JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY
First and Last Volume

Automatic Query Expansion: a Structural Linguistic Perspective

Michael Symonds[1,3], Peter Bruza[1], Guido Zuccon[2],
Bevan Koopman[1,2], Laurianne Sitbon[1], Ian Turner[1]

[1]Queensland University of Technology, Brisbane, QLD 4001
[2]Australian e-Health Research Centre, CSIRO, Brisbane, QLD 4001
michael.symonds@qut.edu.au; p.bruza@qut.edu.au; guido.zuccon@csiro.au
bevan.koopman@csiro.au; laurianne.sitbon@qut.edu.au; i.turner@qut.edu.au
[3] Corresponding author (phone: +614 23 239 013)

A user's query is considered to be an imprecise description of their information need. Automatic query expansion is the process of reformulating the original query with the goal of improving retrieval effectiveness. Many successful query expansion techniques model the syntagmatic associations that exist between words in natural language. However, structural linguistics relies on both syntagmatic and paradigmatic associations to deduce the meaning of a word. Given the success of dependency-based approaches to query expansion and the reliance on word meanings in the query formulation process, we argue that modeling both syntagmatic and paradigmatic information in the query expansion process will improve retrieval effectiveness.

This article develops and evaluates a new query expansion technique that is based on a formal, corpus-based model of word meaning that models syntagmatic and paradigmatic associations. We demonstrate that when sufficient statistical information exists, as in the case of longer queries, including paradigmatic information alone provides significant improvements in retrieval effectiveness across a wide variety of data sets. More generally, when our new query expansion approach is applied to large-scale web retrieval it demonstrates significant improvements in retrieval effectiveness over a strong baseline system, based on a commercial search engine.

Effectively retrieving relevant information from large document collections, such as those found on-line, poses many challenges that has seen strong research interest in the field of information retrieval. At the base of this problem is the ability to judge the relevance of a document for a user's query. Since the Cranfield experiments in document retrieval (Cleverdon, Mills, & Keen, 1966) it has been accepted that a user's query is an imprecise description of their information need. For this reason there is a strong interest in the use of query expansion techniques to augment the query, to arguably be a more precise representation of the information need, and allow more relevant documents to be retrieved. Such techniques have been shown to significantly increase average retrieval effectiveness (Lavrenko & Croft, 2001; Zhai & Lafferty, 2001). Although there have been a wide variety of query expansion approaches proposed in the literature, improvements in retrieval effectiveness have often been derived from explicitly modeling term dependency information within the query expansion process (Metzler & Croft, 2007; Lv & Zhai, 2010; Xue & Croft, 2013). Many of these dependency-based approaches use the intuition that helpful terms for expansion can be computed based on term statistics drawn from the corpus or query log. A natural assumption is that useful expansion terms co-occur in context with the query terms, where the context is often defined as a whole document or perhaps a window of words of a given length. For example, Billhardt, Borrajo, and Maojo (2002) employ term context vectors based on how terms co-occur in a document and then expand the query by computing semantic associations from these vector representations.

At its very heart, query expansion is a means of addressing the vocabulary mismatch problem between query and document representations, which is in turn defined in terms of synonymy (two or more terms with the same meaning) and polysemy (one term with more than one meaning). Defined in this way, we see that query expansion fundamentally deals with word meaning, which is not often in the foreground of query expansion models, particularly those of a statistical nature. This is because probabilistic approaches do not directly model the meaning(s) of a term, but rather focus of ways to compute probabilistic associations between terms. For example, Chung and Jae (2001) evaluate six probabilistic term association measures for query expansion without ever addressing how the meaning or semantics of the terms are involved, and this trend has basically continued to this day. This is not to say that term semantics has not

intersected with probabilistic approaches. For example, a prominent probabilistic query expansion model called “Latent concept expansion" (LCE) developed by Metzler and Croft (2007), who reflect “the use of phrase and proximity features within the model captures syntactic dependencies [between terms], whereas LCE captures query-side semantic dependence". Similarly, Fang and Zhai (2006) provide an axiomatic basis for semantic term matching and demonstrate how this theory improves query expansion. In a similar vein, Bai, Song, Bruza, Nie, and Cao (2005) augment a standard statistical language modeling approach to query expansion with term relationships computed from a semantic space model derived from the underlying document corpus. This article aligns with these works by placing word meaning in the foreground and then develops an account of associations for use in query expansion. This will be achieved by drawing inspiration from the field of structural linguistics.

Structural linguistics states that the meaning of a word can be induced from its syntagmatic and paradigmatic associations. Syntagmatic associations exist between words that co-occur with each other above chance. Typical examples include *hot - sun* or *JASIST - article*. We have argued that syntagmatic associations lie at the basis of current query expansion approaches (Symonds, Bruza, Zuccon, Sitbon, & Turner, 2012) because syntagmatic associations depend on how terms co-occur in context. However, within natural language, there exists another fundamental type of relationship between words, known as a *paradigmatic association*. The association between two words is deemed paradigmatic if they can substitute for one another in a sentence (or context) without affecting the acceptability of the sentence (Lyons, 1968). Typical examples are synonyms like *paper - article*, or related verbs like *eat - drink*. Syntagmatic and paradigmatic associations underpin a *differential view of meaning* (Pavel, 2001), which has been adopted by a number of prominent linguists, including work on sense relations by Lyons (1968). The differential view of meaning, presented by structural linguistics, has been argued to form a relatively clean theory, free of psychology, sociology and anthropology (Holland, 1992). As a consequence, structural linguistics provides a relatively unobstructed path toward developing computational models of word meaning, and hence query expansion.

Given the theoretical setting offered by structural linguistics, the ability to model word meaning becomes heavily dependent on identifying statistical relationships between words. The premise behind the distributional hypothesis states that words with similar meaning will tend to co-occur with similar words (Harris, 1954). By way of illustration, according to the distributional hypothesis "doctor" and "nurse" are semantically similar because they both tend to co-occur with words like "hospital", "sick" etc. The distributional hypothesis underpins a number of computational models of word meaning that have an established track record of replicating human word association norms in cognitive studies, e.g.,. LSA (Latent Semantic Analysis (Landauer & Dumais, 1997)) and HAL (Hyperspace to Analogue of Language (Burgess, Livesay, & Lund, 1998)). Such models are relevant for query expansion as this process naturally involves establishing associations between terms and from the user point of view this process is cognitively situated. The task of an automatic query expansion system is arguably to replicate those associations.

A more recent distributional model of word meaning, known as the *Tensor Encoding* (TE) model, demonstrated robust performance on a wide variety of semantic tasks, including synonym judgement (Symonds, Bruza, Sitbon, & Turner, 2011a), semantic categorization (Symonds, Bruza, Sitbon, & Turner, 2012) and similarity judgements of medical concepts (Symonds et al., 2011a) and importantly for this research, formally combines measures that model both syntagmatic and paradigmatic associations between words.

The first premise of this paper is:

> **H1:** As users rely heavily on word meanings when formulating their queries, modeling the meaning of words by incorporating the TE model within the query expansion process will improve retrieval effectiveness.

This hypothesis is tested by developing a new, formal, query expansion approach, based on the TE model and called *tensor query expansion* (TQE). The approach is evaluated on ad hoc retrieval tasks for a wide variety of data sets, including short and long queries, and newswire and web-based document collections.

Approaches that model word associations in the query expansion process have had mixed success in the past (Voorhees, 1994; Bruza & Song, 2002). However, these attempts have not used a formal model of word meaning that *explicitly* combines information about both syntagmatic and paradigmatic associations between words.

The second premise of this paper is:

> **H2:** As state-of-the-art query expansion techniques primarily model syntagmatic associations, which probe only half the associations underpinning word meaning, the inclusion of paradigmatic information will provide the other half of the associations and improve retrieval effectiveness.

This hypothesis will be tested by controlling the influence of syntagmatic and paradigmatic information within the TQE approach compared to a strong benchmark using the same source of syntagmatic associations, and other model parameter values.

This article makes four major contributions: first, a novel framework for modeling query expansion, in which the

original query representation is expanded using information about syntagmatic and paradigmatic associations of the query terms; second, an implementation of this framework that does not rely on any external linguistic resources; third, a rigorous evaluation of the benefits of including paradigmatic information in the query expansion process; fourth, this novel query expansion technique is evaluated in an industry setting and compared to a strong benchmark model.

The remainder of this article is structured as follows: The related work section outlines relevant approaches to query expansion along with an overview of the tensor encoding model of word meaning. The experimental methodology section details the data sets, benchmark models and tasks on which the TQE approach will be evaluated. The experimental results of these evaluations are then discussed before concluding remarks and future work are presented.

Related Work

The related work for this paper includes: (i) an overview of popular and successful query expansion techniques, and (ii) an overview of past efforts to use information about word associations to augment query representations.

Query Expansion

The query expansion process is often achieved using relevance feedback, which relies on the user indicating which of the top k returned documents were relevant. To reduce the burden on the user the top k documents can be assumed to be relevant, and in this case, the relevance feedback setting is referred to as pseudo relevance feedback or blind feedback.

Query expansion within a (pseudo) relevance feedback setting has been shown to provide significant improvements in retrieval effectiveness (Lv & Zhai, 2010; Metzler & Croft, 2007; Lavrenko, 2004). However, this process is often sensitive to model parameter tuning, and does not consistently assist retrieval effectiveness for *all* queries (Collins-Thompson, 2009; Billerbeck & Zobel, 2004).

We will now present a brief discussion of a number of successful and relevant query expansion approaches. This discussion motivates the choice of benchmark models used in this work.

Rocchio. The Rocchio method (Rocchio, 1971) is designed for working with geometric representations, such as those found within vector space models (Salton, Wong, & Yang, 1975; Buckley, 1995). Rocchio updates the query vector weights using relevance information, such that the query vector is moved closer in space to the vectors representing the relevant documents and away from those representing non-relevant documents. The most common form of the Rocchio algorithm modifies the initial query weights of the query vector Q, according to:

$$q_j(1) = \alpha q_j(0) + \beta \frac{1}{|R|} \sum_{D_i \in R} d_{ij} - \gamma \frac{1}{|NR|} \sum_{D_i \in NR} d_{ij}, \quad (1)$$

where $q_j(0)$ is the initial weight of term j, R is the set of relevant documents in the collection, d_{ij} is the weight of term j in document D_i, NR is the set of non-relevant documents in the collection, and α, β and γ are parameters that control the effect of each component in the equation. In particular, β influences the amount of positive feedback used and γ influences the amount of negative feedback used.

The Relevance Modeling Framework. The ideas behind the Rocchio approach have been used to create a model-based feedback technique that minimizes the divergence between the query distribution and those of the (pseudo) relevant documents (Zhai & Lafferty, 2001).

Another popular technique, that formally augments query representations within the language modeling framework, is known as the relevance modeling approach (Lavrenko & Croft, 2001). This approach is robust (Lv & Zhai, 2009) and is regarded as a benchmark in query expansion research (Metzler & Croft, 2007; Lv & Zhai, 2010) and hence will be used as the reference approach for comparison with techniques developed in this work.

The relevance model augments the query representations used within the retrieval process by estimating the probability of observing a word w given some relevant evidence for a particular information need, represented by the query Q:

$$\begin{aligned} P(w|Q) = P(w|R) &= \int_D P(w|D)P(D|Q), \\ &\approx \frac{\sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}{\sum_w \sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}, \end{aligned} \quad (2)$$

where $\mathcal{R}_Q$ is the set of (pseudo) relevant documents for query Q, D is a document in $\mathcal{R}_Q$, $P(D|Q)$ is the document score of D given Q produced by the underlying language model, and $P(w|D)$ is estimated based on document statistics.

The estimate in Equation (2) is often interpolated with the original query model, to form a final estimate:

$$P(w|Q) = \alpha P_o(w|Q) + (1 - \alpha)P(w|R), \quad (3)$$

where α is the feedback interpolation coefficient that determines the mix with the original query model estimate $P_o(w|Q)$. This updated query representation is then used to re-rank documents.

The Unigram Relevance Model. In the unigram variant of the relevance model, $P(w|D)$ in Equation (2) is often estimated using the Dirichlet smoothed query likelihoods:

$$P(w|D) = \frac{tf_{w,D} + \mu \frac{cf_w}{|C|}}{|D| + \mu}, \quad (4)$$

where $tf_{w,D}$ is the frequency of a term w in D, $|D|$ is the length of the document, cf_w is the frequency of term w in the collection C, $|C|$ is the total number of terms in the collection, and μ is the Dirichlet smoothing parameter. Within a (pseudo) relevance feedback setting, the estimate in Equation (4) is based on document frequencies in the set of (pseudo) relevant documents and hence models syntagmatic associations for query terms (Symonds, Bruza, Zuccon, et al., 2012). In this research, the unigram based relevance model using the Dirichlet smoothed estimate of Equation (4) is referred to as **RM3**.

Even though the unigram relevance model has demonstrated significant improvements in retrieval effectiveness over a unigram language model, recent research has shown that significant improvements can be made over the unigram relevance model by explicitly modeling information about term dependencies in the expansion process. These approaches include the *positional relevance model* (PRM) (Lv & Zhai, 2010) and *latent concept expansion* (LCE) (Metzler & Croft, 2007).

The Positional Relevance Model. Lv and Zhai (2010) found that using positional information of terms within the relevance modeling framework can significantly improve retrieval effectiveness over a unigram approach. Based on the intuition that topically related content is grouped together in text documents, the positional relevance model (PRM) uses proximity and positional information to form expansion term estimates to update the query model. The estimate of observing an expansion term w given a query Q is computed as:

$$P(w|Q) = \frac{P(w,Q)}{P(Q)} \propto P(w,Q) = \sum_{D \in \mathcal{R}_Q} \sum_{i=1}^{|D|} P(w,Q,D,i), \quad (5)$$

where i indicates a position of w in D.

The estimate of $P(w,Q,D,i)$ relies on computing the conditional probability of observing word w (the expansion term) given the document and position of w in the document, i.e., $P(w|D,i)$.

$$P(w|D,i) = (1-\lambda)\frac{c'(w,i)}{\sqrt{2\pi\sigma^2}} + \lambda P(w|C), \quad (6)$$

where

$$c'(w,i) = \sum_{j=1}^{|D|} c(w,j) \exp\left[\frac{-(i-j)^2}{2\sigma^2}\right], \quad (7)$$

and $c(w,j)$ is the *actual* count of term w at position j, λ is a smoothing parameter and σ is used to parameterize the Gaussian kernel function ($f(i) = a \exp\left[\frac{-(i-b)^2}{2c^2}\right]$).

Latent Concept Expansion. *Latent concept expansion* (LCE) (Metzler & Croft, 2007) was developed as a query expansion approach within the framework of the *Markov Random Field* document ranking model (Metzler & Croft, 2005). LCE formally combines various likelihood based measures that effectively model syntagmatic associations for query terms, ordered bigrams and unordered bigrams (Xue & Croft, 2013).

As LCE computes the likelihoods of ordered bigrams, and unordered bigrams the complexity of the model increases exponentially with the length of the query, as is the case on verbose queries.

Query Expansion using Word Associations

The query expansion approaches presented thus far use estimation techniques that can be argued to rely heavily on information about syntagmatic associations (Symonds, Bruza, Zuccon, et al., 2012). The (pseudo) relevance feedback process itself naturally models syntagmatic associations as words that co-occur more often with the query terms are more likely to exist within the set of (pseudo) relevant documents from which the expansion term estimates are derived. However, explicit modeling of dependency information, such as through positional information or bigram likelihoods, as used in PRM and LCE respectively, can also be argued to primarily model syntagmatic associations.

A number of past techniques have taken a more linguistic approach to expanding query representations, by using information about word associations (Voorhees, 1994; Greenberg, 2001; Bai et al., 2005; Hoenkamp, Bruza, Song, & Huang, 2009; Grefenstette, 1992; Bruza & Song, 2002; Xu & Croft, 1996). One approach, known as *Local Context Analysis* (Xu & Croft, 1996) demonstrated analysis of context and phrase structure can be used to help improve retrieval effectiveness. Another approach, relying solely on paradigmatic information to estimate expansion terms, incorporated a linguistic resource, WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), and was unable to produce consistent improvements in retrieval effectiveness (Voorhees, 1994).

Other linguistic attempts, including the information flow model (Bruza & Song, 2002; Bai et al., 2005) that rely on some mix of syntagmatic and paradigmatic information, have provided some improvements in retrieval effectiveness on small data sets (i.e., small newswire document collections). However, these approaches have not been evaluated on larger datasets. These approaches make use of *corpus-based, distributional models* which model the semantic associations between words directly from the co-occurrence patterns of words found in streams of natural language. Therefore, they do not rely on external linguistic resources, and hence are considered to provide a relatively cheap, language independent method of accessing information about word associations. However, for this research, their most attractive feature is their ability to model both syntagmatic and paradigmatic word associations.

Corpus-based Distributional Models. Researchers have argued that relationships between words can be modeled by comparing the distributions of words found

within streams of natural language (Schütze, 1993). Similar to the language development in young children (Jones & Mewhort, 2007), these models build up word distributions by identifying frequently co-occurring words in natural language. Instead of storing these distributions in neural networks, as the brain does, a powerful alternative is to represent these distributions within high-dimensional vector spaces (Turney & Pantel, 2010).

Creating vector representations of words allows techniques from linear algebra to be used to model relationships between objects, including syntagmatic and paradigmatic associations. These approaches are often referred to as *semantic space models* (SSMs), as the distance between words in the space often reflect various semantic groupings, i.e., words related through some semantic association. Spatial representations of semantic associations have been used within psychology for many decades to model affective (emotional) meaning of words (Osgood, Suci, & Tannenbaum, 1957).

There have been a number of successful corpus-based SSMs that have emulated human performance on tasks (including synonym judgement) by learning semantic associations directly from text, including HAL (Hyperspace Analogue to Language (Lund & Burgess, 1996)) and LSA (Latent Semantic Analysis (Landauer & Dumais, 1997)). More recent models (Jones & Mewhort, 2007; Symonds et al., 2011a) have incorporated advances that have addressed issues in earlier SSMS, including the lack of structural information stored in the representations, and the ability to support higher-order tensor representations.[1] Of these more recent models, the *Tensor Encoding* (TE) model (Symonds et al., 2011a) has demonstrated robust effectiveness across a range of semantic tasks (Symonds, Bruza, Sitbon, & Turner, 2012; Symonds, Zuccon, Koopman, Bruza, & Nguyen, 2012) and more importantly for this research, explicitly models syntagmatic and paradigmatic associations.

The TE model efficiently builds high-dimensional, tensor representations of words through a formal binding process and the use of a novel compression technique. These representations are then used to underpin measures of syntagmatic and paradigmatic information, which are combined within a formal framework to provide a probability estimate $P(w|q)$ that words w and q share similar semantic associations within the vocabulary.

Because of the centrality of the TE model to this paper, we discuss how these representations are built and how this impacts the computational complexity of any query expansion technique that is based on the TE model. The TE model's process for creating a representation for each term, known as a *memory tensor*, involves a geometric binding process that uses fixed dimension environment vectors. Each term's environment vector corresponds to a unique unit vector, so that an orthonormal basis is formed (the canonical basis in this case). To illustrate, consider the TE model's binding process for the example sentence, *a dog bit the mailman*, and the resulting vocabulary terms and environment vectors in Table 1.[2]

Table 1
Example vocabulary for: A dog bit the mailman

Id	Term	Environment vector
1	dog	$\boldsymbol{e}_{\text{dog}} = (1 \quad 0 \quad 0)^T$
2	bit	$\boldsymbol{e}_{\text{bit}} = (0 \quad 1 \quad 0)^T$
3	mailman	$\boldsymbol{e}_{\text{mailman}} = (0 \quad 0 \quad 1)^T$

The memory tensor for each term in the vocabulary is constructed by summing the proximity-scaled Kronecker products of the environment vectors within a sliding context window over the text. The number of environment vectors bound using Kronecker products impacts the order of the memory tensors. For example, the binding process that would capture word order and co-occurrence information of 2-tuples within second-order tensor (matrix) representations:

$$\boldsymbol{M}_w = \sum_{k \in \{C|k<w\}} (R - d_k + 1).\boldsymbol{e}_k \otimes \boldsymbol{e}_w^T + \sum_{k \in \{C|w<k\}} (R - d_k + 1).\boldsymbol{e}_w \otimes \boldsymbol{e}_k^T, \tag{8}$$

where C is a totally ordered set of terms created by the sliding context window, containing two order relations $k \prec w$ and $w \prec k$, where $w \in C$ is the target term, $k \in C$ is a non-stop word found within the context window, $k \prec w$ indicates that term k appears before term w in C, R is the radius of the sliding context window, and d_k is the distance between term k and target term w.[3]

For the example vocabulary in Table 1, the resulting memory tensors are effectively $n \times n$ matrices (where $n = 3$ the size of the vocabulary), having elements equal to the co-occurrence frequency of the 2-tuples formed by the target term and the terms found within the context window. For example, consider the memory matrices created for the vocabulary terms using a sliding context window of radius 2 and with the target term shown in square brackets.

Binding Step 1: $\overbrace{A_s \quad [dog] \quad bit \quad the_s}$ $\quad mailman$

$$\begin{aligned} \boldsymbol{M}_{dog} &= 2 \times \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T \\ &= 2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0 \quad 1 \quad 0) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{9}$$

[1] Tensors are the set of geometric objects including vectors (first-order tensors), matrices (second-order tensors) and higher-order tensors (Kolda & Bader, 2009).

[2] *A* and *the* are considered to be stop-list words (noisy, low information terms that are ignored) and hence are not included in the vocabulary in Table 1.

[3] Stop-list words are counted when calculating d_k in Equation (8).

Binding Step 2: $\overbrace{A_s \quad dog \quad [bit] \quad the_s \quad mailman}$

$$\begin{aligned} \boldsymbol{M}_{bit} &= 2 \times \boldsymbol{e}_{dog} \otimes \boldsymbol{e}^T_{bit} + \boldsymbol{e}_{bit} \otimes \boldsymbol{e}^T_{mailman} \\ &= 2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0\ 1\ 0) + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0\ 0\ 1) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{10}$$

Binding Step 3: $A_s \quad dog \quad \overbrace{bit \quad the_s \quad [mailman]}$

$$\begin{aligned} \boldsymbol{M}_{mailman} &= \boldsymbol{e}_{bit} \otimes \boldsymbol{e}^T_{mailman} \\ &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0 \quad 0 \quad 1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{11}$$

The example demonstrates how this binding process results in all non-zero elements being situated on the row or column corresponding to the target term's id. If this vocabulary building process was performed over the entire corpus the general form of a *memory matrix* would be:

$$\boldsymbol{M}_w = \begin{pmatrix} 0, & \ldots, 0, & f_{1w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{(w-1)w}, & 0, & \ldots, 0 \\ f_{w1}, & \ldots, f_{w(w-1)}, & f_{ww}, & f_{w(w+1)}, & \ldots, f_{wn} \\ 0, & \ldots, 0, & f_{(w+1)w}, & 0, & \ldots, 0 \\ & & \ldots & & \\ 0, & \ldots, 0, & f_{nw}, & 0, & \ldots, 0 \end{pmatrix}, \tag{12}$$

where f_{iw} is the value in row i column w of the matrix representing the proximity-scaled, co-occurrence frequency of term i before term w, and n is the size of the vocabulary.

Due to the sparseness of the TE model's memory tensors and their elements having values equal to the proximity-scaled, co-occurrence frequencies of the terms, the construction and storage of these memory tensors can be efficiently achieved using relatively low-dimension storage vectors. For example, the memory matrix for *bit* in Equation (10) can be stored in the following *storage vector* (SV):

$$\mathrm{SV}_{\mathrm{bit}} = [(-1 \quad 2) \quad (3 \quad 1)], \tag{13}$$

where parenthesis have been added to illustrate implicit grouping of $(T \quad CF)$ pairs, where T is the term-id of the co-occurring term and CF is the cumulative, proximity-scaled, co-occurrence frequency of T with term w (*bit* in this example). The sign of T (term-id) indicates the word order of T with w. The information in this vector can be used to reconstruct the memory matrix using the following process:

> If the term Id (T) in the ($T \quad CF$) pair is positive, the CF value is located at row w, column T in the memory tensor. Otherwise, the CF value is located at row T, column w.

The ability of the TE model to efficiently store tensor representations, that capture order and co-occurrence information about n-tuples, increases the flexibility of the model, while preserving its formalism (Symonds et al., 2011a). A computational complexity analysis illustrating the efficiency of the TE model within the TQE approach developed in this work is provided in Appendix A.

Corpus-based distributional models used in past query expansion approaches have (i) only been evaluated in the query expansion process on small data sets (likely due to the computational complexity of the models or availability of more recent large data sets), and have (ii) not explicitly modeled and combined measures of syntagmatic and paradigmatic associations, and hence not allowed the influence of each type of association on retrieval effectiveness to be more fully understood (Bai et al., 2005; Hoenkamp et al., 2009; Bruza & Song, 2002). Therefore, we argue that the TE model's efficiency and ability to separately model syntagmatic and paradigmatic associations, makes it a superior choice of corpus-based distributional model to underpin a new query expansion technique.

A final point of difference that our new query expansion approach has from previous attempts to use corpus-based distributional models, is that our use of the TE model will be formalized within the relevance modeling framework. Positioning the model within a formal framework allows the possible implications of any future enhancements to be more readily predicted and understood.

Tensor Query Expansion

As highlighted in the review of current query expansion techniques, state-of-the-art techniques primarily rely on information about syntagmatic associations between words. Information about syntagmatic associations only make up half of the associations responsible for giving words their meaning; structural linguistics posits that the other half are provided by paradigmatic associations (Lyons, 1968). Intuitively, the user's query (re-)formulation process relies heavily on word meanings. Therefore, we propose the working hypothesis that explicitly combining information about both syntagmatic and paradigmatic associations when estimating query expansion terms will lead to improved retrieval effectiveness.

To illustrate how each association can be used to enhance the query representation to be more like the user's real information need, consider the query: *Best coffee machine*. The user's information need may rely on words such as "*lowest, price, tasting, espresso, maker*". These words 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 (highlighted in bold).

Given the potential for these associations to suggest effective query expansion terms, we provide a formal method for combining the TE model's syntagmatic and paradigmatic measures within the relevance modeling framework for query expansion. The relevance modeling framework is chosen as it provides a formal method for query expansion within the language modeling framework, and has been shown to produce robust effectiveness (Lavrenko & Croft, 2001).

The formalism for the relevance model (Equation (2)) includes estimating the probability $P(w|R)$, from a multinomial distribution. $P(w|R)$ estimates the probability of observing a word w given some relevant evidence (R) often in the form of a set of (pseudo) relevant documents produced by a document ranking model for a particular query Q. Our aim will be to create an analogous distribution to estimate $P(w|R)$, which is based on word meanings formed by explicitly combining measures of syntagmatic and paradigmatic associations between words. These measures will be based on distributional information found within the vocabulary created by the TE model when trained on the same set of (pseudo) relevant documents (i.e., produced from the same underlying document ranking model). We call this technique the *Tensor Query Expansion* (TQE) approach.

To formally estimate the probability of observing a word w given a vocabulary (V_k) built from a set of k (pseudo) relevant documents for a given query Q, we use a Markov random field approach. A Markov random field is an undirected graph combining a number of random variables. The formalism starts by letting an undirected graph G contain nodes that represent random variables, and the edges define the independence semantics between the random variables. Within the graph, a random variable is independent of its non-neighbors given observed values of its neighbors.

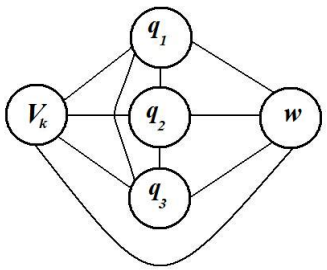

Figure 1. Example of the TQE graphical model for a three term query.

Figure 1 shows a graph G that consists of query nodes q_i, expansion term node w, and a vocabulary node V_k. Term w is constrained to exist within the vocabulary V_k, which is built from a set of k documents considered (pseudo) relevant to Q. We parameterize the graph based on clique sets to provide more flexibility in encoding useful features over cliques in the graph. The joint distribution over the random variables in G is defined by:

$$P_{G,\Gamma}(Q, w, V_k) = \frac{1}{Z_\Gamma} \prod_{c \in cl(G)} \varphi(c; \Gamma), \tag{14}$$

where $Q = q_1, \ldots, q_p$, $cl(G)$ is the set of cliques in G, each $\varphi(.; \Gamma)$ is a non-negative *potential function* over clique configurations parameterized by Γ, and $Z_\Gamma = \sum_{Q,w} \prod_{c \in cl(G)} \varphi(c; \Gamma)$ normalizes the distribution.

This distribution for this graph can be used to estimate a conditional probability of observing w given q (see Symonds, Bruza, Sitbon, and Turner (2011b)), and can be expressed as:

$$P_{G,\Gamma}(w|Q) \propto \gamma s_{\text{par}}(Q, w) + (1 - \gamma) s_{\text{syn}}(Q, w), \tag{15}$$

where $\gamma \in [0, 1]$, mixes the amount of paradigmatic $s_{\text{par}}(Q, w)$ and syntagmatic $s_{\text{syn}}(Q, w)$ measures used in the estimation.

The estimate in Equation (15) is produced from a multinomial distribution akin to those in the unigram and positional relevance models and can be used to augment the query representations within the language modeling framework. Using the relevance models feedback interpolated form, shown in Equation (3), the final conditional probability becomes:

$$P(w|Q) = \alpha P_o(w|Q) + (1 - \alpha) P_{G,\Gamma}(w|Q). \tag{16}$$

The construction of the TQE approach in this way ensures that modifying the mixing parameter (γ) in Equation (15) will add paradigmatic information to the query expansion process in a controlled manner relative to the RM3 benchmark. Assuming that the other parameters in the system are systematically controlled, we argue this approach will allow us to robustly evaluate the impact of paradigmatic information on retrieval effectiveness in a pseudo relevance feedback setting. The following sections outline the measures chosen to model each type of association within the query expansion process.

Modeling Paradigmatic Associations

The measure used by the original TE model to estimate the strength of paradigmatic associations between vocabulary terms can be extended to estimate the strength of paradigmatic associations between a sequence of terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w, as follows:

$$s_{\text{par}}(Q, w) = \frac{1}{Z_{\text{par}}} \sum_{j \in Q} \sum_{i \in V_k} \frac{f_{\overline{ij}} . f_{\overline{iw}}}{\max(f_{\overline{ij}}, f_{\overline{iw}}, f_{\overline{wj}})^2}, \tag{17}$$

where $f_{\overline{ij}} = (f_{ji} + f_{ij})$ is the unordered co-occurrence frequency of terms i and j, and Z_{par} normalizes the distribution of scores, such that $\sum_{w \in V_k} s_{\text{par}}(Q, w) = 1$. The measure in Equation (17) contains a normalizing factor $\frac{1}{f_{\overline{wj}}}$ that reduces the paradigmatic score if w has strong syntagmatic associations with the query terms. This is aimed at enhancing words

that are predominantly paradigmatically related to the query terms.

Effective modeling of paradigmatic associations is often achieved when using a very narrow sliding context window to build up the representations of each vocabulary word in the TE model. This result has been highlighted by a number of models performing tasks that rely heavily on paradigmatic information (e.g., synonym judgement (Bullinaria & Levy, 2007; Symonds et al., 2011a)). Therefore, the size of the sliding context window used to build representations in TQE is set to one (i.e., $R = 1$ in Equation (8)), as also done in previous work (Symonds et al., 2011b).

Modeling Syntagmatic Associations

It has been shown that in a pseudo relevance feedback setting, the Dirichlet smoothed query likelihoods of the unigram relevance model (Equation (4)) effectively estimates the strength of syntagmatic associations of query terms (Symonds, Bruza, Zuccon, et al., 2012).

Given this finding, basing a measure of syntagmatic associations on the estimation technique used within the unigram relevance model (known as RM3) has two significant advantages. Firstly, from a computational complexity perspective, there is no need to build a semantic space to underpin the syntagmatic measure, as the Dirichlet smoothed estimates can be made from frequency information stored in the existing document index. Secondly, from an empirical stand point, the advantage of modeling syntagmatic associations within TQE in the same way as RM3 comes from the potential improved variable control. One of the key aims of our research is to measure the influence of paradigmatic associations on retrieval effectiveness (recall, the second hypothesis, H2). This is best achieved by ensuring the method of modeling syntagmatic information is the same in the benchmark model as our paradigmatically enhanced model. For these reasons RM3 was chosen as the benchmark model, and the syntagmatic measure was based on the same information that underpins RM3's estimate, effectively making the TQE approach a unigram relevance when $\gamma = 0$ in Equation (15).

The resulting measure of syntagmatic associations between a sequence of query terms Q and a vocabulary term w used within our TQE approach will be defined as:

$$s_{\text{syn}}(Q, w) = \frac{1}{Z_{\text{syn}}} \sum_{D_i \in V_k(Q)} P(D_i|Q)P(Q|w)$$
$$= \frac{1}{Z_{\text{syn}}} \sum_{D_i \in V_k(Q)} s(D_i, Q)\frac{tf_{w,D}}{|D_i|}, \qquad (18)$$

where $s(D_i, Q)$ is the document relevance score of the (pseudo) relevant document D_i given query Q. The smoothing feature seen in the Dirichlet estimate of Equation (4) was removed so as to reduce the number of free parameters used in our TQE approach. This reduces the possibility that any improvements in retrieval effectiveness may be due to any differences in degrees of freedom between RM3 and TQE.

Equations (17) and (18) define the two measures that will be used to explicitly model paradigmatic and syntagmatic associations, respectively, within our TQE approach. The time complexity of TQE is linear with the size of the vocabulary created from the set of (pseudo) relevant documents, as detailed in the Appendix.

Experimental Setup and Results

A major premise behind using the TE model within the query expansion process stems from the fact that existing approaches primarily use syntagmatic information, and hence employ only half the associations reported to give rise to word meanings. We have hypothesized that accessing information about both syntagmatic and paradigmatic information within the query expansion process may more effectively augment query representations resulting in improved retrieval effectiveness.

The TQE approach formally places the TE model within the relevance modeling framework. This section details a number of ad hoc retrieval experiments aimed at evaluating the benefits of using the TQE approach, with respect to strong benchmark relevance models, and provides a detailed examination of the improvements in retrieval effectiveness gained by including information about syntagmatic and paradigmatic associations.

These experiments represent different contexts in which the effectiveness of TQE and the importance of paradigmatic information can be evaluated, including:

1. **Short queries:** These experiments will use relatively short queries (often only 2 or 3 words in length), to simulate the context often found within traditional web search engines.

2. **Verbose queries:** These experiments will use relatively long queries, generally greater than 10 words in length. The long queries, also termed verbose queries, often form sentences seen within natural language, and are commonly found when performing question-answer tasks. Therefore, the results of these experiments will not only provide insight into the benefit of using syntagmatic and paradigmatic information to expand long queries, but also may provide insights into the potential value of using TQE within a question-answering context. Given the growing robustness of speech recognition systems and the increased prevalence of query suggestion functionality in search engines, it is expected that the use of verbose queries will be a growing trend in information retrieval research (Allan, Croft, Moffat, & Sanderson, 2012).

In addition, we will present an example where the TQE approach is placed within an industry setting to perform web search.

Data Sets

Evaluation of all models was performed on the TREC[4] data sets outlined in Table 2. All collections were stopped with the default 418 words Lemur stop list and stemmed using a Porter stemmer (Porter, 1980).[5] The experiments in this research were carried out using the Lemur Toolkit.[6] The Lemur implementation of the original positional relevance model is made available online by the original authors.[7]

Queries. The queries used within the short and verbose experiments involve the title and description components of the TREC topics, respectively. The average length of title and descriptions for each data set are shown in Table 2 along with the standard deviation of each set of queries, which provides an idea of the range of query lengths for each data set.[8]

Baseline and Benchmark Models. TQE was evaluated on an ad hoc retrieval task using pseudo relevance feedback, also known as blind feedback. The TQE approach, in Equation (15), was compared to a baseline unigram language model (i.e., with no pseudo relevance feedback) and is denoted as **noFB**, a benchmark unigram relevance model (**RM3**) and a positional relevance model (using *iid* sampling) (**PRM**).

RM3 was chosen as a benchmark model primarily because it is a formal approach that fits within the language modeling framework, is efficient and robust (Lv & Zhai, 2009) and has been used heavily as a benchmark for past query expansion research (Lv & Zhai, 2010; Metzler & Croft, 2007). Even though the unigram relevance model does not explicitly model term dependencies, it was shown earlier, that when used within a pseudo relevance setting it effectively models syntagmatic associations for query terms, and hence RM3's estimation technique was chosen as the TQE's syntagmatic feature. This decision was seen as an effective way to control the influence of paradigmatic information on retrieval effectiveness. This is because, if all other TQE and RM3 model parameters, except the mix of syntagmatic and paradigmatic information in TQE (i.e., γ in Equation (15)) are fixed, then any differences in retrieval effectiveness between TQE and RM3 can reliably be attributed to the influence of paradigmatic information.

A query expansion approach that explicitly models term dependencies was also chosen as a benchmark model. The choice was primarily between LCE and PRM, as these have been shown to significantly outperform RM3 (Lv & Zhai, 2010; Metzler & Croft, 2007). However, PRM fits within the relevance modeling framework, unlike LCE which is based on the Markov random field document ranking model. This means that the set of pseudo relevant documents used by RM3, PRM and TQE for each query will be the same, as they all use the unigram language model. This is important to ensure any difference in retrieval effectiveness between techniques can be attributed to their estimation techniques, rather than the differences in documents on which the estimates are based.

One aim of this research is to evaluate the effect of paradigmatic information on retrieval effectiveness on short and verbose queries. Neither PRM nor LCE have been evaluated on verbose queries, likely due to the complexity of the models. From examining the estimation techniques used in PRM (Lv & Zhai, 2010) and LCE (Metzler & Croft, 2007), PRM can be argued to be a more computationally efficient approach, especially for verbose queries, and has been used in recent strong evaluations of query expansion techniques (Xue & Croft, 2013). A final point for choosing PRM over LCE, relates to its lower count of free model parameters, which means that any improvements in retrieval effectiveness are less likely to be due to an increased number of degrees of freedom (Metzler & Zaragoza, 2009).

Parameter Settings

The baseline unigram language model that underpins all three relevance models being evaluated was obtained using the Lemur default parameters. To avoid the criticism that any model performs better due to an increased number of parameters and to control for the influence of paradigmatic information on retrieval effectiveness all common model variables in our experiments were fixed. To this end, all expansion approaches were evaluated using 30 feedback documents, 30 expansion terms and a feedback interpolation coefficient α=0.5 in Equation (3). These settings have been shown to provide reasonable effectiveness for the RM3 and PRM benchmark models (Lavrenko & Croft, 2001; Lv & Zhai, 2010).

Even though it is common to fix one or more of these (pseudo) relevance feedback parameters (Bendersky, Metzler, & Croft, 2011), it is acknowledged that the success of query expansion has been shown to be sensitive to the number of (pseudo) relevant documents and expansion terms (Billerbeck & Zobel, 2004; Ogilvie, Voorhees, & Callan, 2009). However, if the models under evaluation produce significant improvements in retrieval effectiveness over the baseline unigram language model when these parameters are fixed, then it follows that greater improvements could be achieved if they were tuned.

For each of the query expansion techniques the free model parameters were trained using 3-fold cross validation on the

[4]http://trec.nist.gov/

[5]The Clueweb document index used in these experiments was produced using a Krovetz stemmer.

[6]The Lemur toolkit for language modeling and information retrieval: http://www.lemurproject.org

[7]http://sifaka.cs.uiuc.edu/ ylv2/pub/prm/prm.htm

[8]Topics 1-50 in the ClueWeb data set were not used as their relevance judgments were produced for the estimated AP metric (Yilmaz, Kanoulas, & Aslam, 2008), which is not conceptually equivalent to those used for the MAP metrics.

Table 2
Overview of TREC collections and topics. $\overline{|q|}$ 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.

	Description	# Docs	Topics	title $\overline{\|q\|}$	description $\overline{\|q\|}$	$\overline{\|D\|}$
WSJ	Wall Street Journal 87-92 off TREC Disks 1,2	173,252	1-200	4.8 (3)	19 (7.6)	468
AP	Assoc. Press 88-90 off TREC Disks 1,2,3	242,918	1-200	4.8 (3)	19 (7.6)	494
ROB	Robust 2004 data TREC Disks 4,5 -CR	528,155	301-450 601-700	2.6 (0.7)	16 (5.5)	561
G2	2004 crawl of .gov domain	25,205,179	701-850	2.28 (0.87)	11 (4.1)	1,721
CW	Clueweb09 Category B	50,220,423	Web Track 51-150	2.72 (1.38)	9 (3.3)	804

MAP metric. This includes training the Dirichlet smoothing parameter μ in the unigram relevance model of Equation (4). The free parameters trained for the positional relevance model included both σ and λ in Equation (6). For the TQE approach, the only free parameter was γ in Equation (15).

Experimental Results for Short Queries

Traditional web search often involves users entering very short, two or three word queries. To evaluate the impacts of including information about syntagmatic and paradigmatic associations to augment short query representations within the information retrieval process a retrieval experiment was carried out on the data sets and topic titles outlined in Table 2. The mean average precision (MAP) and precision at 20 (P@20) for the top ranked 1000 documents for all models are reported in Table 3. The significance of the results was evaluated using a one-sided t-test with $\alpha = 0.5$.

The results show that for short queries, the TQE approach can provide significant improvements over the baseline (noFB) on all data sets, except for CW. The finding that RM3 and PRM are unable to achieve consistently significant retrieval effectiveness over the baseline is likely due to the fixing of all other pseudo relevance feedback parameters in these experiments, so that the impact of paradigmatic information on retrieval effectiveness could be rigorously evaluated. This is consistent with past research (Billerbeck & Zobel, 2004) that highlighted the sensitivity of query expansion approaches to these parameters. This means that the results from these experiments will be a conservative estimate of maximum retrieval effectiveness of RM3, PRM and TQE.

To understand how the retrieval effectiveness of TQE compares on a *per query basis* to RM3 and PRM, a robustness analysis is presented next.

Robustness for short queries

Robustness includes considering the ranges of relative increase/decrease in average precision and the number of queries that were improved/degraded, with respect to the baseline unigram language model (noFB). The graphs in Figure 2 and Figure 3 illustrate the relative increase/decrease of average precision scores when compared to the baseline, for the RM3, PRM and TQE approaches evaluated on the ROB and G2 data sets, respectively. The central bars in each figure show how many queries had their baseline average precision score (noFB) improved by between zero and 25 percent. The bars to the right of centre correspond to the number of queries whose average precision scores were improved by even greater percentages, while those to the left of centre indicate the number of queries whose baseline average precision scores were reduced by the indicated percent range. The model which provides the most robust improvements will have a distribution located further to the right (i.e., having helped improve the retrieval effectiveness for a greater proportion of the queries for the intervals chosen) when compared to the other distributions.

Figure 2 and Figure 3 demonstrate there is no consistent difference in robustness between approaches on short queries on the ROB and G2 data sets. A similar result was found on the other data sets. Insight into inconsistencies within the short query results, such as RM3 appearing more robust than TQE on G2 (Figure 3) while TQE was able to achieve a greater average effectiveness on the same data set (Table 3), can be gained by considering the reliance on paradigmatic information in the TQE approach.

Parameter sensitivity for short queries

A parameter sensitivity analysis was performed to understand the role that syntagmatic and paradigmatic informa-

Table 3
Retrieval results on short queries for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM) and tensor query expansion (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. % improvement over noFB shown in parentheses.

	Metric	noFB	RM3	PRM	TQE
WSJ	MAP	0.2686	0.3089^{np} (15%)	0.3061^{n} (13.9%)	$\mathbf{0.3090^{np}}$ (15%)
	P@20	0.4074	0.4423^{n} (8.6%)	0.4413^{n} (8.3%)	$\mathbf{0.4434^{n}}$ (8.8%)
AP	MAP	0.1793	0.2144^{n} (19.6%)	0.2131^{n} (18.8%)	$\mathbf{0.2145^{n}}$ (19.6%)
	P@20	0.2300	0.2723^{n} (18.4%)	0.2788^{n} (22%)	$\mathbf{0.2825^{n}}$ (22.8%)
ROB	MAP	0.2500	0.2700^{n} (8%)	0.2707^{n} (8.3%)	$\mathbf{0.2783^{nrp}}$ (11.3%)
	P@20	0.3558	0.3688^{n} (3.7%)	0.3639 (2.2%)	$\mathbf{0.3741^{nrp}}$ (5.1%)
G2	MAP	.2941	0.3049^{n} (3.6%)	0.3069^{n} (4.3%)	$\mathbf{0.3085^{n}}$ (4.9%)
	P@20	0.5050	0.5013 (-0.7%)	0.5078 (0.5%)	$\mathbf{0.5179^{nr}}$ (2.5%)
CW	MAP	0.0768	0.0778 (1.3%)	**0.0822** (7.1%)	0.0796 (3.7%)
	P@20	0.1872	0.1995 (6.5%)	**0.2031** (8.4%)	0.1995 (6.5%)

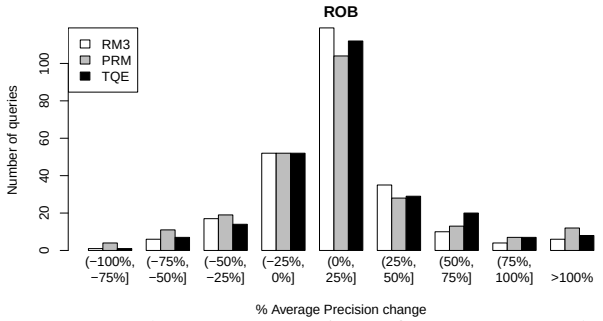

Figure 2. Robustness comparison of RM3, PRM and TQE on the ROB data set for short queries, showing the distribution of change in average precision when compared to the baseline average precision.

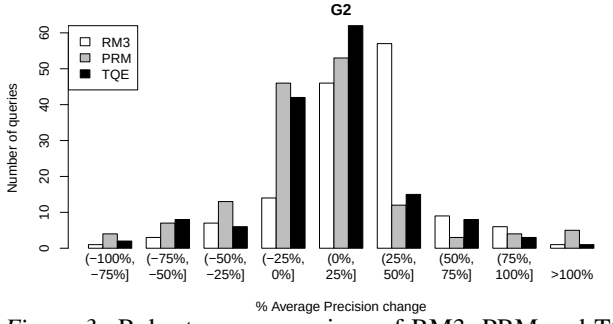

Figure 3. Robustness comparison of RM3, PRM and TQE on the G2 data set for short queries.

tion plays in achieving the reported retrieval effectiveness of TQE. This analysis, shown in Table 4, displays the mix of syntagmatic and paradigmatic information based on the γ value (see Equation (15)) that provides the best retrieval effectiveness for each data set using short queries.[9] Recall that when $\gamma = 0$ in Equation (15), TQE is effectively a unigram relevance model (i.e., RM3) and relies solely on syntagmatic information. A reliance on paradigmatic information in some form is indicated when $\gamma > 0$. The results in Table 4 indicate that the information about paradigmatic associations do not play a major role in providing the best MAP or P@20 scores on small newswire document collections like WSJ, AP and ROB. However, the role of paradigmatic information appears to become more important when searching large web collections, such as those used within the G2 and CW data sets.

Table 4
Parameter sensitivity analysis showing the value of γ in TQE that produces the highest overall MAP and P@20 scores for the TREC collections using short queries. Recall, $\gamma = 0$ means TQE bases estimates solely on the syntagmatic measure (i.e., effectively a unigram relevance model), and $\gamma = 1$ means TQE bases estimates solely on the paradigmatic measure.

	WSJ	AP	ROB	G2	CW
γ for opt. MAP	0	0	0.1	0.2	0.4
γ for opt. P@20	0.2	0 / 0.4	0.2	0.8	0.4

It is hypothesized that information about paradigmatic associations becomes more important on noisy collections. This idea stems from the increased likelihood that query terms and effective expansion terms will co-occur within the same document for small collections that have little noise, and hence syntagmatic associations can be very effectively modeled. However, for larger, noisy collections this likelihood is reduced, and hence syntagmatic associations become less effective. On larger, noisy collections, the modeling of paradigmatic associations may be more effective as the associations are formed between words that do not need to occur

[9]The analysis in Table 4 does not use a train/test split.

in the same (pseudo) relevant document. This may explain the increased reliance on paradigmatic information for the G2 and CW data sets (Table 4).

The increased reliance on paradigmatic information by TQE for the G2 collection may help explain why the TQE outperforms RM3 (Table 3), yet does not appear as robust as RM3 (Figure 3). To illustrate why the effect of paradigmatic information may increase the variance in effectiveness, while still providing a greater average, we provide a linguistically motivated example. Consider a query that contains words that are likely to generate vocabulary mismatch, like TREC Topic 191 from the AP data set: *Efforts to improve U.S. schooling*, may benefit more from using paradigmatic information (c.f., syntagmatic) to expand the query, as terms such as: *attempts, research, enhance, lift, united states, american, teaching, academic results*; are likely to be suggested.

To provide empirical support for this type of linguistic argument, we report that the best retrieval effectiveness of TQE (MAP = 0.334) for this example (TREC Topic 191) was achieved when purely paradigmatic information ($\gamma = 1$) was used to expand the query. In comparison to the effectiveness of TQE (MAP = 0.211) achieved on this query when the trained γ (using 3-fold cross validation, Table 3) is used, this is a substantial improvement (58%). Being able to reason linguistically about why syntagmatic or paradigmatic information may assist the expansion of some queries more than others may provide motivation for the development of an adaptive TQE approach, which would predict the value of γ depending on query characteristics.

Experimental Results for Verbose Queries

Long queries make up a smaller yet important proportion of web queries submitted to search engines, and are common in collaborative question answering (QA) (Huston & Croft, 2010; Bendersky & Croft, 2009; Balasubramanian, Kumaran, & Carvalho, 2010). A recent report produced by the information retrieval community also identified *conversational answer retrieval* as one of six important topics for future information retrieval research (Allan et al., 2012).

The mean average precision (MAP) and precision at 20 (P@20) for the top ranked 1000 documents for all models evaluated on verbose queries (i.e., taken from the topic descriptions in Table 2) are reported in Table 5. The significance of the results were evaluated using a one-sided t-test with $\alpha = 0.5$. Results indicate that TQE can provide significant improvement over the baseline and benchmark models on all data sets (except for CW). These results also indicate that the improvements in retrieval effectiveness of RM3 and PRM are not always significantly better than the baseline (noFB). However, this is likely due to the fixing of all other pseudo relevance feedback parameters, including the number of feedback documents and expansion terms, so that the impact of paradigmatic information on retrieval effectiveness could be rigorously evaluated. Therefore, the results of this experiment are a conservative estimate of maximum retrieval effectiveness of RM3, PRM and TQE for the data sets being considered.

Figure 4 illustrates the percent increase in MAP of the benchmark models over noFB with respect to the query length for the G2, ROB4 and AP/WSJ datasets. As TQE and RM3 use the same source of syntagmatic information, the relatively constant epoch in improved effectiveness of TQE over RM3 can be attributed to the inclusion of paradigmatic information within the TQE. This graph also suggests a link between gains in retrieval effectiveness and the query length may exist.

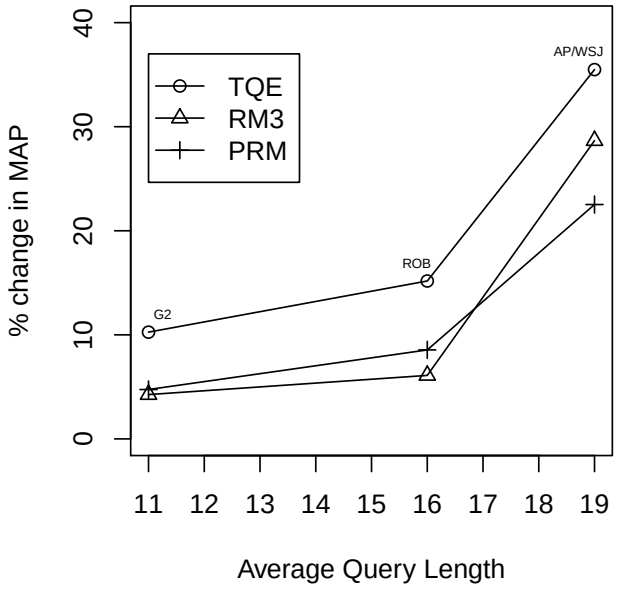

Figure 4. Percent improvement in MAP of RM3, PRM and TQE over the unigram language model (noFB) for the average query lengths of the G2 ($|\bar{q}|$=11), ROB ($|\bar{q}|$=16) and AP/WSJ ($|\bar{q}|$=18) data sets listed in Table 5.

Using average effectiveness to infer a link between the gains in retrieval effectiveness and query length can be problematic, as outliers may have a significant effect on the averages. Therefore, Figure 5 shows the correlation on a per-query basis of the gain in MAP of TQE over noFB for various query lengths.[10] This graph shows a relatively weak link between improvements in MAP and the query lengths ($r = 0.16$), and hence further investigation is required to test this hypothesized link. To this aim we examine the effectiveness of TQE on a modified CW dataset, created by considering TREC topic descriptions with a length greater than 10

[10]Figure 5 was produced with outliers (% MAP increase greater than 150% or less than -50%) removed.

Table 5
Retrieval results for verbose queries, for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM) and tensor query expansion (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. Brackets indicate percent improvement over noFB.

	Metric	noFB	RM3	PRM	TQE
WSJ	MAP	0.2121	0.2682^{n} (26.4%)	0.2589^{n} (22.1%)	$\mathbf{0.2865^{nrp}}$ (35.0%)
	P@20	0.3480	0.3891^{n} (11.8%)	0.3795^{n} (9.1%)	$\mathbf{0.4149^{nrp}}$ (19.2%)
AP	MAP	0.1511	0.1991^{n} (31.8%)	0.1861^{n} (23.2%)	$\mathbf{0.2056^{nrp}}$ (36.1%)
	P@20	0.2300	0.2600^{n} (13.0%)	0.2458 (6.8%)	$\mathbf{0.2738^{nrp}}$ (19.0%)
ROB	MAP	0.2491	0.2643^{n} (6.1%)	0.2704^{n} (8.5%)	$\mathbf{0.2869^{nrp}}$ (15.1%)
	P@20	0.3373	0.3414 (1.2%)	0.3504^{nr} (3.9%)	$\mathbf{0.3650^{nrp}}$ (9.1%)
G2	MAP	0.2466	0.2571^{n} (4.3%)	0.2583^{n} (4.8%)	$\mathbf{0.2719^{nrp}}$ (10.3%)
	P@20	0.4594	0.4620 (0.6%)	0.4732 (1.1%)	$\mathbf{0.4842^{nrp}}$ (5.4%)
CW	MAP	0.0530	0.0558 (5.2%)	$\mathbf{0.0614^{n}}$ (16.3%)	0.0574 (8.3%)
	P@20	0.1561	0.1566 (0.3%)	$\mathbf{0.1724^{n}}$ (10.5%)	0.1607 (2.9%)

words. The CW dataset was chosen as it has the shortest average topic description lengths and was the only data set on which TQE was unable to achieve significant improvement in retrieval effectiveness.

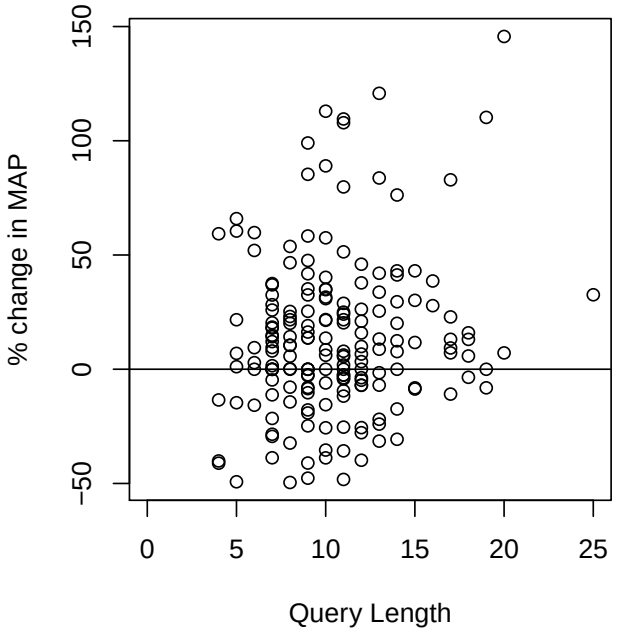

Figure 5. Percent improvement in MAP of TQE over the unigram language model (noFB) for various lengths of TREC topic descriptions found in the G2 and CW data sets.

Our choice of minimum query length (i.e., 11) was based on providing a balance between previous research, including work by Bendersky and Croft (2009) where verbose queries were defined as having a length greater than 12, and choosing a query length which would ensure sufficient data samples for a meaningful analysis. For the CW data set, the number of topics with $|q| > 10$ was 30, with $|q| > 11$ was 16 and with $|q| > 12$ was 11. Therefore, queries (i.e., topic descriptions in Table 2) with length greater than 10 were chosen for this evaluation, as indicated by the CW_v data set in Table 2. The retrieval effectiveness results on the CW_v data set are shown in Table 6 and demonstrate that TQE achieves significant improvement in retrieval effectiveness over the baseline and benchmark models for this data set.

Table 6
Retrieval results for verbose queries ($|q| > 10$) on the CW_v data set, for the unigram language model (noFB), unigram relevance model (RM3), positional relevance model (PRM), and TQE (TQE). Statistically significant results ($p < 0.05$) are indicated by superscripts using the first letter of the baseline over which significant improvement was achieved (n=noFB, r=RM3, p=PRM, t=TQE). Bold indicates the best result for each dataset and metric. Brackets indicate percent improvement over noFB.

	Metric	noFB	RM3	PRM	TQE
CW_v	MAP	.0681	0.0816^{n} (19.7%)	0.0827^{n} (21.4%)	$\mathbf{0.0882^{nrp}}$ (29.4%)
	P@20	0.2267	0.2417 (6.6%)	0.2423^{n} (6.9%)	$\mathbf{0.2500^{nrp}}$ (10.3%)

To understand how the retrieval effectiveness of the TQE approach compares on a *per query basis* to that of RM3 and PRM, a robustness analysis is required.

Robustness for verbose queries

The graphs in Figure 6 and Figure 7 illustrate the relative increase/decrease of average precision scores for the RM3, PRM and TQE approaches over the unigram language model

(noFB) when evaluated on the ROB and G2 data sets, respectively. Recall that the model which provides the most robust improvement in MAP will have a distribution located further to the right when compared to the other distributions.

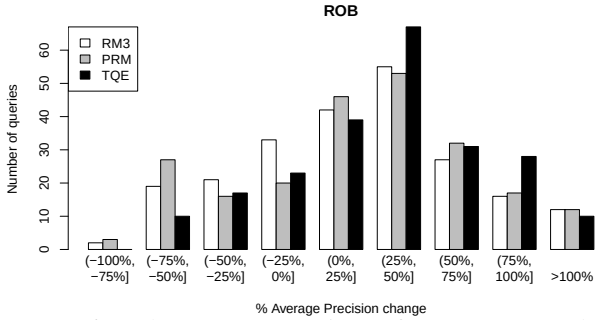

Figure 6. Robustness comparison of RM3, PRM and TQE on the ROB data sets for verbose queries.

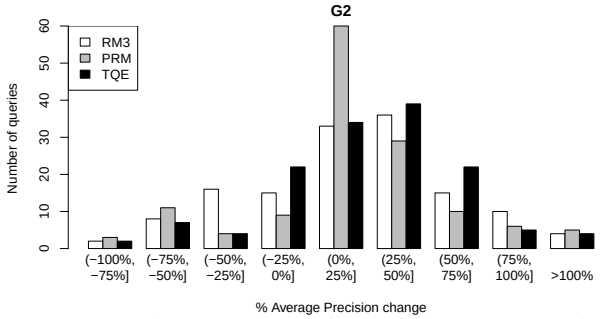

Figure 7. Robustness comparison of RM3, PRM and TQE on the G2 data sets for verbose queries.

This analysis suggests that TQE provides more consistent improvements over the baseline unigram language model (noFB) than RM3 and PRM. The graphs for the other data sets were omitted for space reasons, however, a similar result was observed.

Parameter Sensitivity for verbose queries

To understand the role that syntagmatic and paradigmatic information plays in achieving the reported retrieval effectiveness of TQE, a parameter sensitivity analysis was performed. This analysis, shown in Table 7 displays the value of γ that provides the best retrieval effectiveness of TQE for each data set.

The results (Table 7) show that for verbose queries information about paradigmatic associations are consistently contributing to the ability to achieve maximum retrieval effectiveness within the TQE approach. This differs from a similar analysis on short queries (Table 4), and indicates that the effectiveness of modeling paradigmatic associations is improved the longer the queries (i.e., when more statistical information about query term associations exist).

The increased reliance on the paradigmatic feature for the AP data set, when compared to the WSJ data set, which uses the same topics, may suggest that fewer of the initially retrieved documents from the AP collection were relevant and therefore less within document co-occurrences of query terms and effective expansion terms existed, leading to ineffective modeling of syntagmatic associations. This is supported by the relatively low MAP of the unigram language model (noFB) on the AP data set when compared to the WSJ data set. Therefore, using information about paradigmatic associations may be more effective at improving retrieval effectiveness of difficult verbose queries (i.e., those that have poor initial retrieval).

A final finding from the parameter sensitivity analysis on short and verbose queries is that using solely paradigmatic information for query expansion was shown to produce a lower average retrieval effectiveness on all data sets, when compared to methods relying solely on syntagmatic information (i.e., RM3 and PRM). This is in-line with Voorhees (1994).

Expansion Term Comparison

As an example of the types of terms being produced by each of the query expansion techniques, Table 8 lists the top 10 query terms and estimates for TREC topic 148: *Find information about Martha Stewarts insider trading case;* on the ClueWeb09 CategoryB document collection. Table 8 shows that the top 10 expansion terms for RM3 are identical and in the same order as those produced by the TQE syntagmatic feature, $s_{\text{syn}}(,)$. This adds support to the design claim that the TQE approach behaves as a unigram language model when $\gamma = 0$.

A final interesting point raised by observing the syntactic class of the expansion terms produced by $s_{\text{par}}(,)$ across the CW data set, is that these paradigmatic associations do not often manifest as synonyms or antonyms (commonly adjectives). The $s_{\text{par}}(,)$ expansion terms appear more likely to be related verbs, like *trade-invest* in Table 8. This result is seen as an attractive feature of the TQE approach and may help explain why ontological based attempts at using paradigmatic information within the query expansion process have not been overly successful, like those using WordNet (Voorhees, 1994).

Industry Application of TQE

The evaluation of the TQE approach in this work has been carried out under very controlled conditions, as the focus was on measuring the effect of paradigmatic information on successful query expansion. However, given the sensitivities exhibited by the TQE approach on short queries, an initial investigation into its applicability to more industrial-use would be valuable. To achieve this a system using the TQE approach to augment query representations was entered into

Table 7
Parameter sensitivity for verbose queries, showing the value of γ in TQE that produces the maximum MAP and precision at 20 scores for the TREC collections. Recall, $\gamma = 0$ means TQE bases estimates solely on the syntagmatic measure (i.e., effectively a unigram relevance model), and $\gamma = 1$ means TQE bases estimates solely on the paradigmatic measure.

	WSJ	AP	ROB	G2	CW	CW_v
γ for optimum MAP	0.1	0.4	0.2	0.6	0.4	0.4
γ for optimum P@20	0.2	0.5	0.3	0.6	0.2	0.5

Table 8
*Top 10 expansion terms and their estimates for TREC Web Track topic 148 (*Find information about Martha Stewarts insider trading case*) on the Clueweb09 CategoryB document collection for RM3, PRM, TQE and the paradigmatic and syntagmatic features. The scores in brackets indicate the respective model estimate P(w|q) of each expansion term (w). The values listed in the last row indicate the change in MAP (ΔMAP) achieved by each model, when compared to the baseline (i.e., noFB), and using the top 30 expansion terms of each approach.*

PRM	RM3	TQE	$s_{par}()$	$s_{syn}()$
martha (.0842)	martha (.0510)	martha (.0295)	find (.0016)	martha (.0728)
stewart (.0686)	stewart (.0412)	stewart (.0233)	information (.0015)	stewart (.0563)
new (.0121)	insider (.0402)	insider (.0204)	trade (.0015)	insider (.0503)
com (.0081)	trade (.0131)	trade (.0075)	timeline (.0014)	trade (.0165)
site (.0081)	new (.00945)	new (.0046)	case (.0013)	new (.0115)
live (.0071)	com (.0058)	com (.0033)	stewart (.0013)	com (.0077)
insider (.0057)	site (.0053)	site (.0025)	theme (.0008)	site (.0058)
home (.0050)	home (.0046)	home (.0024)	lawyer (.0007)	home (0.0057)
official (.0049)	article (.0040)	article (.0021)	invest (.0007)	article (.0052)
photo (.0048)	stock (.0037)	information (.0020)	martha (.0007)	stock (.0047)
ΔMAP -54%	ΔMAP +18%	ΔMAP +16%	ΔMAP +23%	ΔMAP +16%

the TREC 2012 Web Track (Symonds, Zuccon, Koopman, & Bruza, 2013). The TREC forum provides an opportunity to evaluate information retrieval systems on various retrieval tasks, using a consistent set of data sets, including very large web collections.

Within this investigation the set of training documents used to build the TE model's vocabulary used within the TQE approach is based on the k top ranked pseudo relevant documents produced by a strong baseline model. The baseline submission, referred to as **QUTParaBline** and depicted in Figure 8, is created using the following approach:

The *ClueWeb09-Category B* documents are indexed using the *indexing without spam* approach ((Zuccon, Nguyen, Leelanupab, & Azzopardi, 2011); and a threshold of 0.45), the standard INQUIRY stop-word list (Allan et al., 2000) and Krovetz stemmer (Krovetz, 1993). Each query is then issued to the Google retrieval service.[11]. and the top 60 retrieved documents are filtered using the spam filtered ClueWeb09-Category B index.[12] This filtered list is then padded, to create a list of 10,000 documents, based on the list of documents returned from a search on the spam filtered index using a unigram language model. The use of Google as the search engine for the top ranked results and the filtering of spam web pages are likely to translate into a strong baseline. This also allows us to understand what potential improvements could be made using TQE on real-world, commercial search engines.

The TQE approach was also applied on top of the baseline, to produce the system depicted in Figure 9. This system expands the original TREC 2012 Web Track topics using TQE based on the k top ranked pseudo-relevant documents produced by the baseline system. The ranked results produced by this system were submitted as run **QUTParaTQEg1** to TREC 2012 Web Track.

Training TQE. The data set used for this experiment is shown in Table 9. In this experiment all parameters in the pseudo relevance feedback setting were trained. These include the number of feedback documents (fbDocs), number of expansion terms (fbTerms), the mix of original and new query models (α) and the mix of syntagmatic and paradigmatic information (γ). Tuning of the QUTParaTQEg1 system parameters was achieved by maximizing ERR@20 on the TREC Web Track data sets from 2010 and 2011. The test topics were those provided by TREC organizers for the

[11] `http://www.google.com`

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

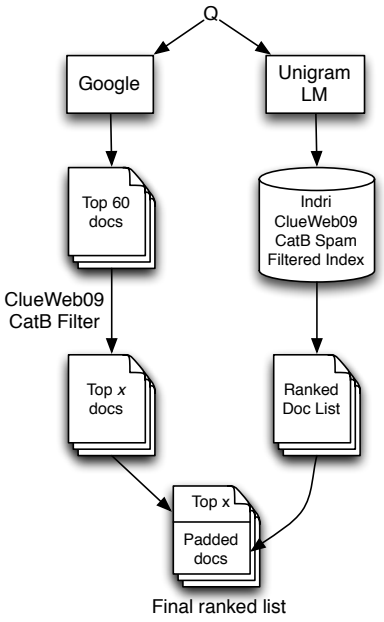

Figure 8. The baseline system: QUTParaBline.

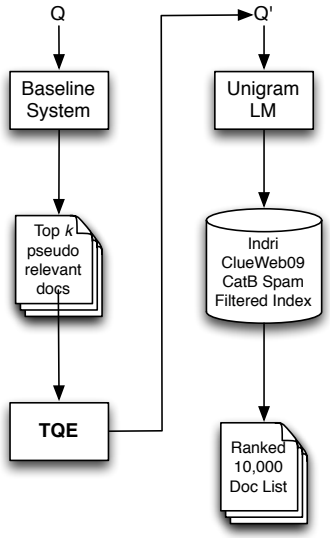

Figure 9. TQE on top of the baseline: QUTParaTQEg1.

2012 Web Track. Participants only receive the topic titles for producing their submissions and thus we did not tune parameters with respect to these test topics. Details regarding descriptions and further relevancy information are only provided after all submissions have been evaluated.

The test parameter values for the QUTParaTQEg1 submission were *Number of feedback documents* equal to 19, *number of expansion terms* equal to 14, *original query weight* equal to 0.4 and *TE model mixing parameter* (γ) equal to 0.1. A value of $\gamma = 0.1$ demonstrates that some combination of both syntagmatic and paradigmatic information provides optimal retrieval effectiveness. 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.

TQE Results on the TREC 2012 Web Track

Table 10 compares the retrieval effectiveness of QUTparaBline and QUTparaTQEg1 along with the average effectiveness of all 48 TREC 2012 Web Track submissions (MeanWT2012), and a baseline unigram language model (noFB). These results show that expanding the query representations using TQE can provide significant improvements over the Google baseline on the binary metrics of MAP and P@20. 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, i.e., between 0 and 4. In addition, measures that use graded judgements, such as ERR (Chapelle, Metlzer, Zhang, & Grinspan, 2009), often 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 say). This causes a heavy discounting to occur for relevant documents ranked lower in the list (Moffat, Scholer, & Thomas, 2012).

Given that Google's rankings are likely based on click-through data and editorial choice, the QUTParaBline system is able to ensure relevant documents are ranked high in the returned list. However, as the QUTParaTQEg1 system performs its final ranking using a unigram language model, which does not use such information, it is not surprising that the QUTParaTQEg1 model is unable to achieve significant improvements over QUTParaBline on the graded metrics ERR@20 and nDCG@20. Recall, that navigational relevant papers are given higher relevance scores than relevant pages in the graded relevance assessment framework.

As the QUTParaTQEg1 system achieved significant improvements over QUTParaBline on the P@20 metric (Table 10) it is returning many more relevant documents in the top 20 results when compared to QUTParaBline. Therefore, it could be argued that significant improvements on graded metrics, such as ERR and nDCG, may be achieved by QUTParaTQEg1 if the final document ranking model was enhanced to take into account graded relevance.

Robustness on Web Track. The graph in Figure 10 illustrates the relative increase/decrease of P@20 scores for QUTParaBline and QUTParaTQEg1 over MeanWT2012 when evaluated on the test topics (151-200) of the CW

Table 9
TREC collections and topics used creating the QUT_Para TREC submissions. $\overline{|q|}$ 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.

	Description	# Docs	Topics	title $\overline{\lvert q\rvert}$	description $\overline{\lvert q\rvert}$	$\overline{\lvert D\rvert}$
CW	Clueweb09 Category B	50,220,423	Web Track 51-200	2.72 (1.38)	9 (3.3)	804

Table 10
Comparison of retrieval performance on TREC 2012 Web Track ad hoc retrieval task. The superscripts u, m, b and t indicate statistically significant differences (calculated using a one-sided t-test $p < 0.05$) over the unigram language model (noFB), the average performance of all TREC Web Track participants (MeanWT2012), our baseline (QUTparaBline) and the TQE approach (QUTparaTQEg1), respectively. The best results for each evaluation measure appear in boldface. Brackets indicate the percentage change between QUTparaTQEg1 and QUTparaBline. Note that no value of MAP was provided for the average of all TREC 2012 Web Track submissions (MeanWT2012).

	Graded Metrics		Binary Metrics	
	ERR@20	nDCG@20	P@20	MAP
noFB	0.160	0.112	0.254	.107
MeanWT2012	0.187	0.123	0.284^{u}	_
QUTparaBline	$\mathbf{0.290}^{um}$	0.167^{um}	0.305^{um}	0.117^{u}
QUTparaTQEg1	0.249^{um}	$\mathbf{0.192}^{um}$	$\mathbf{0.396}^{umb}$	$\mathbf{0.158}^{ub}$
	(-14.2%)	(+15%)	(+29.8%)	(+35%)

data set.[13] This graph suggests that the QUTParaTQEg1 system provides more consistent improvements over the MeanWT2012.

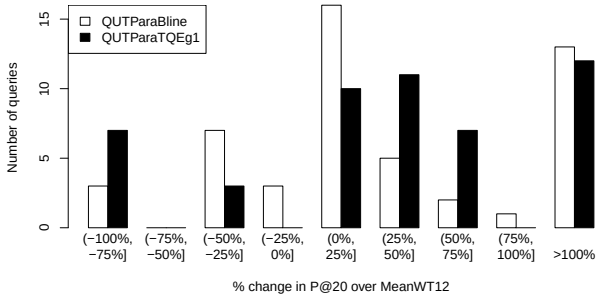

Figure 10. Robustness comparison of the QUTParaBline and QUTParaTQEg1 systems.

The increased variance of the TQE distribution shown in Figure 10 indicates that the use of the same mix of syntagmatic and paradigmatic information on all test queries can have remarkably different impacts on retrieval effectiveness. This may indicate that for some queries insufficient vocabulary statistics exist to allow effective modeling of both syntagmatic and paradigmatic associations. A similar result was found when using the TE model to perform similarity judgement of medical concepts (Symonds, Zuccon, et al., 2012).

Discussion and Final Remarks

The experiments on short queries have demonstrated that the inclusion of paradigmatic information within the query expansion process *does not* consistently enable significant improvements in retrieval effectiveness over syntagmatic information alone. We hypothesize that this result is related to previous TE model research that found the modeling of paradigmatic associations can be unreliable when insufficient statistical information is available (Symonds, Zuccon, et al., 2012).

The experiments on verbose queries have demonstrated that for queries considered to be verbose (i.e., $|q| > 10$), the inclusion of paradigmatic information within the query expansion process *does* provide significant improvements in retrieval effectiveness over methods relying on syntagmatic information alone. Our hypothesis that short queries do not provide sufficient statistical information to make reliable estimates of paradigmatic associations is supported by the increased reliance on paradigmatic information to achieve superior effectiveness on verbose queries.

The application of the TQE approach to an industry setting, tested within the 2012 TREC Web Track forum, demonstrated that when all TQE parameters are trained significant improvements in retrieval effectiveness can be achieved

[13] P@20 was used as no MAP for MeanWT2012 was available, and given the use of a unigram language model within QUTParaTQEg1 to perform the final ranking, ERR@20 or nDCG@20 are unlikely to be meaningful.

over a strong baseline. This indicates that the sensitivity associated with modeling paradigmatic associations on short queries can be overcome.

Summary Of Contributions

This work contributes to the field in the following ways:

1. **The development of a novel query expansion technique grounded in structural linguistic theory that formally synthesizes information about both syntagmatic and paradigmatic associations.** Current query expansion models primarily rely on only one form of word association that is only partly responsible for forming the meaning of words within structural linguistics. For the first time, the TQE approach brings both a linguistic grounding and a formal framework for modeling and combining information about syntagmatic and paradigmatic associations within the query expansion process. These associations being responsible for the formation of word meanings within structural linguistics.

2. **A rigorous evaluation of the impact on retrieval effectiveness of explicitly modeling and combining information about syntagmatic and paradigmatic associations within the query expansion process.** This paper demonstrates that significant improvements in retrieval effectiveness can be made by explicitly modeling both syntagmatic and paradigmatic associations within the query expansion process. The theoretical motivation, based on structural linguistics, makes this an intuitive step given the reliance on word meanings when the user formulates their query.

Conclusion and Future Work

The lack of both syntagmatic and paradigmatic information within existing query expansion techniques, and the reliance on word meanings by a user when formulating their information need, provided motivation for the use of a novel computational model of word meaning, known as the *Tensor Encoding* (TE) model, within the document retrieval process. The TE model formally combines information about the syntagmatic and paradigmatic associations that underpin the meaning of a word based on structural linguistic theories.

Within this research, the TE model was formally applied within the relevance modeling framework. When only the mix of syntagmatic and paradigmatic information was tuned within the TQE approach, significant improvements in retrieval effectiveness were observed on longer queries (verbose queries) for a wide range of data sets. However, when the TQE approach was used to expand shorter queries, modifying only this mix of word associations was unable to reliably produce significant improvements in retrieval effectiveness. This result was attributed to the sensitivity in estimating the strength of syntagmatic and paradigmatic associations between words when insufficient vocabulary statistics are available. However, when all model parameters were tuned on a industry task significant improvements in retrieval effectiveness were observed on short queries, when compared to a state-of-the-art baseline.

The demonstrated effectiveness and efficiency of the TQE approach, combined with its (i) formal framework, (ii) theoretical grounding in linguistics theories, and (iii) purely-corpus based approach, makes it a potentially fruitful approach for future application. Finally, it is hoped that this work provides a significant contribution to the substantive dialogue between the fields of cognitive science and information retrieval.

References

Allan, J., Connell, M. E., Croft, W. B., Feng, F. F., Fisher, D., & Li., X. (2000). INQUERY and TREC-9. In *Proceedings of the 19th Text REtrieval Competition (TREC '00).*

Allan, J., Croft, B., Moffat, A., & Sanderson, M. (2012, May). Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the 2nd Strategic Workshop on Information Retrieval in Lorne. *SIGIR Forum*, *46*(1), 2–32.

Bai, J., Song, D., Bruza, P., Nie, J.-Y., & Cao, G. (2005). Query Expansion using Term Relationships in Language Models for Information Retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)* (pp. 688–695). New York, NY, USA: ACM.

Balasubramanian, N., Kumaran, G., & Carvalho, V. R. (2010). Exploring Reductions for Long Web Queries. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR'10)* (pp. 571–578). New York, NY, USA: ACM.

Bendersky, M., & Croft, W. B. (2009). Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the 2009 workshop on Web Search Click Data (WSCD'09)* (pp. 8–14). New York, NY, USA: ACM.

Bendersky, M., Metzler, D., & Croft, W. B. (2011). Parameterized Concept Weighting in Verbose Queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information (SIGIR'11)* (pp. 605–614). New York, USA: ACM.

Billerbeck, B., & Zobel, J. (2004). Questioning Query Expansion: An Examination of Behaviour and Parameters. In *Proceedings of the 15th Australasian Database Conference (ADC'04)* (Vol. 27, pp. 69–76). Darlinghurst, Australia: Australian Computer Society, Inc.

Billhardt, H., Borrajo, D., & Maojo, V. (2002). A Context Vector Model for Information Retrieval. *Journal of the American Society for Information Science and Technology*, *53*(3), 236–249.

Bruza, P. D., & Song, D. (2002). Inferring Query Models by Computing Information Flow. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)* (pp. 260–269). New York, NY, USA: ACM.

Buckley, C. (1995). Automatic Query Expansion Using SMART. In *Proceedings of the 3rd Text REtrieval Conference (TREC'95)* (pp. 69–80).

Bullinaria, J., & Levy, J. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, *39*, 510-526.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*, *25*(2/3), 211–257.

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected Reciprocal rank for Graded Relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 621–630). New York, NY, USA: ACM.

Chung, Y. M., & Jae, Y. L. (2001, Feb 15). A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, *52*(4), 283-296.

Cleverdon, C., Mills, J., & Keen, M. (1966). *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results*. UK: College of Aeronautics, Cranfield.

Collins-Thompson, K. (2009). Reducing the Risk of Query Expansion via Robust Constrained Optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)* (pp. 837–846). New York, NY, USA: ACM.

Fang, H., & Zhai, C. (2006). Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)* (pp. 115–122). New York, NY, USA: ACM.

Greenberg, J. (2001, April). Optimal Query Expansion (QE) Processing Methods with Semantically Encoded Structured Thesauri Terminology. *Journal of the American Society for Information Science and Technology*, *52*(6), 487–498.

Grefenstette, G. (1992). Use of Syntactic Context to Produce Term Association Lists for Text Retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)* (pp. 89–97). New York, NY, USA: ACM.

Harris, Z. (1954). Distributional Structure. *Word*, *10*(23), 146–162.

Hoenkamp, E., Bruza, P., Song, D., & Huang, Q. (2009). An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval (ICTIR'09)* (Vol. 5766, p. 116-127). Springer Berlin / Heidelberg.

Holland, N. N. (1992). *The Critical I*. New York, USA: Columbia University Press.

Huston, S., & Croft, W. B. (2010). Evaluating Verbose Query Processing Techniques. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR'10)* (pp. 291–298). New York, NY, USA: ACM.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, *114*, 1–37.

Kolda, T. G., & Bader, B. W. (2009, September). Tensor Decompositions and Applications. *SIAM Review*, *51*(3), 455–500.

Krovetz, R. (1993). Viewing Morphology as an Inference Process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)* (pp. 191–202). New York, NY, USA: ACM.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*, 211-240.

Lavrenko, V. (2004). *A Generative Theory of Relevance*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Lavrenko, V., & Croft, W. B. (2001). Relevance-Based Language Models. *In Proceedings of the 24th Annual ACM Conference of Research and Development in Information Retrieval (SIGIR'01)*, 120-127.

Lund, K., & Burgess, C. (1996). Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior research methods, instruments and computers*, *28*, 203-208.

Lv, Y., & Zhai, C. (2009). A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM'09)* (pp. 1895–1898). New York, NY, USA: ACM.

Lv, Y., & Zhai, C. (2010). Positional Relevance Model for Pseudo-relevance Feedback. In *Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)* (pp. 579–586). New York, NY, USA: ACM.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. London: Cambridge University Press.

Metzler, D., & Croft, W. B. (2005). A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)* (pp. 472–479). New York, NY, USA: ACM.

Metzler, D., & Croft, W. B. (2007). Latent Concept Expansion using Markov Random Fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)* (pp. 311–318). New York, NY, USA: ACM.

Metzler, D., & Zaragoza, H. (2009). Semi-parametric and Non-parametric Term Weighting for Information Retrieval. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval (ICTIR'09)* (pp. 42–53). Berlin, Heidelberg: Springer-Verlag.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990, December 21). Introduction to WordNet: an Online Lexical Database. *International Journal of Lexicography*, *3*(4), 235–244.

Moffat, A., Scholer, F., & Thomas, P. (2012). Models and Metrics: IR Evaluation as a User Process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium* (pp. 47–54). New York, NY, USA: ACM.

Ogilvie, P., Voorhees, E., & Callan, J. (2009). On the Number of Terms used in Automatic Query Expansion. *Information Retrieval*, *12*(6), 666–679.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The Measurement of Meaning*. University of Illinois Press. Paperback.

Pavel, T. C. (2001). *The Spell of Language: Poststructuralism and Speculation*. University of Chicago Press.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137.

Rocchio, J. (1971). Relevance Feedback in Information Retrieval. In *The SMART Retrieval System* (pp. 313–323). Prentice-Hall.

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a Means to Encode Order in Word Space. In *Proceedings of the*

30th annual meeting of the cognitive science society (p. 23-26).

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620.

Schütze, H. (1993). Word Space. In *Advances in Neural Information Processing Systems 5* (pp. 895–902). Morgan Kaufmann.

Symonds, M., Bruza, P., Sitbon, L., & Turner, I. (2011a). Modelling Word Meaning using Efficient Tensor Representations. In *Proceedings of the 25th Pacific Asia Conference on Language, Information, and Computation (PACLIC'11)* (pp. 313–322).

Symonds, M., Bruza, P., Sitbon, L., & Turner, I. (2011b). Tensor Query Expansion: A Cognitive Based Relevance Model. In *Proceedings of the 16th Australasian Document and Computing Symposium (ADCS'11)* (pp. 87–94).

Symonds, M., Bruza, P., Zuccon, G., Sitbon, L., & Turner, I. (2012). Is the Unigram Relevance Model Term Independent? Classifying Term Dependencies in Query Expansion. In *Proceedings of the 17th Australasian Document Computing Symposium (ADCS'12)* (pp. 123–127). New York, NY, USA: ACM.

Symonds, M., Bruza, P. D., Sitbon, L., & Turner, I. (2012). A Tensor Encoding Model for Semantic Processing. In *Proceedings of the 21st acm international conference on information and knowledge management (cikm'12)* (pp. 2267–2270). New York, NY, USA: ACM.

Symonds, M., Zuccon, G., Koopman, B., & Bruza, P. (2013). QUT_Para at TREC 2012 Web Track: Word Associations for Retrieving Web Documents. In *The 21st Text Retrieval Conference Proceedings (TREC'12)*. NIST.

Symonds, M., Zuccon, G., Koopman, B., Bruza, P., & Nguyen, A. (2012). Semantic Judgement of Medical Concepts: Combining Syntagmatic and Paradigmatic Information with the Tensor Encoding Model. In *Proceedings of the 10th Australasian Language Technology Workshop (ALTA'12)* (pp. 15–22).

Turney, P. D., & Pantel, P. (2010, January). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Voorhees, E. M. (1994). Query Expansion using Lexical-semantic Relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)* (pp. 61–69). New York, NY, USA: Springer-Verlag, Inc.

Xu, J., & Croft, W. B. (1996). Query Expansion using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)* (pp. 4–11). New York, NY, USA: ACM.

Xue, X., & Croft, W. B. (2013, May). Modeling Reformulation using query distributions. *ACM Transactions on Information Systems*, *31*(2), 1–34.

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and nDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)* (pp. 603–610). New York, NY, USA: ACM.

Zhai, C., & Lafferty, J. (2001). Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)* (pp. 403–410). New York, NY, USA: ACM.

Zuccon, G., Nguyen, A., Leelanupab, T., & Azzopardi, L. (2011). Indexing without Spam. In *Proceedings of the 16th Australasian Document and Computing Symposium (ADCS'11)*.

Appendix
Computational Complexity Analysis

The TQE technique combines two semantic features that measure the strength of syntagmatic and paradigmatic associations. The creation of the memory matrices in Equation (8) provides a formalism for capturing the co-occurrences and encoding word order. However, the original TE model research (Symonds et al., 2011b) demonstrated that the word order and co-occurrence information is efficiently captured within low dimension *storage vectors* (SV) due to the unique structure of the memory matrices. The dimensionality of the storage vectors required depends on the final size of the vocabulary and the radius of the context window used in the vocabulary binding process.

For example, on a synonym judgement task using a vocabulary of 134,000 terms, the TE model's best performance was achieved using the paradigmatic measure, a context window of radius one and storage vectors of 1,000 dimensions (Symonds et al., 2011b). This supports previous research (Sahlgren, Holst, & Kanerva, 2008) that showed paradigmatic associations are most effectively modeled when a very small context window is used. A small context window means less co-occurrences are contained within the TE model representations. Given, the vocabulary of top 30 (pseudo) relevant documents in our experiments contained less than 20,000 terms for all queries, we chose to use storage vectors of 20 dimensions to underpin the TE model representations.

The worst case time complexity of the paradigmatic measure in Equation (17) is $T(n) = O(\frac{D^2_{SV_{\text{par}}}}{4}.|Q|)$, where $D_{SV_{\text{par}}}$ is the dimensionality of the storage vector, and $|Q|$ is the length of the query. Thus, keeping the dimensionality of the storage vector small is important. Given our decision to set $D_{SV_{\text{par}}} = 20$, the time complexity to estimate the updated query model using the paradigmatic measure would be $T(n) = O(100|Q||V_k|)$.

When considering the time complexity of the syntagmatic measure of Equation (18), it can be seen that this estimate is much quicker to compute. This is due to the expressions within the estimate existing in document indexes (i.e., $\frac{df_w}{|D_i|}$), or being already computed by the underlying document model, (i.e., $s(D_i, Q)$ in Equation (18)). Therefore, the time complexity to estimate the updated query model using the syntagmatic measure is $T(n) = O(|V_k|)$.