DOCUMENT REPRESENTATION METHODS FOR NEWS EVENT DETECTION IN CROATIAN

Nikola Ljubešić, Željko Agić, Nikola Bakarić

Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{nljubesi, zagic, nbakaric}@ffzg.hr

ABSTRACT

Constant increase in the amount of available data in the world in general demands new organizational and representational ideas and approaches. Document clustering as a method for event detection uses, supplements and upgrades existing information retrieval methods in order to improve knowledge management and representation. This article describes the research done in order to determine the impact of various methods of document representation on cluster analysis. Several statistical and linguistic NLP morphological normalization methods of document representation are tested in an event detection scenario. Event detection was conducted using online newspaper articles issued on a single day. A cluster analysis was done using the various document representation methods and a clustering algorithm. The results were then compared against a human evaluated golden standard. The results show that both statistical and linguistic methods simplify the representational complexity and minimally improve the results which lead to the conclusion that for this task statistical methods should be preferred.

1. Introduction

The ever-increasing amount of various data in the world has reached a point where standard methods of knowledge management and information retrieval are no longer adequate and need to be complemented with other methods. Contemporary processing of large amounts of data is primarily supervised because they provide better results. Due to the acquisition bottleneck problem the challenge is to create automated, unsupervised methods which can then deal with large sets of data without expensive and tedious human efforts. Therefore, we decided to investigate the applicability and effectiveness of natural language processing methods for automated document clustering.

Cluster analysis is a well-established method for creating order in large sets of data (not only textual). Document clustering as a method for organizing documents into clusters is commonly used in combination with the following information retrieval (IR) methods (Forster 2006):

- support for presentation of information retrieval results - Main-stream information retrieval methods, even plain text search (e.g. finding documents containing a certain word or words), are powerful and proven methods. However, they are often plagued by a large number of irrelevant results, especially in large data sets. Using document clustering it is possible to expand this simple query by selecting a relevant document from the list of results and use it as the query. The results of this method will be more relevant and considerably fewer in number.
- support for document retrieval - Document retrieval using document clustering is based on the same principles as classic IR but searches organized clusters instead of unorganized document collections. Therefore, the query results are not clouded by polysemy and ambiguous terms.
- direct access to documents – These methods are generally based on tracking or mediating user's actions and queries in a limited environment and do not rely directly on text retrieval.

Document clustering consists of two main components: document representation and cluster analysis. Document representation deals with 'translating' documents (articles, web pages, etc.) into structures suitable for clustering (Forster 2006). This is usually done by representing documents numerically as vectors and matrices. Cluster analysis includes methods for creating meaningful data clusters from the data structure produced by methods of document representation. Clusters are created after comparing (measuring the distance) the numerical representations of the documents.

Natural language processing is a large field with many applications. Here it is used to offer a linguistic and statistical view on document representation in order to determine if representing a document as more than a collection of random characters

has any impact on clustering. NLP methods used in this research include tokenization, stemming, lemmatization and n-grams.

Our goal is to compare the results of document clustering using several linguistic and statistical NLP methods and to determine their effectiveness against a human evaluated gold standard. The final goal of the investigated methods is to organize online newspaper articles into clusters which report on a singular event.

2. Problem

This article investigates the effects of linguistic and statistical preprocessing of data as part of document representation in the document clustering task. Our document clustering task aims at organizing online news articles into clusters where every cluster corresponds to a specific event. This task in literature is often called event detection (Yang et al. 1998, Papka et al. 1999). One of the most known online resources that organize information in such manner is Google News (Google 2008).

There are several problems concerning data clustering, the most prominent being combinatorial explosion, a common issue when dealing with large data sets. Most commonly used piece of information aimed at reducing the combinatorial explosion present in data clustering is the time of publishing of the article. Namely, to cluster data, the similarity function between every two data points has to be calculated. Almost every document clustering system uses this crucial piece of information, i.e. event detection calculates clusters on documents published in a specific time frame. In this research we calculate document clusters on articles published during a single day.

Another piece of information often used in online news event detection is the origin of the article, i.e. the publisher. The assumption is that one online publisher will not produce more than one article on a singular event. There is, of course, an objective possibility of erroneous mapping of reports from different publishers pertaining consecutive events (article A1 covering events another publisher covered in more than one article), but this problem is not of interest in this paper.

The evaluation of the efforts described earlier is most often carried out through a gold standard - a data sample organized by hand. In this research we used a small sample of online news published in a single day.

The results of this procedure are aimed at the representation of the data collected by crawlers where all articles covering one event would be presented as one entity.

3. Experiment

The data used in this experiment are obtained from the Institute for Business Intelligence (Zapi 2008) and their web crawler which collects online news on a 10-minute basis. The clustered data consists of 1028 news documents published on May 5 2008. After removing identical documents published on same domains, 1000 articles collected from 18 different domains remained in the data set.

As stated before, the emphasis in this experiment was to test the value of document representation methods as feature extraction methods by comparing their impact on the end result.

The cluster analysis was performed using the cosine similarity measure with a modified single-link hierarchical agglomerative algorithm and a defined threshold. The algorithm starts with every data point forming its own cluster. It merges clusters on the nearest-neighbor principle, merging closest clusters together. When merging clusters, the distance between them is considered as the distance between the two closest data points. The clustering threshold is the criterion that stops the clustering task when there are no clusters as close as the criterion defines. It ensures that the clustering task will not produce just one cluster. A special constraint while performing the clustering task is the fact that one cluster cannot consist of two articles published on the same domain. The clustering algorithm is implemented as follows:

1. Calculate the similarity function between documents
2. Build triples with id-s of documents and their similarity
3. Remove triples whose similarity is lower than the clustering threshold
4. Sort the remaining triples in descending order

5. Move with two nested iterations through combinations of triples
6. Form a new cluster from the first triple and add all following triples, i.e. IDs if they:
    - satisfy the threshold condition
    - are not allocated in any other cluster
    - no article from the same domain is in the cluster already
7. If an article does not meet the third condition from the previous step, do not add any following articles that have a stronger similarity with the article not meeting the condition

No emphasis is put on the weighting method of a selected feature but only the popular tf-idf measure is used (Jones 1973). In order to use the tf-idf measure, the distribution of features over documents has to be known. Therefore a corpus of 30,000 news articles and 6,985,242 tokens is constructed and the distribution of interest is calculated. Also the document space with its corresponding space complexity is defined using this data.

In this research there are six different document representation techniques investigated: TOKEN, STEM, LEMMA, 3-GRAM, 4-GRAM and 5-GRAM.

In the TOKEN representation method the lowercased corpus is tokenized by the python-like regular expression `r'['+ln+']+(?:[-.,@/]['+ln+']+)*'` where the variable `ln` contains all letters and numerals, i.e. token is defined as a sequence of letters and numerals with the characters `'-.,@/'` occurring isolated inside that sequence ('požeško-slavonska', '16.2' 'nick@127.0.0.1' etc.). The TOKEN representation method is considered the baseline of this research.

In the STEM representation method a stemming algorithm still under construction is used on the lowercased corpus since there is no other stemming algorithm for Croatian available for that purpose (the algorithm described in (Ljubešić et al. 2007) is used for normalizing basic word forms in query expansion and the algorithm described in (Šnajder 2006) has not been made publicly available yet). The stemming algorithm used deals with inflectional morphology only and the rules are primarily focused on noun and adjective paradigms.

In the LEMMA representation method the POS-tagging algorithm described in (Agić, Tadić 2006) and (Agić et al. 2008) is used.

In the remaining three representation methods a character n-gram morphological normalization approach as described in (Šilić et al. 2007) is used. In our case, character n-grams are calculated from tokens using the TOKEN representation method (e.g. the token 'imaju' is described through 4-grams '_ima', 'imaj', 'maju' and 'aju_').

These document representation methods are evaluated in the clustering process aiming at event detection. The event detection task is evaluated using a gold standard – 1,000 news articles are manually organized into clusters. Software under development is used for this task. Out of 1,000 articles, 396 of them describe events not covered by any other article, i.e. they form clusters containing just one article. The remaining 604 documents are organized into 144 clusters with an average of 4.19 elements per cluster. The median of the non-one cluster size distribution is 3 and the maximum is 18.

The basic evaluation measures used are precision and recall. They are both calculated through the best-case intersection of the gold standard and the clustering result. Additional evaluation measure is the F0.5 which favors precision twice as much as recall. Namely, the results are used for supporting information retrieval results which makes precision more important than recall.

4. Results/Discussion

The vector space complexity of different document representation methods is shown in Table 1. STEM simplifies the space complexity by 1.3 and LEMMA by almost 2. The highest space simplification is obtained by 3-GRAM, 4-GRAM is equivalent to LEMMA whilst 5-GRAM increases the space complexity.

	Number of dimensions	Simplification coefficient
TOKEN	249,136	1.0
STEM	191,676	1.3
LEMMA	125,406	1.99
3-GRAM	22,264	11.19
4-GRAM	115,693	2.15
5-GRAM	287,325	0.87

Table 1: Space complexity and simplification coefficient regarding document representation methods

In Table 2 the maximum F0.5 evaluation measures with the space simplification coefficient and corresponding clustering threshold criterion concerning the representation method is shown. The data shows no obvious difference regarding the document representation methods with STEM, LEMMA and 4-GRAM outperforming slightly the TOKEN baseline and 3-GRAM and 5-GRAM producing lower results. Taking the simplification factor into account, it would be advisable to use the 3-GRAM method for computationally intensive problems since the space simplification is large. LEMMA and 4-GRAM obtain same results and a similar simplification factor, but the 4-GRAM method is much simpler and language independent. When comparing results of language dependent methods, STEM outperforms LEMMA with a higher F0.5 measure, a lower space simplification factor, but a much simpler and faster method.

	F0.5	Simplification coefficient	Clustering threshold
TOKEN	0.858	1.0	0.25
STEM	0.868	1.3	0.35
LEMMA	0.859	1.99	0.4
3-GRAM	0.853	11.19	0.5
4-GRAM	0.860	2.15	0.45
5-GRAM	0.857	0.87	0.45

Table 2: Maximal F0.5 measure, simplification coefficient and clustering threshold regarding the document representation method

Figure 1 shows precision, recall and the F0.5 measures concerning the clustering threshold criterion (CTC) from the STEM data. As expected, with the threshold rising, precision rises and recall falls. F0.5 experiences its maximum at CTC=0.3.

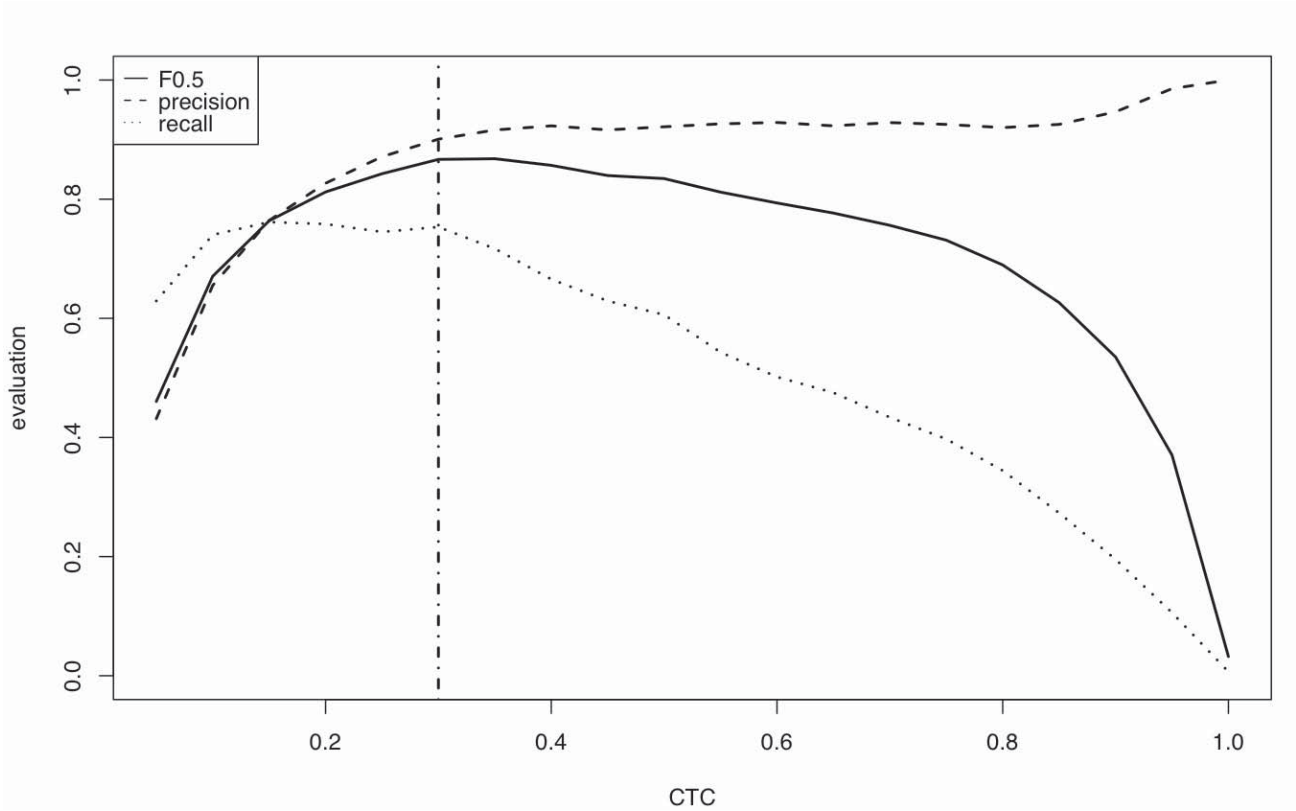

Figure 1: Precision, recall and F0.5 concerning clustering threshold criterion (CTC)

Concerning the variable of the number of elements in clusters with more than one element, it correlates strongly positively with recall (0.993) and negatively with precision (-0.531). The F0.5 measure shows its quality in measuring the overall performance of the clustering task with correlating rather highly with the described variable (0.756).

F0.5 regarding the clustering threshold criterion and ignoring the document representation method is shown in table 3. The criterion reaches its stable maximum at the value of 0.35 and could be recommended for usage regardless of the document representation method.

CTC	Average F0.5
0.05	0.416
0.1	0.580
0.15	0.687
0.2	0.758
0.25	0.806
0.3	0.833
0.35	0.845
0.4	0.851
0.45	0.846
0.5	0.841

CTC	Average F0.5
0.55	0.826
0.6	0.811
0.65	0.796
0.7	0.772
0.75	0.751
0.8	0.714
0.85	0.661
0.9	0.575
0.95	0.417
1.0	0.024

Table 3: Average F0.5 regarding clustering threshold criterion (CTC)

5. Conclusion

In this research we have analyzed the impact of different statistical and linguistic morphological normalization methods in a document clustering i.e. news event detection task. As the baseline we used pure tokenization. No significant improvement was observed when using normalization methods. Possible reason for such results could be the nature of problem because singular events tend to be described by different sources using same word forms. Highest level of the document space simplification was obtained using the character 3-grams with a slight decline in the evaluation measures and is therefore recommended when dealing with computationally demanding tasks. When comparing linguistic methods, stemming slightly outperformed lemmatization. The reason for such results could be low quality of the processed data (HTML escape sequences and such).Lemmatization yielded a higher level of document space simplification. When comparing statistical and linguistic method, they both managed to achieve similar document space simplification and evaluation measures which leads to the conclusion that the statistical methods should be preferred due to their simplicity, higher speed and language independence. Further research will include larger evaluation sets for validation and testing as well as experimenting with collocations and syntactic and semantic processing.

References

Agić, Ž.; Tadić, M.; Dovedan, Z. 2008. Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian. // Proceedings of IS-LTC (in press)

Agić, Ž.; Tadić, M. 2006. Evaluating morphosyn-tactic tagging of croatian texts. In LREC2006 Proceedings, Genoa-Paris. ELRA.

Forster, R. 2006. Document Clustering in Large German Corpora Using Natural Language Processing. Thesis presented to the Faculty of Arts of the University of Zürich for the degree of Doctor of Philosophy.

Google News. 2008. Google. http://news.google.com/.

Ljubešić, N.; Boras, D.; Kubelka, O. 2007. Retrieving Information in Croatian: Building a Simple and Efficient Rule-based Stemmer // Digital information and heritage / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu. Str. 313-320.

Papka, R.;Croft, B.W.;Barto, A.G.;Danai, K.;Kurose, J.F. 1999. On-line New Event Detection, Clustering and Tracking

Sparck Jones, K. 1973. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation

Šnajder, J. 2006. Rule-Based Automatic Acquisition of Large-Coverage Morphological Lexicons for Information Retrieval. Technical Report MZOS 2003-082, Department of Electronics, Microelectronics, Computer and Intelligent Systems, FER, University of Zagreb

Šilić, A.; Chauchat, J.; Dalbelo Bašić, B.; Morin, A. 2007. N-Grams and Morphological Normalization in Text Classification: A Comparison on a Croatian-English Parallel Corpus. // Lecture Notes in Artificial Intelligence. 4874; 671-682

ZAPI. 2008. Institute for business intelligence, http://www.zapi.hr.

Yang, Y.; Pierce, T.; Carbonell, J. 1998. A Study on Retrospective and On-line Event Detection. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM press