Term Associations in Query Expansion: a Structural Linguistic Perspective

Michael Symonds* mikesymo@gmail.com
Guido Zuccon† guido.zuccon@csiro.au
Bevan Koopman† bevan.koopman@csiro.au
Peter Bruza* p.bruza@qut.edu.au
Laurianne Sitbon‡ l.sitbon@qut.edu.au

ABSTRACT

Many successful query expansion techniques ignore information about the term dependencies that exist within natural language. However, researchers have recently demonstrated that consistent and significant improvements in retrieval effectiveness can be achieved by explicitly modelling term dependencies within the query expansion process. This has created an increased interest in *dependency*-based models.

State-of-the-art dependency-based approaches primarily model term associations known within structural linguistics as syntagmatic associations, which are formed when terms co-occur together more often than by chance. However, structural linguistics proposes that the meaning of a word is also dependent on its paradigmatic associations, which are formed between words that can substitute for each other without effecting the acceptability of a sentence. Given the reliance on word meanings when a user formulates their query, our approach takes the novel step of modelling both syntagmatic and paradigmatic associations within the query expansion process based on the (pseudo) relevant documents returned in web search. The results demonstrate that this approach can provide significant improvements in web retrieval effectiveness when compared to a strong benchmark retrieval system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval
Terms: Algorithms, Experimentation
Keywords: Query Expansion, Term Dependencies

*School of Information Systems, Queensland University of Technology, Brisbane, Qld, Australia, 4001
†Australian e-Health Research Centre, CSIRO, Brisbane, Qld, Australia, 4001
‡Department of Computer Science, Qld University of Technology, Brisbane, Qld, Australia, 4001

CIKM 2013 San Francisco, USA

http://dx.doi.org/10.1145/2505515.2507852.

1. INTRODUCTION

Dependency-based models of information retrieval have demonstrated superior retrieval effectiveness over models that ignore term dependencies, like $tf.idf$ and language modelling approaches [8, 9, 6, 2]. These approaches often use the intuition that terms that co-occur in context with the query terms within a document are likely to assist retrieval effectiveness. These dependencies are similar to those defined as *syntagmatic associations* used within structural linguistics.

Before Chomsky's theories of generative grammar, linguistics was dominated by the structuralist theories of Ferdinand de Saussure (1916). Saussure proposed that the meaning of a word was created from its syntagmatic and paradigmatic associations. Syntagmatic associations are formed between words that co-occur above chance within natural language [7]. Typical examples include, *sun - hot* and *coffee - taste*. The association between two words is considered paradigmatic if they can substitute for one another in a sentence without effecting the acceptability of the sentence [7]. Typical examples include *article - paper* and *dog - cat*. These definitions indicate that syntagmatic and paradigmatic associations can be modelled solely from occurrence patterns of words observed in natural language.

The motivation for modelling both syntagmatic and paradigmatic information within the information seeking process stems from the reliance on word meanings when a user formulates their query. Consider the example query (*best coffee machine*). The user's information need may rely on associations to words like "*lowest, price, tasting, espresso, maker*". These associations can be argued to have syntagmatic: (***best***-*price*; *tasting*-***coffee***; *espresso*-***machine***); and paradigmatic: (***best***-*lowest*; ***coffee***-*espresso*; ***machine***-*maker*) associations with the original query terms.

Given state-of-the-art dependency-based approaches primarily model syntagmatic information and the reliance of the information seeking process on word meanings, it was hypothesised that: modelling both syntagmatic and paradigmatic associations in the information retrieval process would provide *significant* improvements in retrieval effectiveness. Preliminary evaluations testing this hypothesis incorporated an efficient, computational model of word meaning, known as the *Tensor Encoding* (TE) model [10], within the query expansion process. This approach, known as *Tensor Query Expansion* (TQE), demonstrated significant improvements in ad hoc retrieval effectiveness over the unigram relevance model, on small newswire collections [11].

However, our research aims to assess this approach on the TREC 2012 Web track (i.e., on a large web collection) compared to a much stronger benchmark system (based on the Google retrieval service).

2. RELATED WORK

Dependency-based query expansion techniques, such as *Latent Concept Expansion* (LCE), are growing in popularity and have demonstrated superior effectiveness over those that ignore term dependencies [9, 5]. Most are based on the likelihood estimates of terms, or the co-occurrence information of a possible expansion term with a query term, and hence within a (pseudo) relevance feedback setting can be argued to model syntagmatic associations [12]. However, paradigmatic information is modelled by looking at the vocabulary terms that co-occur often with a query term and the potential expansion term (i.e., not the co-occurrence between the query term and potential expansion term itself).

The TE model allows the TQE approach to model paradigmatic associations between a sequence of terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w, using a novel estimation technique:

$$s_{\text{par}}(Q, w) = \frac{1}{Z_{\text{par}}} \sum_{j \in Q} \sum_{i \in V_k} \frac{f_{\overline{ij}} . f_{\overline{iw}}}{\max(f_{\overline{ij}}, f_{\overline{iw}}, f_{\overline{wj}})^2}, \quad (1)$$

where $f_{\overline{ij}}$ is the unordered co-occurrence frequency of terms i and j seen within a sliding context window moved across the set of (pseudo) relevant documents, V_k is the vocabulary created from the set of k (pseudo) relevant documents), and Z_{par} normalizes the distribution. The context window size is often set to 1, as this has been shown to effectively model paradigmatic associations [10].

In a (pseudo) relevance feedback setting, the Dirichlet smoothed likelihoods estimates of query terms within the (pseudo) relevant documents have been shown to efficiently and effectively model syntagmatic information [12], and hence was chosen as the syntagmatic measure in this work. This means that the TQE approach becomes a unigram relevance model when relying solely on the syntagmatic measure.

Past research, using primarily paradigmatic information sourced from WordNet[1] to expand query representations was unable to achieve consistent improvements in retrieval effectiveness [13]. Corpus-based, query expansion techniques, that *implicitly* model syntagmatic and paradigmatic associations have been presented in the past and have demonstrated significant improvements in retrieval effectiveness on small newswire collections [5, 1]. The features of the TQE approach that separates it from previous corpus-based, query expansion techniques is (i) its ability to *explicitly* model and combine measures of syntagmatic and paradigmatic information within a single, formal framework, and (ii) its superior efficiency.

Given the TQE approach has only been evaluated on small newswire data sets and against models based within the unigram language modelling framework, an evaluation on web-scale retrieval tasks compared to strong benchmark systems is required before wider conclusions can be drawn. The TREC WebTrack provides such an opportunity. LCE, which is based on the MRF document ranking model is often considered to be a strong benchmark model [5]. However, the TREC forum allows systems to be compared (i.e., not just models). Therefore, a stronger benchmark is possible. Based on industry reputations, the Google web service is chosen to underpin our benchmark model. This choice is supported by our benchmark submission achieving an ERR=0.29 compared to an ERR=0.313 for the best TREC 2012 WebTrack submission [4]. The average of all TREC 2012 Web Track submissions was ERR=0.187.

[1]A hand-crafted ontology of English words grouped into cognitive synonyms, http://wordnet.princeton.edu/

3. METHOD

3.1 Benchmark System

The benchmark model is created in the following way. The *ClueWeb09-Category B* documents are indexed using the 'indexing without spam' approach [14]. Each query is then issued to the Google retrieval service[2] and the top 60 retrieved documents are filtered using the spam filtered ClueWeb09 Category B index[3]. This filtered list is then padded, to create a ranked list of 10,000 documents, using the ranked documents returned by a unigram language model on the spam filtered index. These rankings form our benchmark system (**GBline**) and this process is depicted in Figure 1.

3.2 Extending the Benchmark with TQE

To augment query representations within the TQE approach, an estimate of the conditional probability $P(w|Q)$, i.e. the probability of selecting a vocabulary term w as an expansion term given the query Q, is provided by:

$$P(w|Q) = \frac{1}{Z} \left[\gamma s_{\text{par}}(Q, w) + (1 - \gamma) s_{\text{syn}}(Q, w)\right], \quad (2)$$

where w is any term in the TE vocabulary (formed from the set of k pseudo relevant documents returned by the benchmark model - GBline), Q is the sequence of original query terms, $s_{\text{par}}(Q, w)$ is the paradigmatic measure shown in Equation (1), $s_{\text{syn}}(Q, w)$ is the syntagmatic measure (i.e., the Dirichlet smoothed likelihood estimates), $\gamma \in [0, 1]$ mixes the paradigmatic $s_{\text{par}}()$ and syntagmatic $s_{\text{syn}}()$ measures, and Z normalises the resulting distribution.

From these estimates an augmented query representation Q' is created, as shown in Figure 2. In our study, this updated query representation is passed to a unigram language model to perform the final search on the spam filtered ClueWeb09 Category B index. A unigram language model was used as Google often treats long queries in a reductive approach, resulting in no documents matching the search. The system depicted in Figure 2 is referred to as **GTQE** in the remainder of this paper.

4. AD HOC WEB RETRIEVAL

4.1 Experimental Setup

The ClueWeb09 Category B dataset and the TREC Web Track 51-200 topics were used to evaluate the approaches presented in Section 3; collection statistics are reported in Table 1. The documents were stopped using the standard INQUIRY stop-word list and stemmed using a Krovetz stemmer, as implemented by Indri toolkit [4]. Queries were formed from the title components of the TREC topics.

[2]http://www.google.com

[3]We limited the number of documents retrieved with Google to 60 because of Google's policies regarding the retrieval service at the time.

[4]Available at http://sourceforge.net/projects/lemur

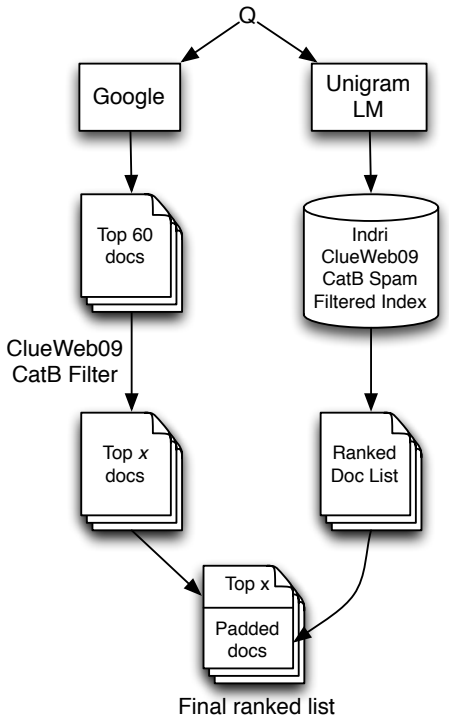

Figure 1: Benchmark System (GBline).

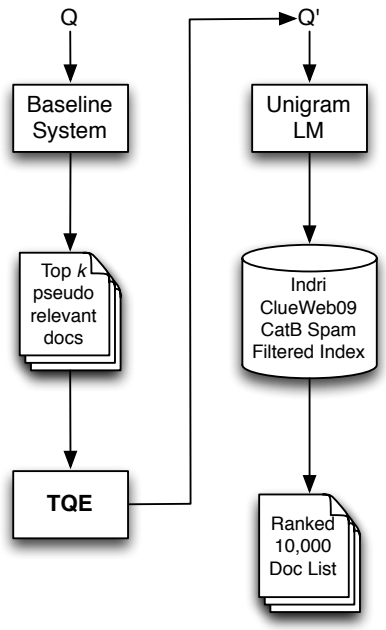

Figure 2: System using TQE (GTQE).

Description	# Docs	Topics	$\overline{\lvert q \rvert}$	$\overline{\lvert D \rvert}$
ClueWeb09 Category B	50,220,423	Web Track 51-200	2.72 (1.38)	804

Table 1: Collection statistics for the TREC ClueWeb09 Category B collection. $\overline{|q|}$ represents the average length of the queries, the value in brackets is the standard deviation of the query lengths, and $\overline{|D|}$ is the average document length.

Documents were indexed using the 'indexing without spam' method; the Waterloo spam list with threshold of 0.45 was used to estimate spam-likelihood of documents [14]. Indexing and retrieval approaches were implemented using the Indri toolkit. The parameters used within the unigram language model were based on the Indri defaults.

4.2 Training the GTQE System

Tuning of the GTQE system parameters was achieved by training on ERR@20 using the TREC Web topics from 2010 and 2011 (i.e., 51-150). The test runs were performed on the 2012 TREC Web track topics (151-200). The test parameter values used by the GTQE system were *Number of feedback documents* equal to 19, *number of expansion terms* equal to 14, and *TE model mixing parameter* (γ in Equation (2)) equal to 0.1. The ERR@20 of the TQE system during training (i.e., on topics 51-150) varied between 0.1201 and 0.1302 for (i) 5 to 25 expansion terms, and (ii) 4 to 30 feedback documents.

4.3 Retrieval Results

In this section we compare the results of the two retrieval systems (GBline and GTQE) on the task of ad hoc web retrieval. MAP, P@20, ERR@20 and nDCG@20 for the top ranked 10,000 documents for both system are reported in Table 2.

	Binary Metrics		Graded Metrics	
	P@20	MAP	ERR@20	nDCG@20
GBline	0.305	0.117	**0.290**	0.167
GTQE	**0.396**^{b} (+29.8%)	**0.158**^{b} (+35%)	0.249 (-14.2%)	**0.192** (+15%)

Table 2: Retrieval performance on the TREC 2012 Web Track ad hoc retrieval task. Superscript b indicates statistically significant differences (calculated using a paired t-test $p < 0.05$) over the benchmark (GBline). The best result for each evaluation measure appears in boldface. Brackets indicate the percentage change from GBline to GTQE.

The results demonstrate that expanding query representations using TQE can provide significant improvements over GBline on binary metrics (i.e. MAP and P@20). Binary metrics are those which use relevance judgements of 0 (non relevant) and 1 (relevant) for each document. However, no significant difference in retrieval effectiveness was noted on the graded metrics (ERR@20 and nDCG@20).

Graded metrics are those that base their effectiveness score on documents that are assigned a relevance judgement in a range, e.g., between 0 and 4. In addition, measures that use graded judgements, such as ERR, bias the scores for systems that return relevant documents toward the very top of the ranked list (i.e., in positions 1, 2 and 3). This causes a heavy discounting to occur for relevant documents ranked lower in the list, as seen from the expression used to calculate ERR at rank k [3]. Given Google rankings are likely augmented with click through data and editorial choice, the GBline system (Figure 1) is able to ensure highly relevant documents

are ranked in the top few positions. However, as the GTQE system (Figure 2) performs its final ranking using a unigram language model, which does not use such information, it is not surprising that the GTQE system is unable to achieve significant improvements over GBline on the graded metrics (i.e., ERR@20 and nDCG@20).

Note that the GTQE system achieved significant improvements over GBline on the P@20 metric (Table 2), indicating that many more relevant documents were returned in the top 20 by GTQE than GBline. It may be therefore reasonable to deduce that significant improvements on ERR@20 and nDCG@20 may be achievable if a final re-ranking step, that took into account these graded relevance judgements, was added to the GTQE system (Figure 2).

To provide a comparison with an alternate query expansion approach, it is worth noting that our implementation of TQE becomes a unigram relevance model when $\gamma = 0$ (i.e., uses the syntagmatic measure to produce estimates). The effectiveness of TQE when $\gamma = 0$ was reported as ERR=0.241, 3% less than TQE.

4.4 Robustness Analysis

Robustness analysis includes considering the ranges of relative increase/decrease in effectiveness and the number of queries that were improved/degraded, with respect to some baseline. Figure 3 illustrates the relative increase/decrease of P@20 scores for GBline and GTQE over the average of all TREC 2012 Web track submissions (MeanWT12)[5].

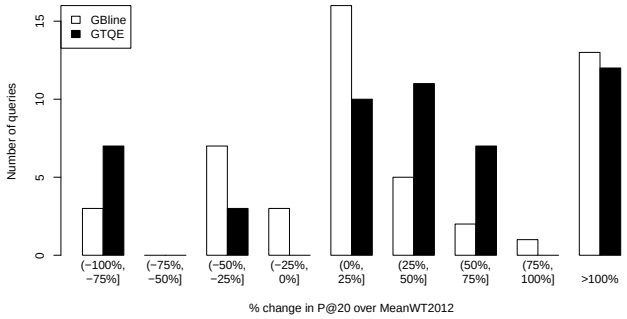

Figure 3: Robustness comparison of the GBline and GTQE systems when compared with MeanWT12.

Figure 3 shows that the GTQE system provides more consistent improvements over MeanWT12 than the GBline system - as indicated by the distribution of GTQE being positioned more to the right than GBline.

This graph also highlights the adverse impact GTQE has on seven queries - as seen at the extreme left of the graph. Initial investigations indicate that this effect may be due to the very short nature of the effected queries, which have an average length of 1.8 (c.f., 2.7 for the test set), which may impact the effectiveness of modelling the word associations. More detailed investigations into this effect are left for future work.

5. CONCLUSION

Dependency-based models of information retrieval primarily use information about word associations known as syntagmatic associations. Within structural linguistics, word meanings are induced from syntagmatic and paradigmatic associations. Given the reliance on word meanings in the information seeking process it was hypothesised that modelling both syntagmatic and paradigmatic information within a dependency-based approach would provide significant improvements in retrieval effectiveness.

The TQE approach provides a formal framework in which to achieve this. When the TQE approach is used to expand query representations on the TREC 2012 WebTrack significant improvements in retrieval effectiveness are achieved when compared to a strong benchmark system created from the Google retrieval service.

[5]P@20 was used as Section 4.3 suggests that a comparison on ERR@20 or nDCG@20 is unlikely to be meaningful.

6. REFERENCES

[1] J. Bai, J.-Y. Nie, G. Cao, and H. Bouchard. Using Query Contexts in Information Retrieval. In *SIGIR'07*, pages 15–22, New York, NY, USA, 2007. ACM.

[2] C. Carpineto and G. Romano. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50, Jan. 2012.

[3] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In *CIKM '09*, pages 621–630, NY, USA, 2009. ACM.

[4] C. Clarke, N. Craswell, and E. Voorhees. Overview of the TREC 2012 Web Track. In *TREC '12*. NIST, 2012. Special Publication.

[5] Y. Hou, Z. Zhao, D. Song, and W. Li. Mining Pure High-order Word Associations via Information Geometry for Information Retrieval. *ACM TOIS*, pages In–press, 2013.

[6] Y. Lv and C. Zhai. Positional Relevance Model for Pseudo-relevance Feedback. In *SIGIR '10*, pages 579–586, New York, NY, USA, 2010. ACM.

[7] J. Lyons. *Introduction to Theoretical Linguistics*. Cambridge University Press, London, 1968.

[8] D. Metzler and W. B. Croft. A Markov Random Field Model for Term Dependencies. In *SIGIR '05*, pages 472–479, New York, NY, USA, 2005. ACM.

[9] D. Metzler and W. B. Croft. Latent Concept Expansion using Markov Random Fields. In *SIGIR '07*, pages 311–318, New York, NY, USA, 2007. ACM.

[10] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Modelling Word Meaning using Efficient Tensor Representations. In *PACLIC '11*, pages 313–322. In: Helena Hong Gao and Minghui Dong (eds), 2011.

[11] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Tensor Query Expansion: a cognitive based relevance model. In *ADCS '11*, pages 87–94. RMIT University(Melbourne), 2011.

[12] M. Symonds, P. Bruza, G. Zuccon, L. Sitbon, and I. Turner. Is The Unigram Relevance Model Term Independent?: Classifying Term Dependencies in Query Expansion. In *ADCS '12*, pages 123–127, New York, NY, USA, 2012. ACM.

[13] E. M. Voorhees. Query Expansion Using Lexical-semantic Relations. In *SIGIR '94*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[14] G. Zuccon, A. Nguyen, T. Leelanupab, and L. Azzopardi. Indexing without Spam. In *ADCS '11*, pages 87–94. RMIT University(Melbourne), 2011.