Linguistic processing in lattice-based taxonomy construction

Anastasia Novokreshchenova, Maria Shabanova, Dmitry Zaytsev and Nina Belyaeva

Higher School of Economics, Moscow, Russia **

Abstract. Building a lattice-based taxonomy over a text corpus with formal concept analysis (FCA) methods requires preliminary text processing that would enable construction of a context. We consider several natural language processing methods aimed at automatic attribute acquisition from texts. In particular, we derive attributes of three types: frequent words, latent topics and named entities. Afterwards, we construct a context for each type taking documents in the corpus as a set of objects. Then the corresponding concept lattices are built and pruned with the help of stability index in order to improve the readability of the diagrams. The proposed technique is illustrated on a collection of 26 texts in English dealing with political domain. In this case, the technique serves as a tool for deeper understanding of the interests of different political actors producing political texts by clarifying the connections between notions they use in them.

1 Introduction

Constructing a taxonomy over a text corpus requires preliminary text processing. In this paper we explore capabilities of several natural language processing methods for extracting a set of attributes from the documents each taken as an object. The first method is keyword extraction which is based on the Vector Space Model [1]. The second technique is Latent Dirichlet Allocation, an extension of probabilistic latent semantic analysis (pLSA) [2], which allows one to extract latent topics from word distributions. These two methods are based on the "bag-of-words" assumption—that the order of the words in documents can be beglected. The third technique we use for attribute extraction is Named Entity Recognition (NER), which involves linguistic analysis and part-of-speech-tagging (POS-tagging).

Here, we apply these three methods to a collection of 26 texts dealing with political domain. After constructing an object-attribute matrix and building a lattice, fitering based on the stability index is applied in order to improve the readability of the diagrams.

** This work is supported by project 10-04-0017 of the Scientific Foundation of the State University-Higher School of Economics "Discrete mathematical models for political analysis of democratic institutions and human rights".

Our primary data consists of 26 texts corresponding to speeches of European and American leaders and spokepersons during the 2007-2010 period addressing their relations with Russia. These documents were collected by experts in political science as part of an interdisciplinary research project. From the viewpoint of political studies, the aims were (1) to define the context in which Russia is addressed in the speeches of Western leaders and international organizations, (2) to analyze the role and importance of democracy and human rights agenda. However, this text corpus is used here primarily as the testing ground for different methods of text analysis, and will be expanded in further research in order to reach a higher level of validity for the conclusions.

2 Building a context with frequent words

In The Vector Space Model (VSM) [3] documents are represented as vectors of features consisting of the weights of the terms that occur within the collection. The term weighing we perform here is based on word frequences. This factor is called *term frequency* of a term j within a document i:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{ik}},$$

where n_{ij} is the number of occurrences of the word j in document i; the denominator is the total number of words in the document i.

Constructing the Context and Stability-based Pruning Formal Concept Analysis (FCA) methods have been proven useful for representing a meaningful structure of a given knowledge community (such as a scientific community [5]) in a form of a lattice-based taxonomy (see [4] for FCA preliminaries). While constructing the context we let the documents be a set of objects and words that are most frequently used within each document be a set of attributes. The context construction involved two common techniques: stemming (using the Porter stemmer [6]) and elimination of stop words.

Using the database of 26 political texts we have built a context made of documents and terms mentioned in each document most frequently. In particular, we took 20 most popular terms for each document according to *tf* measure. The resulting context contains 26 documents and 249 terms which yields a lattice of 453 formal concepts.

The number of concepts is too large to be shown in a diagram. In order to obtain intelligible diagrams, we apply the pruning technique based on the notion of concept stability [7]. For a formal context $\mathbb{K} = (G, M, I)$ and a formal concept (A, B) of the context $\mathbb{K}$ the stability index is defined as follows:

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}}$$

The basic stability-based pruning method is to remove all concepts with stability below a fixed threshold. Of course, stable concepts (i.e., satisfying the

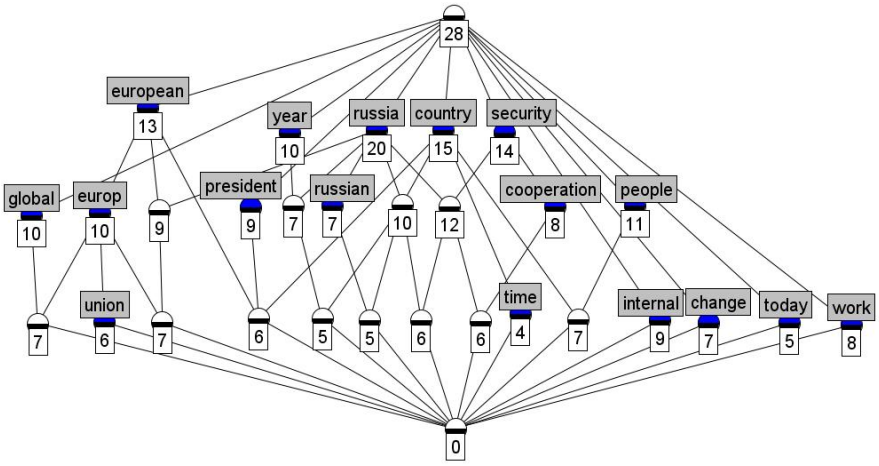

Fig. 1. The most stable concepts in the political dataset. Figures in squares show the sizes of concept extents.

chosen stability threshold) do not always form a lattice. However, this fact does not influence interpretation of results.

The reduced substructure featuring the 31 most stable concepts is presented in Fig. 1. The diagrams were produced with Concept Explorer.[8] From this diagram it is possible to provide the following description concerning the area of discourse by European countries and the US the situation with Russia. The term "russia" is obviously a central issue—this word is one of the most frequent in 20 documents out of 26. It is also a parent for several associated subtopics such as concepts with intents {"security", "russia"} and {"european", "russia"}. On the whole, it may be concluded that according to word frequencies security issues and the relationships between Russia and Europe, as well as some global problems, are the most discussed topics.

3 Building a context with latent topics

In this section we address an issue of probabilistic modeling of text. Its basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Such probabilistic topic approach to document modeling was introduced as probabilistic Latent Semantic Analysis (pLSA) [2]. Here we apply an extended probabilistic model Latent Dirichlet Allocation (LDA) [9] to our dataset of 26 political texts taking the number of latent topics equal to 20.[1] As previously, before running the model, all words were stemmed with Porter algorithm and stop-words were eliminated.

[1] The experiments were done with Matlab Topic Modeling Toolbox: http://psiexp.ss.uci.edu/research.

Table 1 presents word distributions for several of the obtained topics. The words in the columns are arranged according to probability assignment of a word to a certain topic.

Table 1. Several Topics Obtained with LDA Model

Economics	Democracy	America	Georgia	Energy	Global	Security	Europe
crisi	right	nation	georgia	russia	global	secur	europ
presid	human	unit	russian	russian	polici	today	european
finance	govern	nuclear	intern	interest	institut	member	union
econom	peopl	america	georgian	energy	issu	challeng	idea
system	democraci	american	territori	medvedev	effort	afganistan	point
govern	work	interest	south	issu	respons	strateg	thing
reform	women	futur	order	rule	achiev	face	bring
propos	democrat	weapon	process	trust	approach	matter	global

After applying the LDA model to the text corpora we construct a context taking topics as attributes for the documents—a topic is assigned to a document if the total number of times that words in this document were assigned to a particular topic exceeds a fixed threshold. The resulting substructure of the corresponding pruned lattice is presented in Fig. 2. Lists of words in squares represent the first six words assigned to a topic with the highest probability.

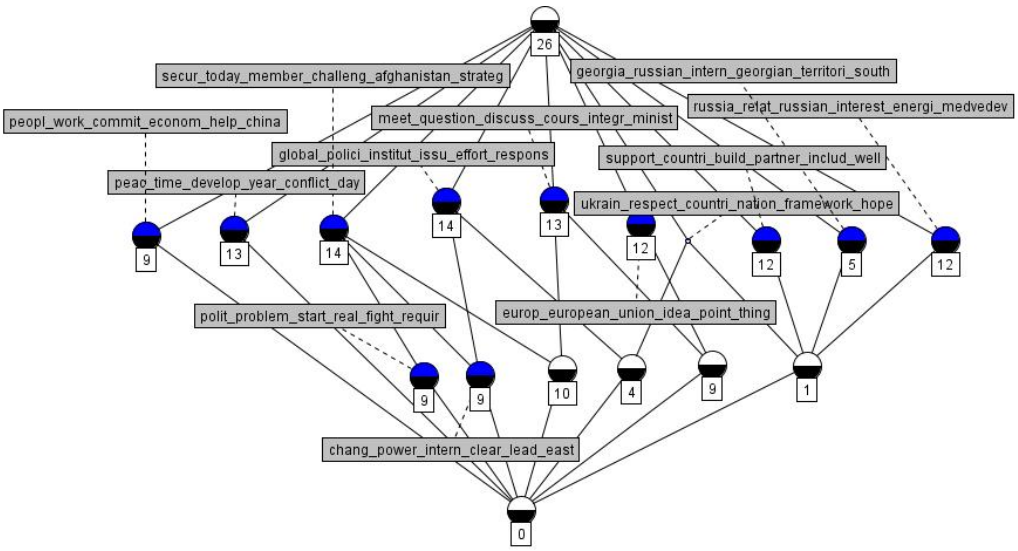

Fig. 2. The lattice of the 17 most stable concepts of a context built from 26 documents and 20 topics.

According to this lattice the most actual topics are those connected with European Union (topic represented by terms “europ”, “european”, “union”, “idea” and assigned to 12 documents), global problems (“global”, “polici”, “institut”, “issu”, “effort”, 14 documents) and security issues (“secur”, “today”, “member”, “challeng”, “afganistan”, “starteg”, 14 documents), as well as energy resources (“russia”, “relat”, “russian”, “interest”, “energi”, “medvedev”, 12 documents). There is a concept corresponding to the topic of Russian-Georgian conflict which contains 5 documents. In addition there is an isolated concept related to economic development of China (“peopl”, “work”, “commit”, “econom”, “help”, “china”, 9 documents).

4 Building a Context with Named Entities

Name Entity Recognition (NER) is the process of finding mentions of fixed types of information in human language. From the 26 texts, we extracted 38 paragraphs that touch issues related to Russia. The paragraphs were processed with GATE[2] system and three types of named entities were extracted—names of persons, organizations and geographical objects. We construct a context taking paragraphs as objects and organizations, persons and locations mentioned in a certain paragraph more than a fixed number of times as attributes. The resulting pruned lattice is presented in Fig. 3.

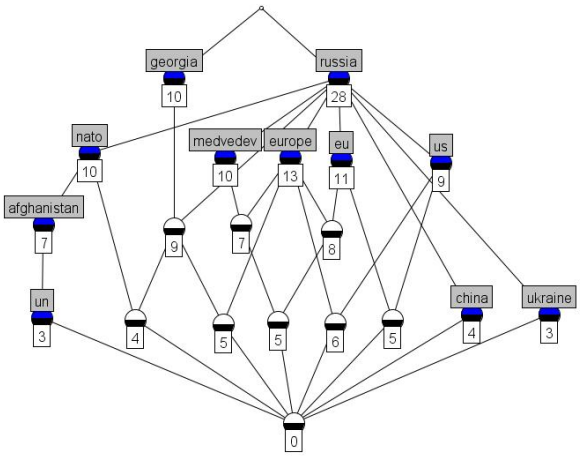

Fig. 3. The lattice of the 21 most stable concepts of a context built from paragraphs and named entities.

From this lattice it can be noticed that Europe and European Union are the most discussed topics as it appeared in previous results. It is worth noticing that

[2] GATE: General Architecture for Text Engineering: http://www.gate.ac.uk/

the United Nations (UN) is mentioned only in the context of Afghanistan, which in its turn is mentioned solely in the context of NATO. Topics corresponding to China and Ukraine form isolated concepts.

5 Conclusion

The first method based on frequent words allowed us to identify what questions are raised most frequently by European and American leaders while talking about Russia, whereas latent topic modeling allowed us to specify and describe these issues more thoroughly. The results obtained with named-entity lattice are rather similar. We concluded that it could be more useful to merge NER with the LDA model, for instance, by taking named entities as a set of objects and topics as a set of attributes. Expanding the corpus of the texts and combining NER with the LDA model can be useful in testing the hypotheses addressed in political studies.

On the whole, building lattice-based taxonomies using various language processing techniques is promissing for obtaining additional knowledge regarding open and hidden intentions of political actors. What is important for social sciences is that the knowledge is obtained without collecting any additional information and by a "neutral" tool—through deriving deeper connections between notions used by the actors in their texts.

References

1. Salton G., Wong A., and Yang C. A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, pp. 613–620 (1975)
2. Steyvers M., Griffits T. Probabilistic Topic Models. Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum (2005)
3. Jurafsky D. and Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall (2000)
4. Ganter B. and Wille R. Formal Concept Analysis, Mathematical Foundations. Springer Verlag, Berlin (1999)
5. Roth, C., Obiedkov, S., Kourie, D.G. Towards concise representation for taxonomies of epistemic communities. In: Yahia, S.B., Nguifo, E.M. (eds.) Proc. CLA 4th Intl. Conf. on Concept Lattices and their Applications. LNCS/LNAI, vol. 4923, pp. 240–255. Springer (2006)
6. Porter M. An algorithm for suffix stripping. Program, vol. 14, pp.130–137 (1980)
7. Kuznetsov, S., Obiedkov, S., Roth, C. Reducing the representation complexity of lattice-based taxonomies. In: Priss, U., Polovina, S., Hill, R. (eds.) Conceptual Structures: Knowledge Architectures for Smart Applications: 15th Intl Conf on Conceptual Structures, ICCS 2007, Sheffield, UK. LNCS/LNAI, vol. 4604, pp. 241–254. Springer (July 2007)
8. Yevtushenko, S. System of data analysis "Concept Explorer". (In Russian). Proceedings of the 7th national conference on Artificial Intelligence KII-2000, p. 127–134, Russia, 2000
9. Blei D., Ng A., Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research, vol. 3: pp. 993–1022 (2003)