Generalizing Translation Models in the Probabilistic Relevance Framework

Navid Rekabsaz*, Mihai Lupu, Allan Hanbury
Vienna University of Technology
$family\_name$@ifs.tuwien.ac.at

Guido Zuccon
Queensland University of Technology
g.zuccon@qut.edu.au

ABSTRACT

A recurring question in information retrieval is whether term associations can be properly integrated in traditional information retrieval models while preserving their robustness and effectiveness. In this paper, we revisit a wide spectrum of existing models (Pivoted Document Normalization, BM25, BM25 Verboseness Aware, Multi-Aspect TF, and Language Modelling) by introducing a generalisation of the idea of the translation model. This generalisation is a de facto transformation of the translation models from Language Modelling to the probabilistic models. In doing so, we observe a potential limitation of these generalised translation models: they only affect the term frequency based components of all the models, ignoring changes in document and collection statistics. We correct this limitation by extending the translation models with the statistics of term associations and provide extensive experimental results to demonstrate the benefit of the newly proposed methods. Additionally, we compare the translation models with query expansion methods based on the same term association resources, as well as based on Pseudo-Relevance Feedback (PRF). We observe that translation models always outperform the first, but provide complementary information with the second, such that by using PRF and our translation models together we observe results better than the current state of the art.

Keywords

IR Models; translation model; word embeddings; related terms

1. INTRODUCTION

In Information Retrieval, terms are still the fundamental building blocks for establishing topical relevance relationships between documents and queries. This is not a limitation of the research, nor of the machines, but rather a fact of human communication. We count terms because we cannot otherwise quantify meaning.

*This work is funded by: Self-Optimizer (FFG 852624) in the EUROSTARS programme, funded by EUREKA, the BMWFW and the European Union, and ADMIRE (P 25905-N23) by FWF. Thanks to Joni Sayeler and Linus Wretblad for their contributions in the SelfOptimizer project.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983833

This is not for lack of trying. Distributional semantics became popular with Latent Semantic Analysis/Indexing (LSA/LSI) [6] in the early 1990s. Probabilistic Latent Semantic Indexing [10] (pLSI), Latent Dirichlet Allocation [4] (LDA), Random Indexing [28], and most recently, neural-network based learning methods [23] are its successors.

Nevertheless, the "basic" models based on the Probabilistic Retrieval (PR) Framework [30], and language modelling [27] have maintained a respectable command of the research and practice of IR. Despite their differences in the event spaces on which probabilities are being considered, they are all fundamentally based on term frequency (tf) as a representation of the importance of a term within a document, and document frequency (df) as a representation of the specificity of a term, potentially normalised, pivoted, or smoothed by collection statistics (e.g. average document length, average term frequency, collection frequency).

The extension of these models with some form of semantic matching receives continuous attention in our community. Recently, Li and Xu [20] published a survey on the topic, grouping the various approaches into 5 categories:

1. Matching by Query Reformulation
2. Matching with Translation Model
3. Matching with Term Dependency Model
4. Matching with Topic Model
5. Matching with Latent Space Model

Both Topic Modelling and Latent Space Models are still to be conclusively proven competitive in terms of both efficiency and effectiveness with probabilistic and language models.

Term Dependency Models address one of the fundamental assumptions in IR and attempt to go beyond the Bag-of-Words model. Recently, Hudson and Croft [11] presented a systematic comparison of such models. This line of research is complementary to the current study.

Of the five categories, we focus here on Translation Models and Query Reformulations. The two are in fact related, because one may argue that a translation model acts as if the query had been reformulated. Both have a considerable history behind them. Considering Pseudo-Relevance Feedback (PRF) as a form of query reformulation, we can trace this back to the the late 60s [31], while translation models have appeared immediately after the introduction of language models in the late 90s [3].

Essentially, translation models as introduced by Berger and Lafferty extend the probability of a query q given the model M_d of a document d by including a translation probability P_T between all the terms t_d of the document d and each term t_q of the query:

$$P(q|M_d) = \prod_{t_q \in q} \left(\sum_{t_d \in d} P_T(t_q|t_d) P(t_d|M_d) \right) \quad (1)$$

While translation models have been further investigated in the context of language modelling (most recently in the work of Karimzadehgan et al. [12, 13]), the idea has not been considered in the context of the Probabilistic Relevance Framework.

In the context of the current popularity of neural-network based methods for distributional semantics, it is therefore interesting to revisit these models, and potentially extend them towards the probabilistic models.

To address this, we propose to expand the PR Framework-based IR models in a way that does not affect their core tenets, but still takes advantage of the newly available, high-quality results in term-term similarity.

As before, we consider the terms as the representations of meaning. A query "information management" is the composition of the two concepts denoted by the two terms. When to compute a tf/df score we count occurrences, we implicitly assume that a document containing the term "information" will be to some extent (proportional to the tf) about the concept denoted by this term. Equally, if the term "information" appears in many documents, we implicitly assume that it is not a discriminative term (proportional to df). A document containing the term "knowledge" however, is also related to the concept "information", yet it does not contribute to the sense of a document not containing "knowledge". If we think of "information" however not as a term, but as a concept, we are entitled to replace the term "knowledge" with "information" and assign it a lower weight. It is here that the term-term similarity comes into play: the similarity is used to compute such a weight. Essentially, we are not even expanding the meaning of term frequency, because there was always an implicit assumption that we are counting concepts (this is why we normally do stemming). We propose to simply give tf the possibility to have values below 1, when terms are conceptually related but are not the same.

This change, while coming from a different perspective on the nature of text documents, can be viewed as a generalization of the translation model idea from LM to the PR Framework.

However, when observed from the PR Framework perspective, this change has some implications on the other statistics used in IR models: document length, document frequency, collection frequency. For instance, if we change the tf, then the length of a document, which is the sum of the tf values of its terms, changes as well. We set out to investigate the effects of these changes as well.

In summary, the main contributions of the current study are:

1. a generalisation of the idea of translation models into the PR framework models (we consider four models: Pivoted Document normalization, BM25, BM25 Verboseness Aware, and Multi-Aspect TF)
2. an extension of the translation models in PR Framework by considering the effects of changing tf on all other term, document, and collection statistics.
3. extensive experimental results comparing the traditional translation model, the newly proposed ones, as well as query expansion methods, including Pseudo-Relevance Feedback.

The proposed models go beyond the state of the art in experimental results, and maintain the simplicity and robustness of the existing models, despite the fact that, at least in this paper, we do not perform any optimisation on existing parameters (e.g. b, k_1 in BM25).

The remainder of this work is structured as follows: First, we review related work in Section 2. We introduce the translation as well as extended IR models in Section 3. We present our experimental setup in Section 4, followed by discussing the results in Section 5. Section 6 summarises our observations and concludes the paper.

2. BACKGROUND AND RELATED WORK

We should start by noting that fundamentally this study addresses the problem of synonymity in information retrieval and therefore the following page will not do justice to the amount of related work in this area. As pointed out in the introduction, we investigate the benefits of a newly available resource in IR (neural-network based term similarities), revisit a classic method to use term-term relations (translation model) and expand it to the probabilistic relevance framework. We also compare with query expansion methods. This section is therefore structured along these lines.

2.1 Word embeddings

Before talking about the retrieval models or the query expansion methods, a few notes about the term-term similarity method used. We are, from the onset, omitting any consideration of manually created resources. There are well known reasons to do so, primarily concerning their creation costs and extensibility to new domains and languages.

In terms of automatically created similarity resources, there are several variants of considering co-occurrence, or context similarity. The currently undisputed state of the art is neural network based. We use the method proposed by Mikolov et al. [23]: skip-gram with negative-sampling training (SGNS) method in the Word2Vec framework.

While this is not the newest method in this category (e.g. Pennington et al. [26] introduced GloVe and reported superior results), independent benchmarking provided by Levy et al. [19] shows that there is no fundamental performance difference between the recent word embedding models. In fact, based on their experiments, they conclude that the performance gain observed by one model or another is mainly due to the setting of the hyper-parameters of the models. Their study also motivates our decision to use SGNS: *"SGNS is a robust baseline. While it might not be the best method for every task, it does not significantly underperform in any scenario."*

2.2 Retrieval models

Berger and Lafferty [3] introduced translation models almost two decades ago as an extension to the language modelling, specifically the Query Likelihood model. In the Query Likelihood model, the score of a document d with respect to a query q is considered to be the probability of generating the query with a model M_d estimated based on the document: $score(q, d) = P(q|M_d)$

The method to estimate $P(q|M_d)$ is therefore the essence here. This implies two issues: defining what kind of model M_d should be, and estimating the probability of q given the chosen model type. Typically, the model is a multinomial distribution and the probability is computed with a maximum likelihood estimator, together with some form of smoothing. This smoothing, while not being part of the original idea, is in the practice of LM-based methods of paramount importance. However, this not being the focus of this study, we use Dirichlet smoothing [38], as many others have done, successfully, before us [12, 34, 41].

A translation model introduces in the estimation of $P(q|M_d)$ a translation probability P_T (see Eq. 1) defined on the set of terms, always used in its conditional form $P_T(t|t')$ and interpreted as the probability of observing term t, having observed term t'.

This adds a third issue to the two above. Berger and Lafferty had used for computing P_T the scan Expectation Maximisation approach inspired by machine translation approaches. Karimzadehgan and Zhai [12] explored translation models using mutual information. Zuccon et al. [41] extend the work by evaluating word embeddings on translation language models. The work shows potential improvement in applying word embedding. We reproduce some of the experiments in these last two studies.

Other recent studies have combined language modelling and distributional semantics: Ganguly et al. [8] expand the classic language models through word embedding-based noisy channels which aim to discover the hidden dependencies between terms. Vulić and Moens [34] essentially provide a linear combination between language modeling and word embedding-based scores, calculated by generating an aggregated vector for the query. Both of these methods are ad-hoc insertions of the term-term dependence in the language modelling framework. With respect to the results presented by Ganguly et al., our experiments show significant improvement. Vulić and Moens focus on multilingual data and provide results only on one, custom, test collection.

In general, while query likelihood models have demonstrated excellent performance in standardized benchmarking, a recurrent critique has been that they do not model the concept of relevance. Lafferty and Zhai [17] introduced a formal way to relate language modelling to relevance, but this has been disputed by Robertson [29] and others. Research in the context of the Probabilistic Relevance Framework has continued in parallel to that on language modelling, with recently introduced models like the Multi Aspect TF (MATF [24]), BM25 Verboseness Aware (BM25VA [22]), and Significant Words Language Model [7] demonstrating state of the art results.

There have been repeated efforts to expand methods of the probabilistic relevance framework with information about term-term relatedness (also referred to as dependence or similarity). For instance, Zheng and Callan [40] address query term weighting by exploiting word embedding as a feature vector to train a model for the optimal term weights. However, keeping the changes limited to the set of terms in the original query significantly limits the impact of their method.

Zhao et al. [39] defined a set of methods for distance-based cross term dependence and use them to modify the IR components i.e. document term frequency, and document frequency for boosting retrieval. However, the focus of their study was terms appearing in proximity of each other in terms of their locations in the documents, not in terms of their semantic representations. More recently, Lioma et al. [21] further explored the issue of co-occurring terms as indication of divergence from the compositional assumption (i.e. that a phrase composed of n terms has a meaning that can be explained by the composition of the meaning of the n terms). These studies, addressing fundamentally the disadvantages of the unigram bag-of-words models are complementary to the present study.

2.3 Query expansion

Expanding existing retrieval models with term-term similarity using translation models has an intuitive connection with direct query expansion methods, where terms are actually added to the query and/or weights are being recalculated. Xu and Croft [36], in one of the earlier papers in this area divides query expansion methods in *global techniques* and *local feedback*. That is, we can either use general knowledge about the terms, extracted from external resources such as logs [5, 9], manual or automatic knowledge-bases [16, 35, 37], or we can use some form of pseudo-relevance feedback (PRF) [1, 18, 31]. The two methods actually use complementary information sources, so we will consider them both separately and combined, in our experimental section.

For the global techniques, more than for the local feedback methods, attention has to be paid to the proper weighting of the new terms, as they come from outside the model used to rank documents. Cui et al. [5] and later Gao and Nie [9] used a logarithm to weight a term with respect to the query and the term-term similarities:

$$w(t) = ln\left(\prod_{t_q \in Q} P_T(t|t_q) + 1\right) \tag{2}$$

Another way to normalize the weights on some set $\mathcal{T}$ of candidate terms to be added to the query is:

$$w(t) = \frac{P_T(t|t_q)}{\sum_{t' \in \mathcal{T}} P_T(t'|t_q)} \tag{3}$$

This was done for instance by Xiong and Callan [35] when considering Freebase as a source of external knowledge.

In terms of Pseudo Relevance Feedback (PRF) the probabilistic relevance framework has a built-in concept of relevance and therefore can naturally incorporate information provided through feedback, be it from users or pseudo [30]. For language modelling, Lavrenko and Croft [18] introduced the Relevance Model (RM), which selects expansion terms from top ranked documents and weights them based on the score of document ranking.

The divergence from randomness (DFR) framework [2] also allows a relatively straight-forward inclusion of feedback information: Amati et al. [1] study the robustness of QE by two factors: divergence of the distribution of the query term in the retrieved documents from a random distribution and the frequency of the term in the whole document. In this paper, we use one of the variants explored in Amati's study (the Bose-Einstein 1), via its implementation in Terrier. As the authors pointed out, it had been shown to perform well in previous TREC tasks. Nevertheless, we specifically decided not to chose a QE method based on either the probabilistic relevance framework or language modelling, as we wanted to minimize the possibility of introducing unknown effects in our experiments when we compare these two models.

3. NOVEL TRANSLATION MODELS

We now introduce our approach to integrate the ideas of the translation model in the Probability Relevance Framework, referring to as Generalized Translation Model. We put the focus of this study on four models: two classical: Pivoted Length Normalization [33] and BM25, and two state-of-the-art schemes: Multi Aspect Term Frequency [24] and BM25 Verboseness Aware [22].

While translation models only focus on changing the tf components, when we consider the relation between tf and other document and collection statistics in the probabilistic relevance framework, a valid hypothesis to investigate is that simultaneously changing the other components (e.g. df, document length) would further improve the final models. Our assumption is that these new models benefit from semantic relations of the terms while the robustness of the original models has been preserved. We call this approach Extended Translation Model and integrate it in the probabilistic relevance as well as the language modelling framework.

In what follows, first we explain the approach to extend the basic components of the models (tf, df) and then use the extended components to introduce the translation models in the four probability relevance models as well as in language modelling. Finally, we briefly revisit query expansion, explaining the approach for combining it with any translation model.

3.1 Basic Components

The fundamental idea of the introduced translation methods is, for each term t of a query q, to replace any existing related terms t' in a document d with the term itself, but counting its occurrence as less than 1. Consequently, a set of changes will appear in the definitions of tf_d, df, and T_d (term frequency, document frequency, and the set of terms in a document).

In order to define the new components, we first denote the set *related terms* to a given term as $R(t)$. The similarity value of each term in this set is expected to be between 0 and 1. In this work, we calculate the value by using the Cosine function of the vector representations of the terms from a word embedding model. While the use of Cosine may be arguable in this context, it is the current practice and an investigation in this sense is outside the scope of this study. To create this set, we follow two approaches: 1. using the top-N most similar terms and 2. filtering the terms with similarity values higher than a threshold. The details of each will be discussed in the next sections.

Let us start with T_d: the set of terms associated with a document d changes with respect to a query q by replacing each related term with the term of the query to which it is related:

$$\widehat{T}_d = T_d \setminus \bigcup_{t \in q} \{t' \in R(t)\} \cup \{t \in q \colon R(t) \cap T_d \neq \emptyset\} \quad (4)$$

As a consequence of this redefinition of the documents, we must change the document frequency statistic accordingly:

$$\widehat{df}_t = \left|\{d \in D \colon t \in T_d \vee \exists t' \in R(t), t' \in T_d\}\right| \quad (5)$$

where D is the set of the documents in the collection. As defined here, the extended document frequency $\widehat{df}_t$ considers the documents containing sufficiently similar words in addition to the ones with the term itself. The hypothesis is that it prevents over-scoring of the documents that have terms with many similar terms in the query.

Finally, and most importantly, given the set of the related terms to the query, we define the extended term frequency as follows:

$$\widehat{tf}_d(t) = tf_d(t) + \sum_{t' \in R(t)} P_T(t|t') tf_d(t') \quad (6)$$

As defined, the new $\widehat{tf}_d(t)$ extends the basic $tf_d(t)$ by similar terms and therefore rewards the documents with more related terms.

Given the above three fundamental building blocks, the other remaining components are defined as follows:

$\widehat{L}_d = \sum_{t \in \widehat{T}_d} \widehat{tf}_d(t)$	document length
$\widehat{avgdl} = \frac{1}{\lvert D \rvert} \sum_{d \in D} \widehat{L}_d$	average document length
$\widehat{tf}_c(t) = \sum_{d \subseteq D} \widehat{tf}_d(t)$	term collection frequency
$\widehat{L}_c = \sum_{t \in T} \widehat{tf}_c(t)$	collection size
$\widehat{avgtf}_d = \frac{1}{\lvert \widehat{T}_d \rvert} \sum_{t \in \widehat{T}_d} \widehat{tf}_d(t)$	average term frequency
$\widehat{mavgtf} = \frac{1}{\lvert D \rvert} \sum_{d \in D} \widehat{avgtf}_d$	mean average term frequency

where their original forms are denoted as L_d, $avgdl$, $tf_c(t)$, L_c, $avgtf_d$, and $mavgtf$ respectively.

3.2 Generalized and Extended Translation Models

Based on the extended factors we just defined, we revisit the IR models and replace their components with the introduced extended ones. Since the logarithm function is regularly used as the dampening function, we use $\Lambda(x) = log(1 + x)$ to shorten notations.

3.2.1 Pivoted Length Normalization

Singhal et al. [33] identify a bias in the Cosine normalization as it favors long documents in retrieval. They then propose the pivoted length normalization (PL) schema by introducing a correction factor on the document length normalization. By replacing the elements of the original model, we define the Generalized Translation model (GT) and Extended Translation (ET) model as follows:

$$PL_{GT}(q,d) = \sum_{t \in \widehat{T}_d \cap T_q} \frac{\Lambda(\Lambda(\widehat{tf}_d(t)))}{1 - s + s\frac{L_d}{avgdl}} tf_q(t) log \frac{|D| + 1}{df_t} \quad (7)$$

$$PL_{ET}(q,d) = \sum_{t \in \widehat{T}_d \cap T_q} \frac{\Lambda(\Lambda(\widehat{tf}_d(t)))}{1 - s + s\frac{\widehat{L}_d}{\widehat{avgdl}}} tf_q(t) log \frac{|D| + 1}{\widehat{df}_t} \quad (8)$$

We should note that the original formulation uses $1 + log(1 + log(tf_d)$ in the numerator, while in our formula above we use $log(1+log(1+tf_d))$. For values of $tf_d > 1$ there is little difference between the two, and they have both been used in the literature. In our case, as it is theoretically possible that $tf_d < 1$, the formulation $1 + log(tf_d)$ may give negative values, hence we prefer the $log(1 + tf_d)$ variant.

3.2.2 BM25

BM25 is a widely popular and well-studied weighting model, rooted in the 2-Poisson probabilistic model of term frequencies in documents [30]. Due to the lack of space, we only show the Extended Translation model in Eq. 9 and 10. The Generalized Translation model ($BM25_{GT}$) follows the same approach while only replacing the $\widehat{tf}_d(t)$ and $\widehat{T}_d$ components in the classical version.

$$BM25_{ET}(q,d) = \sum_{t \in \widehat{T}_d \cap T_q} \frac{(k_1+1)\overline{\widehat{tf}_d(t)}}{k_1 + \overline{\widehat{tf}_d(t)}} \frac{(k_3+1)tf_q(t)}{k_3 + tf_q(t)} \log \frac{|D| + 0.5}{\widehat{df}_t + 0.5} \quad (9)$$

with

$$\overline{\widehat{tf}_d(t)} = \frac{\widehat{tf}_d(t)}{\widehat{B}(d)}, \quad \widehat{B}(d) = (1 - b) + b\frac{\widehat{L}_d}{\widehat{avgdl}} \quad (10)$$

3.2.3 Multi Aspect TF

Recently, Paik [24] addressed the limitations of the pivoted length normalisation by exploiting new statistical factors in the Multi Aspect TF (MATF) schema. The first component is Term Frequency Factor (TFF) which consists of two factors: Relative Intra-document tf (RI) measures the importance of a term regarding to the average tf of the document and Length Regularised tf (LR) that considers the length of the document in relation to the average document length in the collection. Paik [24] then mentions the different tendency of the factors to long and short queries and combines them using the parameter ω which promises a reasonable balance between the factors based on the query length. Both the factors are dampened first by the log and then by $f(x) = \frac{x}{1+x}$. We therefore revisit the TFF component as follows:

$$\widehat{RI}(t,d) = \frac{\Lambda(\widehat{tf}_d(t))}{\Lambda(\widehat{avgtf}_d)} \quad (11)$$

$$\widehat{LR}(t,d) = \widehat{tf}_d(t)\Lambda(\frac{\widehat{avgdl}}{\widehat{L}_d}) \quad (12)$$

$$\widehat{TFF}(t,d) = \omega\frac{\widehat{RI}(t,d)}{1 + \widehat{RI}(t,d)} + (1 - \omega)\frac{\widehat{LR}(t,d)}{1 + \widehat{LR}(t,d)} \quad (13)$$

as suggested by the paper, the ω parameter can be estimated by the following function:

$$\omega = \frac{2}{1 + \Lambda(|q|)} \quad (14)$$

where $|q|$ is the length of the query.

The second component is the Term Discrimination Factor (TDC) which uses inverse document frequency as well as average elite set term frequency (AEF) based on the total occurrence of a term in the entire collection. The extension of the factor as formulated as follows:

$$\widehat{AEF}(t) = \frac{\widehat{tf}_c(t)}{\widehat{df}_t} \tag{15}$$

$$\widehat{TDC}(t) = log\frac{|D|+1}{\widehat{df}_t}\frac{\widehat{AEF}(t)}{1+\widehat{AEF}(t)} \tag{16}$$

Finally, the extended version of the MATF's translation model is defined by integrating the two components:

$$MATF_{ET}(q,d) = \sum_{t\in\widehat{T}_d\cap T_q} \widehat{TFF}(t,d)\widehat{TDC}(t) \tag{17}$$

Here again, the generalised translation model ($MATF_{GT}$) is defined by only replacing $\widehat{tf}_d(t)$ and $\widehat{T}_d$ in the original form of the model i.e. only Eq. 11 and 17 are affected.

3.2.4 BM25 Verboseness Aware

Most recently, Lipani et al. [22] addressed the document length normalisation factor of BM25 by proposing a novel parameter-free length normalisation method that removes the need for the b parameter of BM25, called BM25 Verboseness Aware (BM25VA). The method leverages the mean of the average occurrences of a term in the documents to discover and supervise the effect of verboseness in the documents. As the difference between the method and BM25 is only in the factor $B(d)$, we introduce the extended version as follows:

$$\widehat{B}_{VA}(d) = \widehat{mavgtf}^{-2}\frac{\widehat{L}_d}{\widehat{T}_d} + (1-\widehat{mavgtf}^{-1})\frac{\widehat{L}_d}{\widehat{avgdl}} \tag{18}$$

The Generalized Translation and Extended Translation models replace the original (B_{VA}) and extended ($\widehat{B}_{VA}$) form with the $B(d)$ component of the Translation and Extended BM25 model (Eq. 10) respectively.

3.2.5 Language Model

The translation model has been introduced in the framework of language modelling [3], so in this case we only point out that the Generalised Translation model is the original one, as introduced by Berger and Laferty (i.e. it is Generalised from language modelling to the Probabilistic Relevance Framework). For completeness, we also introduce the Extended Translation model for the LM framework.

In order to unify the notation, we can rewrite the translation LM in Eq. 1 as follows:

$$LM_{GT}(q,d) = \prod_{t_q\in q} P_T(t|d) \tag{19}$$

where $P_T(t|d) = \sum_{t_d\in d} P_T(t_q|t_d)P(t_d|d)$ is the translation probability of generating term t in document d. Similar to related studies [12, 41] , we define $P(t_d|M_d)$ as the maximum likelihood estimation and inject $P_T(t|d)$ into a Dirichlet smoothing function obtaining:

$$P_T(t|d) = \frac{L_d}{L_d+\mu}\left[\sum_{t'\in T_d}\frac{P_T(t|t')tf_d(t')}{L_d}\right] + \frac{\mu}{L_d+\mu}p(w|C) \tag{20}$$

Now we can select the alternative terms t' based on the set of related terms $R(t)$ and rewrite the element in the square brackets above by explicitly exposing the term t where its translation probability to itself is one:

$$tf_d(t) + \sum_{t'\in R(t)} P_T(t|t')tf_d(t') \tag{21}$$

Eq. 21 is in fact identical to our definition of $\widehat{tf}_d(t)$ (Eq. 6) and therefore we can formulate the translation language model based on $\widehat{tf}_d(t)$ factor as follows:

$$LM_{GT}(q,d) = \prod_{t_q\in q}\left(\sum_{t_d\in d}\frac{L_d}{L_d+\mu}\widehat{tf}_d(t) + \frac{\mu}{L_d+\mu}\frac{tf_c(t)}{L_c}\right) \tag{22}$$

Finally we define the Extended Translation model by replacing the other components with their extended versions:

$$LM_{ET}(q,d) = \prod_{t_q\in q}\left(\sum_{t_d\in d}\frac{\widehat{L}_d}{\widehat{L}_d+\mu}\widehat{tf}_d(t) + \frac{\mu}{\widehat{L}_d+\mu}\frac{\widehat{tf}_c(t)}{\widehat{L}_c}\right) \tag{23}$$

3.3 Translation Models with Query Expansion

Translation models have an intuitive connection with direct query expansion methods. A natural question arising from generalising translation models into the probabilistic relevance framework is how they compare with query expansion methods and whether they benefit from pseudo relevance feedback (PRF).

Considering a query expansion method ϕ and the new set of terms as $\phi(q)$, the general query expansion models is defined as:

$$S^*(q,d) = \sum_{t\in\phi(q)} wS(t,d) \tag{24}$$

where each of the new terms has a coefficient of w, S^* is the final document score, and S is any scoring schema described above. If ϕ is based on term-term similarity, then S must be one of the basic methods (i.e. not using either the Generalized, nor Extended Translation models) because we would be using the same terms in both cases. If ϕ is PRF, then S can be any of the methods previously described because the set of words would be different.

4. EXPERIMENTAL METHODOLOGY

In order to evaluate the performance of the introduced Generalized Translation (GT) and the Extended Translation (ET) models, we evaluate them based on each of the mentioned relevance models (Section 3.2) on 6 test collections. In addition, we combine and test both the translation models with the PRF query expansion method as described in Section 3.3. We denote the Generalized Translation and Extended Translation models, combined with PRF as *PRF-GT* and *PRF-ET* respectively.

In the following, we introduce our experimental methodology, including test collections, baselines, parameter settings, and evaluation metrics.

Data Resources.

We conduct the experiments on 6 collections: combination of TREC 1 to 3, TREC-6, TREC-7, and TREC-8 of the AdHoc track, TREC-2005 HARD track, and CLEF eHealth 2015 Task 2 User-Centred Health Information Retrieval [25]. For the TREC tasks we always used the title of the queries for retrieval. Table 1 summarises the statistics of the test collections. For pre-processing, we apply

Table 1: Test collections

Name	Collection	# Doc	Topics
TREC 123	Disc1&2	740088	51-200 Adhoc
TREC 6	Disc4&5	551873	301-350 Adhoc
TREC 7, 8	Disc4&5 without CR	523951	351-400, 401-450 Adhoc
HARD	AQUAINT	1033461	2005 Track (50 topics)
eHealth	as defined in [25]	1104337	CLEF-eHealth 2015 Task 2 (67 topics)

Table 2: Baselines and their symbols for the significance tests

Baseline	Tested from	Sig. Test
STD	All the models and baselines	†
LOG	GT, ET	ℓ
NORM	GT, ET	ν
PRF	PRF-GT, PRF-ET	ρ
GT / PRF-GT	ET / PRF-ET	§

the Porter stemmer and remove stop words using a small list of 127 common English terms. We use the NLTK[1] toolkit.

We train the word embedding model for the Adhoc and Hard tracks using the Wikipedia dump file for August 2015. For the eHealth task, similar to Koopman et al. [15] that train word embeddings based on the domain corpora, we use the corpus extracted from the task's collection. For both, we use the Word2Vec SGNS method with vectors of 300 dimensions, sub-sampling parameter set to 10^{-5}, context windows of 5 words, epochs of 25, and word count threshold 20. Our own experiments (not reported here) as well as those reported by Zuccon et al. [41], indicate these parameters as reasonable as a general baseline.

Baselines.

In order to test the performance of the introduced translation models (GT and ET), for each IR schema we define three baselines:

1. *STD*: The original version of the models
2. *LOG*: The query expanded version using the logarithm weighting model, introduced in Eq. 2
3. *NORM*: The query expanded version with the normalization over the expanded terms, formulated in Eq. 3

Both the LOG and NORM expansion models calculate the final score using Eq. 24, while their weighting methods are only for the expanded terms and the weights of the original terms of the query are one ($t \in T_q : w = 1$). In addition to these two methods, we experimented with the direct use of translation probability $P_T(t|t_q)$ as the weight for expansion. However due to the extremely weak performance observed, we removed it from the baselines.

In the experiments with combination to Pseudo Relevance Feedback query expansion (PRF-GT and PRF-ET), we test them against two baselines: original PRF, and original model (STD).

Finally, in both basic and with PRF modes, we test the performance of Extended Translation model (ET/PRF-ET) against the Generalised Translation model (GT/PRF-GT) respectively.

All the baselines as well as their corresponding symbols for the significance test are summarised in Table 2. Statistical significance tests are done using the two sided paired t-test and statistical significance is reported for $p < 0.05$.

Related Terms.

An essential part of all our extended models is the definition of "the set of related terms". In order to find this set for a given term ($R(t)$ in Section 3.1), we consider two approaches: 1. selecting the top-N similar terms in the collection, and 2. selecting the set of terms whose similarity values to the term t are above a threshold θ.

Normally, the first approach is the common method for defining the related terms, used in several studies [8, 41]. However, as shown by Karlgren et al. [14], the distribution of the distances of the neighbouring terms is different for various terms, i.e. some words have more/less neighbours in a specific boundary. Inspired by this study, we observe the neighbouring terms of the term *excursion* in the Word2Vec model and spot the term *tourist* is the 4th most similar (closest) neighbor. However, looking at the top neighbours of *tourist*, the term *excursion* is the 17th one[2]. We assume that as *tourist* is a more frequent term with more contexts in the language, its neighbourhood is richer than *excursion* and in other words has more related terms. This observation motivates us for experimenting the effect of selecting the related terms based on a threshold θ as an alternative for the top-N approach.

Table 3: Example of conceptually related document, found by our approach, while not judged in the TREC-6 AdHoc track

Qry. 311	Industrial Espionage
Doc. FBIS4-23903	... recruited last year by the **intelligence** service once more ... were indicted for high treason in the form of **spying**, including ... **agent** was in particular in charge of financing ...

Parameter Setting.

Since the basic parameters of each model are shared between the extended and original method, the choice of parameters is not explored as part of this study. Therefore, for each method we select a standard set of parameters, suggested in related studies. For BM25 and BM25 Verboseness Aware (BM25VA), we set $b = 0.6$ (only for BM25), $k_1 = 1.2$, and $k_3 = 1000$, for Pivoted Length Normalization (PL), the parameter s is set to 0.05, and for Language Modeling (LM), we set μ to 1000. The Multi Aspect TF (MATF) does not require any parameter setting. For PRF we arbitrarily fixed the number of top-ranked documents to 3 and the number of expanded terms to 10.

In filtering related terms, for the top-N approach, following the related studies, we try N with 2, 5 and 10 and for the threshold approach, we experiment with θ values of $\{0.65, 0.68, 0.70, 0.72, 0.77, 0.82\}$.

Evaluation Metrics.

The evaluation of retrieval effectiveness is done with respect to MAP and NDCG@20, as standard measures. However, our initial experiments showed that the extended methods retrieved a substantial proportion of unjudged documents. Looking at some of these unjudged retrieved results, we find different documents that seem relevant to the query. For example, as shown in Table 3, the document does not contain the term 'espionage' requested by the query, but there are many occurrences of the similar words like 'spy', 'intelligence service', or 'agent'. We assume that it is due to the essential difference between the extended models and the standard term frequency-based methods which contributed to the creation of the relevance assessments used in the collections. Therefore, in order to provide a fairer evaluation framework, we consider MAP and NDCG over the condensed lists [32][3].

[1] http://www.nltk.org

[2] Noted that, the Cosine similarity is symmetric and in this case its values is equal to 0.54.

[3] The condensed lists are used by adding the -J parameter to the trec_eval command parameters

5. RESULTS AND DISCUSSION

We evaluated the performance of the introduced Generalized as well as Extended Translation models on the mentioned IR models (Section 3.2) with different parameters for filtering the related terms, discussed in the previous section. Among the results, we found using the threshold approach with $\theta = 0.7$ as the best performing and also most stable result among various models and therefore used for comparison of the models in the following.

The evaluation results of the MAP and NDCG@20 measures on the 6 test collections are shown in Figure 1. Each line in the plots shows the result of one IR model in two sections: from STD to ET the standalone version, and from PRF to PRF-ET when combined with the Pseudo Relevance Feedback query expansion. Significant differences of the results against the respective baselines are marked on the plots using the symbols defined in Table 2. Table 5 shows the detailed results.

Starting with the results of the MAP measure, we observe that using the Generalised as well as Extended Translation models we gain significantly better performance in 4 of 6 collections, compared to the original models as well as compared with the LOG and NORM expansion methods. Only in the TREC-123 and TREC-7 collections, there is no statistically significant improvement, although there is no deterioration of results either. Looking at the expansion methods, the LOG and NORM models also improve the baseline only slightly.

The results of combining PRF query expansion with the Generalised and Extended Translation models shows significant improvement over the original as well as PRF models (except in TREC-123 and TREC-7), achieved by both translation models. This improvement over PRF is similar to the improvement achieved by the models without PRF over the original models, showing indeed that *global techniques* and *local feedback* can effectively complement each other.

Comparing the Extended Translation model with the Generalised one, in general ET/PRF-ET brings only a slight improvement to GT/PRF-GT. In some cases, notably the eHealth collection, the PRF-ET model provides a significant improvement over all the other models including PRF-GT.

The trends in the results of the NDCG@20 measure are generally similar to the ones of MAP, except in some rare cases such as the LM and PTFIDF methods in the TREC-7 collection.

In order to have an overview on all the models, we calculate the gain of each model over its original form and averaged the gains on the six collections. The results for MAP are depicted in Figure 2a. As mentioned before, we can see the significance improvement of the GT and ET over the baselines. Also, while PRF has improved the baselines, its performance has then significantly been boosted by the generalised and extended translation models. In addition, ET/PRF-ET show overall slight improvement to GT/PRF-GT. In some cases, e.g. for the BM25 and BM25VA models, this is significant.

In order to compare with previously reported results, Table 4 shows the best achieved results in each collection with the normal evaluation (i.e. not considering only the condensed lists, but rather considering the retrieved unjudged documents as non-relevant). In the literature it is not always clear what method the authors have used, so identifying the state-of-the-art for each collection is difficult and potentially controversial. TREC-8 Ad Hoc is however one of the most widely reported benchmarks, and regardless of whether we consider the condensed lists or not, the generalised and extended translation models proposed here show considerable improvements with respect to reports of the most recent experiments in our field [8, 22, 24, 41].

Table 4: The best results per collection

Measure	Collection	Method	Scoring	Value
TREC-123	MAP	PRF	BM25V	0.306
	NDCG@20	ET	BM25V	0.571
TREC-6	MAP	PRF-ET	BM25V	0.270
	NDCG@20	PRF-ET	BM25V	0.455
TREC-7	MAP	PRF-ET	MATF	0.226
	NDCG@20	PRF-ET	MATF	0.424
TREC-8	MAP	PRF-ET	MATF	0.295
	NDCG@20	PRF-ET	MATF	0.481
HARD	MAP	PRF-GT	BM25V	0.241
	NDCG@20	PRF-GT	BM25V	0.375

Threshold or Top-N.

As mentioned before, we considered two approaches for selecting the related terms: threshold-based and top-N. Figure 2b shows the aggregated gain of the best performing top-N approach ($N = 2$) over all the collections. Comparing it with Figure 2a, we see that while the selection of related terms from the top N terms generally improves the baselines, the performance of GT and ET and respectively PRF-GT and PRF-ET models using the threshold method considerably outperforms the top-N approach.

By having a closer look to the number of selected terms per term in the threshold approach with $\theta = 0.7$, we see a wide range of numbers, from 0 (no expansion) in several cases to a maximum of 63 terms. The average number of terms is 1.4, but the standard deviation is 3.7.

On the other hand, the LOG and NORM models are only marginally affected by changing the approaches and keep the results close to the baseline. This is due to their conservative approaches for weighting the expanded terms—aggregating over all weights in NORM and dampening in LOG.

Limitations.

As with any method relying on a numerical value to represent the similarity of two terms, our extended components are limited by the definition of similarity. Analysing the cases where the extended model results were lower than the optimal showed that sometimes the extended terms introduce bias in search as they represent related terms but not similar ones. For example, the word embedding models indicate 'Alzheimer' as highly related to 'Multiple sclerosis (MS)' (as they usually appear in very similar contexts), although they are not similar in the sense that a query on one of them is hardly presumed to be satisfied by a document on the other. However, this is a general issue in query expansion, when the expanded words introduce bias to the original query. We can only consider these points as limitations of the methods and open questions for further research.

Efficiency.

Before concluding, it is worth noting that the Generalised as well as the Extended Translation models do not impose significant query-time overheads on the existing IR engines. Given the threshold, the set of related terms can be precomputed. The overhead of changing the statistics of the collection for the Extended model is computationally similar to one query time which makes it similar to the overhead of using PRF. Further optimization in this area is certainly possible. Our code is open source and available on Github[4].

6. CONCLUSION

We have proposed a generalisation and an extension of translation models in the probabilistic relevance framework models in or-

[4]https://github.com/neds/semanticsim

(a) TREC-Adhoc-123 (b) TREC-Adhoc-6 (c) TREC-Adhoc-7

(d) TREC-Adhoc-8 (e) TREC-2005-HARD (f) CLEF-eHealth15

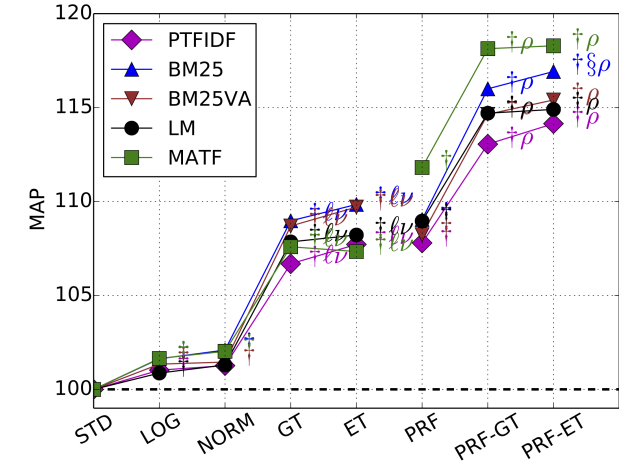
(g) TREC-Adhoc-123 (h) TREC-Adhoc-6 (i) TREC-Adhoc-7

(j) TREC-Adhoc-8 (k) TREC-2005-HARD (l) CLEF-eHealth15

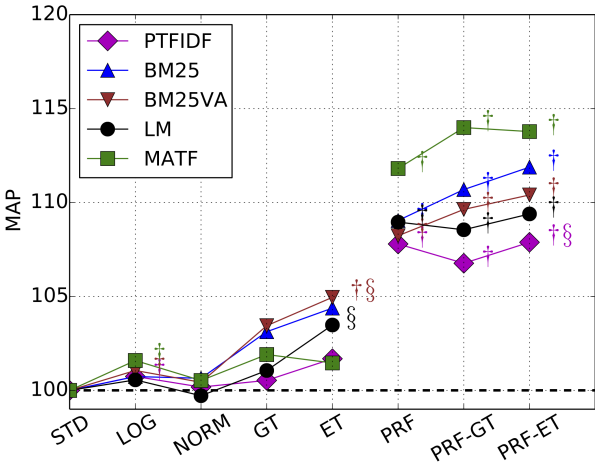
Figure 1: MAP and NDCG@20 evaluation of the TREC-123, TREC-6, TREC-7, TREC-8 Adhoc, TREC-2005 HARD, and CLEF-eHealth 2015 task 2. The baselines and the signs for significance difference tests are shown in Table 2. The related terms are filtered when the similarities of the neighbouring terms are higher than the threshold $\theta = 0.7$

der to take advantage of the new, rich semantic resources provided by recent developments in machine learning.

Concretely, we have introduced changes in the calculation of core elements of probabilistic relevance framework models (term frequency, document frequency), following the implicit assumption that query terms denote concepts and that counting the presence of these terms in the documents and the collection is a surrogate for counting the presence of the concepts. By simply replacing the occurrence of similar terms with that of the query terms we maintain the simplicity and robustness of the existing models, while improv-



(a) Related terms with best performing threshold ($\theta = 0.7$)



(b) Related terms with best performing top-N ($N = 2$)

Figure 2: The gain of the models with the MAP measure regarding to their original versions, aggregated over all the collections.

ing retrieval performance. We compared this approach with query expansion and also combined it with PRF based methods, observing the complementary effect of these two approaches, resulting in boosted performance.

This improvement in retrieval effectiveness is demonstrated on six test collections and five IR models, by achieving state-of-the-art results.

In the process, we also observe the effectiveness of selecting the "related terms" based on similarity boundary around the neighbouring space of a term. This approach shows competitive performance compared with selecting the top-N most similar terms, which is, based on our experiments, conclusively shown to underperform.

7. REFERENCES

[1] G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Query difficulty, robustness, and selective application of query expansion. In *Proc. of ECIR*, 2004.

[2] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *TOIS*, 2002.

[3] A. Berger and J. Lafferty. Information Retrieval As Statistical Translation. In *Proc. of SIGIR*, 1999.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[5] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proc. of WWW*, 2002.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of the American Society of Information Science*, 1990.

[7] M. Dehghani, H. Azarbonyad, J. Kamps, D. Hiemstra, and M. Marx. Luhn revisited: Significant words language models. In *Proc. of CIKM*, 2016.

[8] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word Embedding based Generalized Language Model for Information Retrieval. In *Proc. of SIGIR*, 2015.

[9] J. Gao and J.-Y. Nie. Towards concept-based translation models using search logs for query expansion. In *Proc. of CIKM*.

[10] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR*, 1999.

[11] S. Huston and W. B. Croft. A Comparison of Retrieval Models Using Term Dependencies. In *Proc. of CIKM*, 2014.

[12] M. Karimzadehgan and C. Zhai. Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval. In *Proc. of SIGIR*, 2010.

[13] M. Karimzadehgan and C. Zhai. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Proc. of ECIR*, 2012.

[14] J. Karlgren, A. Holst, and M. Sahlgren. Filaments of meaning in word space. In *Advances in Information Retrieval*. 2008.

[15] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proc. of CIKM*, 2012.

[16] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *Proc. of WSDM*, 2012.

[17] J. D. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In *Language modeling and information retrieval*, 2003.

[18] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of SIGIR*, 2001.

[19] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transaction of the Association of Computational Linguists (ACL)*, 2015.

[20] H. Li and J. Xu. Semantic Matching in Search. *Foundations and Trends in Information Retrieval*, 2014.

[21] C. Lioma, J. G. Simonsen, B. Larsen, and N. D. Hansen. Non-compositional term dependence for information retrieval. In *Proc. of SIGIR*, 2015.

[22] A. Lipani, M. Lupu, A. Hanbury, and A. Aizawa. Verboseness fission for bm25 document length normalization. In *Proc. of ICTIR*, 2015.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[24] J. H. Paik. A novel tf-idf weighting scheme for effective ranking. In *Proc. of SIGIR*, 2013.

[25] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. CLEF, 2015.

[26] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proc. of EMNLP*, 2014.

Table 5: MAP and NDCG@20 evaluation of the TREC-123, TREC-6, TREC-7, TREC-8 Adhoc, TREC-2005 HARD, and CLEF-eHealth 2015 task 2. In the models that need a set of related terms, the set is calculated based on the threshold approach with $\theta = 0.7$. The corresponding baselines for each model and their signs for the test of significance are shown in Table 2.

Collection	Method	PTFIDF		BM25		BM25VA		LM		MATF	
		MAP	**NDCG**	**MAP**	**NDCG**	**MAP**	**NDCG**	**MAP**	**NDCG**	**MAP**	**NDCG**
TREC-123	STD	0.304	0.577	0.303	0.557	0.311	0.572	0.302	0.566	0.290	0.560
	LOG	0.304	0.577	0.304	0.551	0.311	0.569	0.304	0.570	0.295	0.563
	NORM	0.305	0.573	0.306	0.552	0.311	0.564	0.303	0.566	0.296	0.559
	GT	0.310	0.590	0.314	0.577	0.320	0.588	0.316	0.583	0.304	0.572
	ET	0.315	0.593	0.319	0.585$\ell\nu$	0.326	**0.595**	0.320	0.588	0.304	0.574
	PRF	0.322†	0.572	0.326†	0.562	**0.330**†	0.564	0.324†	0.572	0.328†	0.560
	PRF-GT	0.313	0.565	0.322	0.547	0.326	0.560	0.320	0.559	0.326†	0.559
	PRF-ET	0.320	0.571	0.326†	0.557§	0.330	0.564	0.327†	0.560	0.323†	0.559
TREC-6	STD	0.240	0.424	0.250	0.440	0.254	0.449	0.250	0.428	0.260	0.460
	LOG	0.246	0.425	0.258†	0.443	0.265†	0.457	0.253	0.438	0.267	0.455
	NORM	0.247	0.422	0.258	0.442	0.264	0.451	0.256	0.434	0.267	0.450
	GT	0.265†	0.429	0.279†	0.457	0.287†ν	0.470	0.275†$\ell\nu$	0.439	0.288†$\ell\nu$	0.477
	ET	0.268†$\ell\nu$	0.434	0.283†$\ell\nu$	0.461	0.290†$\ell\nu$	0.474	0.275†$\ell\nu$	0.440	0.287†ℓ	0.479
	PRF	0.258†	0.436	0.268†	0.451	0.273†	0.458	0.271†	0.448	0.274	0.464
	PRF-GT	0.276†	0.439	0.288†	0.457	0.293†ρ	0.468	0.289†	0.449	0.295†ρ	0.476
	PRF-ET	0.281†ρ	0.444	0.292†ρ	0.459	**0.299**†ρ	0.472	0.288†	0.447	0.298†ρ	**0.479**
TREC-7	STD	0.190	0.397	0.194	0.404	0.201	0.418	0.200	0.399	0.194	0.415
	LOG	0.191	0.395	0.198†	0.410	0.204	0.417	0.202	0.405	0.197	0.419
	NORM	0.189	0.381	0.197	0.396	0.202	0.405	0.200	0.392	0.196	0.399
	GT	0.199	0.374	0.211	0.398	0.217	0.402	0.209	0.376	0.209	0.403
	ET	0.200	0.376	0.213	0.400	0.219	0.407	0.208	0.373	0.209	0.404
	PRF	0.213†	0.408	0.214†	0.402	0.221†	0.422	0.222†	0.410	0.226†	0.422
	PRF-GT	0.221†	0.388	0.231†	0.408	0.235†	0.420	0.231†	0.398	**0.239**†	0.419
	PRF-ET	0.222†	0.393	0.230†	0.406	0.234†	0.421	0.229†	0.396	**0.239**†	**0.424**§
TREC-8	STD	0.243	0.428	0.253	0.446	0.261	0.463	0.261	0.446	0.262	0.468
	LOG	0.245	0.437	0.257†	0.450	0.263	0.461	0.262	0.454	0.264	0.476
	NORM	0.243	0.430	0.256	0.446	0.262	0.458	0.262	0.449	0.263	0.472
	GT	0.257†$\ell\nu$	0.441	0.274†$\ell\nu$	0.453	0.281†$\ell\nu$	0.472	0.277$\ell\nu$	0.457	0.275ν	0.474
	ET	0.255ν	0.438	0.273†$\ell\nu$	0.451	0.280†$\ell\nu$	0.473	0.274	0.455	0.272	0.471
	PRF	0.256	0.427	0.272†	0.456	0.277	0.461	0.274	0.461	0.289†	0.475
	PRF-GT	0.274†ρ	0.440	0.298†ρ	0.469	0.300†ρ	0.472	0.295†ρ	0.473	0.311†ρ	0.484
	PRF-ET	0.274†ρ	0.440	0.299†ρ	0.476ρ§	0.300†ρ	0.478ρ§	0.292†ρ	0.469	**0.311**†ρ	**0.485**
HARD	STD	0.221	0.336	0.214	0.308	0.222	0.330	0.218	0.330	0.214	0.350
	LOG	0.224	0.340	0.218	0.318	0.223	0.337	0.223	0.340	0.217	0.349
	NORM	0.227	0.344	0.220	0.321	0.226	0.341	0.227	0.348	0.220	0.355
	GT	0.253†$\ell\nu$	0.362	0.252†$\ell\nu$	0.341	0.260†$\ell\nu$	0.361	0.259†$\ell\nu$	0.368†ℓ	0.244†ℓ	0.379†ℓ
	ET	0.255†$\ell\nu$	0.362†	0.253†$\ell\nu$	0.341	0.261†$\ell\nu$	0.365†	0.259†$\ell\nu$	0.365†ℓ	0.242†ℓ	0.376†ℓ
	PRF	0.250†	0.345	0.253†	0.339	0.257†	0.356	0.255†	0.357	0.258†	0.356
	PRF-GT	0.279†ρ	0.370	0.285†ρ	0.360†	0.290†ρ	0.379†	**0.291**†ρ	**0.380**†	0.286†ρ	0.377
	PRF-ET	0.278†ρ	0.365	0.285†ρ	0.361†	0.289†ρ	0.373†	0.287†ρ	0.375†	0.285†ρ	0.378
eHealth	STD	0.302	0.361	0.296	0.354	0.293	0.352	0.296	0.342	0.299	0.356
	LOG	0.304	0.363	0.299	0.358	0.297	0.355	0.296	0.339	0.304	0.361
	NORM	0.307	0.365	0.304	0.365	0.299	0.362	0.299	0.344	0.308	0.364
	GT	0.310	0.370	0.306	0.366	0.303	0.360	0.306	0.353	0.308	0.372†
	ET	0.317†ℓ	0.381†	0.309†	0.370†	0.307†	0.369†	0.311†ℓ	0.362†ℓ	0.310†§	0.373†
	PRF	0.309	0.362	0.306	0.363	0.304	0.356	0.310	0.358	0.316	0.371
	PRF-GT	0.316	0.371	0.310	0.372	0.307	0.362	0.310	0.358	0.322	0.378
	PRF-ET	0.322†ρ	0.380ρ	0.317ρ§	0.376	0.312†ρ	0.366	0.319†ρ§	0.367	**0.325**†§	**0.381**§

[27] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of SIGIR*, 1998.

[28] G. Recchia, M. Jones, M. Sahlgren, and P. Kanerva. Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. In *Proceedings the Cognitive Science Society Conference*, 2010.

[29] S. Robertson. On Event Spaces and Probabilistic Models in Information Retrieval. *Information Retrieval*, 8, 2005.

[30] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[31] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System– Experiments in Automatic Document Processing*, 1971.

[32] T. Sakai. Alternatives to bpref. In *Proc. of SIGIR*, 2007.

[33] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. of SIGIR*, 1996.

[34] I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. of SIGIR*, 2015.

[35] C. Xiong and J. Callan. Query expansion with Freebase. In *Proc. of ICTIR*, 2015.

[36] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proc. of SIGIR*, 1996.

[37] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proc. of SIGIR*, 2009.

[38] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proc. of SIGIR*, 2001.

[39] J. Zhao, J. X. Huang, and Z. Ye. Modeling term associations for probabilistic information retrieval. *TOIS*, 2014.

[40] G. Zheng and J. Callan. Learning to reweight terms with distributed representations. In *Proc. of SIGIR*, 2015.

[41] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In *Proc. of Australasian Document Computing Symposium*, 2015.