessarily the immediately next one, as it is depicted in Figure **1** (dotted arrows suggest different paths that the system supports). Depending on individual needs or on the existence of specific resources (manual annotated corpora for a specific language), different language processing chains can be created. The entire system is designed in such a way that each individual module brings an extra annotation to the text therefore, when building a processing chain, some modules can be skipped.

2.1 Named Entity Recognizer

Named Entity Recognition (NER) is a computational linguistic task that seeks to classify sequences of words in predefined categories. In this approach the categories are organized under four top level classes (PERSON, LOCATION, ORGANIZATION and MISC) and a total of nine subclasses.

In order to identify the type of entities a voting system is implemented, being meant to decide between different heuristics, which use automatically calibrated weights for different features, where high scores are given for the entities within gazetteers. Examples of features are: context bi/tri grams for different classes; appearance of a definite article; partial matches with gazetteers or within the same text.

2.2 Anaphora Resolution

The AR module used in MultiDPS is based on the work done in Anechitei et al (2013), and improved by adding a classifier, to predict whether there is a relation between each pair of noun phrases, resulting in a hybrid approach. Examples of features used to decide if there is a co-referential chain between two noun phrases are: number agreement, gender agreement, and morphological description, implementing on the head noun; similarity between the two noun phrases, both at lemma level and text level implemented on the head noun and also on the entire noun phrase; condition if the two noun phrases belong to the same phrase or not.

If the matching score given by the two methods is greater than an automatically computed threshold, then the actual noun phrase is added to already existing chain of referential expressions attached to the noun phrase, and all the features are copied onto the list of features of the new referential expression. If there is no previous noun phrase, for which the matching score to be greater than the threshold, then a new co-referential chain is created containing only the actual noun phrase along with its features.

2.3 Clause Splitter

A clause is a grammatical unit comprising a predicate and an explicit or implied subject, and expresses a proposition. For the present work, the delimitation of clauses follows the work done in Anechitei et al (2013) and starts from the identification of verbs and verb compounds. Verb compounds are sequences of more than one verb in which one is the main verb and the others are auxiliaries ("is writing", "like to read"). Examples of features used to build the model of compound verbs are: distance between the verbs; the existence of punctuation or markers between them; the lemma and the morphological description of the verbs, etc.

The semantics of the compound verbs makes it necessary to take the whole construction together not putting boundary in the interior, so that the clause does not lose its meaning. Clause boundaries are looked between verbs and compound verbs which are considered the pivots of clauses. The exact location of a boundary is, in many cases, best indicated by discourse markers. A discourse marker is a word, or a group of words, that also have the function to indicate a rhetorical relation between two clauses. The features used to build the marker's model are: the lemma and the context of the marker expressed as configurable length sequences of POS tags and the distance from the verb in front of it.

When markers are missing, boundaries can still be indicated by statistical methods, trained on explicit annotations. The weights of the features are tuned like in previous examples, by running the calibration system on the manual annotated corpora and creating the models using *MaxEnt*[1] library.

[1] The Maximum Entropy Framework: http://maxent.sourceforge.net/about.html

2.4 Discourse Parser

The approach to discourse parsing implemented in MultiDPS follows the one described in Anechitei et al (2013) and is a symbolic approach rooted on (Marcu, 1999). The generated discourse trees put in evidence only the nuclearity of the nodes, while the name of relations is ignored. The discourse parser adopts an incremental policy in developing the trees and it is constrained by two general principles, well known in discourse parsing: sequentiality of the terminal nodes (Marcu, 2000) and attachment restricted to the right frontier (Cristea, 2005). The algorithm involves a *generate-rank-evaluate* method by generating a forest of developing trees at each step, followed by heuristics for ranking and evaluating the trees. The heuristics are suggested by both Veins Theory (Cristea et al, 1998) and Centering Theory (Grosz et al, 1995). The aim of these heuristics is to assign scores to the developing trees and also to master the exponential explosion of the developing structure.

2.5 The Summarizer

For the summarization purpose, the discourse structure gives more information than properly needed. The summary is achieved by trimming unimportant clauses/sentences on the basis of the relative saliency, cohesion and coherence properties. For each discourse unit, a score is attached and reflects the properties mentioned above. Each component of MultiDPS contributes to the calculation of this score.

3 Implementation of the modules

The main idea behind the system architecture is that, if a module is fuelled with appropriate language resources, it can be put to work on any language. For the Romanian language, the input for MultiDPS is obtained using a deep noun phrase chunker (Simionescu, 2011) and for the English language using the Stanford Parser (Socher et al, 2013). All the resources (manually annotated corpora for English and Romanian) are available for download.

The clear benefit of this system architecture using web services is that if an improvement is made in a certain module, the results will be propagated through the others, without the need of human intervention. Figure **2** illustrates the web interface for the discourse parser, where the XML annotations are mapped in a visual mode.

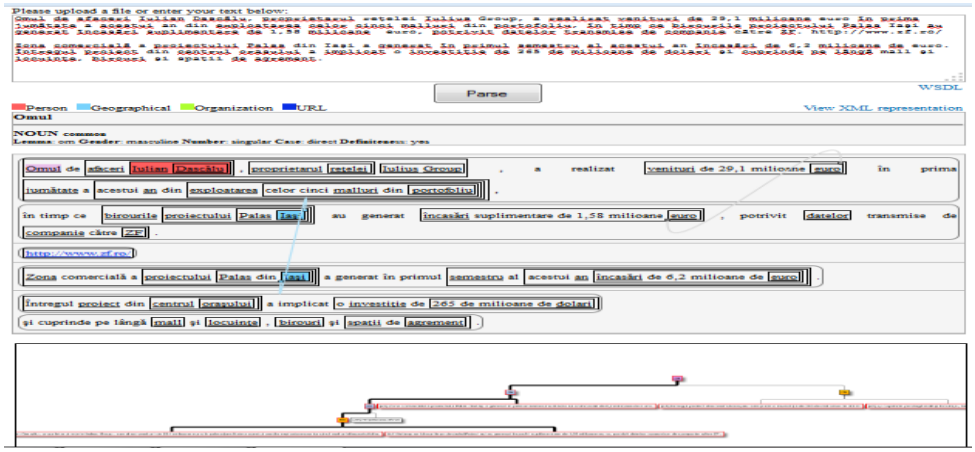

Figure 2: View of the Discourse Parser web application that illustrates all annotations.

In addition to the web applications and the core of the system (each module can be used as a library), what is made available is a wide range of free additional tools like online annotation services and calibration systems for each individual module. MultiDPS was easily adapted for other languages where there was input provided for the system entry and training corpus for each module.

4 Experiments and results

In this paper I present the results obtained after combining all the modules to create a multilingual summarization system. The results were obtained after attending an international workshop on summarization (Kubina et al., 2013), where the objective of each participant was to compute a maximum

250 words summary for each document for at least two of the dataset languages (30 documents per language). The submitted summaries were evaluated using ROUGE metric (Lin, 2004) and presented in the next table, where the oracle in the table represents the "perfect summary":

Language	System							
	baseline	s1	s2	s3	s4	s5	s6	oracle
bg	0.2854	**0.3190**	0.2955	0.2969	0.2974			0.3966
de	0.2529	**0.3414**	0.3198	0.3341	0.3203			0.3675
el	0.2899	**0.3229**	0.2777	0.2747	0.2698			0.3775
en	0.4113	0.3273	0.2781	0.2799	0.2765	**0.3638**	0.3411	0.5554
ro	0.3125	**0.3337**	0.29048	0.3006	0.2985			0.4361

Table 1: ROUGE-1 average for all five languages

(Bulgarian, German, Greek, English and Romanian)

Nevertheless, the results are encouraging for this complex system (**s1** is the id of the system presented in this paper).

5 Conclusions

MultiDPS's strength is manifested through its online availability and the existence of the online services for creating corpora for each module. Moreover, considering that the results obtained by putting together all the modules are similar for different languages, the system can be regarded as having language-wide validity.

Reference

Barbara J. Grosz, Aravind K. Joshi and Scott Weinstein. 1995. *Centering: A framework for modeling the local coherence of discourse*. Computational Linguistics, 21(2), pages 203–226.

Chin-Yew Lin. 2004. *Rouge: A package for automatic evaluation of summaries*. In Proceedings of the ACL Workshop on Text Summarization Branches Out, Barcelona, Spain.

Dan Cristea, Nancy Ide, Laurent Romary. 1998. *Veins theory: A model of global discourse cohesion and coherence*. In Proceedings of the 17th international conference on Computational linguistics, pages 281-285, Montreal.

Dan Cristea. 2005. *The Right Frontier Constraint Holds Unconditionally*. In Proceedings of the Multidisciplinary Approaches to Discourse (MAD'05), Chorin/Berlin, Germany.

Daniel A. Anechitei, Dan Cristea, Ioannidis Dimosthenis, Eugen Ignat, Diman Karagiozov, Svetla Koeva, Mateusz Kopeć, Cristina Vertan. 2013. *Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context*. In Neustein, A., Markowitz, J.A. (eds.), Where Humans Meet Machines: Innovative Solutions to Knotty natural Language Problems. Springer Verlag, Heidelber/New York.

Daniel Marcu. 1999. *Discourse trees are good indicators of importance in text*. In I. Mani and M. Maybury (eds.), Advances in Automatic Text Summarization, pages 123-136, The MIT Press.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Jeff Kubina, John M. Conroy, Judith D. Schleisinger. 2013. *ACL 2013 MultiLing Pilot Overview*. In Proceedings of MultiLing 2013 Workshop on Multilingual Multi-document Summarization, Sofia, Bulgaria, pages 29-38, workshop in conjunction with the 51th Annual Meeting of the Association for Computational Linguistics (ACL 2013).

Radu Simionescu. 2011. *Romanian Deep Noun Phrase Chunking Using Graphical Grammar Studio*. In Proceedings of The International Conference on Resources and tools for Romanian Language.

Richard Socher, John Bauer, Christopher D. Manning, Andrew Y. Ng. 2013. *Parsing with Compositional Vector Grammars*. In Proceedings of ACL.