AUTOMATED TAG EXTRACTION & CLUSTERING IN DOCUMENTS CONTAINING COMPOSITIONAL PHRASEMES

Vera Danilova, Xavier Blanco, Dmitry Stefanovskiy

Abstract: *This article aims to present the results of clustering in documents, extracted from Internet and related to compositional phrasemes (pragmatemes). We are studying conditions (situation, context), which can stipulate presence of these units in a text. Pragmateme's structure and functioning particularities are taken into consideration. An important objective of the work is selection of an adequate algorithm for tag extraction and clustering, so that we can further compare and apply the results, obtained for different languages.*

Keywords: *pragmateme, compositional phraseme, tag extraction, clustering analysis*

ACM Classification Keywords: *I.2.7. Natural Language Processing*

Introduction

The present work is dedicated to defining context/situations, characteristic of compositional phrasemes (pragmatemes) usage, by means of tag extraction and clustering results analysis.

The description of the term "pragmateme" is taken from [Mel'čuk, 1995] and [Blanco, 2010]: the meaning (or signified) of a pragmateme is not freely built from a specific conceptual representation, though it can be a regular sum of meanings for lexemes A and B. Thus, the meaning of these structures isn't free and cannot be replaced by any other meaning. There are two types of phonetic representation of the phrase (or signifier): it is not freely built from the signified and regular (in such case the meaning and the form of the phrase are totally limited by the situation) or it is relatively freely built (there are several synonymic forms of phonetic representation, regulated by the rules of the language).

A pragmateme presents a complex semiotic sign: text, which is often accompanied by a correspondent image. This combination is used to communicate a message of certain content (prohibition, indication etc.).

Such pragmatemes as routine or conversational formulae are used for stereotypical social interaction. They include discourse formulae, opening and closing conversations, psycho-social formulae [Nunes, 2007]. Some proverbs also can be understood as pragmatemes, according to [Pastor, 1995] point of view.

This work is a part of the project "Compositional pragmatic phraseology" of MCuT being accomplished at the Department of Romance Languages at the Autonomous University of Barcelona [MCyI, 2010-2012]. Our research has been carried out within the framework of R programming language system [R, http].

Problem setting

At this stage is selection of an adequate algorithm to carry out clustering processes so that we can obtain an appropriate distribution of context words.

The general purpose of defining contexts for pragmatic phrasemes is improvement of automatic translation for texts, containing these units.

In prospect we plan to dedicate our study to the functioning of pragmatic phrasemes in intercultural context, because major part of them is related to the fundamental realities of different countries and certain situational/contextual regularities, relevant to the correspondent country, may be revealed.

Within the limits of this experiment only those pragmatisms, which can be represented by both text and graphic sign (e.g., road signs, indicators, such as *"No camping", "No parking"*), will be examined.

Indexing & clustering

To reveal the topics related to the given pragmatic phrasemes we use the following technique:

1) Extraction of documents from Internet depositories using the well-known Google search machine.
2) Construction of term list as the basis for document set indexing
3) Performing clustering applying selected terms

There are many approaches for constructing list of terms to be used in clustering/classification process. Entropy based methods select the most informative words in the corpus. Statistics based methods use non-uniformity of word distribution among the documents. We use criterion of term specificity. Namely we select words whose frequency in a given document set exceeds their frequency in the General Lexis by K times. In our situation the General Lexis was a list of words taken from the British National Corpus of documents. Coefficient K is titled as a word specificity. The higher the K is the fewer words will be extracted from a given document set. This procedure is easily performed by the program LexisTerm [Lopez, 2011]. We suppose that such an approach to term selection is relevant to the goal of the research.

At present there is a large variety of clustering methods, which belong to different groups: 1) hierarchy based methods where number of clusters is not fixed 2) exemplar based methods where number of clusters is given in advanced and 3) density based methods where number of clusters is determined automatically [Alexandrov, 2007]. The methods of the second group are the most common and simple, but they cause more errors and critical comments (the reason is the fixed number of clusters). Nevertheless we use K-means within the framework of the procedure where K varies. Namely we increase K since K=2 till the moment when Dunn criterion (measure of cluster validity) reaches its maximum. This type of approach is popular in Machine Learning.

Experiment

Our corpus consists of 40 documents, containing English pragmateme "Camping prohibited" (all documents represent two topics by default: rules for campers and general information for tourists). We used this amount of documents, because at this stage it's essential to be able to check the results manually.

The procedure of indexing was implemented with the program LexisTerm two times with the coefficients of specificity K=10, 50, 100, 500. The coefficient of K=100 was selected as optimal and the resultant number of terms was 254. This result was adjusted and applied for the further clustering (non-relevant terms were excluded and some relevant ones were added).

Our assumption is that the documents shall form cluster for each context basing on key-words. Clustering was accomplished in R programming language system. There were created several interrelated scripts, setting the variables for each text and stop-words and two main scripts for clustering and key-word test. A term document matrix was compiled on the basis of the resulting data set. Punctuation and stop-words were eliminated. The value of each matrix point was divided by the sum of the values in a correspondent row to normalize the vectors. The optimal number of clusters (two) for the given set of documents was obtained using Dunn's partition coefficient. The information on the resulting clusters is given in the below.

	size	max_diss	av_diss	diameter	separation
[1]	261	1.0613199	0.75311746	1.4142136	0.1732051
[2]	59	0.6123724	0.09783235	0.6123724	0.1732051

Graphic representation of clustering results confirmed the aforementioned calculations (Fig. 1): the first cluster contained 261 elements (rules and regulations for campers) and the second - 59 elements (information for tourists on different types of accommodation).

The key-words were first manually selected from the matrix, according to their degree of occurrence in the given documents. The result was compared with the sequence of interrelated key-words, obtained automatically: a formula was derived, applying coefficients, obtained in calculation of polynomial regression (coefficients, expressing the dependencies between key-word frequencies). There were found two sets of key-words, marking the first cluster: x2:x3 (property: parking) = 1.34120743 and x2:x4 (property: trail) = 13.00345224. The presence of each set in a document stipulates the assignment of the latter to the first cluster (regulations (code) for campers). The rest of the documents shall be automatically assigned to the second cluster (general information for tourists) respectively.

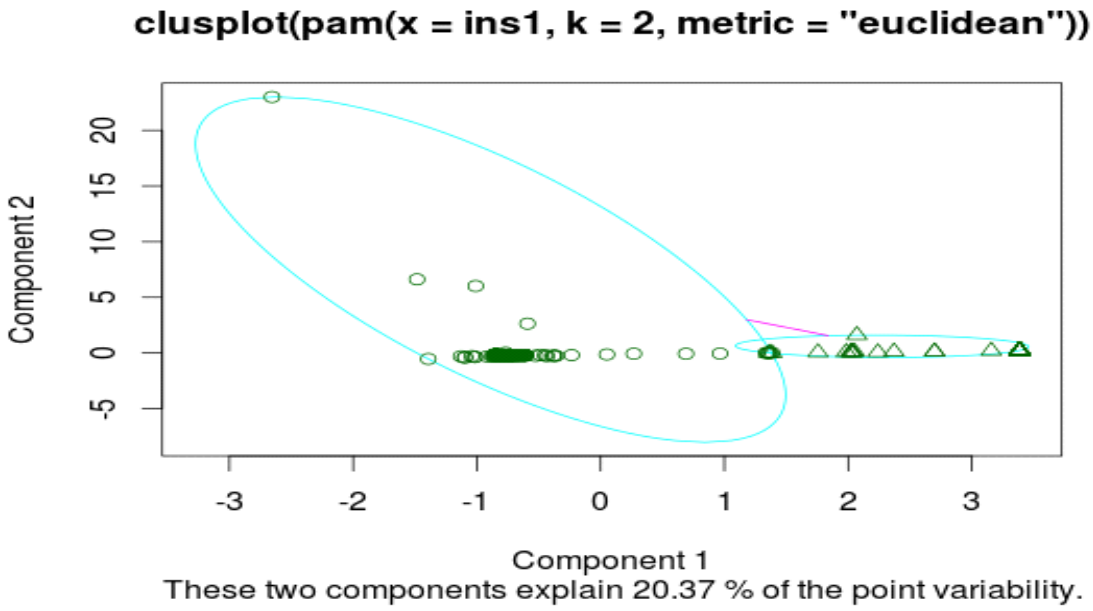

Fig.1 Results of clustering

Conclusions

The results of the present pilot experiment represent a successful automatic assignment of the given documents to clusters (and contexts respectively) by means of R programming language system. They coincide with the results of manual distribution.

In prospect we plan to study the distribution of pragmateme contexts in different languages so that we can reveal regularities in the use of these units in different countries. The key-words will be presented not only by words, but also by fixed word combinations, because they tend to be more useful for the interpretation of a context. Also we plan to take into account the synonymy of words and collocations.

Bibliography

[Alexandrov, 2007] M. Alexandrov, P. Makagonov. Introduction to Technique of Clustering. Proc. of 3-rd Summer School on Comp. Biology, Brno, pp. 55-80, 2007.

[Blanco, 2010] X. Blanco. Los frasemas composicionales pragmáticos. Mejri, S. Et Mogorrón, P. (dir.): Opacité, Idiomaticité, Traduction, Universitat d´Alacant, 2010.

[MCyI, 2010-2012] Compositional pragmatic phraseology. Project, Spanish Ministry of Science and Innovation (MCyI), Reg.No. FFI2010-15229, 2010-2012.

[Lopez, 2011] R. Lopez, M. Alexandrov, D. Barreda, J. Tejada. Lexisterm – The Program For Term Selection by the Criterion of Specificity. Proc. of 4th Intern. Conf. on Intell. Inform. and Eng. Systems, Poland, ITHEA Publ. House, 6 pp., 2011.

[Mel'čuk, 1995] I. Mel'čuk. "Phrasemes in Language and Phraseology in Linguistics" In Idioms: Structural and Psychological Perspectives. Everaert, M. et al. ed. Hillsdale, New Jersey and UK: Lawrence Erlbaum Associates. 167-232, 1995.

[Nunes, 2007] S.C. Nunes. The nature of fixed language in the subtitling of a documentary film. (Doctorate thesis). Escola Superior de Educação – Instituto Politécnico de Bragança , Universitat Rovira i Virgili, 2007.

[Pastor, 1995] C.G. Pastor. Un estudio paralelo de los sistemas fraseológicos del inglés y del español. (Doctorate thesis), Universidad Complutense de Madrid. Málaga: Secretariado de Publicaciones de la Universidad de Málaga, 1995.

[R, http] R-study http://www.rstudio.org/

Authors' Information

Vera Danilova *– Autonomous University of Barcelona, the Department of Romance Languages, PhD program student;*

e-mail: maolve@gmail.com

Major Fields of Scientific Research: Automated translation, Pragmatic structures

Xavier Blanco – *Cathedratic University Professor (Full Professor), fLexSem Research Laboratory (Fonètica, Lexicologia i Semàntica), Department of French and Romance Philology, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: Xavier.Blanco@uab.cat

Major Fields of Scientific Research: lexicology, lexicography, machine translation

Dmitry Stefanovskiy *- Assoc. Prof., Ph.D, the Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571;*

e-mail: dstefanovskiy@gmail.com

Major Fields of Scientific Research: Modelling for Analysis of Socio-Economic Processes