R. Delmonte, D. Gîfu, R. Tripodi

A LINGUISTICALLY-BASED ANALYZER OF PRINT PRESS DISCOURSE

Abstract. The success of a newspaper article for the public opinion can be measured by the degree in which the journalist is able to report and modify attitudes, opinions, feelings and political beliefs. We present a linguistically-based system for Italian, derived from GETARUNS, which integrates a range of natural language processing tools with the intent to characterize the print press discourse. The method could help journalists by evidencing hidden aspects of the linguistic abilities of politicians.

1. Introduction

The aim of an interdisciplinary approach such as analyzing the language of political discourse with NLP tools is to define and explain different discursive contexts, in this case, reflected by the online media. Content analysis, which is based on[1] requires an extremely laborious methodology for objective interpretations. In this paper, we discuss paradigms for evaluating linguistic interpretation of discourses as applied by the system GETARUNS (General Text And Reference Understanding System)[2], which addresses the needs to restrict access to extra linguistic knowledge of the world by contextual reasoning. We focus on three aspects critical to a successful evaluation: creation of large quantities of reasonably good training data, semantic and pragmatic analysis. We assume that in order to properly capture

[1] *Osgood C.E.* The representational model and relevant research methods // De Sola Pool I. (Ed.), Trends in Content Analysis. University of Illinois Press, 1959.

[2] *Bos J., Delmonte R.* (eds.), Semantics in Text Processing (STEP), Research in Computational Semantics, vol.1, College Publications, London, 2008; *Delmonte R.* Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution. New York, Nova Science Publishers, 2009.

opinion and sentiment[3] expressed in a text or dialog any system needs a deep text processing approach. This distinction is obtained by searching for factivity markers at propositional level[4]. In order to produce this output, the system makes use of a flat syntactic structure and a vector of semantic attributes associated to the verb compound at propositional level and memorized. Important notions required by the computation of opinion and sentiment are also the distinction of the semantic content of each proposition into two separate categories: objective vs. subjective. In particular we take into account: modality operators like intensifiers and diminishes, modal verbs, modifiers and attributes adjuncts at sentence level and lexical type of the verb. More on this below.

The paper is structured as follows. Section 2 shortly describes the system. Section 3 discusses a user case from Italian press. Finally, section 4 highlights interpretations anchored in our analysis and presents conclusions.

2. The system GETARUNS

In this section we will present a short description of the system for Italian that we used in this experiment. The system is derived from GETARUNS, a multilingual system for deep text understanding, that works for English, German and Italian. The current version (see 3.0) of the system can be used with unlimited text and vocabulary, again for English and Italian. The two versions – the deep and "shallow" bottomup version – work in a pipeline in order to prevent failures of the deep version. They can also work separately to produce less

---

[3] *Wiebe J., Wilson T., Cardie C.* Annotating expressions of opinions and emotions in language // Language Resources and Evaluation, 2005, 39(2). P. 165–210.

[4] *Saurì R., Pustejovsky J.* Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text // Computational Linguistics, 2012, 38, 2. P. 261–299.

constrained interpretations of the text at hand. This second version has been used for the RTE challenges and for TAC summarization tasks[5].

In order to proceed to the semantic level, each nominal expression is classified at first on the basis of the assigned tag and multiword expressions are built on a lexical basis receiving a NER-like classification. The remaining nominal expressions are classified using the classes derived from ItalWordNet (*Italian WordNet*)[6]. In addition to that, we have compiled specialized terminology databases for a number of common domains including: medical, political, economic, and military. These lexica are used to add a specific class label to the general ones derived from ItalWordNet. To produce sentiment analysis, we created an ad hoc lexicon for the majority of politically related concepts (some 3000) contained in the text we analysed, in order to reduce the problem of ambiguity. This was done labelling only those concepts which were uniquely intended as one or the other sentiment, restricting reference to the domain of political discourse. The output of this semantic classification phase is a vector of features associated to the word and lemma, together with the sentence index and sentence position.

3. A comparative study

Here we have to recognize the huge relevance of semantics and pragmatics in analyzing of text.

A. The corpus

For the elaboration of preliminary conclusions on the process of the change of the Italian government and president of government, we collected, stored and processed – partially manually, partially automatically,– relevant texts published by three national on-line

---

[5] *Delmonte R.*, *Vincenzo P.* Opinion Mining and Sentiment Analysis Need Text Understanding // Advances in Distributed Agent-based Retrieval Tools, 2011 Springer. P. 81–96.

[6] **http://www.ilc.cnr.it/iwndb/iwndb_php/**

newspapers having similar profiles[7]. For analytical results to be comparable to those taken so far by second author[8], we needed a big corpus[9], especially considering the rigorous criteria that we list below:

- Type of message (type of opinions circulated by the editorial: pro, against Berlusconi and impartial as follows: Corriere della Sera – also called The People Newspaper – impartial; Libero, pro Berlusconi; and La Republica, against Berlusconi).
- Period of time (a month before the resignation of Berlusconi (12 November 2011), abbreviated to OMBB, the period between the presentation of Berlusconi's resignation and the appointment of Mario Monti (16 November 2011) as premier of the Italian Government, abbreviated with PTMB, and a month after the resignation of Berlusconi, abbreviated with OMAB).

B. The semantic and pragmatic analysis

We show in this section the results outputted by GETARUNS when analysing the streams of textual data belonging to the three sections of the corpus. In Fig. 1 below, we present comparative semantic polarity and subjectivity analysis.

[7] **www.corriere.it, www.liberoquotidiano.it, www.repubblica.it**

[8] *Gîfu D., Cristea D.* Multi-dimensional analysis of political language // Proceedings of The 7th FTRA International Conference on Future Information Technology, Application, and Service – FutureTech-2012, vol. 1, Springer, James J. (Jong Hyuk) Park, Victor Leung, Taeshik Shon, Cho-Li Wang (eds.). P. 213–221.

[9] *Kennedy G.* An Introduction to Corpus Linguistics. London & New York: Longman, 1998.

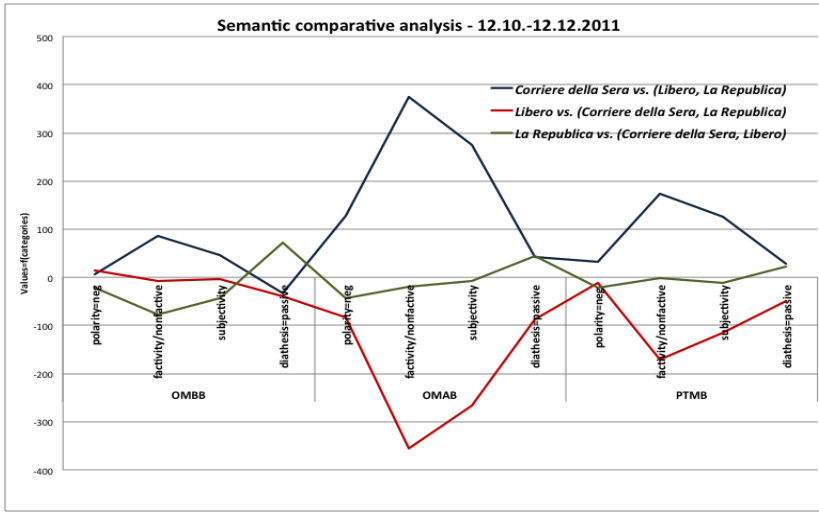

Fig. 1. Comparative semantic analysis

On the graph we show differences in values for four linguistic variables: they are measured as percent value over the total number of semantic linguistic variables selected from the overall analysis and distributed over three time periods on X axis. To display the data we use a simple difference formula, where Difference value is subtracted from the average of the values of the other two newpapers for that class. Differences may appear over or below the 0 line. In particular, values above the 0x axis mean they assume positive or higher than values below the 0x axis, which have a negative import. The classes chosen are respectively: 1. propositional level polarity with NEGATIVE value; 2. factivity or factuality computed at propositional level, which contains values for non factual descriptions; 3. subjectivity again computed at propositional level; 4. passive diathesis. We can now evaluate different attitudes and styles of the three newspapers with respect to the three historical periods: in particular we can now appreciate whether the articles report facts objectively without the use of additional comments documenting the opinion of the journalist. Or if it is rather the case that the subjective opinion of the journalist is present only in certain time spans and not

in others. So for instance, *Corriere*, the blue or darker line, has higher nonfactive values than the other two newspapers; *Repubblica* is the most balanced newspaper, in all values chosen. It has the lowest values in OMBB, and *Libero* has the absolute lowest values in OMAB but also in PTMB. Subjectivity is distributed very much in the same way as factuality, in the three time periods. *Libero* is the most factual newspaper, with the least number of subjective clauses. Similar conclusion can be drawn from the use of passive clauses, where we see again that *Libero* has the lowest number. The reasons for *Libero* having the lowest number of nonfactive clauses in OMAB, needs to be connected with the highest number of NEGATIVE polarity clauses, which is related to the nomination of Monti instead of Berlusconi, and is felt and is communicated to its readers as less reliable, trustable, trustworthy. Uncertainty is clearly shown in the intermediate period, PTMB, where *Corriere* has again the highest number of nonfactual clauses.

In Fig. 2 we represent comparative differences between the three newspaper in the use of three linguistic variables for each time period. In particular, we plotted the following classes of pragmatic linguistic objects: 1. references to Berlusconi as entity (Silvio, Silvio_Berlusconi, Berlusconi, Cavaliere, Caimano); 2. references to Monti as entity (Monti, prof_Monti, professore, Mario_Monti, super_Mario); 3. negative words or overall negative content words. To capture coreference mentions to the same entity we built a specialized coreference algorithm. With one month before Berlusconi's resignation (OMBB), we can highlight the opinions of the three dailies as follows: *Corriere della Sera* and *Libero* are concerned mostly with Berlusconi (see *Berlusconi occurrences*), with a remarkable difference however in terms of positive – *Libero* – vs negative – *Corriere* – comments. After Berlusconi resigned (OMAB) *Corriere* is more concerned than the other two newspapers on Monti: negative appreciation is always higher with *Corriere* and not with the other two. Finally in the intermediate period, both *Libero* and *Corriere* seem to be the most concerned with the new government, again with the highest number of negative comments.

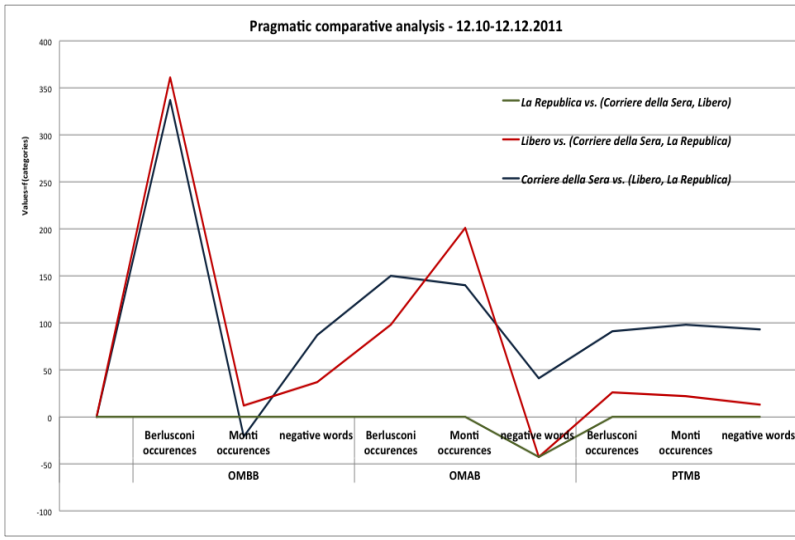

Fig. 2. Comparative pragmatic analysis of Italian newspapers

5. Conclusions

The analysis we proposed in this paper aims at testing if a linguistic perspective anchored in natural language processing techniques (in this case, GETARUNS system) could be of some use in evaluating political discourse in print press. However, we are aware of the dangers of false interpretation. For instance, if we take as example the three newspapers we used in our experiments, differences at the level of pragmatic, which we have highlighted as differentiating them, should be attributed only partially to their idiosyncratic rhetorical styles, because these differences could also have editorial roots. It remains yet to be decided the impact that the use of certain pragmatic structures could have over a wider audience of political discourse.

The system helps to outline distinctive features which bring a new and, sometimes, unexpected vision upon the discursive feature of journalists' writing.