Multi-word Term Extraction from Domain-specific documents

Extended Abstract

ABSTRACT

This research presents a combination of automatic methods to extract multi-word term from a domain-specific collection of documents. Exploiting a combination of linguistic and statistical methods provides the basis to obtain a better quality of results in machine learning algorithms. A hypothesis of evaluation extracted candidates using associative words obtained from word2vec had been developed. The candidates were obtained by applying the topic modeling algorithms.

KEYWORDS

Term Extraction, Topic Model, Word2Vec, Thesaurus

1 INTRODUCTION

The development of a high quality terminological database is the first step to create a comprehensive domain-specific thesaurus. The quality of semantic relation between terms in the thesaurus depends to a great extent upon the high coverage of the terminological database. The most common way to solve this problem is to enrich the terminological database with time consuming manual terms extraction by domain assessors. However, this approach may cause a coverage loss. Automatic terms extraction can be exploited for developing domain-specific thesauri and ontology, entity and fact extraction, information retrieval. The observation shows, that single-word terms appear frequently, but automatic single-word term extraction is not enough for developing high quality thesauri. This research focuses on multi-word terms. Researchers are interested in getting a universal method with the highest percentage of probability of extracting terms.

Kiselev and co-authors [1] present the method of extracting the hyponymy/hypernymy relations from a dictionary definitions in Russian using lexico-syntactic patterns. We decide to apply the structure to extract terms increased the length of terms from 1-word to multi-word terms. In addition, the vector representation [2] for words, which is the state-of-the-art word embeddings for data experiments. We apply word2vec algorithm as an indicator of reliability in the automatic terms extraction. This overview research would not be complete if we did not consider the topic modeling, so we carry out the experiment, relying on a paper [3].

2 AN OVERVIEW OF METHODS AND EXPERIMENTAL RESULTS

2.1 An overview of methods

Linguistic methods. The accurate development of linguistic patterns leads to high accuracy term extraction. Templates can be designed with a number of features regarding morphology, punctuation, and the rules of sentence-construction, for instance, the development of templates for extracting terms from explanatory dictionaries. Terms can be extracted with high accuracy by using this method, but the coverage will remain low due to the lack of template flexibility.

Methods based on machine learning algorithms. In this case Machine learning algorithms for term extraction intersect *the classification task*, in particular, the separation of words and phrases into two groups: term-candidates and other words or phrases. *Topic models* are designed to identify the topic groups presented in the document collection, as well as the extraction of a list of terms belonging to each topic group.

2.2 The extraction approach

Manual labeling. At the first stage of work a list of poetics terms was compiled manually by domain assessors. This list contains 1 544 unique domain-specific terms. The area is poetics. Part-of-speech tagging was performed to identify phrases patterns. We have collected data in statistics, using the manually created list of terms and have defined the most frequently occurring multi-word terms.

Linguistic method. Series of patterns for a single word and multi-word terms were iteratively developed by using the pattern "entry – definition".

Automatic extraction. The current stage is divided into two parts:

(1) Exploiting the topic model for terms extraction;
(2) Checking the relevance of candidates applying for associative words getting from the Word2Vec.

To research the poetics area, the experts selected 31 domain-specific sources, converting them into electronic form. The sources were applied to test models. Morphological analysis of text collections and words lemmatization were performed at the stage of pre-processing. The templates obtained in the first step were used as extraction patterns. Training of Word2Vec model was conducted on the Russian Wikipedia data because it includes a connected terminology graph of domain-specific area. We obtained a vector representation of the main word and associative words with a measure of similarity, which is presented in cosine similarity between the vectors of a main word and associate words. Further, associate words were built for each candidate, obtained in step 1. The associative words are indicators of an adoption of the candidate in the list of domain terms. An indicative measure is positive when the terms from list of terms or new added terms have an exact match with a candidate.

REFERENCES

[1] Yuri Kiselev, Sergey Porshnev, and Mikhail Mukhin. 2015. Method of Extracting Hyponym-Hypernym Relationships for Nouns from Definitions of Explanatory Dictionaries (*in Russian*). *Software Engineering* 10 (2015), 38–48.

[2] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 746–751.

[3] Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 697–702.