ScaleText: The Design of a Scalable, Adaptable, and User-Friendly Document System for Similarity Searches

Digging for Nuggets of Wisdom in Text

Jan Rygl[1] Petr Sojka[2] Michal Růžička[2] Radim Řehůřek[1]

[1]RaRe Technologies, {jimmy,radim}@rare-technologies.com

[2]Faculty of Informatics, Masaryk University, Brno, Czech Republic
sojka@fi.muni.cz and mruzicka@mail.muni.cz

December 2nd, 2016

Introduction

- Search as a gateway and primary access method for information in documents.
- From keyword based search to meaning based.
- From keyword based search to phrase/free question/paragraph based search.
- Topic modeling in large documents.
- Scalability—problem even with linear complexity.

Design Imperatives

- **S**calability: with the size of today's document collections, efficiency is a primary concern, allowing low latency responses.
- **A**daptability: since no size fits all, the system should be easily customizable and tunable for any given application purpose.
- **R**elevance: search precision could be improved by clever semantic representations of the meanings of indexed texts. It is both necessary and desirable to find highly relevant document chunks.
- **I**mplementation **C**larity: the implementation should be written with ease of maintenance in mind.
- **S**implicity: keep it simple stupid, yet provide the functionality needed.

Approaches

- *a discrete representation* of meaning, which can be based on knowledge-based representations such as WordNet, BabelNet, Freebase or Wikipedia, or
- *a smooth representation* in vector spaces based on a distributional hypothesis, e.g. representing meanings as word, phrase, sentence, . . . embeddings (Mikolov, 2013) which are learned from the language used in big corpora by unsupervised, deep learning approaches, or by topic modeling (Blei, 2012)

Software Systems to support Semantic Similarity Search

- based on Gensim (Rehurek, Sojka, 2010)
- Kvasir
- Similarity search based on trees (M-tree) et al.

Indexing I

Storing Document Chunks as Points in Vector Space

ScaleText introduces a flexible data processing pipeline for document indexing, leading to semantic document representations in a vector space. The overall scheme of document transformations in the indexing workflow is depicted in Figure 1.

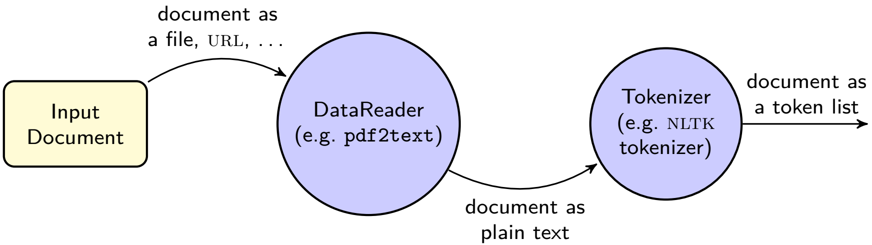

Figure 1: Data flow diagram of document indexing in ScaleText

Indexing II

Storing Document Chunks as Points in Vector Space

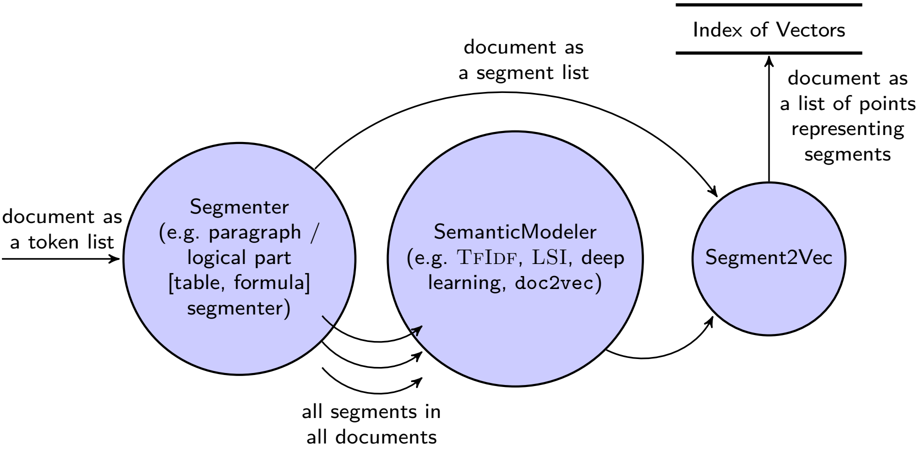

Figure 1: Data flow diagram of document indexing in ScaleText

Document Similarity Search I

Digging for Nuggets of Wisdom

The indexed dataset is used for similarity searching. To pursue the gold mining metaphor, gold nuggets are washed with different gold mining techniques. The overall schema of the search procedure is depicted in Figure 2.

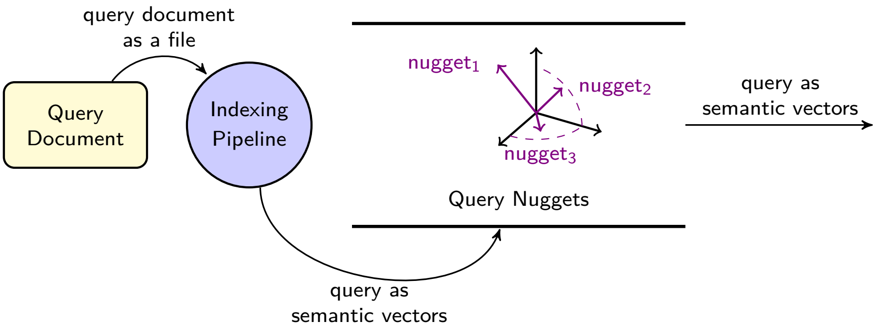

Figure 2: Data flow diagram of document similarity search in ScaleText. q is the number of query nuggets, K is the number of best nugget candidates for each query nugget, and k is the number of desired results.

Document Similarity Search III

Digging for Nuggets of Wisdom

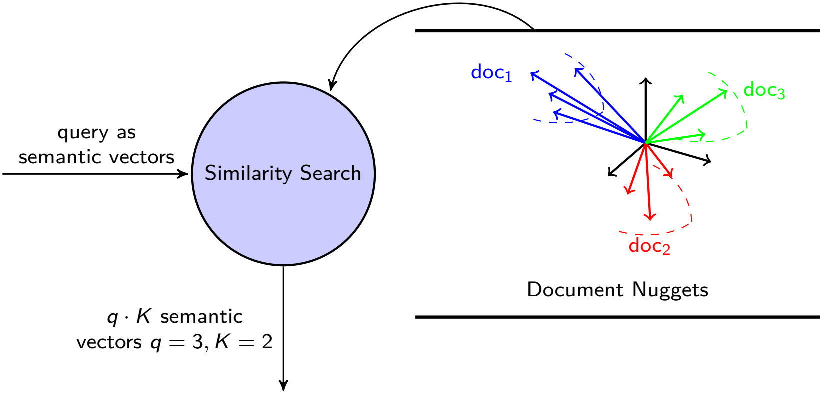

Figure 2: Data flow diagram of document similarity search in ScaleText. q is the number of query nuggets, K is the number of best nugget candidates for each query nugget, and k is the number of desired results.

Document Similarity Search IV

Digging for Nuggets of Wisdom

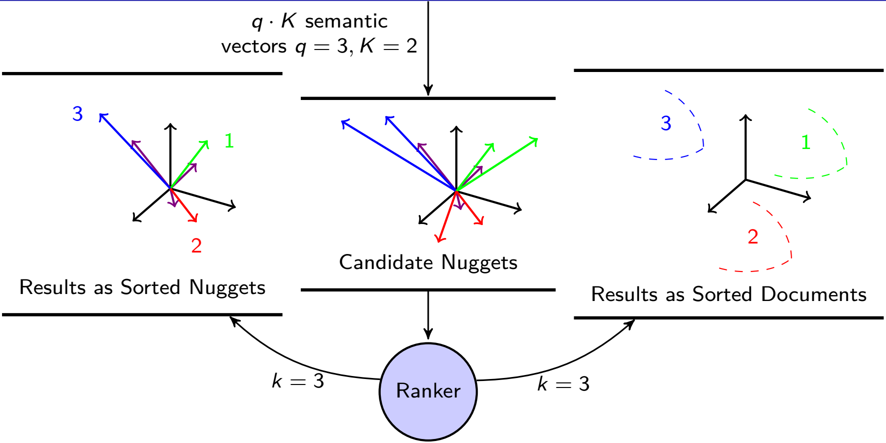

Figure 2: Data flow diagram of document similarity search in ScaleText. q is the number of query nuggets, K is the number of best nugget candidates for each query nugget, and k is the number of desired results.

Automatic Evaluation Framework for System Modules I

We implemented Bpref@k as follows:

$$\text{Bpref@}k = \frac{1}{\min(R,k)} \sum_r \left(1 - \frac{\min\left(\text{number of } n \text{ ranked higher than } r, R\right)}{\min(N,R)}\right),$$

where

- R is the number of documents relevant to the topic,
- N is the number of documents irrelevant to the topic,
- k is the maximal number of inspected results, and
- "number of n ranked higher than r" is the number of irrelevant documents (according to the judgment) ranked higher than the relevant (according to the judgment) document r that is being processed in the step.

Automatic Evaluation Framework for System Modules II

Table 1: ScaleText prototype evaluation on the Enron dataset via Bpref. The single metric value is the average of Bpref@100 over all the queries

doc. model	document ranking strategy	#feat.	avg Bpref@100
TfIdf	maximum nugget score	100	0.0451
TfIdf+LSI	maximum nugget score	50	0.0460
TfIdf+LSI	maximum nugget score	100	0.0565
TfIdf+LSI	maximum nugget score	500	0.0358
TfIdf	average nugget score	100	0.0451
TfIdf+LSI	average nugget score	50	0.0460
TfIdf+LSI	average nugget score	100	0.0548
TfIdf+LSI	average nugget score	500	0.0358
TfIdf	normalized sum of nugget scores	100	0.0451
TfIdf+LSI	normalized sum of nugget scores	50	0.0460
TfIdf+LSI	normalized sum of nugget scores	100	0.0534
TfIdf+LSI	normalized sum of nugget scores	500	0.0358

Conclusion and Future Work

We have several research questions in our sights:

- **Word disambiguation in context:** current methods represent a word in the vector space as the centroid of its different meanings. We want to evaluate an approach based on random walks through texts so as to distinguish the representation of words in context.

Conclusion and Future Work

We have several research questions in our sights:

- **Word disambiguation in context:** current methods represent a word in the vector space as the centroid of its different meanings. We want to evaluate an approach based on random walks through texts so as to distinguish the representation of words in context.
- **Compositionality of segment representation:** semantic vectors representing the meaning of segments should reflect compositionality of meaning of its parts, e.g. words, phrases and sentences.

Conclusion and Future Work

We have several research questions in our sights:

- **Word disambiguation in context:** current methods represent a word in the vector space as the centroid of its different meanings. We want to evaluate an approach based on random walks through texts so as to distinguish the representation of words in context.
- **Compositionality of segment representation:** semantic vectors representing the meaning of segments should reflect compositionality of meaning of its parts, e.g. words, phrases and sentences.
- **Representation of narrativity:** we may represent narrative text qualities [5] as a *trajectory* of words or nuggets in vector space, e.g. document representation may be a trajectory instead of a point.

References I

- Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (Apr 2012), http://doi.acm.org/10.1145/2133806.2133826

- Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 25–32. SIGIR '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/1008992.1009000

- Cormack, G.V., Grossman, M.R., Hedin, B., Oard, D.W.: Overview of the TREC 2010 legal track. In: Proc. 19th Text REtrieval Conference. pp. 1–45. National Institute of Standards and Technology, Gaithersburg, MD (2010), http://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf

References II

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)

Hoenkamp, E., Bruza, P., Song, D., Huang, Q.: An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model. In: Azzopardi, L., Kazai, G., Robertson, S.E., Rüger, S.M., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR. Lecture Notes in Computer Science, vol. 5766, pp. 116–127. Springer (2009), `http://dx.doi.org/10.1007/978-3-642-04417-5_11`

Kdorff: Bpreftreceval2006 (2007), `http://icb.med.cornell.edu/wiki/index.php/BPrefTrecEval2006`, Accessed: 2016-10-29

References III

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013), http://arxiv.org/abs/1310.4546, arXiv:1310.4546

- Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. pp. 45–50. Valletta, Malta (2010), software available at http://nlp.fi.muni.cz/projekty/gensim

- TREC version of EDRM Enron Dataset, version 2, Accessed: 2016-10-29, http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set-v2

References IV

Wang, L., Tasoulis, S., Roos, T., Kangasharju, J.: Kvasir: Seamless integration of latent semantic analysis-based content provision into web browsing. In: Proc. of the 24th International Conference on World Wide Web. pp. 251–254. WWW '15 Companion, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2740908.2742825