DE GRUYTER OPEN
DOI 10.1515/cplbu-2015-0023
3rd International Engineering and Technology Education Conference & 7th Balkan Region Conference on Engineering and Business Education
Sibiu, Romania, November, 1st – 4th, 2015

Innovative Methods for Business Education using Isotope Linking on Anonymous Readers' Comments

Daniela GÎFU
"Alexandru Ioan Cuza" University of Iaşi, Romania
daniela.gifu@info.uaic.ro

Marius CIOCA
"Lucian Blaga" University of Sibiu, Romania
marius.cioca@ulbsibiu.ro

ABSTRACT

The paper presents the importance of analysis isotopes on anonymous readers' comments as an important part of deep interpretation of texts. Furthermore, we describe a classification methodology of the anonymous readers' comments on online articles, through the overlapping of isotopes, which completed the traditional analytical methods. Automatic recognition of isotopes is an important topic in Natural Language Processing (NLP), especially in the semantic disambiguation. The aim of this article is the automatic comparative analysis of the identified isotopes in articles and comments, which reveals an important part of online behavior. Moreover, we present a new tool that classifies the online commentators based on existing resources, open-source or freely available for research purposes. This study is intend to help direct beneficiaries (journalists, business, education, managers, PR specialists), but also specialists and researchers in the field of natural language processing, linguists, psychologists, etc.

Keywords: *isotopes, natural language processing, online commentators' classes, anonymous readers' comments, education, business*

INTRODUCTION

In our context, isotopes in writing refer to how public consumers express a personal opinion of their experience about different subjects proposed by journalists. The motivation for this topic is to clarify and describe the civic online profiles (Cioca et al., 2013), (Gîfu & Cioca, 2013), (Gîfu, Stoica, & Cristea, 2013) by comparing isotopes, found in texts on different *platforms* (articles and comments). These isotopes are influenced by the amount of media texts regardless of their nature and purpose. We considered this as an important point of view to understand the media consumer behavior. The isotopy[1] concept was introduced by A.J. Greimas (Greimas, 1970), in structural semantics, describing the coherence and homogeneity of texts. The isotopy had a major

[1] The concept was borrowed the term from nuclear physics.

impact on the field of semiotics, and was redefined multiple times (Rastier, 1972) (Ricœur, 1995) (Plett, 1983).

This paper considers the fact the discursive isotopy can disclose an important part of the online civic behavior.

The paper is structured in five sections. After a brief introduction about the importance of this study, the section 2 mentions some important works focused on discursive isotopy. The section 3 shortly describes a new tool for automatic recognition of isotopes and the classification of the online commentators, and section 4 presents a case study on Romanian print press and the statistical results. The last section highlights conclusions and mentions for the future work, in order to improve the automatically isotopes recognizing.

BACKGROUND

If we understand the text as a manifestation of freedom of expression, assuming the constitutive ambivalence of the sign-text, it can be treated as eminentely verbal entity and also as a part of a complex semiotic process (Vlad, 2000, p. 22). The isotopy is an essence of the text, and because of that it was long time a subject of debate involving the critical literature in the age of reproducibility (Riva, 2011). Furthermore, by using a computational technology (Ciotti & Crupi, 2015) we can rethink the methods in textual hermeneutics, observing the macrostructural and microstructural results (styles, lexemes, isotopes) of critical analysis. The authors propose a hermeneutic template that allows semantic indexing of isotopes.

The emergence of a large and increasingly number of increasingly large tools and technologies - which allow the textual data storage and the electronic editions in different formats (RTF, PDF, etc.) and thus their analysis (quantitative, mostly) - decreases the computerized hermeneutical potential, when the text is divided into atoms with the same meaning (Trevisan, 2008). Moreover, the textual criticism often has no historical dimension, a solution being TEI (*Text Encoding Initiative*), which encodes some semantic features in the modern texts. For instance, the Crilet Laboratory of the Faculty of Arts of the "Sapienza" University of Rome, proposes the interpretation of documents extending (Mordenti, 2007), using digital transcription and reformulation through semantic markup. In other words, it is possible that a narrative corpus to belong to several semantic families. Thus, it could be analyzed: *vertical*, by studying the lexical sorting at maximum frequency hapax-legomena; *semantic*, by analyzing frequency and position of the isotopes selected in text (Greimas, 1983); *alphabetically*, by generating an alphabetical order to identify meanings of families. Having built a system based on text, it is useful to start the critical thinking adding marking XML for links to sites with historical references. TEI model explains the text coding principles (Cummings, 2007) (Romary, 2009) (Vanhoutte & Van den Branden, 2010).

This study aims to demonstrate the potentiality of textual analysis, highlighting the interdisciplinary nature of the methodological approach which that will be described below.

IARC TOOL DESCRIPTION

IARC (*Izotopes of Anonymous Readers' Comments*) is an application implemented recently at the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași (UAIC) which has a simple functionality, but very useful, given the huge volume of comments on online newspapers forums.

This tool is able to automatically detect and to compare the isotopes from texts (article vs. comment) and to classify the online commentators taking into account these. This tool is based on information like labelling of parts of speech, extracting of isotopes and classify the anonymous readers' comments as we describe below:

1. We access the newspaper website which has a section for anonymous readers' comments. The tool extracts only the text (article and comments). Each text is passed through an extraction module keyword (*topical extraction*), which consists in the automatic pre-processing chain applied on our corpus and includes the following tasks, executed in sequence:

- Segmentation (splitting the text in sentences);
- Part-of-speech tagging (identifies morpho-syntactic information of tokens);
- Noun phrase chunking (Simionescu, 2012) (recognizing the chunks that consist of noun phrases (NPs)).

2. The words found are passed through a filter cleaning (*cleaning*) being eliminated the connecting words (conjunction and preposition), and adverbs and pronouns. Each keyword (isotopy) found in article is checked with other keywords found in each comment individually (one-to-one). In case they are synonyms, their weights are aggregated, being retained only the keyword with the highest frequency.

3. In order to categorize the commentators, the formulas (3 in this moment) were made after some tests that can be improved as the corpus will increase.

a) Each isotopy from a comment is checked if it belongs to the isotopes list extracted from the article. If one of the isotopy is retained in the article list or is a synonym with this -> is added to the final result.

We defined five commentators' classes: *none*, *low*, *medium*, *high* and *expert* and they have the weights assigned to the following intervals:

1 ->[0-4] -> none
2 ->(4-10] -> low
3 ->(10-14] -> medium
4 ->(14-20] -> high
5->(20-100] -> expert

b) Check the number of isotopes that appear in both lists (article and comment), then the weight 100% will be the total number of words in the article.

Ex.: in a comment we identify x, y, z isotopes, and in article x, y, z, t. The result will be:

3*100/4=75% (1)

The new intervals for commentators' classes:

1 ->[0-20] -> none
2 ->(20-30] -> low
3 ->(30-50] -> medium

4 ->(50-75] -> high
5 ->(75-100] -> expert

c) For each isotopy that appears in both lists (article and comment), we added the article weight, the 100% weight is given by the sum of the weights of the article.
Ex: If we have the weights: x = 5%, y = 6%, z = 10%, t = 3% and in comment the isotopes y, z, t its weight will be calculated with the formula:

(6+10+3)*100/5+6+10+3) = 86% (2)

In this case, the intervals for commentators' classes:
1 ->[0-20] -> none
2 ->(20-30] -> low
3 ->(30-50] -> medium
4 ->(50-75] -> high
5 ->(75-100] -> expert

The final result will be the average of these formulas[2].

CASE STUDY AND STATISTICAL METHODES

We present a short case study and its results, starting from a communication crisis between the Prime Minister and President about the Silvic Code to show the tool's utility for the signatures'' article (journalists, in this case).

In this sense, we monitored stored and pre-processed all articles from the newspaper *Adevărul* in the period 13-15 May 2015, structured as follows (see Fig. 1):

- 13 May (3 before) – 12 comments (720 words) and 2 articles (1248 words);
- 14 May (the crisis in course) - 233 comments (32296 words) and 11 article (6451 words);
- 15 May (after) – 197 comments (4746 words) and 3 article, one without comments (859 words).

Like we said, the work corpus was automatically and manually annotating (75%, training corpus, and 25%, testing corpus). All automatic results were checked manually (Table 1) in order to calculate the statistical parameters (Table 2): Precision, Recall and F-measure, important to improve our formulas for the classification of online commentators.

With these values, the Precision (3), Recall (4) and F-measure (5) could be computed, for categorization and classification of online commentators (CCi) using isotopes.

$$R = \frac{\#correctly\_identified\_CCi}{\#manually\_annotated\_CCi} (3)$$

[2] Note that these calculations have resulted from a few tests (manual annotation vs. automatic annotation) that disadvantage the short and very short comments.

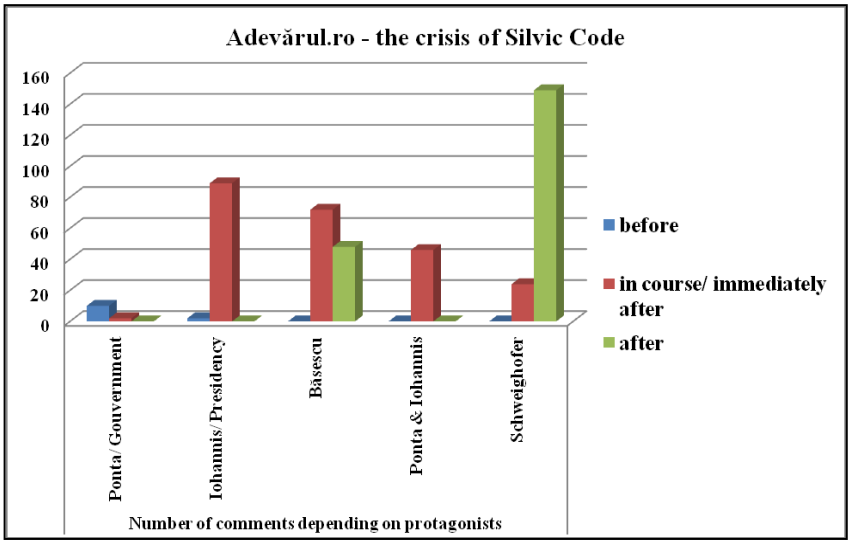

Figure 1: The number of online comments depending on protagonists

Table 1: Automatic and manual annotation results

Total Number of Isotopes	478
manual isotopes for commentator class *None* (mICN)	271
automatic isotopes for commentator class *None* (aICN)	366
manual isotopes for commentator class *Low* (mICL)	83
automatic isotopes for commentator class *Low* (aICL)	49
manual isotopes for commentator class *Medium* (mICM)	67
automatic isotopes for commentator class *Medium* (aICM)	38
manual isotopes for commentator class *High* (mICM)	32
automatic isotopes for commentator class *High* (aICM)	11
manual isotopes for commentator class *Expert* (mICM)	25
automatic isotopes for commentator class *Expert* (aICM)	14

$$R = \frac{\#correctly\_identified\_CCi}{\#manually\_annotated\_CCi} \qquad (4)$$

$$F - measure = \frac{2 * P * R}{P + R} \qquad (5)$$

The values are given in Table 2.

Table 2: Statistical results for the detection commentator class

Commentator class by isotopy	Recall	Precision	F-measure
None	74,04%	42,54%	54,03%
Low	70,94%	62,87%	66,66%
Medium	69,79%	63,80%	66,66%
High	60,37%	74,41%	66,65%
Expert	69,44%	64,10%	66,66%

As shown in Table 2 the results for the automatic detection of isotopes online commentators classes, which could be better. For instance, the fact that the High class is scored better than the None class, could be due to the special attention that we paid on annotating isotopes.

CONCLUSIONS AND FUTURE WORK

This paper presents an automatic method able to detect the discursive isotopies and online commentators' classes. Although the statistics are satisfactory, it is premature to advance some firm conclusions about the accuracy of the data obtained. We intend to consider more than syntactic threshold (the frequency of words) and to pass to the discursive semantic registers. Actually we talk about the syntactic and semantic isotopes. We believe the results will be better if we will analyse the semantic isotopes (synonyms, the words / phrases that are part of the same semantic register). For instance: Forester Code includes forester, trees, silviculture, logs, etc.

ACKNOWLEDGMENTS

I am grateful to the NLP-Group@UAIC-FII, especially to Silviu Pantilimon, for supporting me in the development of IARC tool for the automatic commentators' categorization using isotope concept.

REFERENCES

Cioca, M. Ghete, A., Cioca, L.I., & Gifu, D. (2013). Machine Learning and Creative Methods used to Classify Customers in a CRM Systems, *Applied Mechanics and Materials*, pp 769-773.

Ciotti, F., & Crupi, G. (2015). *Dall'Informatica umanistica alle culture digitali. Atti del Convegno di studi in memoria di Giuseppe Gigliozzi* . Roma.

Cummings, J. (2007). *The Text Encoding Initiative and the Study of Literature.* Ray Siemens and Susan Schreibman (eds.), Blackwell Companion to Digital Literary Studies (Blackwell: Malden), 451-476.

Gîfu, D., & Cioca, M. (2013). Online Civic Identity. Extraction of Features. *Procedia – Social and Behavioral Sciences* , 366-371.

Gîfu, D., Stoica, D., & Cristea, D. (2013). Virtual Civic Identity. *9th International Conference Linguistic Resources and Tools for Processing The Romanian Language* (pp. 139-148). Iasi: "Alexandru Ioan Cuza" University Publishing House.

Greimas, A. J. (1970). *Del senso.* Milano: Bompiani.

Greimas, A. J. (1983). *Structural Semantics: An Attempt at a Method.* University of Nebraska Press.

Mordenti, R. (2007). *L'altra critica. La nuova critica della letteratura fra studi culturali, didattica e informatica.* Roma: Meltemi.

Plett, H. F. (1983). *Ştiinţa textului şi analiza de text.* Bucuresti: Editura Univers.

Rastier, F. (1972). *Systématique des isotopies.* Paris: Essais de sémiotique poétique, Larousse.

Ricœur, P. (1995). *De la text la acţiune. Eseuri de hermeneutică.* Bucuresti: Editura Humanitas.

Riva, M. (2011). *Il futuro della letteratura. Lopera letteraria nell'epoca della sua.* (ri)producibilità digitale, Scriptaweb.

Romary, L. (2009). *Questions & Answers for TEI Newcomers.* Retrieved 05 07, 2015, from Jahrbuch für Computerphilologie: http://computerphilologie.de/jg08/ romary.pdf

Simionescu, R. (2012). Romanian deep noun phrase chunking using graphical grammar studio. *Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language"* (pp. 135–143). Moruz, M. A., Cristea, D., Tufiș, D., Iftene, A., Teodorescu, H. N. (eds.).

Trevisan, M. (2008). *A natural language generator for QueryTool.* (KRDB Research Centre Technical Report). Bozen, IT: Free University of BozenBolzano.

Vanhoutte, E., & Van den Branden, R. (2010). *The Text Encoding Initiative.* Marcia J. Bates and Mary Niles Maack (eds.), Encyclopedia of Library and Information Sciences .

Vlad, C. (2000). *Textul aisberg.* Cluj-Napoca: Casa Cărții de Știință.