IJCTA, 9(30), 2016, pp. 33-40

Identifying Semantic Relations Between Named Entities by Latent Semantic Analysis with Different Methods of Constructing a Vector Representation of Text Documents in Russian

Nikolai Valerievich Bradis* and Dmitrii Aleksandrovich Sytnik*

Abstract : The high-priority task of the information news flow processing system is to identify the named entities, namely the persons, organizations and geographical objects as well as the relations between them. The types of defined relations depend on the task that the extraction system and the researcher face. It is possible to identify the frequency links based on the information about the number of documents, in which a single object occurs in conjunction with another one, the semantic links that show the relationship between objects, are allocated based on the results of text parsing, and use the specified parameters of possible relationships in the form of rules. This paper will focus on latent semantic analysis as a way to extract the latent semantic relations between named entities at various ways of preparing data for analysis.

Keywords : News flow, named entity, latent semantic analysis, LSA, singular value decomposition, text analysis, TF-IDF, automatic text processing, occurrence frequency.

1. INTRODUCTION

This work is dedicated to the identification of latent semantic relations between named entities identified from the texts of news flows. This paper does not consider the process of identifying named entities from texts. You can learn about this process in more detail from works [1-2]. The process of identifying latent semantic relations consists in using latent semantic analysis (LSA) and transiting from the vector-specific to the semantic space of documents.

The data of news flows from the information online portals gazeta.ru, lenta.ru, rbk.ru, ria.ru and vedomosti.ru were used as the initial data for the study. An unstructured message published at a certain time (date and time of publication) and taken from a particular source (a mass media outlet) is understood as a unit of news flow (a piece of news). Every news item should describe any past event and include information about the participants of this event. After loading, a pure text without html-tags and any other control footing is extracted from each news article. Then a morphological analysis of text words is conducted, and named entities are allocated. Then, the selected text is saved into a database for further processing. All the news from different feeds are chosen at the same time interval, from which we can conclude that they belong to the same political events that took place during this period.

The latent semantic analysis [3] is used worldwide to determine the semantic similarity of documents, terms (words and key phrases), as well as to index and cluster document corpuses. In this paper, the LSA is only used to determine the degree of semantic similarity between the named entities. However, the method itself is statistical and

* Complex Systems LLC Russia, 170021, Tver, Skvortsova-Stepanova Street, 83

is based on the words occurrence frequency within documents, that is, it is based on the hypothesis of a "bag of words", the meaning of which is that the order of documents in a corpus and of words in a document is not important for determining the theme of the document.

The purpose of this paper is to use the LSA method for detecting latent semantic relations between named entities identified from news streams, and to test this method at different types of input data construction with different levels of their detailed elaboration. To solve this problem, it is required to conduct a primary processing of news texts in order to allocate the array of word forms, to remove stop words and stop characters, and to conduct a morphological analysis of text word forms. After the initial processing, it is necessary to identify the named entities and build a vector representation of documents.

After applying the method of singular value decomposition of matrix [4] of the vector representation of text corpus (the "terms to documents" matrix) and reducing the dimension of semantic space, it is necessary to calculate the strength value of semantic relations between the named entities using the method for determining the cosine distance between the respective vectors.

2. THE METHOD OF LATENT SEMANTIC ANALYSIS

General description of the method

The latent semantic analysis consists of three components: the first is a plurality of words and phrases (terms), the second is a plurality of documents, and the third is a hidden component that connects the first one with the second one, that is, a plurality of terms with a plurality of documents. The method consists in transiting from the vector space of terms and documents to their semantic space and in defining for each term and document their coordinates within the resulting semantic space.

The vector representation of text corpus is the "terms to documents" matrix, the rows of which correspond to the identified terms, and the columns — to the corpus documents. A cell of this matrix records the occurrence frequency of a term that corresponds to the row within a document that corresponds to the column. In addition, the matrix cells can include weights that take into account both the occurrence frequency of terms in the document as well as the occurrence frequency of this document within the whole corpus of documents. These weights may be obtained through the TF-IDF normalization [5]. It may be necessary to determine the importance of a word within the context of a document taken from the document corpus. It can reduce the weight of words that occur in the vast majority of the corpus documents, that is, the noise words for the given text corpus, and, on the contrary, increase the importance of rare words that are unique to a relatively small pluralities of documents from the corpus.

To transit from the vector representation to the semantic space and to obtain the hidden component, the LSA method uses the singular value decomposition of the "terms to documents" matrix [4]:

$$A = U \times S \times V^T \quad (1)$$

where matrices U and V are orthogonal, and S is a diagonal matrix consisting of singular values of the A matrix in decreasing order.

This decomposition is a very useful feature. If to leave in the diagonal S matrix only the *k* number of the largest singular values, and to leave in the U and V matrices only the columns corresponding to the selected values, then the product of the resulting matrices gives the best approximation to the original A matrix of krank. Given the provided singular value decomposition property, it becomes possible to interpret the results of the original "terms to documents" matrix decomposition as follows. The U matrix is a set of rows, each of which is a term vector within the semantic space. The V matrix is a set of document vectors. The S matrix is a set of components that connects matrices U and V, wherein the more the value of the component is, the stronger is the influence of the corresponding elements of the term vectors and documents on the final result.

This property of the singular value decomposition can reduce the dimensions of the semantic space by omitting small values and the corresponding thereto elements of term vectors and documents. When identifying connections between terms and documents, the reduction in the semantic space allows identifying the key components of the "terms to documents" matrix and ignoring the noise, as well as reducing costs fordefining relations between terms and documents.

The *k* value is selected experimentally and depends on the task set and the number of original documents. A too high ratio value is meaningless for the LSA, since the method loses its strength and becomes too sensitive to the differences between terms and documents. At a small k value, the ability to detect differences between the terms or documents worsens.

To determine the strength of the semantic relation between the terms or documents, it is possible to use a variety of methods: scalar product of vectors, cosine distance, Euclidean distance and Manhattan (chess) distance. Selecting a method for determining the strength of relation does not affect the essence of the LSA method.

Pre-processing of texts

Before performing the singular value decomposition of the "terms to documents" matrix, it is necessary to split the analyzed text into words, remove stop words and punctuation, conduct the remaining morphological analysis of word forms, define a list of terms, and build the "terms to documents" matrix.

The boundaries of words in the text are defined by spaces and punctuation marks, after which all the words are checked on entry into the pre-formed stop word dictionary. Stop-words are the text words, which do not carry the semantic and thematic load. They include prepositions, participles, interjections, particles, punctuation marks, and parenthetical words. The words found in the stop-words dictionary are removed from the text and do not take part in the future work of the method.

After removing stop-words, it is required to conduct morphological analysis of each word of the text in order to identify it uniquely and independently of the form, in which it is represented, as well as to count its occurrence frequency within the documents. For these purposes, the morphological analyzer pymorphy2 was applied [6], which uses the lemmatization method, the essence of which is to reduce the word forms to the lemma, that is, to the normal dictionary form.

Stemming is another method of morphological analysis, the essence of which consists in cutting off all variable parts of a word and extracting the unchangeable part thereof. This method is well suited to the English language, but is of little use for languages that use complex word formation, including the Russian language. The most famous stemming algorithm was proposed by Martin Porter in 1980 [7] and later adapted by him for various Indo-European languages, including the Russian language.

Comparing the correctness of the stemming and lemmatization algorithms work was provided by Mikhail Korobov through comparing Mystem 3.0 and pymorphy2 [8]. The comparison showed that depending on the type of original data and words for parsing, both implementations have their advantages and disadvantages.

To determine the list of named entities described in one word, the pymorphy 2 morphological analyzer is used, which in addition to the normal form of the word and the accompanying morphological information about it (part of speech, gender, number and case), provides information on the possible word's pertain to one of the types of named entities: geography object, organization, surname, name, or patronymic. It is very convenient in terms of performance, because there is no need to search for words in various dictionaries of organizations and geography objects, but rather to find a word only once in the dictionary used by the morphological analyzer. To determine the named entities described by more than one word, the statistical method of identifying significant phrases based on the calculation of the MI-measure (Mutual Information) [9] is used:

$$\mathrm{MI} = \log_2 \frac{f(a,b) \times \mathrm{N}}{f(a) \times f(b)}, \qquad (2)$$

where :

a, b are terms,

$f(a)$, $f(b)$ are the absolute occurrence frequencies of a and b terms within the text corpus,

$f(a, b)$ is the occurrence frequency of a term in a pair with the b term,

N is the total number of word forms within the text corpus.

After calculating this measure for the phrases consisting of 2 or more words, those of them are truncated, the measure value of which is below the threshold, and the rest are considered to be significant. A unity is selected as a threshold value. For each word of the significant phrase, there is information about the possibility of belonging to a named entity of a particular type, and on the basis of this information, the possible relation of the significant phrase to the named entity of a particular type is determined. In addition, to identify the multi-word named entities of the organization and geography object type, the lists of words-pointers are used. For organizations, these may be words and abbreviations of ownership patterns and words that indicate the type of organization: joint-stock company, Ltd., company, corporation, ministry, department, etc. For geographic names, the words-pointers of the following kind can be used: river, city, street, state, island, country, etc.

Preparation of the "terms to documents" matrix

The software implemented within the framework of this paper uses two options for building the "terms to documents" matrix. In the first building option, only the identified named entities are included as terms into the matrix, and therefore, only they are involved in constructing semantic space, while the remaining news text words are considered as the noise ones. In the second option, the matrix includes all the named entities and all the words that are not the stop-words. It is obvious that in the first option of matrix building, its dimension is much smaller than in the second, which should provide for a significant boost in the method performance. The second option makes it possible to obtain the more detailed semantic space with the corresponding losses in performance. Since the news streams tend to describe certain events occurring during a fixed period of time in specific locations and with specific participants, the first approach can produce acceptable quality analysis with a considerable acceleration of data processing.

For further use of the "terms to documents" matrix, it can be normalized using the TF-IDF normalization [5] to determine the weight of terms in specific documents, but from the viewpoint of determining the weight of a named entity, it may not be entirely correct. For example, if we talk about the Russian news feeds, it is clear that in an overwhelming number of news texts, the references to the country's president and geographical entity "Russia" will occur. However, it is impossible to reduce the weight of these terms, because they are not the noise ones and are related to specific events and another named entities. This may be true for the news feeds of other countries as well.

Application of LSA to various "terms to documents" matrices

Regardless of the method for constructing the "terms to documents" matrix, the method of singular value decomposition and semantic space dimension reduction is applied thereto, which were described in detail in the above-mentioned general description of the method. As a result, a vector of coordinates in the semantic space is determined for each term and document.

When constructing the "terms to documents" matrix, which contains only the identified named entities as terms, the strength of semantic relation between all terms is calculated by the method of cosine distance calculation (a cosine of the angle between the vectors):

$$\cos \phi = \frac{\vec{a}.\vec{b}}{|\vec{a}|.|\vec{b}|} \tag{3}$$

The cosine of the angle between the vectors is determined by the ratio of the scalar product of vectors to the product of their lengths.

The choice of the cosine distance calculation is due to the lack of the necessity in result normalization.

When constructing the "terms to documents" matrix, which contains as terms not only the identified named entity, but all the words except for the stop-words, the strength of semantic relation is calculated only between the named entities according to the formula (3), and the vectors of all other terms are not taken into account.

The difference between the above methods consists in that when restoring the hidden component (the S matrix in formula (1)) and the semantic space vector, the first approach uses only the named entities, and the second approach uses the named entities and all other terms.

3. RESULTS

The work result is a software implementation of the described method and its testing on real news streams. To evaluate the LSA method performance at various methods of constructing the vector space of documents and with/without the TF-IDF normalization, six random news articles published on August 4, 2016 were selected. Such a small number of articles is due to the necessity in visual evaluation of the method performance results on a small amount of data. Table 1 provides a list of news articles that were used for the analysis.

Table 1. List of news articles for analysis

N	*Article titles*	*Source*
1.	The Syrian group demands prisoners exchange for the Mi-8 crew bodies	http://www.rbc.ru/politics/04/08/2016/57a3a2df9a794727cfcce351
2.	The United Russia party called on the Ministry of Education to stop scaring the nation with the reduction in budget places and scientists	https://lenta.ru/news/2016/08/04/minobr/
3.	Klintsevich predicted Russia's refusal to replace the candidacy of Ambassador to Ukraine	https://lenta.ru/news/2016/08/04/klinc/
4.	The petition for the resignation of Medvedev gained 100 thousand signatures per day	http://www.gazeta.ru/politics/news/2016/08/04/n_8959577.shtml
5.	Peskov commented on the size of salaries among teachers in connection with the Medvedev's statement	https://lenta.ru/news/2016/08/04/peskov_medved/
6.	Zurabov was removed from the post of Putin's representative for economic relations with Ukraine	https://lenta.ru/news/2016/08/04/zurabov/

In total, 28 named entities were identified from the given news articles. The number 3 (*a* 2-fold reduction in space) was chosen as the value for the coefficient of the semantic space dimension reduction *k*. All relations between the named entities, the strength of which (a cosine of the angle between vectors within the semantic space) exceeds 0.81, were taken for the consideration.

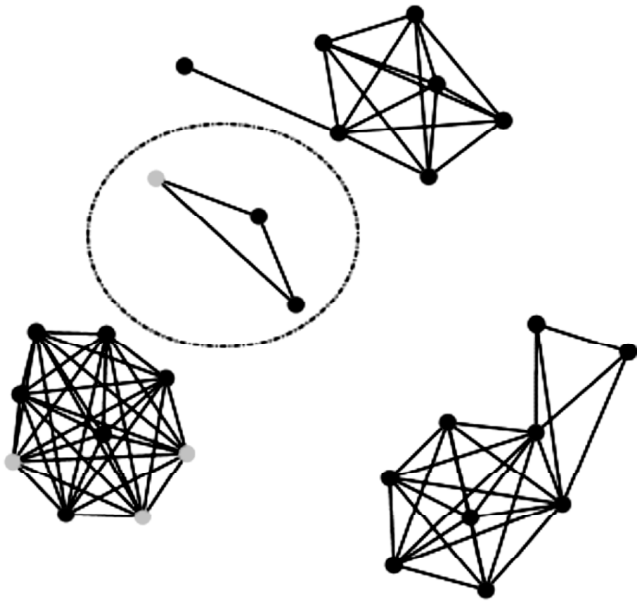

Fig. 1. The LSA method results, when only the identified named entities without using TF-IDF normalization are introduced into the "terms to documents" matrix. The required subgraph is highlighted in red.

To verify the correctness of the algorithm at different methods for constructing a matrix and using the TF-IDF normalization, three named entities were selected: Lebanon (a state on the eastern shore of the Mediterranean Sea), Damascus (the capital of Syria bordering with Lebanon) and Hezbollah (the Shiite organization and the political party advocating for the creation of an Islamic state in Lebanon by analogy with Iran, which was recognized as a terrorist group by many countries). All three named entities have similar features — they occur once only in the first article, wherein this article never mentions the named entities that occur in the texts of other news articles. Besides, the essence and the participants of this article do not interact with the other listed articles. In these circumstances, it is logical to assume that the output is supposed to be a separate from the above graph with the given named entities.

Figure 1 shows that it is possible to conclude that the assumption on that the output graph will contain separate subgraph that connects only the named entities of Lebanon, Damascus and Hezbollah was confirmed when applying the LSA method to the matrix containing only the named entities.

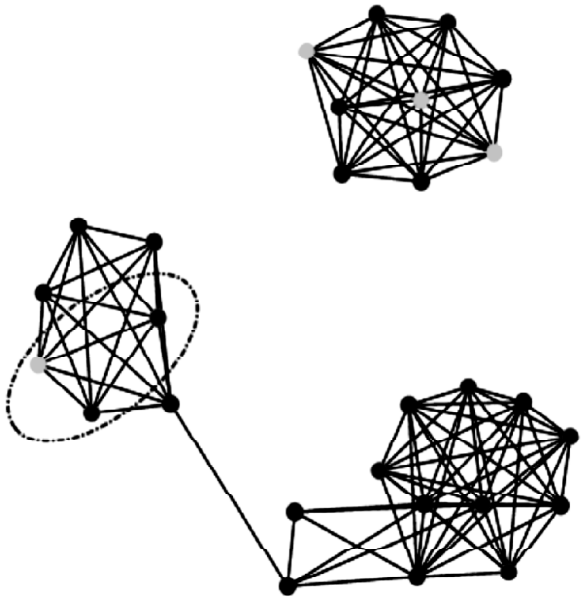

Fig. 2. The LSA method results, when the identified named entities and all words except for the stop-words without TF-IDF normalization were introduced into the "terms to documents" matrix. The required subgraph is highlighted in red.

Figure 2 makes it clear that the named entities selected for verification were associated with the named entities from other news reports, because the analysis used the "terms to documents" matrix that contained all terms, rather than the named entities only. At closer examination, it becomes clear that the formation of new relations is affected by the following words: business, read, statement, and document. These words are present in the majority of these documents and carry no general semantic load.

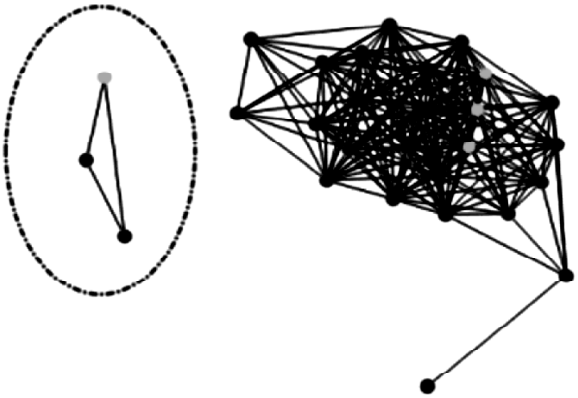

Fig. 3. The LSA method results, when the identified named entities and all words except for the stop-words with TF-IDF normalization were introduced into the "terms to documents" matrix. The required subgraph is highlighted in red.

Figure 3 makes it possible to conclude that when introducing the TF-IDF normalization, the problem of the noise words influence was eliminated, however provided that the identified named entities occur only once and in one article. For the named entities that occur in several articles, their weights within the "terms to documents" matrix are reduced and become commensurate with the weights of the remaining words on such scanty data, which is why the strength of relations between these objects increases due to the equivalence of their interrelations to the relations with the rest of the words. At sufficiently big data, this problem can be eliminated due to a rare use of named entities as compared to the rest of the words, but it can be correct not for all entities, which was already discussed in the introduction.

4. CONCLUSION

The proposed method can detect semantic relations between named entities from the flows of unstructured text information. Testing this method gave acceptable results on the news flows for a specific period of time. In addition, the analysis results were compared using different methods for constructing the "terms to documents" matrix, and a conclusion was made on that using only named entities in the news section as the analysis input data ensures quite acceptable results and an increase in the method performance as compared to the of all terms for analysis.

The main disadvantage of the LSA method is the assumption that the words in the documents are distributed according to the laws of normal distribution, which is not always the case. For this reason, in the future we should consider using a probabilistic method of latent semantic analysis and its modifications [10].

This method takes into account the exact number of references to a named entity in the text of the document, even though the entity can be determined by a pronoun or indirectly (he, him, she, this person, in this country, etc.). This problem cannot be solved without the application of parsing [11], grammar bonds [12-13], which will significantly reduce the method performance, but it will significantly improve the results of the analysis, and provide many other useful results.

The used method can be applied in the future to the latent semantic indexing of news texts and semantic search for the documents relevant (texts close in the strength of semantic relations) to the request. However, this work had the task of finding relations between the named entities.

5. ACKNOWLEDGEMENTS

The Ministry of Education and Science of the Russian Federation (Grant Agreement 14.579.21.0088, a unique identifier RFMEFI57914X0088 for Applied Scientific Research) financially supported this work.

6. REFERENCES

1. Kuznetsov, I.P. (2012) The Methods of Discovery of Objects and Their Links Presented Implicitly in Texts. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue 2012"*.

2. Khokhlova, M.V. (2010). *Issledovanieleksiko-sintaksicheskoysochetaemosti v russkomyazyke s pomoshch'yustatisticheskikhmetodov: dissert at siyana so i skanieuchenoy stepenikandid at afilologicheskikhnauk* [The Study of Lexical and Syntactic Compatibility in the Russian Language with the Help of Statistical Methods (Ph.D. Thesis)].St. Petersburg:St. Petersburg State University.

3. Landauer, T., Foltz, P.W., &Laham, D. (1998). Introduction to Latent Semantic Analysis.*Discourse Processes, 25*, 259-284.

4. Press, W.H., Teukolsky, S.A., Vetterling, W.T., &Flannery, B.P. (1992). Singular Value Decomposition. In W.H. Press, S.A. Teukolsky, W.T.Vetterling, & B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed.). Cambridge: Cambridge University Press.

5. Vlavatskaya, M.V. (2015).Kombinat ornayaleksikologiya: funktsional'no-semanticheskayaklassifikatsiya kollokatsiy [Combinatorial Lexicology: Functional and Semantic Classification of Collocations].*Filologicheskienauki. Voprosyteoriiipraktiki, 11*(53), Part 1.

6. Kilgarriff, A. (2006). Collocationality (And How to Measure It).In *Proceedings of the Euralex International Congress*. Torino.

7. Salton, G.,&Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management, 24*(5): 513-523.

8. *Morphological Analyzer Pymorphy2*. (2013, February 14). Retrieved September 14, 2015, from https://pymorphy2.readthedocs.io/en/latest/index.html.

9. Porter, M.F. (1997). *An Algorithm for Suffix Stripping*. San Francisco: Morgan Kaufmann Publishers Inc.

10. Korobov, M. (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages. In *Proceedings of the Fourth International Conference on Analysis of Images, Social Networks and Texts, AIST 2015* (pp. 320-332).

11. Magerman, D.M.,&Marcus, M.P. (1990). Parsing a Natural Language Using Mutual Information Statistics. In *Proceedings of the Eighth National conference on Artificial Intelligence, August, 1990*.

12. Gaussier, E., Goutte, C., Popat, K., & Chen, F. (2002).A Hierarchical Model for Clustering and Categorizing Documents. In *Advances in Information Retrieval – Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02)*.

13. Meng, H., Luk, P., Xu, K., &Weng, F. (2002).GLR Parsing with Multiple Grammars for Natural Language Queries. *ACM Transactions on Asian Language Information Processing (TALIP), 1*(20), 123-144.

14. *Link Grammar*. (2004). Retrieved June 2, 2016, from http://www.link.cs.cmu.edu/link.

15. Protasov, S.V. (2006). Obuchenie s nulyagrammatikesvyazeyrusskogoyazyka[Learning from Scratch the Grammar Relations of the Russian Language]. In *10-ya natsional' nayakon ferentsiya poiskusstvennomu intellektu s mezhdunarodnymuchastiem KII-2006, Obninsk, Rossiya, 25-28 sentyabrya 2006 goda* [Tenth National Conference on Artificial Intelligence with International Participation KII 2006, Obninsk, Russia, September 25-28, 2006].