Recognition of a person named entity from the text written in a natural language

A V Dolbin¹, V L Rozaliev¹, Y A Orlova¹

¹ Volgograd State Technical University, 28, Lenina ave., Volgograd, 400131, Russia

E-mail: vladimir.rozaliev@gmail.com

Abstract. This work is devoted to the semantic analysis of texts, which were written in a natural language. The main goal of the research was to compare latent Dirichlet allocation and latent semantic analysis to identify elements of the human appearance in the text. The completeness of information retrieval was chosen as the efficiency criteria for methods comparison. However, it was insufficient to choose only one method for achieving high recognition rates. Thus, additional methods were used for finding references to the personality in the text. All these methods are based on the created information model, which represents person’s appearance.

1. Introduction

Named entity recognition belongs to the category of information retrieval tasks. Nowadays, there are many methods for extracting knowledge from the text. Despite the fact that this problem is relatively new, it is already one of the most important directions in the field of computer technology. A large amount of information in the text is described implicitly. For example, if extracted information relates to the elements of human appearance, maybe it becomes possible to conclude about person’s professional activities or habits. However, this task is not always easy as it seems because of the features of natural languages. Usually, the text is written in a free form and, therefore, it is quite difficult to interpret the meaning of words. [1]

This paper presents an investigation which method is better for recognition of the appearance elements of a man from the text, composed in a natural language. These two methods were selected for investigation since they are very common and show very good results in practice. In particular, the study was carried by the example of the Russian language. However, it is necessary to bear in mind that latent semantic analysis and latent Dirichlet allocation can be used for a large variety of natural languages.

2. The information model of person’s appearance

First of all, it was necessary to create an information model for providing a research basis. This model should met the following requirements:

- Scalability. Requirements could be changed after experiments.
- Model should fit easily into one of the common programming paradigms.
- Visibility. It should be easy to find necessary information.
Fullness. The model should contain detailed description of person’s appearance so that it was possible to visualize a human.

The frame representation language (FRL) was chosen as a formal description of person’s appearance since it has all necessary parameters. Figure 1 shows the information model of person’s appearance in FRL. Each slot of the frame has a gender determination procedure, which starts after the slot has been filled. The gender determination process relies heavily on morphological aspects of slot’s value. “M” is a predefined set of possible parameters for the appearance model.

```plaintext
{frame Human_Appearance
  (Height (value M) (IF_ADDED(detect_gender)))
  (Body (value M) (IF_ADDED(detect_gender)))
  (Head (value M) (IF_ADDED(detect_gender)))
  (Hair (value M) (IF_ADDED(detect_gender)))
  (Face (value M) (IF_ADDED(detect_gender)))
  (Forehead (value M) (IF_ADDED(detect_gender)))
  (Eyebrows (value M) (IF_ADDED(detect_gender)))
  (Eyes (value M) (IF_ADDED(detect_gender)))
  (Eyelashes (value M) (IF_ADDED(detect_gender)))
  (Nose (value M) (IF_ADDED(detect_gender)))
  (Lips (value M) (IF_ADDED(detect_gender)))
  (Chin (value M) (IF_ADDED(detect_gender)))
  (Teeth (value M) (IF_ADDED(detect_gender)))
  (Neck (value M) (IF_ADDED(detect_gender)))
  (Shoulders (value M) (IF_ADDED(detect_gender)))
  (Chest (value M) (IF_ADDED(detect_gender)))
  (Back (value M) (IF_ADDED(detect_gender)))
  (Legs (value M) (IF_ADDED(detect_gender)))
  (Arms (value M) (IF_ADDED(detect_gender)))
}
```

Figure 1. The person entity information model.

3. Latent semantic analysis and latent Dirichlet allocation

Latent semantic analysis (LSA) is a method for processing text information in a natural language, which examines the dependency of keywords on collection of documents (or sentences). The main purpose is to find a vector of documents, which is as close as possible to the specified keyword. [2]

The algorithm of latent semantic analysis is as follows:

- Select specified keywords. These keywords should describe the appearance of a person.
- Construct frequency matrix $A$, in which columns are documents and rows are keywords. Each element of the matrix should represent the occurrence of the keyword in the document.
- Apply a TF-IDF method on the basis of a frequency matrix to ensure that results are relevant. [3]
- It is necessary to use singular value decomposition on the basis of a resulting matrix. SVD separates the frequency matrix into a matrix of documents ($U$), a matrix of keywords ($V_t$) and a diagonal matrix ($S$).

$$A = U \cdot S \cdot V^T$$  \hspace{1cm} (1)

- A number of rows from matrix $V_t$ and columns from matrix $U$ can be dropped. This statement follows from the feature of singular value decomposition. Thus, less significant values will not take part in future computations. According to experiments, the best variant is to use $V_t$ with 2 rows and $U$ with 2 columns.
- Matrix $U$ contains the coordinates of keywords and $V_t$ – coordinates of documents. It is now possible to obtain the nearest documents, which has the same semantic meaning as a specified keyword. [4]
Latent Dirichlet allocation (LDA) serves the same purpose as LSA – it finds the dependency between the specified keyword and related documents, but in another form. LDA provides the information about how likely each keyword refers to each topic. On the other hand, the result contains probability of the document belonging to one of the topic. [5]

LDA takes at least three parameters as an input:

- Number of iterations.
- Number of topics.
- Text in a natural language.
- Set of keywords.

The algorithm of this method is as follows:

- Compose the list of keywords. These keywords should describe the appearance of a person.
- Assign topic for every word in each document. This assignment can be arbitrary or based on any external signs (in this case, it can be the same frequency matrix as in the LSA method).
- Calculate the percentage of topics, which relates to specified document \( p(t|d) \) and the probability that the word relates to current topic \( p(w|t) \). A specified word relates to the topic with probability:

\[
p = p(t|d) \times p(w|t)
\]

(2)

- Repeat the previous iteration of a predetermined number of times.

Both methods have much in common. The performance of these methods is low for the large document collection. Latent semantic analysis treats the text as a set of words that do not have any semantic links; so does latent Dirichlet allocation. However, this approach has an advantage in partial disambiguation from words with several meanings.

4. Person entity recognition

A person is one the most difficult named entity to recognize in the text. It is very easy to recognize the person if the text contains the name or surname of the entity. Nevertheless, such cases are extremely rare and references to an individual occurs differently. It is necessary to perform pre-processing of the text to increase the accuracy of latent semantic methods. LSA and LDA should perform searching of person’s appearance only in documents, which contain such description. [6, 9]

One common approach is to use contextual rules. Every rule represents a standard regular expression, which is constructs from the training sample as follows:

- Any mention of a person should be replaced with a special word \{PERSON\}.
- If the word in a training sample is an element of a human appearance, it should be set in its initial form.
- All other words should be replaced with its parts of speech.
- After processing the entire training sample, similar contextual rules should be combined using special characters. The ‘?’ symbol means that this position can be omitted, the ‘+’ symbol means that the position can be repeated one or more times in a row and the ‘|’ symbol represents logical “or”.
- Optionally, some phrases can be listed at the end of the training sample, which indicates that the sentence excludes the possibility of containing the entity.
Thus, the outputs are kind of regular expressions, which are applied to the text to define a named entity with it.

The next step is to resolve the reference of pronouns in the third person. Reference resolution of pronouns in the third form is one of the most common, but at the same time the simplest case, but completeness of information retrieval receives a great boost. This task can be considered as a problem of binary classification. Therefore, it is a good opportunity to use support vector machines (SVM). The simplest case of SVM was programmed using the linear classifier. Thereby the training sample must be compiled carefully to avoid the case of linear inseparability. [7, 8]

The list of parameters for support vector machines which were used for training:

- Number of sentences between antecedent and anaphora.
- Whether the antecedent is in the nominative.
- Position of the anaphora in the sentence.
- Position of the antecedent in the sentence.
- Number of nouns and pronouns, which are located in sentences with antecedent and anaphora.
- Is the antecedent and anaphora case matches.
- Is the antecedent and anaphora genus matches.
- Is antecedent and anaphora both in a plural or singular form.

5. Latent semantic methods comparison

All methods, which were mentioned in this paper, were written in Python 3. Pymorphy2 library was used for any morphological operations. It uses the “OpenCorpora” dictionary that counts about million and a half tokens at this moment.

A training set for support vector machines consists of five hundred samples. The training set for contextual rules also counts five hundred sentences. Latent semantic analysis and latent Dirichlet allocation do not need any training data.

Let \( N \) be a total number of person’s appearance elements, and \( N_r \) – the number of items found by one of the method. These two methods are compared by calculation of the completeness of information retrieval, which is as follows: [10]

\[
CIR = \frac{N_r}{N}
\]  

Experiment results are shown in table 1. Natural language texts used in the experiment were taken from areas that had been used to compile OpenCorpora, such as imaginative literature, blogs, Wikipedia and legal texts. Table 1 shows that if the search is performed only based on text passages that contain person references, the accuracy of recognition will increase greatly. In turn, the latent semantic analysis shows slightly better results for the recognition of human appearance in comparison with latent Dirichlet allocation regardless of the volume of the input information.

Table 1. The results of the comparison of latent semantic analysis and latent Dirichlet allocation.

| Number of documents | The approximate number of words per document | LSA without person recognition | LDA without person recognition | LSA with person recognition | LDA with person recognition |
|---------------------|---------------------------------------------|-------------------------------|--------------------------------|-----------------------------|-----------------------------|
| 1                   | 100                                         | 0.7                           | 0.68                           | 0.85                        | 0.83                        |
| 2                   | 100                                         | 0.69                          | 0.67                           | 0.84                        | 0.82                        |
| 5                   | 200                                         | 0.69                          | 0.67                           | 0.84                        | 0.82                        |
| 7                   | 200                                         | 0.67                          | 0.65                           | 0.83                        | 0.81                        |
6. Acknowledgments
The study was financially supported by RFBR research projects №15-07-05440, 15-07-07519, 15-37-70014, 16-07-00453.

7. Conclusion
Completeness of information retrieval on the subject of human appearance has been investigated using the methods of latent semantic analysis and latent Dirichlet allocation. The completeness of information retrieval was chosen as the efficiency criteria for methods comparison. Both methods have shown quite good results in practice. However, LSA is slightly better than LDA in case of recognition the appearance of a person. Preprocessing the text with finding references to the person greatly increases the accuracy of both methods.

The program was written in Python 3 with the usage of pymorphy2 library for morphological analysis. OpenCorpora was chosen as a main dictionary, with the help of which words were put in their normal forms.

References
[1] Dmitriev A, Zaboleeva-Zotova A, Orlova J and Rozaliev V 2013 Automatic Identification of Time and Space Categories in the Natural Language Text App. Com. 187-90
[2] Soloshenko A, Orlova J, Rozaliev V and Zaboleeva-Zotova A 2015 Establishing Semantic Similarity of the Cluster Documents and Extracting Key Entities in the Problem of the Semantic Analysis of News Texts Mod. Appl. Sci. 9(5) 246-68
[3] Ramos J 2003 Using TF-IDF to Determine Word Relevance in Document Queries Dep. Of Comp. Sci.
[4] Deerwester S, Sumais S, Furnas G, Landauer T and Harshman R 1990 Indexing by Latent Semantic Analysis Jour. of the Amer. Soc. For Inf. Sci. 41 391-407
[5] Blei D, Ng A and Jordan M 2003 Latent Dirichlet Allocation The Jour. Of. Mach. Lear. Res. 3 993-1022
[6] Ritter A, Clark S, Mausam and Etzioni O 2011 Named Entity Recognition in Tweets: an Experimental Study Proc. of the Conf. on Emp. Met. In Nat. Lan. Proc. 1524-34
[7] Keerthi S, Chapelle O and DeCoste D 2006 Building Support Vector Machines with Reduced Classifier Complexity Jour. Of Mac. Lear. Res. 7 1493-515
[8] Lee H, Chang A, Peirsman Y, Chambers N, Surdeanu M and Jurafsky D 2013 Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules Com. Lin. 39(4) 885-916
[9] Minkov E, Wang R and Cohen W 2005 Extracting Personal Names From Email: Applying Named Entity Recognition to Informal Text Proc. of the Conf. on Emp. Met. In Nat. Lan. Proc. 443-50
[10] Zhou B and Yao Y 2010 Evaluating information retrieval system performance based on user preference Jour. of Int. Inf. Sys. 34(4) 227-48