Integrating Query Performance Prediction in Term Scoring for Diachronic Thesaurus

Chaya Liebeskind, Ido Dagan
Department of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
liebchaya@gmail.com
dagan@cs.biu.ac.il

Abstract

A diachronic thesaurus is a lexical resource that aims to map between modern terms and their semantically related terms in earlier periods. In this paper, we investigate the task of collecting a list of relevant modern target terms for a domain-specific diachronic thesaurus. We propose a supervised learning scheme, which integrates features from two closely related fields: Terminology Extraction and Query Performance Prediction (QPP). Our method further expands modern candidate terms with ancient related terms, before assessing their corpus relevancy with QPP measures. We evaluate the empirical benefit of our method for a thesaurus for a diachronic Jewish corpus.

1 Introduction

In recent years, there has been growing interest in diachronic lexical resources, which comprise terms from different language periods. (Borin and Forsberg, 2011; Liebeskind et al., 2013; Riedl et al., 2014). These resources are mainly used for studying language change and supporting searches in historical domains, bridging the lexical gap between modern and ancient language.

In particular, we are interested in this paper in a certain type of diachronic thesaurus. It contains entries for modern terms, denoted as *target terms*. Each entry includes a list of ancient *related terms*. Beyond being a historical linguistic resource, such thesaurus is useful for supporting searches in a diachronic corpus, composed of both modern and ancient documents. For example, in our historical Jewish corpus, the modern Hebrew term for *terminal patient*[1] has only few verbatim occurrences, in modern documents, but this topic has been widely discussed in ancient periods. A domain searcher needs the diachronic thesaurus to enrich the search with ancient synonyms or related terms, such as *dying* and *living for the moment*.

Prior work on diachronic thesauri addressed the problem of collecting relevant related terms for given thesaurus entries. In this paper we focus on the complementary preceding task of collecting a relevant list of modern target terms for a diachronic thesaurus in a certain domain. As a starting point, we assume that a list of meaningful terms in the modern language is given, such as titles of Wikipedia articles. Then, our task is to automatically decide which of these *candidate terms* are likely to be relevant for the corpus domain and should be included in the thesaurus. In other words, we need to decide which of the candidate modern terms corresponds to a concept that has been discussed significantly in the diachronic domain corpus.

Our task is closely related to term scoring in the known Terminology Extraction (TE) task in NLP. The goal of corpus-based TE is to automatically extract prominent terms from a given corpus and score them for domain relevancy. In our setting, since all the target terms are modern, we avoid extracting them from the diachronic corpus of modern and ancient language. Instead, we use a given candidate list and apply only the term scoring phase. As a starting point, we adopt a rich set of state-of-the-art TE scoring measures and integrate them as features in a common supervised classification approach (Foo and Merkel, 2010; Zhang et al., 2010; Loukachevitch, 2012).

Given our Information Retrieval (IR) motivation, we notice a closely related task to TE, namely Query Performance Prediction (QPP). QPP methods are designed to estimate the retrieval quality of search queries, by assessing their relevance to the text collection. Therefore, QPP scoring measures

[1] The examples in this paper refer to Hebrew terms that were literally translated.

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 89–94,
Beijing, China, July 30, 2015.

seem to be potentially suitable also for our terminology scoring task, by considering the candidate term as a search query. Some of the QPP measures are indeed similar in nature to the TE methods, analyzing the distribution of the query terms within the collection. However, some of the QPP methods have different IR-biased characteristics and may provide a marginal contribution. Therefore, we adopted them as additional features for our classifier and indeed observed a performance increase.

Most of the QPP methods prioritize query terms with high frequency in the corpus. However, in a diachronic corpus, such criterion may sometimes be problematic. A modern target term might appear only in few modern documents, while being referred to, via ancient terminology, also in ancient documents. Therefore, we would like our prediction measure to be aware of these ancient documents as well. Following a particular QPP measure (Zhou and Croft, 2007), we address this problem through Query Expansion (QE). Accordingly, our method first expands the query containing the modern candidate term, then calculates the QPP scores of the expanded query and then utilizes them as scoring features. Combining the baseline features with our expansion-based QPP features yields additional improvement in the classification results.

2 Term Scoring Measures

This section reviews common measures developed for Terminology Extraction (Section 2.1) and for Query Performance Prediction (Section 2.2). Table 1 lists those measures that were considered as features in our system, as described in Section 3.

2.1 Terminology Extraction

Terminology Extraction (TE) methods aim to identify terms that are frequently used in a specific domain. Typically, linguistic processors (e.g. POS tagger, phrase chunker) are used to filter out stop words and restrict candidate terms to nouns or noun phrases. Then, statistical measures are used to rank the candidate terms. There are two main terminological properties that the statistical measures identify: *unithood* and *termhood*. Measures that express unithood indicate the collocation strength of units that comprise a single term. Measures that express termhood indicate the statistical prominence of the term in the target domain corpus. For our task, we focus on the second property, since the candidates are taken from a key-list of terms whose coherence in the language is already known. Measures expressing termhood are based either on frequency in the target corpus (1, 2, 3, 4, 9, 11, 12, 13)[2], or on comparison with frequency in a reference background corpus (8, 14, 16). Recently, approaches which combine both unithood and termhood were investigated as well (7, 8, 15, 16).

2.2 Query Performance Prediction

Query Performance Prediction (QPP) aims to estimate the quality of answers that a search system would return in response to a particular query. Statistical QPP methods are categorized into two types: pre-retrieval methods, analyzing the distribution of the query term within the document collection; and post-retrieval methods, additionally analyzing the search results. Some of the pre-retrieval methods are similar to TE methods based on the same term frequency statistics.

Pre-retrieval methods measure various properties of the query: specificity (17, 18, 24, 25), similarity to the corpus (19), coherence of the documents containing the query terms (26), variance of the query terms' weights over the documents containing it (20); and relatedness, as good performance is expected when the query terms co-occur frequently in the collection (21).

Post-retrieval methods are usually more complex, where the top search results are retrieved and analyzed. They are categorized into three main paradigms: clarity-based methods (28), robustness-based methods (22) and score distribution based methods (23, 29).

We pay special attention to two post-retrieval QPP methods; *Query Feedback* (22) and *Clarity* (23). The Clarity method measures the coherence of the query's search results with respect to the corpus. It is defined as the KL divergence between a language model induced from the result list and that induced from the corpus. The Query Feedback method measures the robustness of the query's results to query perturbations. It models retrieval as a communication channel. The input is the query, the channel is the search system, and the set of results is the noisy output of the channel. A new query is generated from the list of search

[2]The numbers in parentheses correspond to the numbers in Table 1.

Terminology Extraction measures			
1	Term Frequency (TF)	9	Relative Frequency
2	Document Frequency	10	N-gram Length
3	Residual Inverse Document Frequency (Manning and Schütze, 1999)	11	TF-Inverse Document Frequency (TF-IDF) (Witten et al., 1999)
4	Average Term frequency	12	Term Contribution (Liu et al., 2003)
5	Term Variance (Liu et al., 2005)	13	Term Variance Quality (Liu et al., 2005)
6	TF-Disjoint Corpora Frequency (Lopes et al., 2012)	14	Weirdness (Ahmad et al., 1999)
7	C-value (Frantzi and Ananiadou, 1999)	15	NC-value (Frantzi and Ananiadou, 1999)
8	Glossex (Kozakov et al., 2004)	16	TermExtractor (Sclano and Velardi, 2007)
Query Performance Prediction measures			
17	Average IDF (He and Ounis, 2004)	24	Average ICTF (Inverse collection term frequency) (Plachouras et al., 2004)
18	Query Scope (He and Ounis, 2004)	25	Simplified Clarity Score (He and Ounis, 2004)
19	Similarity Collection Query (Zhao et al., 2008)	26	Query Coherence (He et al., 2008)
20	Average Variance (Zhao et al., 2008)	27	Average Entropy (Cristina, 2013)
21	Term Relatedness (Hauff et al., 2008)	28	Clarity (Cronen-Townsend et al., 2002)
22	Query Feedback (Zhou and Croft, 2007)	29	Normalized Query Commitment (Shtok et al., 2009)
23	Weighted Information Gain (Zhou and Croft, 2007)		

Table 1: Prior art measures considered in our work

results, taking the terms with maximal contribution to the Clarity score, and then a second list of results is retrieved for that second query. The overlap between the two lists is the robustness score. Our suggested method was inspired by the Query Feedback measure, as detailed in the next section.

3 Integrated Term Scoring

We adopt the supervised framework for TE (Foo and Merkel, 2010; Zhang et al., 2010; Loukachevitch, 2012), considering each candidate target term as a learning instance. For each candidate, we calculate a set of features over which learning and classification are performed. The classification predicts which candidates are suitable as target terms for the diachronic thesaurus. Our baseline system (*TE*) includes state-of-the-art TE measures as features, listed in the upper part of Table 1.

Next, we introduce two system variants that integrate QPP measures as additional features. The first system, *TE-QPP*$_{Term}$, applies the QPP measures to the candidate term as the query. All QPP measures, listed in the lower part of Table 1, are utilized except for the Query Feedback measure (22) (see below). To verify which QPP features are actually beneficial for terminology scoring, we measure the marginal contribution of each feature via ablation tests in 10-fold cross validation over the training data (see Section 4.1). Features which did not yield marginal contribution were not included[3].

The two systems, described so far, rely on corpus occurrences of the original candidate term, prioritizing relatively frequent terms. In a diachronic corpus, however, a candidate term might be rare in its original modern form, yet frequently referred to by archaic forms. Therefore, we adopt a query expansion strategy based on Pseudo Relevance Feedback, which expands a query based on analyzing the top retrieved documents. In our setting, this approach takes advantage of a typical property of modern documents in a diachronic corpus, namely their temporally-mixed language. Often, modern documents in a diachronic domain include ancient terms that were either preserved in modern language or appear as citations. Therefore, an expanded query of a modern term, which retrieves only modern documents, is likely to pick some of these ancient terms as well. Thus, the expanded query would likely retrieve both modern and ancient documents and would allow QPP measures to evaluate the query relevance across periods.

Therefore, our second integrated system, *TE-QPP*$_{QE}$, utilizes the Pseudo Relevance Feedback Query Expansion approach to expand our modern candidate with topically-related terms. First, similarly to the Query Feedback measure (measure (22) in the lower part of Table 1), we expand the candidate by adding terms with maximal contribution (top 5, in our experiments) to the Clarity score (Section 2.2). Then, we calculate all QPP measures for the expanded query. Since the expan-

[3]Removed features from *TE-QPP*$_{Term}$: 17, 19, 22, 23, 24, 25.

sions that we extract from the top retrieved documents typically include ancient terms as well, the new scores may better express the relevancy of the candidate's topic across the diachronic corpus. We also performed feature selection, as done for the first system[4].

4 Evaluation

4.1 Evaluation Setting

We applied our method to the diachronic corpus is the Responsa project Hebrew corpus[5]. The Responsa corpus includes rabbinic case-law rulings which represent the historical-sociological milieu of real-life situations, collected over more than a thousand years, from the 11^{th} century until today. The corpus consists of 81,993 documents, and was used for previous NLP and IR research (Choueka et al., 1971; Choueka et al., 1987; HaCohen-Kerner et al., 2008; Liebeskind et al., 2012; Zohar et al., 2013; Liebeskind et al., 2013).

The candidate target terms for our classification task were taken from the publicly available key-list of Hebrew Wikipedia entries[6]. Since many of these tens of thousands entries, such as person names and place names, were not suitable as target terms, we first filtered them by Hebrew Named Entity Recognition[7] and manually. Then, a list of approximately 5000 candidate target terms was manually annotated by two domain experts. The experts decided which of the candidates corresponds to a concept that has been discussed significantly in our diachronic domain corpus. Only candidates that the annotators agreed on their annotation were retained, and then balanced for equal number of positive and negative examples. Consequently, the balanced training and test sets contain 500 and 200 candidates, respectively.

For classification, Weka's[8] Support Vector Machine supervised classifier with polynomial kernel was used. We train the classifier with our training set and measure the accuracy on the test set.

4.2 Results

Table 2 compares the classification performance of our baseline (*TE*) and integrated systems, (*TE-QPP*$_{Term}$) and (*TE-QPP*$_{QE}$), proposed in Section 3.

[4]Removed features from *TE-QPP*$_{QE}$: 20, 21, 22, 26.
[5]http://www.biu.ac.il/jh/Responsa/
[6]http://he.wikipedia.org/wiki/
[7]http://www.cs.bgu.ac.il/ nlpproj/hebrewNER/
[8]http://www.cs.waikato.ac.nz/ml/weka/

Feature Set	Accuracy (%)
TE	61.5
TE-QPP$_{Term}$	65
TE-QPP$_{QE}$	**66.5**

Table 2: Comparison of system performance

In general, additional QPP features increase the classification accuracy. Even though the improvement of the term-based QPP over the baseline is not statistically significant according to the McNemar's test (McNemar, 1947), on our diachronic corpus it seems to help. Yet, when the QPP score is measured over the expanded candidate, and ancient documents are utilized, the performance increase is more notable (5 points) and the improvement over the baseline is statistically significant according to the McNemar's test with $p<0.05$.

We analyzed the false negative classifications of the baseline that were classified correctly by the QE-based configuration. We found that their expanded forms contain ancient terms that help the system making the right decision. For example, the Hebrew target term for *slippers* was expanded by the ancient expression corresponding to *made of leather*. This is a useful expansion since in the ancient documents slippers are discussed in the context of fasts, as in two of the Jewish fasts wearing leather shoes is forbidden and people wear cloth-made slippers.

5 Conclusions and Future Work

We introduced a method that combines features from two closely related tasks, terminology extraction and query performance prediction, to solve the task of target terms selection for a diachronic thesaurus. In our diachronic setting, we showed that enriching TE measures with QPP measures, particularly when calculated on expanded candidates, significantly improves performance. Our results suggest that it may be worth investigating this integrated approach also for other terminology extraction and QPP settings.

We plan to further explore the suggested method by utilizing additional query expansion algorithms. In particular, to avoid expanding queries for which expansion degrade retrieval performance, we plan to investigate the selective query expansion approach (Cronen-Townsend et al., 2004).

References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *Proceedings of the eighth Text REtrieval Conference, TREC 1999*.

Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 41–61. Springer Berlin Heidelberg.

Yaacov Choueka, M. Cohen, J. Dueck, Aviezri S. Fraenkel, and M. Slae. 1971. Full text document retrieval: Hebrew legal texts. In *Proceedings of the International ACM SIGIR conference on Information Storage and Retrieval, SIGIR 1971*, pages 61–79.

Yaacov Choueka, Aviezri S. Fraenkel, Shmuel T. Klein, and E. Segal. 1987. Improved techniques for processing queries in full-text systems. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1987*, pages 306–315, New Orleans, USA. ACM.

Haiduc Sonia Cristina. 2013. *Supporting Text Retrieval Query Formulation In Software Engineering*. Ph.D. thesis, Wayne State University.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2002, pages 299–306, New York, NY, USA. ACM.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2004. A framework for selective query expansion. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM 2004, pages 236–237, New York, NY, USA. ACM.

Jody Foo and Magnus Merkel. 2010. Using machine learning to perform automatic term recognition. In *Proceedings of the LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, pages 49–54.

Katerina T Frantzi and Sophia Ananiadou. 1999. The c-value/nc-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Yaakov HaCohen-Kerner, Ariel Kass, and Ariel Peretz. 2008. Combined one sense disambiguation of abbreviations. In *Proceedings of ACL 2008: HLT, Short Papers*, pages 61–64.

Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM 2008, pages 1419–1420, New York, NY, USA. ACM.

Ben He and Iadh Ounis. 2004. Inferring query performance using pre-retrieval predictors. In Alberto Apostolico and Massimo Melucci, editors, *String Processing and Information Retrieval*, volume 3246 of *Lecture Notes in Computer Science*, pages 43–54. Springer Berlin Heidelberg.

Jiyin He, Martha Larson, and Maarten de Rijke. 2008. Using coherence-based measures to predict query difficulty. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and RyenW. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 689–694. Springer Berlin Heidelberg.

L Kozakov, Y Park, T Fin, Y Drissi, N Doganata, and T Confino. 2004. Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and he 2nd Asian Semantic Web Conference (Se-mantic Web Challenge Track)*.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2012. Statistical thesaurus construction for a morphologically rich language. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 59–64, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2013. Semi-automatic construction of cross-period thesaurus. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 29–35, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. 2003. An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning, ICML 2003*, volume 3, pages 488–495.

Luying Liu, Jianchu Kang, Jing Yu, and Zhongliang Wang. 2005. A comparative study on unsupervised feature selection methods for text clustering. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE 2005*, pages 597–601, Oct.

Lucelene Lopes, Paulo Fernandes, and Renata Vieira. 2012. Domain term relevance through tf-dcf. *ICAI–International Conference in Artificial Inteligence*, pages 1–7.

Natalia Loukachevitch. 2012. Automatic term recognition needs multiple evidence. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry

Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Vassilis Plachouras, Ben He, and Iadh Ounis. 2004. University of glasgow at trec 2004: Experiments in web, robust, and terabyte tracks with terrier. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*.

Martin Riedl, Richard Steuer, and Chris Biemann. 2014. Distributed distributional similarities of google books over the centuries. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1401–1405, Reykjavik, Iceland.

Francesco Sclano and Paola Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications, I-ESA 2007*, Funchal (Madeira Island), Portugal, March.

Anna Shtok, Oren Kurland, and David Carmel. 2009. Predicting query performance by query-drift estimation. In Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rger, Milad Shokouhi, Dawei Song, and Emine Yilmaz, editors, *Advances in Information Retrieval Theory*, volume 5766 of *Lecture Notes in Computer Science*, pages 305–312. Springer Berlin Heidelberg.

Ian H. Witten, Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes (2Nd Ed.): Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Xing Zhang, Yan Song, and A.C. Fang. 2010. Term recognition using conditional random fields. In *Natural Language Processing and Knowledge Engineering, NLP-KE 2010*, pages 1–6, Aug.

Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and RyenW. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 52–64. Springer Berlin Heidelberg.

Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2007, pages 543–550, New York, NY, USA. ACM.

Hadas Zohar, Chaya Liebeskind, Jonathan Schler, and Ido Dagan. 2013. Automatic thesaurus construction for cross generation corpus. *Journal on Computing and Cultural Heritage (JOCCH)*, 6(1):4:1–4:19, April.