Is the Unigram Relevance Model Term Independent? Classifying Term Dependencies in Query Expansion

Mike Symonds[1], Peter Bruza[1], Guido Zuccon[2], Laurianne Sitbon[1], Ian Turner[3]

michael.symonds@qut.edu.au, p.bruza@qut.edu.au, guido.zuccon@csiro.au, l.sitbon@qut.edu.au, i.turner@qut.edu.au

[1] School of Information Systems, Queensland University of Technology
[2] Australian e-Health Research Centre, CSIRO
[3] School of Mathematical Sciences, Queensland University of Technology

Brisbane, Australia

ABSTRACT

This paper develops a framework for classifying term dependencies in query expansion with respect to the role terms play in structural linguistic associations. The framework is used to classify and compare the query expansion terms produced by the unigram and positional relevance models. As the unigram relevance model does not explicitly model term dependencies in its estimation process it is often thought to ignore dependencies that exist between words in natural language.

The framework presented in this paper is underpinned by two types of linguistic association, namely syntagmatic and paradigmatic associations. It was found that syntagmatic associations were a more prevalent form of linguistic association used in query expansion. Paradoxically, it was the unigram model that exhibited this association more than the positional relevance model. This surprising finding has two potential implications for information retrieval models: (1) if linguistic associations underpin query expansion, then a probabilistic term dependence assumption based on position is inadequate for capturing them; (2) the unigram relevance model captures more term dependency information than its underlying theoretical model suggests, so its normative position as a baseline that ignores term dependencies should perhaps be reviewed.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Experimentation, Theory

Keywords

Query expansion, relevance models, linguistic associations

ADCS '12 December 5-6, 2012, Dunedin, New Zealand.

1. INTRODUCTION

Within the information retrieval community it is understood that a user's query is often an imprecise description of their real information need. Therefore, there is a strong interest in the use of query expansion techniques. These techniques have been shown to provide significant improvements in retrieval effectiveness [1, 2, 3].

Early query expansion techniques did not explicitly use information about the dependencies that exist between terms in natural language [1, 9]. More recent approaches that explicitly model term dependencies have shown significant improvements in retrieval effectiveness over earlier techniques, and this improvement is often attributed to the *explicit* modelling of term dependencies [8, 3, 2]. For example, the unigram relevance model is often thought to ignore term dependencies. This assumption has been used to explain why dependency based approaches, like the positional relevance model, can significantly outperform the unigram relevance model.

In this paper we develop a novel framework to test this claim by comparing the extent to which linguistic associations are used within a unigram and positional relevance model. The framework is based on a recent query expansion approach, known as *tensor query expansion* (TQE), that uses features which have been shown to effectively measure the strength of linguistic associations.

A second contribution of this work is the discovery that although the unigram relevance model does not *explicitly* model term dependencies, unlike the positional relevance model, the estimation technique more effectively uses a form of term dependency underpinning most traditional dependency based query expansion approaches.

It is important to observe that this paper does not try to evaluate the effectiveness of each query expansion technique, but provides insight into the validity of claims relating to the cause of differences in retrieval effectiveness. Specifically those attributed to the explicit modelling of term dependencies within the query expansion process.

Section 2 introduces the relevance modelling framework, and outlines the unigram and positional relevance models. The novel framework for evaluating the types of linguistic associations within these query expansion techniques is presented in Section 3. This framework is applied in Section 4 to provide an empirical evaluation of the linguistic associations modelled within the unigram and positional relevance models, before concluding remarks are presented in Section 5.

2. RELEVANCE MODELS

Most state-of-the-art document retrieval models are probabilistic in nature. These include the formally grounded family of language models [4]. In the language modelling framework, the query representations can be formally augmented using a relevance modelling approach [1]. This approach has been put forward as a strong benchmark in past information retrieval research investigating the effectiveness of query expansion approaches [3, 2].

2.1 Unigram relevance model

A relevance model estimates the probability of observing a word w given some relevant evidence for a particular information need, represented by the query Q. The relevance model $P(w|R)$ is sampled from a multinomial distribution and is approximated as:

$$P(w|Q) = P(w|R) = \int_D P(w|D)P(D|Q)$$

$$\approx \frac{\sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}{\sum_w \sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}, \quad (1)$$

where $\mathcal{R}_Q$ is the set of documents pseudo-relevant or relevant to query Q, and D is a document in $\mathcal{R}_Q$. To simplify the estimation, $P(D)$ is assumed uniform over this set of documents. In the unigram variant of the relevance model, where term dependencies are ignored, the estimate for $P(w|D)$ is often based on the Dirichlet smoothed term likelihood scores:

$$P(w|D) = \frac{df_w + \mu \frac{cf_w}{|C|}}{|D| + \mu}, \quad (2)$$

where df_w is the document frequency of term w, cf_w is the collection frequency of term w, $|C|$ is the word count in the collection, $|D|$ is the word count of the document and μ is the Dirichlet smoothing parameter. This form of the unigram relevance model will be referred to as **RM3** for the remainder of this paper.

2.2 Dependency based approaches

Dependency based query expansion approaches, such as the *positional relevance model* (PRM) [2] and *latent concept expansion* (LCE) [3], explicitly model term dependencies when producing expansion term estimates and this has been credited with producing significantly improved retrieval effectiveness over the unigram relevance model [2].

The LCE approach is a feature based approach set atop of the Markov random field (MRF) document ranking model. Since the MRF model uses term dependencies itself, and has been shown to be a stronger baseline than a unigram language model, it is not an appropriate comparison model for use in this investigation of the unigram relevance model. However, the PRM is set within the relevance modelling framework and hence uses the unigram language model for document ranking. This makes an appropriate comparison model for our investigation.

2.2.1 The Positional Relevance Model

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 produce expansion term estimates in the following way:

$$P(w|Q) = \frac{P(w,Q)}{P(Q)} \propto P(w,Q) = \sum_{D \in \mathcal{R}_Q} \sum_{i=1}^{|D|} P(w,Q,D,i), \quad (3)$$

where i indicates a position in document D, and $\mathcal{R}_Q$ is the set of feedback documents (assumed to be relevant). Two sampling methods were proposed to estimate $P(w,Q,D,i)$. The first uses independent and identically distributed (*iid*) sampling, such that:

$$P(w,Q,D,i) \propto \frac{P(Q|D,i)P(w|D,i)}{|D|}. \quad (4)$$

The second approach to estimating $P(w,Q,D,i)$ uses conditional sampling, such that:

$$P(w,Q,D,i) = P(Q)P(D|Q)P(i|Q,D)P(w|D,i). \quad (5)$$

Both approaches are based on the following estimate:

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

where

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

and $c(w,j)$ is the *actual* count of term w at position j, $|D|$ is the length of the document, λ is a smoothing parameter and σ is used to parameterize the Guassian Kernel function.

The modelling of term dependencies within query expansion approaches are rarely linguistically motivated and often involve increasing the degrees of freedom of a model by adding free parameters, as seen in PRM. Being able to classify the types of linguistic associations modelled within a query expansion process could improve the understanding of the role term dependencies play in improving retrieval effectiveness. To this end, we aim to develop a framework for classify linguistic word associations used by query expansion techniques and test it by comparing the types of associations modelled within the RM3 and PRM approaches.

3. MODELLING WORD ASSOCIATIONS

To be able to classify the different types of linguistically meaningful word associations modelled within a query expansion technique, we use a recent corpus based model of word meaning, known as the *tensor encoding* (TE) model [5]. The TE model is grounded in structural linguistic theory, which states that that a meaning of a word is based on its relationships with other words. The two types of linguistic relationships underpinning meaning are: (i) syntagmatic and (ii) paradigmatic associations.

A *syntagmatic association* exists between two words if they co-occur more frequently than expected from chance. Some common examples may include “dog-bit” and “weather-rain”. A *paradigmatic association* exists between two words if they can substitute for one another in a sentence without affecting the grammaticality or acceptability of the sentence. Some common examples may include “bit-chased” and “book-article”.

3.1 The Tensor Encoding Model

The TE model provides a formal framework for combining measures of syntagmatic and paradigmatic associations that can be used to estimate the probability of observing a word w given a priming word q, and can be stated as:

$$P(w|q) = \frac{1}{Z_\Gamma}\left[\gamma s_{\text{syn}}(q,w) + (1-\gamma)s_{\text{par}}(q,w)\right], \qquad (7)$$

where $\gamma \in [0,1]$, mixes the amount of syntagmatic and paradigmatic features used in the estimation, and $Z_\Gamma = \sum_{w \in V_k}\left[\gamma s_{\text{syn}}(q,w) + (1-\gamma)s_{\text{par}}(q,w)\right]$, is used to normalise the distribution.

The TE model has been used to underpin an effective query expansion technique, known as *tensor query expansion* (TQE) [6]. For query expansion, the estimate in Equation (7) is extended to estimate the probability of observing a word w given a sequence of query terms $Q = (q_1, \ldots, q_p)$, and can be expressed as:

$$P(w|Q) = \frac{1}{Z_\Gamma}\left[\gamma s_{\text{syn}}(Q,w) + (1-\gamma)s_{\text{par}}(Q,w)\right]. \qquad (8)$$

To model syntagmatic and paradigmatic associations, the TE model is underpin by geometric representations of words that are automatically built from word order and co-occurrence information found in a set of training documents. The binding process that builds these geometric representations for each word involves moving a sliding context window across the text. The binding process for the second order TE model is defined as:

$$\boldsymbol{M}_w = \sum_{t \in CW}^{t \prec w} (r - d_t).\boldsymbol{e}_t \otimes \boldsymbol{e}_w^T + \sum_{t \in CW}^{t \succ w} (r - d_t).\boldsymbol{e}_w \otimes \boldsymbol{e}_t^T, \quad (9)$$

where w is the focus term, t is a non-stop word found within the sliding context window (CW), $k \prec w$ indicates that term t appears before term w in the context window, $k \succ w$ indicates that term k appears after term w, r is the radius of the sliding context window, and d_k is the number of terms separating term k and term w within the context window. A context window is often referred to by its length. However, in the TE model the term *radius* is used to define the context window, as it better highlights the symmetric nature of the window and it also makes the equations behind the model less cumbersome in notation.

In a (pseudo) relevance feedback setting, the training documents, which the context window is slid across, refers to the top k (pseudo) relevant documents. Once the text has been bound, each term is represented as a matrix in which the element values are proportional to the term-term co-occurrence frequencies. The generalised form of the matrix for term w will be similar to:

$$\boldsymbol{M}_w = \begin{pmatrix} 0, \ \ldots, 0, \ f_{1w}, \ 0, \ \ldots, 0 \\ \ldots \\ 0, \ \ldots, 0, \ f_{(w-1)w}, \ 0, \ \ldots, 0 \\ f_{w1}, \ldots, f_{w(w-1)}, f_{ww}, f_{w(w+1)}, \ldots, f_{wN} \\ 0, \ \ldots, 0, \ f_{(w+1)w}, \ 0, \ \ldots, 0 \\ \ldots \\ 0, \ \ldots, 0, \ f_{Nw}, \ 0, \ \ldots, 0 \end{pmatrix},$$

where f_{iw} is the value in row i column w of the matrix which represents the proximity scaled co-occurrence frequencies of term i before term w, f_{wj} is the value in row w column j of the matrix that represents the proximity scaled co-occurrence of term j after term w, and N is the number of unique terms in the vocabulary. This sparse representation is efficiently stored in low dimension storage vectors, that allow for computationally efficient similarity measures to be used on the terms.

Intuitively, in a (pseudo) relevance feedback setting strong syntagmatic associations between query terms and the other terms in the set of (pseudo) relevant documents are likely to exist. This is because most document ranking models, such as the unigram language model, score documents higher if they contain many instances of the query terms. Therefore, the top k (pseudo) relevant documents will contain terms seen often around the query terms. This suggests that the expansion terms used to update the query representation within a (pseudo) relevance feedback setting, even those produced by the unigram relevance model will have strong syntagmatic associations with the query. To test this prediction, we can compare the sets of expansion terms produced by the unigram relevance model and syntagmatic measure of the TQE approach.

Within the TQE approach the strength of syntagmatic associations between a sequence of query terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w can be measured using the cosine metric (i.e., the normalised dot product of the matrix representations), and simplifies to $s_{\text{syn}}(Q,w) =$

$$\frac{\sum_{\substack{j=1 \\ w \in Q}}^{N} s_w^2 f_{jw}^2 + \sum_{\substack{j=1 \\ j \neq w \\ w \in Q}}^{N} s_w^2 f_{wj}^2 + \sum_{\substack{i=q_1 \\ i \neq w}}^{q_m} (s_i^2 f_{iw}^2 + s_i^2 f_{wi}^2)}{\sqrt{\sum_{i=q_1}^{q_m}\left[\sum_{j=1}^{N} s_i^2 f_{ji}^2 + \sum_{\substack{j=1 \\ j \neq i}}^{N} s_i^2 f_{ij}^2\right]}\sqrt{\sum_{j=1}^{N} f_{jw}^2 + \sum_{\substack{j=1 \\ j \neq w}}^{N} f_{wj}^2}}, \qquad (10)$$

where $q_1, \ldots, q_m$ are the unique query terms in Q having $m \leq p$; s_i is the number of times term q_i appears in Q; f_{ab} is the co-occurrence frequency of term a appearing before term b in the vocabulary; f_{ba} is the co-occurrence frequency of term a appearing after term b. This measure was shown to provide effective estimates for words most likely to succeed or precede another in text [5] and hence was reputed to be a reliable indicator of syntagmatic associations.

To complete the picture on how the other half of word meaning can be modelled within the TQE approach, a measure of the strength of *paradigmatic* associations between a sequence of query terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w, can be expressed as:

$$s_{\text{par}}(Q,w) = \frac{1}{Z_{\text{par}}}\sum_{j=q_1}^{q_p}\sum_{i=1}^{N}\frac{f_{\overline{ij}}.f_{\overline{iw}}}{\max(f_{\overline{ij}}, f_{\overline{iw}}, f_{\overline{wj}})^2}, \qquad (11)$$

where $f_{\overline{ij}} = (f_{ji} + f_{ij})$, being the unordered co-occurrence frequency of terms i and j; N is the size of the vocabulary; max() returns the maximum argument value; and Z_{par} normalizes the distribution. The use of the TE model's paradigmatic measure was shown to outperform human judgement and like models on a benchmark synonym judgement test [5].

Given the demonstrated effectiveness of these measures of syntagmatic and paradigmatic information, they will be used to underpin the framework developed in this paper for classifying linguistic associations modelled within the unigram and positional relevance models. Before applying this framework in an empirical evaluation, the similarities between the estimation techniques used by the syntagmatic feature in Equation (10) and unigram relevance model in Equation (2) will be demonstrated. This is to provide algebraic support to the intuition that syntagmatic associations are modelled within the unigram relevance model.

3.2 Use of Syntagmatic Associations

Research into the use of explicit term dependencies within the query expansion process found that when using information about syntagmatic associations a wider context window can lead to improved retrieval effectiveness [8]. That is, words far apart in a document can display strong syntagmatic associations.

We evaluated the retrieval effectiveness of the TQE approach, using solely the syntagmatic feature, $s_{\text{syn}}(.)$ to estimate query expansion terms within a pseudo relevance feedback setting on two TREC web data sets (Table 1). The results (Figure 1) indicate consistent improvements in retrieval effectiveness can be achieved by using larger context windows when modelling syntagmatic associations. Figure 1 also illustrates the robustness of $s_{\text{syn}}(.)$ for context window radii above 200. As the context window in the TE binding process does not cross document boundaries, it is worth considering the retrieval effectiveness achieved when the context window radius is set to each document's length. Using this radius, the MAP scores achieved by $s_{\text{syn}}(.)$ are 0.2491 and 0.0492 on the GOV2 (G2) and ClueWeb09 Category B (CW) data sets, respectively. This result indicates that superior retrieval effectiveness is achieved when syntagmatic associations between terms across the whole document are considered.

This condition can be modelled by the TE binding process in Equation (9) by setting the context window radius (r) equal to the document length ($|D|$). The resulting binding expression becomes:

$$\boldsymbol{M}_w = \sum_{t \in CW}^{t \prec w} (|D| - d_t).\boldsymbol{e}_t \otimes \boldsymbol{e}_w^T + \sum_{t \in CW}^{t \succ w} (|D| - d_t).\boldsymbol{e}_w \otimes \boldsymbol{e}_t^T. \tag{12}$$

The algebraic form of elements on row w of the matrix M_w in Equation (12) becomes:

$$f_{w,j} = \sum_{\substack{D \in \mathcal{R}_Q \\ w \in D}} df_j(|D| - \bar{d}_{w,j}), \tag{13}$$

and on column w:

$$f_{i,w} = \sum_{\substack{D \in \mathcal{R}_Q \\ w \in D}} df_i(|D| - \bar{d}_{i,w}), \tag{14}$$

where D is a document in the set of pseudo relevant documents $\mathcal{R}_Q$; $|D|$ is the length of document D; df_j is the frequency of term j in document D; $\bar{d}_{w,j}$ is the average number of terms separating term w from term j when w is seen before j in document D; and $\bar{d}_{i,w}$ is the average number of terms separating term w from term i when w is seen after i in document D.

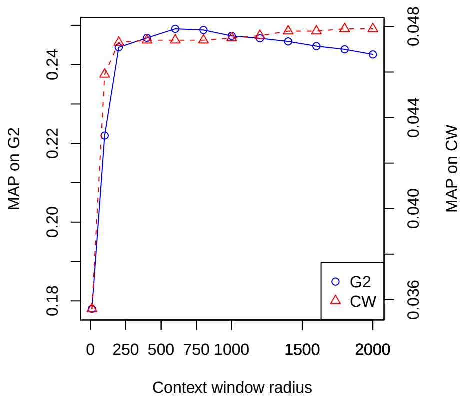

Figure 1: Sensitivity of $s_{\text{syn}}(.)$ to window radius.

	Description	# Docs	Topics	$\lvert q \rvert$
G2	2004 crawl of .gov domain	25,205,179	701-850	11 (4.1)
CW	Clueweb09 Category B	50,220,423	Web Track 51-150	9 (3.3)

Table 1: Overview of TREC collections and topic descriptions. $\overline{|q|}$ represents the average length of the queries, and the value in brackets is the standard deviation of the query lengths.

When Equations (13) and (14) are substituted into Equation (10), the syntagmatic feature $s_{\text{syn}}(.)$ produces higher scores for terms that occur frequently (large df_j and df_i) in the pseudo relevant documents. This result is similar to that produced by the Dirichlet smoothed likelihood estimation in Equation (1), which underpins RM3. However, Equation (1) contains a document normalisation factor. The cosine metric that defines $s_{\text{syn}}(.)$, also uses a form of normalisation that is linked to the document length. Equations (13) and (14) infer that terms that occur in larger documents will likely produce larger Frobenius norms (denominator of Equation (10)), and hence normalise the syntagmatic measure based on document length.

Therefore, the estimation techniques used in RM3 and $s_{\text{syn}}(.)$ (when the binding process in Equation 12 is used), are effectively based on term document frequencies and a document length normalisation factor. This result would lead us to believe that RM3 may be using very similar information to TQE's syntagmatic feature.

4. CLASSIFYING TERM DEPENDENCIES

The following section develops a framework to classify linguistic associations used within query expansion. Given that the TE model's syntagmatic feature has performed effectively on a word priming task and the paradigmatic feature has outperformed human judgement and like models on a benchmark synonym judgement task [5], we argue that they provide two reliable measures of structural linguistic associations.

The expansion terms used in the following analysis are produced during an ad hoc retrieval task carried out in a pseudo relevance feedback setting. Data set details are shown in Table 1. These TREC data sets are large web based collections that may make findings from these experiments relevant to web based applications. *Verbose* queries were chosen as they are long, discourse like queries, likely to provide sufficient term statistics to allow effective modelling of word associations within the TE model [7].

The experiments in this research were carried out using the Lemur Toolkit[1]. The Lemur implementation of the original positional relevance model is made available by the original authors[2]. The comparison of expansion terms is carried out using a Jaccard coefficient analysis and a Spearman's rank correlation coefficient analysis. The Jaccard coefficient analysis measures the average number of expansion terms that are common between two approaches. The Spearman's rank correlation coefficient is a finer grained analysis and measures, on a per query basis, how similar the overlap of

[1] The Lemur toolkit for language modelling and information retrieval: http://www.lemurproject.org
[2] http://sifaka.cs.uiuc.edu/~ylv2/pub/prm/prm.htm

		RM3	PRM	$s_{\text{syn}}(.)$
G2	PRM	.509 (20)	1 (30)	
	$s_{\text{syn}}(.)$	.458 (19)	.362 (16)	1 (30)
	$s_{\text{par}}(.)$	.104 (6)	.108 (7)	.138 (6)
CW	PRM	.634 (23)	1 (30)	
	$s_{\text{syn}}(.)$	.466 (19)	.437 (18)	1 (30)
	$s_{\text{par}}(.)$	.131 (7)	.130 (7)	.144 (8)

Table 2: Average Jaccard co-efficients for the sets of expansion terms produced on the G2 and CW data sets for the best performing RM3, PRM, TQE syntagmatic and paradigmatic features. The average number of expansion terms that overlap between approaches is shown in brackets.

two sets of expansion terms are with a third set.

The models used in the evaluation include RM3 and PRM. PRM was included in the evaluation to provide a benchmark for the amount of linguistic information being used by a technique that *explicitly* models term dependencies.

Given the focus is on comparing the expansion terms produced by each estimation technique, all common model parameters were fixed, including the number of feedback documents (30) and the number of expansion terms (30).

For each of the query expansion techniques, the remaining free model parameters were trained using 3-fold cross validation with the objective function maximising the MAP metric. This includes training the μ in Equation (2) for RM3. The free parameters trained for PRM include both σ and λ in Equation (6). The baseline unigram language model, used as the document scoring technique for all approaches, was run using the Lemur default parameters. The syntagmatic and paradigmatic features were built on a semantic space using a context window radius of 200 and 1 respectively.

Table 2 reports the Jaccard coefficients for the sets of expansion terms produced by RM3, PRM and the TQE syntagmatic and paradigmatic features. When compared to RM3, the syntagmatic feature $s_{\text{syn}}(.)$ has a minimum Jaccard coefficient of 0.458 (Table 2). This means that on average at least 19 out of 30 expansion terms suggested by $s_{\text{syn}}(.)$ are in common with those suggested by RM3.

As a comparison, PRM has a minimum Jaccard coefficient of 0.362 with $s_{\text{syn}}(.)$. This implies that on average at least 16 of 30 expansion terms are in common between PRM and $s_{\text{syn}}(.)$. This suggests that both RM3 and PRM use syntagmatic information when estimating query expansion terms, and that in fact RM3 has a stronger claim to the use of this form of term dependency.

Table 2 also shows that on average RM3 and PRM share a maximum of 7 expansion terms (out of 30) with those produced by TQE's paradigmatic measure $s_{\text{par}}(.)$. This result suggests that both RM3 and PRM use very little information about paradigmatic associations in their estimation process.

To investigate the overlap for each topic, a per-topic Spearman's rank correlation coefficient analysis, along the number of overlapping expansion terms on the $s_{\text{par}}(.)$ feature, was performed for the RM3, PRM and $s_{\text{syn}}(.)$ approaches. The resulting coefficients were, $\rho_{(PAR:SYN,RM3)} = 0.941$, $\rho_{(PAR:SYN,PRM)} = 0.863$ and $\rho_{(PAR:RM3,PRM)} = 0.883$. This result again suggests that RM3 may be using more information about syntagmatic associations than PRM.

The above discussion provides empirical and theoretical evidence to suggest that in augmenting the query model, RM3 uses information about syntagmatic associations. Given the linguistic credentials of TE model's syntagmatic feature, this research raises questions over the claim that dependency based approaches, like PRM and LCE, significantly outperform RM3 due to their use of explicit modelling of term dependencies. The gap in retrieval effectiveness may then be due to other factors.

5. CONCLUSION

The framework outlined in this paper provides a valuable method for classifying linguistic associations used within query expansion. We believe this framework can help information retrieval researchers better understand the types of linguistic term dependencies that may be responsible for differences in retrieval effectiveness.

This was demonstrated by using the framework to compare the strength of syntagmatic and paradigmatic associations displayed in query expansion terms for the unigram and positional relevance models. We found that not only do the best expanded query models for each approach display heavy use of syntagmatic associations, but the unigram relevance model has a stronger reliance on these syntagmatic associations. This leads us to question the claim that the unigram relevance model is outperformed by dependency based query expansion approaches because they use term dependencies.

6. REFERENCES

[1] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, New York, NY, USA, 2001. ACM.

[2] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR '10*, SIGIR '10, pages 579–586, New York, NY, USA, 2010. ACM.

[3] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR '07*, pages 311–318, New York, NY, USA, 2007. ACM.

[4] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.

[5] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Modelling word meaning using efficient tensor representations. In *PACLIC '11*, pages 313–322, 2011.

[6] M. Symonds, P. Bruza, L. Sitbon, and I. Turner. Tensor Query Expansion: a cognitive based relevance model. In *Australasian Document Computing Symposium 2011*, pages 87–94, 2011.

[7] M. Symonds, G. Zuccon, B. Koopman, P. Bruza, and A. Nguyen. Semantic judgement of medical concepts: Combining syntagmatic and paradigmatic information with the tensor encoding model. In *ALTA '12*, pages 87–94, 2012.

[8] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96*, pages 4–11, New York, NY, USA, 1996. ACM.

[9] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, pages 403–410, New York, NY, USA, 2001. ACM.