An Online Tool for Discriminative Keyword Extraction from Large Text Corpora

Jenna Kanerva,[1,2] **Filip Ginter**[2]

[1]University of Turku Graduate School UTUGS
[2]Department of Information Technology
University of Turku, Finland
jmnybl@utu.fi, ginter@cs.utu.fi

1. Introduction

Recently, a large corpus has been released comprising all discussions over the entire history of the Finnish discussion board *Suomi24* (Lagus et al., 2016). This board is one of the largest non-English discussion boards and the corpus totals over two billion tokens. This data is of interest to a diverse group of researchers in the humanities, who are interested in various historical, sociologic and linguistic aspects of the discussions. Given the size of the corpus, these researchers need user-friendly, online tools to explore the data. So far, the Suomi24 corpus has been indexed by two publicly available corpus query services: the *Korp*[1] concordance tool (Borin et al., 2012) maintained at the CSC computing centre and the SETS[2] syntactic query system maintained at the University of Turku (Luotolahti et al., 2015b). Both are browser-based systems needing no local installation.

One of the typical query types that the researchers interested in exploring the Suomi24 corpus have posed can be paraphrased as *"what are the distinguishing features of discussions about a given topic as contrasted to other topics, or the background text"*. For instance, questions such as *"what is typical for discussions about Russia"* or *"what is typical for discussions about competing social media platforms"* have been asked.

The only option realistically available for the researchers without the expertise of large corpus processing to answer such questions is Korp's word sketches. Korp word sketch view shows lists of important pre- and post-modifiers, as well as lists of verbs occurring before and after the searched word. However, as we will demonstrate later, this functionality is insufficient for the task because it is not intended to gather the necessary topic-contrastive statistics. Therefore, to address this particular need, we develop and in this abstract present an online system for extracting contrastive keyword characteristics of different user defined topics or concepts from large text corpora. Given a list of queries, most typically expressed as words defining one or several topics of interest, the system extracts sentences where at least one of the words is mentioned, and based on the sentence context it then provides a list of most representative words for each query topic. For instance, a query for the single topic *Russia* would result in a list of keywords typical for discussions about Russia as opposed to the rest of the corpus, whereas a query for *Russia* and *Sweden* would result in one list of keywords that are specific for discussions about Russia, as opposed to Sweden, and one list of keywords that are specific for discussions about Sweden, as opposed to Russia.

Our approach is to cast the keyword extraction as a classification task, where a multiclass classifier is trained to distinguish among the query topics, using words from sentences mentioning the topics as features. The typical keywords for every topic are then extracted from the classifier model and presented as the result. This is a technique employed, for example, by Vo and Zhang (2016) who also claim that such a discriminative approach is superior to the more common mutual-information based approaches.

2. Functionality

The user can define any number of topics, each specified using a query which most typically would be a simple list of words or lemmas. Note, however, that since the system uses the SETS syntactic tree query service to extract the sentences for every topic, any queries supported by the SETS query language can be used, allowing restrictions not only on the words and lemmas, but also on their syntactic role and context. In practice, though, simple word lists are the most likely to be used.

Given the set of topics and their associated queries, the system collects up to 5,000 sentences from the corpus for each of the given topics, and extracts words which are highly associated with each of the query topics. If only one topic is defined, it is contrasted against a random sample of the corpus, whereas if two or more topics are given, they are contrasted against one another.

3. Case Studies

One actual use case, incidentally mirroring almost exactly that of Vo and Zhang (2016), was to distinguish sentences where positive and negative emoticons are used, where the resulting distinguishing features should comprise a simple sentiment lexicon.[3] Indeed, for the positive emoticon `:)` the list of most representative words returned by the system comprises of positively charged words e.g. *thank you, good, fortunately, wonderful, nice, good luck, congratulations*. Whereas for the negative emoticon `:(` the list is the very opposite, comprised of words like *sad, unfortunately, too bad, be afraid, help, hurt*.

[1] http://korp.csc.fi

[2] http://bionlp-www.utu.fi/dep_search

[3] A direct link to the query: http://bit.ly/2bu2SRH.

Greece	antiquity, loan, collateral, debt restructuring, Turkey, debt, island, money, Portugal, road, Greek, word, bank, financial assistance package, bankruptcy
Russia	Putin, China, to attack, Karelia, Russian, threat, oil, NATO, Soviet Union, army, Ukraine, relationship, history, border, ruble
Italy	mafia, Spain, Berlusconi, France, bella, Mussolini, crime, the, Rome, Milan, trip, crucifix, Ferrari, Sicily, police
Czech	Slovakia, Prague, Poland, republic, Skoda, Canada, match, Hungary, beer, team, to play, factory, to win, game

Table 1: 15 most representative words for Greece, Russia, Italy and Czech, when these four countries are compared against each other.

In a study on mining national stereotypes from Suomi24 discussions, we wanted to know how Finns talk about different countries and nations, and what topics are covered. Here, we collected a list of country codes and for each assigned a list of words used to refer to a particular country or a nation (mostly the official country name and the nationality term). This list was then run through the pipeline to collect words distinguishing a particular country from other countries. These words can be seen as a view the Finns have towards a country. The example results for four countries are shown in Table 1 and the full results for over 150 countries are available online.[4]

To compare our results to those obtained from Korp, we search for a word and use the word sketch view to see words occurring close by. Korp word sketch view shows lists of important pre- and post-modifiers, as well as lists of verbs occurring before and after the searched word. For *Greece* top-5 modifiers are *antiquity, ancient, for example, RUSTET (rüstet, German word), to drift* (pre-modifiers) and *as, e.g., immigration pressure, 10., 9.* (post-modifiers). Top-5 verbs occurring before Greece are *to fall, pay, drift, quit, need*, and after Greece are *to support, fund, help, save, lend money*. As Korp is only able to return words occurring right before or after the searched token, especially the coverage of other nouns occurring nearby is not as wide as in our results. Korp also returns very common modifiers (*as, e.g., for example*) which can be assumed to be frequent in every such query, and thus not showing up in the results of the contrastive approach.

4. Implementation

The sentence texts for the classification are collected using the API[5] of the SETS dependency tree search tool (Luotolahti et al., 2015b), which is able to efficiently find the requested tree structures or lexical items from large pre-indexed datasets. At the moment, there are two large indexed corpora available, the Finnish Internet Parsebank (Luotolahti et al., 2015a) with over 4B tokens of Internet crawled Finnish, and the over 2B token corpus of the abovementioned Suomi24 discussions.

The lists of most representative words for each topics are produced by a linear SVM one-vs-all multiclass classifier trained to separate the topics from each other using either words or lemmas from the topic sentences as features. After the classifier is trained, words/lemmas with the highest positive weights for each topic are presented to the user. We use the LinearSVC implementation from the scikit-learn Python package (Pedregosa et al., 2011).

The web interface is two-stage. Upon submitting the query, the users are immediately given a link where the results will be available once the data extraction and classifier training is complete. Typical runtimes before the query results become available are on the order of several minutes in most cases. The results are persistent and can be revisited at any time. Further, should the same query be submitted again at a later point, the system redirects the user to the already existing page with the results — unless a forced rerun is specified by the user in the submission form. The users thus do not need to remember the link to the results, they can simply submit the query again.

The user interface is accessible at `http://bionlp-www.utu.fi/keywords_webgui/`, and its source code is freely available at `https://github.com/jmnybl/keywords_webgui`. The pipeline is fully language independent and it can be used with all sufficiently sized corpora available via the SETS search system.

5. Conclusion and Future Work

In this abstract we presented an easy to use online service which targets the specific need of studying features that distinguish discussions about particular user-defined topics of interest. The service is currently in active use in the context of the Suomi24 discussion fora research. The system was demonstrated with two different use cases: simple sentiment lexicon extraction using emoticons, and mining country stereotypes from online discussions.

As a future work, the system will be extended with temporal analysis, i.e. the ability of tracing the change of the distinguishing features over time. Further, a common request for extension of the current system is sentiment detection — whereby not only the distinguishing features, but also the overall polarity would be extracted.

The service is freely available and can query any corpus indexed in SETS. A possible future work would be to also allow querying the corpora from Korp using the Korp API. This would substantially extend the corpus coverage, beyond the current ongoing use case of the Suomi24 dataset.

Acknowledgements

This work is supported by the Finnish Academy and the Kone Foundation.

[4] `http://bit.ly/2blP7FD`

[5] `http://bionlp-www.utu.fi/dep_search_webapi/`

References

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, page 474478.

Krista Lagus, Mika Pantzar, Minna Ruckenstein, and Marjoriikka Ylisiurua. 2016. Suomi24 – muodonantoa aineistolle. Technical report.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015a. Towards universal web parsebanks. In *Proceedings of the International Conference on Dependency Linguistics (Depling'15)*, pages 211–220. Uppsala University.

Juhani Luotolahti, Jenna Kanerva, Sampo Pyysalo, and Filip Ginter. 2015b. SETS: Scalable and efficient tree search in dependency graphs. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 51–55. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tin Duy Vo and Yue Zhang. 2016. Don't count, predict! an automatic approach to learning sentiment lexicons for short text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 219–224. Association for Computational Linguistics.