A Cross-lingual Framework for Monolingual Biomedical Information Retrieval

Dolf Trieschnigg[1] Djoerd Hiemstra[1] Franciska de Jong[1] Wessel Kraaij[2]
[1] HMI/DB group, University of Twente, Enschede, The Netherlands
[2] Dept. Computer Science, Radboud University Nijmegen / TNO, Delft, The Netherlands
{trieschn,hiemstra,fdejong}@cs.utwente.nl, kraaijw@acm.org

ABSTRACT

An important challenge for biomedical information retrieval (IR) is dealing with the complex, inconsistent and ambiguous biomedical terminology. Frequently, a concept-based representation defined in terms of a domain-specific terminological resource is employed to deal with this challenge. In this paper, we approach the incorporation of a concept-based representation in monolingual biomedical IR from a cross-lingual perspective. In the proposed framework, this is realized by translating and matching between text and concept-based representations. The approach allows for deployment of a rich set of techniques proposed and evaluated in traditional cross-lingual IR. We compare six translation models and measure their effectiveness in the biomedical domain. We demonstrate that the approach can result in significant improvements in retrieval effectiveness over word-based retrieval. Moreover, we demonstrate increased effectiveness of a CLIR framework for monolingual biomedical IR if basic translations models are combined.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval Models – language models

General Terms: Algorithm, Experimentation, Performance.

Keywords: CLIR framework, Biomedical IR, concepts, TREC Genomics, MeSH, UMLS.

1. INTRODUCTION

A major challenge for information retrieval in the life science domain is coping with its complex, inconsistent and ambiguous terminology [14, 22]. A single biomedical concept is often referred to using multiple terms (synonymy), including long multi-word phrases, ad hoc abbreviations and spelling variations. Shorter terms, in particular abbreviations, can be ambiguous: often the same term is used to refer to different concepts (homonymy).

It is evident that word-based information retrieval in this domain may benefit from knowledge found in terminological resources, such as controlled vocabularies, thesauri and domain-specific databases. These resources are commonly used for query expansion. Experiences during the TREC Genomics benchmarks illustrated, however, that beneficial incorporation of these terminological resources is far from trivial. An out-of-the-box TF.IDF retrieval system outperformed many sophisticated approaches incorporating knowledge from terminological resources [10]. Approaches which do benefit from terminological resources are frequently ad hoc or heavily geared towards the task at hand [21].

In this work, we view the integration of a concept-based representation in biomedical IR as a cross-lingual retrieval problem. We will demonstrate that approaches to traditional cross-lingual IR can be successfully applied for the integration of domain knowledge in biomedical IR.

The structure of this paper is as follows. First, we will describe our 'cross-lingual' framework for biomedical IR. In section 3 we will describe a number of translation models in this framework. In section 4 we will describe how these translation models are used to improve word-based retrieval. In section 5 the experimental setup for evaluating the proposed framework will be described. In section 6 the results of the experiments will be reported and discussed. We will conclude in section 7.

2. A CROSS-LINGUAL PERSPECTIVE ON BIOMEDICAL IR

Traditional cross-language IR (CLIR) is concerned with retrieving documents in a language different from the user's query language. For example, a user can formulate his or her information need in Spanish and the retrieval system retrieves English documents. Some kind of translation has to take place to allow for such cross-lingual matching. The translation can be based on a machine translation system, bilingual lexicons, translation models learned from multilingual document collections to name but a few [17].

Also in the monolingual setting, the mismatch between terms used in a query and terms used in relevant documents can be viewed as a cross-lingual matching problem. Berger and Lafferty [3] formalized this observation by viewing the query formulation process as a noisy translation from the language used in relevant documents. In biomedical IR the vocabulary mismatch can be substantial given the number of synonyms and ambiguous terms. In this paper, we take Berger et al.'s (1999) work a step further by identifying a second concept-based representation language.

The framework we propose is visualized in figure 1.

We identify two representation languages in this frame-

CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.

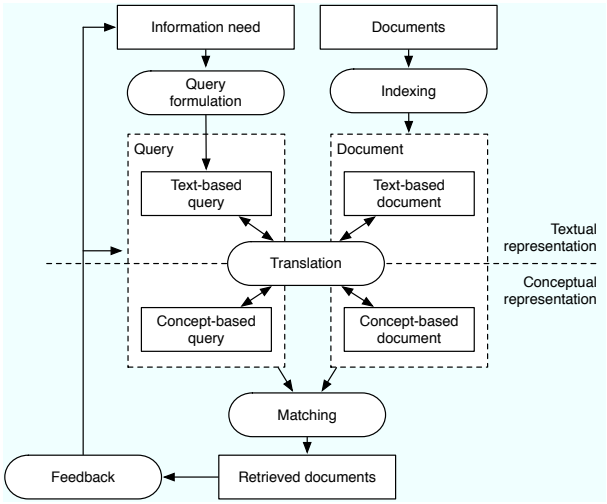

Figure 1: A cross-lingual view on biomedical IR

work. Firstly, a textual language in which queries are formulated by a user in free text and in which documents have been written. Secondly, a conceptual language which is defined by the concepts or synset entries in a terminological resource. For instance, the concept [Mad cow disease][1], which groups synonymous terms such as "mad cow disease" and "BSE". In some cases, documents are already available in such a conceptual representation. The citations in MEDLINE, for instance, have been manually annotated with terms from the Medical Subject Headings (MeSH) thesaurus. Queries are typically not directly available in a concept-based representation; some form of translation has to take place to obtain such a representation. In many other cases, both the query and document concept-based representations have to be obtained automatically. For instance, a biomedical named entity recognizer can be used to tag occurrences of concepts in (document and query) text.

Table. 1 shows an example of a document in three representations. The first column shows the title and the abstract of the citation. The second column lists a concept-based representation which has been manually determined by human indexers. The last column shows a concept-based representation which has been automatically obtained.

The integration of a concept-based representation in biomedical IR is then reduced to translating the query and/or documents, and matching them in the same representation language. Such a cross-lingual perspective gives the opportunity of adopting a large set of established CLIR methods and techniques for this domain.

In theory, a conceptual representation is preferred over a word-based representation. Synonymous (including complex multi-word) terms are mapped to a single conceptual representation. Ambiguous terms are mapped onto the conceptual representation which corresponds to the context in which they appear. IR then simply reduces to matching the conceptual representations of documents to queries.

In practice, however, a concept-based representation has its limitations. Early work by [11] demonstrated that using only a concept-based representation for retrieval was harmful for retrieval effectiveness. One reason was that not all information needs could be represented in terms of the concept vocabulary. Therefore we combine text or word-based retrieval with concept-based retrieval. This is clearly different from traditional CLIR where queries and documents are available only in different languages. In this CLIR-enhanced framework for biomedical IR, retrieval based on a text or word-based representation can be improved with a concept-based representation.

We identify two main translation resources for this type of biomedical CLIR. Firstly, the concept vocabulary itself. The concept vocabulary defines which phrases are used to express a concept, but does not indicate how frequently these terms are actually used or show the ambiguity of these terms. Analogous to traditional CLIR, the concept vocabulary can be used as a dictionary to translate between concepts and phrases and vice versa. Secondly, document corpora in a dual representation of both a text and concept-based representation can be used as a translation resource. In conventional CLIR, such parallel or comparable corpora are used to train translation models. Translation models between a source and target language are obtained from a large corpus of translated sentences or documents in both languages. A similar approach can be used to train translation models between a concept and a word-based representation.

Our main research question is as follows: *How can we adapt CLIR methods and techniques for more effective monolingual biomedical information retrieval?* We are interested in particular in how to build translation models in this domain and how these translation models can be used to improve monolingual (that is, text only) retrieval.

In this paper we will investigate a basic implementation of the framework. The text or word-based representation is restricted to a unigram word-based representation. Moreover, we limit the trained translation models to word-to-concept and concept-to-word translations.

We will investigate two concept languages in our work:

MeSH The Medical Subject Headings thesaurus[2], a controlled vocabulary used for manually indexing MEDLINE citations. We will only use the main headings as concepts, which are around 24,000 concepts using the 2008 edition of the thesaurus.

UMLS++ The Unified Medical Language System (UMLS) metathesaurus[3] extended with several gene and protein dictionaries for four species, referred to as UMLS++. The combined thesaurus consists of around 640,000 concepts from 59 vocabularies.

A major difference between the two is that for MeSH a manually curated document representation is already available. For UMLS++ the document representation is obtained automatically, using Peregrine [19]. Peregrine scans for UMLS entries in the text and performs a lightweight disambiguation strategy to resolve ambiguous terms. Another difference is the alignment of the concept representations with the text. For UMLS++, each concept is aligned to the word or phrase it refers to. For MeSH, such an alignment is not available: MeSH terms are assigned at the citation level. Hence, the representation in UMLS++ can be viewed as a parallel corpus, whereas the representation in MeSH can be viewed as a comparable corpus.

[1] square bracket notation is used to refer to concepts

[2] http://www.nlm.nih.gov/mesh/

[3] http://www.nlm.nih.gov/research/umls/

Table 1: PubMed citation (PMID 10050890) in text and two concept-based representations (shown partially).

Text (title + abstract)	MeSH (manual)	UMLS++ (automatic)
Fatal familial insomnia: a new Austrian family. We present clinical, pathological and molecular features of the first Austrian family with fatal familial insomnia. Detailed clinical data are available in five patients and autopsy in four patients. Age at onset of disease ranged between 20 and 60 years, and disease duration between 8 and 20 months. Severe loss of weight was an early symptom in all five patients. Four patients developed insomnia ...	[Adult] [Austria] [Brain] [Female] [Humans] [Sleep Initiation and Maintenance Disorders] [Male] [Middle Aged] [Pedigree] [Prions] [Blotting, Western] [Fatal Outcome] [PrPSc Proteins].	[Abnormality] [Adrenal Cortex] [Age] [Aging] [Analysis] [Astrogliosis] [Austrians] [Autonomic dysfunction] [Autopsy] [Bos taurus] [Brain Stem] [Brain] [Cattle] [Cell Nucleus] [Cerebellum] [Cerebral cortex] [Codon Genus] [Congenital Abnormality] [Cytoplasmic Granules] [Disease] [Dissociation] [Encephalopathies] ...

3. TRANSLATION MODELS

In the previous section, we mentioned two resources to build translation models. Firstly, a collection of documents in both a text and a concept representation. And secondly, the terminological resource itself, which groups a number of phrases into a concept.

The first translation model we will investigate, based on pseudo-feedback translation, translates a text-based representation as a whole to a concept-based representation, based on the co-occurrence of words and concepts in a comparable corpus. The other five translation models we will investigate are used to translate representations in a term-by-term fashion. They employ different methods to estimate probabilities for $P(w|c)$ (the probability of translating a concept c to the word w) and $P(c|w)$ (the probability of translating the word w to a concept c). On their own these term-by-term translation models are not expected to perform well, since they rely on only very little information for translation. However, they are expected to be useful when combined with the KNN translation model.

We will now describe six models based on these resources.

3.1 Pseudo-feedback translation (KNN)

The first translation model we will discuss is based on pseudo-relevance feedback in a different representation. The representation to translate is used to search a collection in a dual representation, and the translation of the nearest neighboring documents is used as a translation. In conventional CLIR such an approach was proposed by [15]. In work by [20] and [24], a similar feedback mechanism was used for MeSH. We refer to this translation as KNN, since K nearest neighboring documents are used to obtain the translation.

The translation is modeled as follows. We assume to have a document collection $\mathcal{D}$ available in both a conceptual and textual representation. For each document D, we can estimate a textual language model and a conceptual language model, $P(w|\theta_D)$ and $P(c|\phi_D)$ respectively.

We want to translate the text to translate (referred to as Q) to a conceptual language model $P(c|\phi_Q)$. The approximation of the language model is based on the joint probability of observing the concept c with the query Q in the previously introduced document collection $\mathcal{D}$. In words, this approach determines which concepts are most likely to co-occur with the query. Formally:

$$P(c|\phi_Q) \approx \frac{P(c,Q)}{\sum_{c'} P(c',Q)}, \quad (1)$$

where $P(c,Q)$ is the joint probability of observing a concept c with the query Q.

The joint probability of observing the concept with the query is approximated by independently sampling documents from the collection $\mathcal{D}$, followed by independently sampling the concept and the query from each document.

$$P(c,Q) = \sum_{D\in\mathcal{D}} P(D)P(c|\phi_D)P(Q|\theta_D) \quad (2)$$

$$\approx \sum_{D\in\mathcal{D}} P(D)P(c|\phi_D)\prod_{i=1..n} P(q_i|\theta_D) \quad (3)$$

where $P(D)$ is a prior probability of sampling the document D from the collection (assumed to be uniform) and $P(Q|\theta_D)$ is the probability of sampling the query from the document, the query likelihood (assuming term independence $P(Q|\theta_D) = \prod_{i=1..n} P(q_i|\theta_D)$).

Obviously, requiring the complete collection $\mathcal{D}$ to be processed for classifying a piece of text, makes the model infeasible in practice. The contribution of many documents to $P(c,Q)$ is relatively small, however, since they are not likely to generate the query ($P(Q|\theta_D)$ is small). Therefore, following [15], we can safely reduce this document collection to n documents with the highest probability of generating the query $P(Q|\theta_D)$. In practice, these are the top n documents ranked by query likelihood.

3.2 IBM Model 1 (M1)

The second translation model we will investigate is based on IBM Model 1, a statistical model of the translation process commonly used for traditional CLIR. [6] proposed five models for determining statistical translation models based on a bilingual collection of sentences. Central to these models is the estimation of an *alignment* of the sentences in two languages. This alignment connects terms in the sentences in one language to terms in the translated sentence in the other language. An EM-algorithm is employed to iteratively improve the alignment and the parameters of the translation model, respectively.

IBM Model 1 is the simplest of the five models proposed by [6], and does not take word order into account. Models 2 to 5 are increasingly sophisticated, incorporating absolute and relative word reordering and a fertility model. For biomedical CLIR, the concept-based representation does not have a term order. Since we limited our experiments to term-by-term translation models, we will only use Model 1 for our translation models from text to concepts and vice versa.

An advantage of using Model 1 for training biomedical translation models is its theoretical soundness. The subsequent models proposed by [6] illustrate that Model 1 is highly suitable to be extended to more sophisticated models. Disadvantages are that training the translation model is resource intensive and that with new concepts the whole training process has to be repeated.

3.3 Pointwise Mutual Information (PMI)

The third translation model we will investigate is derived from the pointwise mutual information (PMI) between the concept-based and word-based event space [7]. PMI indicates the association of two events based on their joint distribution in comparison to their individual probabilities. PMI and mutual information have frequently been used as an association measure for IR [25] and in particular for filtering ambiguous translations in a CLIR setting [4, 9]. [3] used the mutual information statistic for constructing a distribution function of words over documents to sample queries for documents. We will use such a distribution directly as a translation model. We argue that strongly associated concepts and words can be used as translations of each other.

In the literature, definitions of mutual information and pointwise mutual information are frequently confused. In this work, the following definition will be used for PMI.

$$PMI(w,c) = \log_2 \frac{p(w,c)}{p(w)p(c)} = \log_2 \frac{Nf(w,c)}{f(w)f(c)} \quad (4)$$

where $p(w,c)$ is the probability of encountering the word and concept together in a document collection, and $p(w)$ and $p(c)$ are the probabilities of encountering them separately in the collection. In the subsequent estimation of these probabilities $f(w,c)$ denotes the number of documents in which the words w and c appear together; $f(w)$ and $f(c)$ indicate the number of documents in which the word and concept appear respectively, and N is the size of the collection.

[16] noted that PMI is not an ideal measure for measuring the association between terms, since it is biased towards low-frequency words. Similar to [3], we circumvent this bias towards low-frequency words by introducing an additional factor based on occurrence frequency of the pair:

$$PMI'(w,c) = f(w,c) \log_2 \frac{p(w,c)}{p(w)p(c)} \quad (5)$$

Based on these scores, we create the translation model for a term in an ad hoc fashion: the n translation terms with the highest PMI' scores are selected and normalized by dividing the sum of the top n scores.

3.4 Parsimonious Term Translation (PTT)

The fourth translation model we will investigate is based on the conditional probabilities of encountering the target (translation) term after observing the source term in a large set of documents. Formally:

$$P(w|c) = \frac{f(w,c)}{\sum_{w' \in V} f(w',c)}, \quad (6)$$

where $f(w,c)$ is the number of times a word and a concept term occur together in a document, and the denominator indicates the sum of co-occurrences of the concept with any word in the word vocabulary.

Using this formula, relatively high translation probabilities will be assigned to frequently occurring words or concepts. It is undesirable to assign a high probability to the frequently appearing word 'study' as a translation of the concept [Parkinson's disease] simply because the word 'study' frequently co-occurs with the concept. An Expectation Maximization (EM) algorithm proposed by [12] is employed to prune low probability translations and remove these common terms. In monolingual IR, this approach has been used for query expansion [18] and determining domain models [1].

Table 3: KNN Concept translations for "Ferroportin-1 in humans".

0.095 [Humans], 0.091 [Cation Transport Proteins], 0.079 [Iron], 0.078 [Animals], 0.050 [Membrane Proteins], 0.038 [Enterocytes], 0.038 [Hemochromatosis], 0.036 [Carrier Proteins], 0.034 [Male], 0.021 [Iron-Binding Proteins], 0.021 [Mice], 0.018 [Biological Transport, Active], 0.018 [Cloning, Molecular], 0.018 [Zebrafish], 0.018 [Ferric Compounds], 0.017 [Duodenum], 0.017 [Models, Biological], ...

We use the EM algorithm as follows. After initializing the translation probabilities with the maximum likelihood estimate defined in eq. 6, the EM algorithm will be applied: During the expectation step, the probability mass will be redistributed depending on the global probability of a term. During the maximization step, the probability distribution will be normalized, that is, normalizing the sum of the translations to one.

$$\text{E-step:} \quad e_w = f(w,c)\frac{(1-\lambda)P(w|c)}{\lambda P(w) + (1-\lambda)P(w|c)} \quad (7)$$

$$\text{M-step:} \quad P(w|c) = \frac{e_w}{\sum_{w'} e_{w'}}, \quad (8)$$

where $P(w)$ is the probability of encountering the term w in a large collection and λ determines how parsimonious the translation model will be: a value of 0 results in the maximum likelihood estimate; a value close to 1 results in a translation model in which probability mass has been redistributed to fewer translations.

We will refer to these translation models as *parsimonious term translation* models (PTT).

3.5 Thesaurus (THES and STATTHES)

The last two translation models we will investigate use the thesaurus for determining translation probabilities between concepts and terms. In traditional CLIR, similar approaches have used to machine readable dictionaries to estimate translation models [13].

In the *naive* translation model based on a thesaurus (THES), the translation from words to concepts and vice versa, is estimated by their relative co-occurrence frequencies in entries in the thesaurus. As a formula:

$$P(w|c) = \frac{f(w,c)}{\sum_{w'} f(w',c)}, \quad (9)$$

where f(w,c) is the number of times the word w is used to describe c in the thesaurus. For instance, when the concept [Mice] has synonyms "mice", "house mouse" and "mouse", the probability of P(mouse|[Mice]) is equal to $\frac{2}{1+1+2} = 0.5$.

Similarly, the probability of translating a word to a concept can be approximated ($P(c|w) = \frac{f(w,c)}{\sum_{c'} f(w,c')}$).

The model based on a *statistical thesaurus* (STATTHES), also takes into account how frequently a particular word is used to refer to a concept in a corpus of documents. This requires the text to be tagged with concepts found in a thesaurus. f(w,c) is then defined as the frequency that the word w was tagged with the concept c.

Table 2: Translations and translation probabilities obtained from different translation models.
(a) Translating the MeSH concept [Mad cow disease] to words (word stems)

Model	Translations
M1	0.228 bse, 0.096 spongiform, 0.096 encephalopathi, 0.038 diseas, 0.030 transmiss, 0.028 cattl, 0.027 infect, 0.025 case, 0.020 agent, 0.019 bovin, 0.019 anim, 0.014 mad, 0.012 epidem, 0.011 variant, 0.011 clinic, 0.010 human, 0.009 scrapi, 0.009 prion, 0.300 ...
THES	0.250 spongiform, 0.250 bovin, 0.125 enceph, 0.125 encephalopathi, 0.062 bse, 0.062 cow, 0.062 mad, 0.062 diseas.

(b) Translating the word "ferroportin" to MeSH concepts.

Model	Translations
M1	0.184 [Cation Transport Proteins], 0.085 [Carrier Proteins], 0.076 [Homeostasis], 0.074 [Genetic Heterogeneity], 0.073 [Mutation, Missense], 0.054 [Amino Acid Substitution], 0.052 [Mononuclear Phagocyte System], 0.047 [Membrane Proteins], 0.047 [Receptors, Transferrin], 0.044 [Iron Overload], 0.043 [Italy], 0.042 [Chromosomes, Human, Pair 2], 0.040 [Codon], 0.140 ...
THES	*no translations available*

4. RETRIEVAL MODELS

In this section we will first describe the retrieval model used for integrating word and concept-based retrieval. After that, we will describe a number of extensions of this retrieval model which combine multiple translation models.

4.1 Basic retrieval model

Our basic word and concept-based retrieval system is based on statistical language models. Queries and documents are represented by unigram word and unigram concept language models. As a baseline, only the word language models of queries and documents are matched. Matching is enhanced by matching the (translated) concept query and document language models. Documents are ranked according to the negated cross entropy between query and document word and concept language models:

$$RSV(D,Q) = -\alpha H(\phi_Q|\phi_D) - (1-\alpha)H(\theta_Q|\theta_D) \quad (10)$$

where ϕ_Q and ϕ_D are the concept-based query and document language models respectively; θ_Q and θ_D are the word-based query and document language models; α controls the relative importance of the concept-based representation (for the baseline α is set to 0). Other fusion methods were investigated as well (such as CombMNZ, CombMax and CombSum [8]), but this form of interpolation of document scores turned out to be most effective.

ϕ_D and θ_D are smoothed document language models based on a maximum likelihood estimate. $P(c|\phi_D)$, the probability of generating the concept c from the language model is estimated as follows:

$$P(c|\phi_D) = (1-\lambda_c)\frac{f(c,D_C)}{|D_C|} + \lambda_c P(w|\hat{\phi}_C) \quad (11)$$

where $f(c,D_C)$ is the concept term frequency; $|D_C|$ is the total number of concepts in the document representation; $\hat{\phi}_C$ is a background language model used for smoothing and λ_c is a parameter which controls the amount of smoothing. $P(w|\theta_D)$ is estimated in a similar fashion.

The word-based query language model is based on an (unsmoothed) maximum likelihood estimate from the original query text. The concept-based query language model is obtained through query translation. The KNN translation model translates the word-based query model as a whole, that is, it translates a probability distribution over words directly to a probability distribution over concepts. The other five translation models are used to translate the word-based query word-by-word, similar to [3]. Formally:

$$P(c|\phi_Q) = \sum_{w\in Q} P(c,w|Q) = \sum_{w\in Q} P(c|w,Q)P(w|\theta_Q) \quad (12)$$

$$\approx \sum_{w\in Q} P(c|w)P(w|\theta_Q), \quad (13)$$

where w are the words in the word-based query Q, $P(w|\theta_Q)$ is the probability of w in the word-based query model and $P(c|w)$ is the translation probability as determined using the various translation models (PMI, M1, PTT, THES and STATTHES).

4.2 Combining translation models

In traditional CLIR, combining different translation resources has shown to be an effective way to improve translation quality [2, 5]. In the following sections we will propose a number of retrieval models and strategies which aim for a similar effect in biomedical CLIR.

4.2.1 Pruning

Since the translation based on pseudo-feedback (KNN) is based on documents, it is expected to contain noisy concepts which are only indirectly related to the original query. Indeed, the example translation in table 3 contains concepts which were found in related documents, but could not be directly linked to the text to translate (for example, [Animals], [Mice] and [Zebrafish]).

We propose the use of the term-by-term translation models to prune concepts from the translated concept-based query obtained through feedback (KNN). The concept-based representation obtained by feedback translation is filtered as follows:

$$P(c|\phi_Q) = \begin{cases} \kappa\, P_{\mathrm{KNN}}(c|\phi_Q) & \text{if } \sum_{t\in Q} P(t|c) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where $P_{\mathrm{KNN}}(c|\phi_Q)$ is the conceptual query language model estimated through feedback; $P(t|c)$ is a concept to term translation model; and κ is a query dependent normalization constant, which normalizes $\sum_c P(c|\phi_Q)$ to 1.

Note that this type of pruning based on term-by-term translation models is not very restrictive: concepts are only pruned from the translation when this concept cannot be translated to any of the query words; the translation probability in the concept-to-word translation model is not taken into account.

4.2.2 Reweighting

A well-known drawback of using pseudo-relevance feedback is possible query drift: an expanded query can overemphasize or neglect particular aspects from the original query, or skew towards aspects not mentioned in the original query. In the case of a pseudo-feedback translation to a conceptual representation, the neglect of a particular query aspect can be substantiated by the fact that aspects cannot be represented accurately by the concept vocabulary. As a result, combining a word and concept-based representation based on feedback may understate aspects present in the word-based representation. The goal of the reweighting procedure we will now describe is to prevent that a word-based query combined with a concept-based query (obtained through feedback) neglects aspects found in the word-based query. To achieve this, the word-based query model is reweighted: depending on how well the concept-based representation *covers* the words in the query, the word weights are updated: well-covered words receive a lower weight, whereas poorly covered words receive an increased weight.

The reweighting process is as follows: 1) The feedback translation model (KNN) is used to translate a word-based query model $P(w|\theta_Q)$ to a concept-based query model $P(c|\phi_Q)$. 2) The coverage of the words in the original word-based query model $P_{cov}(w|\phi_Q)$ is determined by translating the concept-based query model using the term-by-term translation models described earlier. 3) An updated word-based query model $P(w|\theta'_Q)$ is based on $P_{cov}(w|\phi_Q)$. The updated word-based query model is combined with the concept-based query model for retrieving documents.

How the coverage and updated word-based query model are determined will now be described.

Determining the coverage of the word-based query

The coverage of a word-based query by a concept-based representation is defined as a probability distribution over the words in the original query. If the word-based query is evenly covered by a concept-based representation this probability distribution is uniform: all query words are covered by concepts in the concept-based representation.

We use a term-by-term translation model to determine this coverage as follows.

$$P_{cov}(w|\phi_Q) = \frac{\sum_c P(w|c,\phi_Q)P(c|\phi_Q)}{\sum_{w'\in Q}\sum_c P(w'|c,\phi_Q)P(c|\phi_Q)} \quad (15)$$

$$\approx \frac{\sum_c P(w|c)P(c|\phi_Q)}{\sum_{w'\in Q}\sum_c P(w'|c)P(c|\phi_Q)}, \quad (16)$$

where $P(c|\phi_Q)$ is the concept language model obtained through pseudo-feedback translation of the original word-based query and $P(w|c)$ is the term-by-term translation probability of translating a concept c to a word w. In the (unlikely) case that none of the concepts can be translated to a query word $P_{cov}(w|\phi_Q)$ is equal to 0 for all w[4].

[4]This can be viewed as a coverage of a *null*-query word with probability 1.

Updating the word-based query language model

The coverage of the original word-based query language model is used to determine an updated word-based query language model.

We assume that all the aspects mentioned in the original text-based query are equally important: when searching with a combined word and concept-based query representation this balance should be maintained. When the concept-based representation does not cover all query aspects this balance is disturbed: some aspects are overemphasized leading to query drift. This query drift of a combined word and concept-based query representation can be prevented by decreasing the weight of words which are well covered by the concept-based representation.

We assume that the aspects of a query can be represented by the original word-based query language model (based on a maximum likelihood estimate). To retain the original query balance, the updated word-based query language model combined with the coverage by the concept-based query language model should approximate the original query word distribution. Formally:

$$P(w|\theta_Q) = \beta_Q\ P_{cov}(w|\phi_Q) + (1-\beta_Q)P(w|\theta'_Q) \quad (17)$$

where $P(w|\theta_Q)$ is the original query word language model, which should be covered by the translation of a conceptual query language model $P_{cov}(w|\phi_Q)$ and by an updated query language model $P(w|\theta'_Q)$. The query dependent parameter β_Q indicates the relative importance of the updated word-based query language model in comparison to the translated concept-based query language model.

To approximate eq. 17, initial estimates of the updated word-based query language model are calculated as follows:

$$e_w = \begin{cases} P(w|\theta_Q) & \text{if } P_{cov}(w|\phi_Q) = 0 \\ \dfrac{P(w|\theta_Q) - \beta_Q P_{cov}(w|\phi_Q)}{1-\beta_Q} & \text{otherwise} \end{cases} \quad (18)$$

The updated query language model is determined by normalizing these initial estimates:

$$P(w|\theta'_Q) = \frac{e_w}{\sum_{w'\in Q} e_{w'}} \quad (19)$$

Note that the second line of the equation is obtained by rewriting eq. 17. The value β_Q has to be restricted to prevent $P(w|\theta'_Q)$ becoming less than zero, formally:

$$0 \leq \beta_Q \leq \min_{w\in\phi_Q} \frac{P(w|\theta_Q)}{P_{cov}(w|\phi_Q)} \quad (20)$$

A β-value of 0 indicates that the updated word-based query language model is exactly the same as the original word-based query model; the largest possible value of β modifies $P(w|\theta_Q)$ as much as possible to retain the original query term balance.

Table 4 illustrates this reweighting in practice for a query consisting of three words (w_1 to w_3). Their original importance weights, based on the original query formulation is found in the second column. The third column indicates to what extent the words are covered by concepts found in the query. w_1 for example, has an original probability of 0.5, but is only covered by the translation with a probability of 0.1. The updated probability should therefore be higher than 0.5. The last three columns of the table show the re-estimated

	Original	Coverage	Updated weight $P(w\|\theta'_Q)$		
	$P(w\|\theta_Q)$	$P_{cov}(w\|\phi_Q)$	$\beta_Q = 0$	$\beta_Q = 0.1$	$\beta_Q = 0.25$
w_1	0.5	0.1	0.5	0.54	0.63
w_2	0.4	0.5	0.4	0.39	0.37
w_3	0.1	0.4	0.1	0.07	0

Table 4: Example of query term reweighting.

weights for three different values of β_Q. The highest possible value of β_Q for this query is 0.25, resulting in a reweighted probability for the word w_3 of 0.

To control the value of β_Q at a global level (that is across different queries), we introduce the parameter α (between 0 and 1) which linearly scales β_Q between its minimum and maximum value. Formally $\beta_Q = \alpha \min_{w \in Q} \frac{P(w|\theta_Q)}{P_{cov}(w|\phi_Q)}$.

4.2.3 Structuring

The last approach we investigate to combine translation models combines the original textual query with a conceptual query based on pseudo feedback into a structure. The approach is motivated by the idea that the translated concepts should be linked to the query words they represent. We hypothesize that such an approach balances the original textual query with its translation, and prevents query drift.

To allow for such an integration we need to model concepts and words in the same event space. We achieve this by simply merging the two representations, that is mixing the identifiers of the concepts with the tokens extracted from the text. From a principled modeling perspective, mixing the representations is not very attractive: concepts and words are different units of information and should therefore be kept separated. On the other hand, the mixed representation is easy to understand and straightforward to implement.

The parameters of the mixed document language model $P(t|\psi_D)$ are again based on a maximum likelihood estimation, smoothed with a background language model.

The initial parameters of the mixed query language model $P(t|\psi_D)$ are based on a linear interpolation of the word-based query model and the concept-based query model:

$$P(t|\psi_Q) = \alpha\ P(t|\theta_Q) + (1-\alpha) P(t|\phi_Q) \tag{21}$$

where α indicates the relative importance of the text-based representation with respect to the concept-based representation.

We will use a translation model $P(w|c)$ to create an *alignment* between the concepts and the words in this mixed query language model. Based on the translation model, each concept is assigned to (at most) one word. Assuming that the l terms in the word-based query are w_1 to w_l, and that the m concepts in the concept-based query are c_1 to c_m, we can define an alignment function between c_i and w_j as follows.

$$\delta(c_i, w_j) \quad = \quad \begin{cases} 1 & \text{if } j = \arg\max_{j'} P(w'_j|c_i) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

In words: the concept c_i is aligned to the word w_j with the highest translation probability. We now define $\sigma(w_j)$ of a word w_j as the set containing the word itself and the concepts which have been assigned to it.

$$\sigma(w_j) = \{w_j\} \cup \{\ c_i\ ; \delta(c_i, w_j) = 1\} \tag{23}$$

Similar to [13, p. 133], we use this set to define an equivalence class of the word and the concepts mapped to it:

$$P(class(w_j)|\psi_D) \quad = \quad \sum_{t \in \sigma(w_j)} \frac{P(t|\psi_Q)}{\sum_{t' \in \sigma(w_j)} P(t'|\psi_Q)} P(t|\psi_D)$$

The query language model of the equivalence class is defined as follows.

$$P(class(w_j)|\psi_Q) \quad = \quad P(w_j|\psi_Q) \tag{24}$$

5. EXPERIMENTAL SETUP

In this section we will describe the experimental setup for comparing the different translation and retrieval models.

The TREC Genomics document collections and topics sets between 2004 and 2007 were used for the evaluation [10]. The 2004 and 2005 topic sets consist of 50 queries and were used to search a document collection of 4,591,008 MEDLINE citations (referred to as the 2004 document collection). The 2006 and 2007 topic sets consist of 28 and 36 queries and were used to search a document collection of 162,259 full-text journal articles from Highwire Press (the 2006 document collection). The TREC Genomics task of 2006 and 2007 were passage retrieval tasks. In this evaluation, however, we only investigated ad hoc document retrieval: documents containing relevant passages were assumed to be relevant to the query.

Mean average precision (MAP) and rank precision (precision at 10) were used as evaluation measures. Due to space limitations, we will only mention MAP in the results section of this paper.

The translation models which required training data (all except for the naive thesaurus translation model), were trained with documents from the TREC Genomics 2004 document collection. Word-based representations of these documents were obtained using a tokenizer adapted to biomedical text [23]. The MeSH-based representations of the documents were based on the major MeSH headings assigned by NLM indexers; subheadings were discarded. The UMLS++-based document representation was obtained using the Peregrine[19]. The document collection in word and UMLS++-based representations was used both as a parallel and a comparable corpus. For training the STATTHES translation models, the explicit alignment between words and concepts (obtained from Peregrine) were used. For the other translation models, the alignment was discarded and the representation was treated as a comparable corpus. The document collection in word and MeSH-based representations was only used as a comparable corpus. Translation models were built for translation between MeSH and words, UMLS++ and words and vice versa.

The translation models for PMI and PTT were based on co-occurrence counts of concepts and words in the complete 2004 document collection. Because of scalability issues, the IBM model 1 translation models were built on a subset of the collection. 1,200,000 randomly selected documents from the collection were used to build the translation models. A slightly modified version of the GIZA++[5] machine translation toolkit was used to train the models based on IBM

[5] http://www.fjoch.com/GIZA++.html

model 1. The default setting of 5 iterations of the EM algorithm was used.

All translation models went through the following post-processing to remove noise: 1) Translations with a probability smaller than 0.001 were removed; 2) Words or concepts which occurred in fewer than 3 documents in the collection were pruned; 3) Single character words and numbers were removed. The remaining translations were normalized for each term (assuring $\sum_{t'} P(t'|t) = 1$).

The Lemur Toolkit[6] was used for indexing and retrieval.

6. RESULTS

This results section is structured as follows. First, we will investigate the effectiveness of the individual translation models. In sections 6.2 to 6.4 we will look into the effectiveness of combining the pseudo-feedback translation model with the term-by-term translation models for pruning, reweighting and structuring respectively.

As a baseline, retrieval using only the word-based representation was used. Each column lists the results of the TREC Genomics query set of that year (2004 to 2007). Results using a statistical thesaurus (STATTHES) are only reported for the UMLS++ representation; only for this vocabulary such a translation model was available.

6.1 Translation models compared

Table 5 lists the retrieval effectiveness in terms of mean average precision when using the combined word and translated concept-based language models for retrieval.

A first observation is that a concept-based representation translated from the textual query can significantly improve word-based retrieval. Using an additional MeSH-based representation leads to (significant) improvements up to 9.5% in MAP. For UMLS++ improvements up to 9.9% can be observed. The precision at 10 (not displayed in the table) shows similar improvements.

As expected, KNN performs best when considering all 4 topic sets. For 6 out of 8 cases, retrieval using a word-based representation combined with the concept translation obtained with KNN results in the highest MAP. The other translation models are all extremely limited in the amount of context they take into account for translation: a concept-based query is obtained by individually translating each word in the query to concepts. Considering the ambiguity of individual words in this domain, it is in fact surprising that this naive term-by-term translation results in improvements in retrieval effectiveness.

Using translations obtained from the naive thesaurus translation model does show slight improvements in mean average precision, but none of the improvements are statistically significant. A possible explanation for this lack of significant improvement is noise in the terminological resource: the resource sometimes mentions terms for concepts which are rarely used. The correct, or most common, translation of a concept or term may therefore receive a low translation probability.

The translation models trained on the comparable corpus (M1, PTT and PMI), performed slightly better than the translation model based solely on thesaurus information (THES). No significant differences were observed, however, between M1, PTT and PMI.

[6] http://www.lemurproject.org/

6.2 Pruning

The effect of pruning obviously depends on how many concepts are in fact pruned. The pruning method described in section 4.2.1 removed many concepts: between 49.9% and 91.5% of the concepts in the KNN translation were removed. The translation models based on PMI and IBM model 1, resulted in the most restrictive pruning (between 49.9% and 79.1%); the models based on PTT and the thesauri (THES and STATTHES) resulted in stronger pruning (between 81.9% and 91.5%).

This indicates that the KNN and term-by-term translations are quite different: for many concepts found in the KNN translation no translation to a word in the original query is indicated by the term-by-term translation models.

Table 6 lists the results of combining the pruned concept language model with a word-based language model for retrieval.

For MeSH, the pruned concept language model can still be used to increase the performance of word-based retrieval. However, for the query sets using the 2004 document collection (consisting of citations with relatively little text), the original concept translation performs better than the pruned translation. Apparently, pruning resulted in the removal of MeSH concepts which were beneficial for retrieval. For searching the full-text article collection (2006 and 2007 topic sets), pruning turned out to be more useful: pruning the KNN translation with the naive thesaurus translation model resulted in the highest retrieval effectiveness.

For UMLS++, interpolating the pruned concept representations with the text representations turned out to be almost as effective as or even more effective than the unpruned representation. Irrespective of the type of translation model used for pruning, significant improvements (up to 10.5% in MAP) over the text-based baseline were observed. For UMLS++ pruning turned out to be very useful: between 50.5% and 91.5% of the terms in the concept-based query could be pruned with the same or improved retrieval effectiveness.

6.3 Reweighting

Table 7 shows the result of reweighting the word-based query model, based on the coverage of the KNN concept language model, determined using the term-by-term translation models. The table shows the results for α set to 0.5. On average 0.19 and 0.17 of the probability mass of word-based query language model was redistributed for MeSH and UMLS++ respectively.

Reweighting based on coverage by MeSH concepts in many cases led to detrimental retrieval effectiveness. For the 2004, 2006 and 2007 not reweighting resulted in a higher effectiveness. For the 2006 and 2007 topic sets, effectiveness even significantly dropped below the word-based baseline. This effect can be explained by the exhaustiveness of the MeSH-based document representation. On average, a document in the 2006 Genomics collection is represented by only 15 MeSH concepts. It is likely that many more MeSH terms are in fact relevant to this document but have not been assigned. The word-based representation of the document is more exhaustive than the concept-based representation. Despite the fact that according to a translation model a concept 'covers' a query word, it is likely that this covering concept reduces the recall in comparison to the query word.

Reweighting based on the UMLS++ representation turned

Table 5: Retrieval effectiveness (MAP) using different translation models for obtaining a concept-based query model. ∧, △ and ▲ indicate significant differences (sign test) with p-levels < 0.05, 0.01 and 0.001 respectively.

(a) Using MeSH concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3868**△	**0.2429**∧	0.3736	**0.2916**
word + M1	0.3621▲	0.2250	0.3895	0.2864
word + PTT	0.3644▲	0.2240	0.3876	0.2844
word + PMI	0.3612▲	0.2235	0.3836	0.2842
word + THES	0.3589	0.2220	**0.3898**	0.2820

(b) Using UMLS++ concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3929**△	0.2285	**0.4048**	**0.2981**
word + M1	0.3743	0.2277▲	0.3928	0.2933∧
word + PTT	0.3751∧	0.2273△	0.4037	0.2930△
word + PMI	0.3630∧	**0.2287**▲	0.3976	0.2909△
word + THES	0.3617	0.2227	0.3925	0.2926
word + STATTHES	0.3652∧	0.2249∧	0.3982	0.2970

Table 6: Retrieval effectiveness after pruning.

(a) Using MeSH concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3868**△	**0.2429**∧	0.3736	0.2916
M1 prune	0.3656▲	0.2282△	0.3917	0.2858
PTT prune	0.3660△	0.2282∧	0.3857	0.2818
PMI prune	0.3660▲	0.2289▲	0.3912	0.2847
THES prune	0.3651▲	0.2277	**0.4011**	**0.2923**

(b) Using UMLS++ concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3929**△	0.2285	0.4048	0.2981
M1 prune	0.3854△	0.2275	0.4114	0.3062▲
PTT prune	0.3801	0.2293	0.4077	0.2947△
PMI prune	0.3853△	**0.2319**△	**0.4179**	0.3005△
STATTHES prune	0.3796△	0.2297	0.4122	0.3063△
THES prune	0.3806△	0.2303	0.4089	**0.3089**△

out to be more effective. Improvements (up to 5.3%) could be observed for the 2005, 2006 and 2007 topic sets. Many of the results are insignificant however. For the 2004 topic set, no improvements could be observed, even with different values of α. The effect of reweighting turned out to be independent of the translation model used.

6.4 Structuring

Considering the original number of concepts in a query (50), the structuring does not result in very large changes to the original query. On average between 1.1 to 3.2 equivalence classes were created, with between 1.8 and 6.2 concepts grouped into a single equivalence class.

Table 8 lists the impact of structuring the query using the term-by-term translation models. Structuring the representations turned out to give strongly varying results, from significant deteriorations (up to 22.7% in MAP) to significant improvements (up to 6.4% in MAP). The decline in performance can to some extent be attributed to a difference in granularity of the word terms which have been grouped with more specific or over general concept terms. For instance, the UMLS++ concept [nicotinic acetylcholine receptor location] is treated as a synonym of the word "nicotin". In other cases, clearly incorrect equivalence classes were formed. For example, the UMLS++ concept [Device breakage] is grouped with the word "break" in the context of "DNA breaks". In this case, the translation through feedback introduced these errors; by mapping these errors to original query words and treating them as equivalent, the impact of the erroneous translation was further emphasized. Improvements were observed when the words and concepts in the same equivalence class were clearly linked and were defined at the same granularity level.

7. CONCLUSION

In this paper we proposed a cross-lingual framework for biomedical IR. We distinguish between a concept and word-based representation language. We hypothesized that the integration of a concept-based representation in biomedical IR could benefit from methods and techniques used in established CLIR. In analogy to what is common in traditional CLIR, we identified three types of translation models for biomedical CLIR: 1) a comparable corpus of documents in both a text and concept-based representation; 2) term-by-term translation models trained on a comparable corpus; and 3) a thesaurus. We used these sources in different cross-lingual retrieval models. Despite the limited context taken into account, word-to-concept translation could still improve word-based retrieval. Translation based on pseudo-feedback using a comparable corpus in both a word and concept-based representation proved to perform best. In the other three retrieval models we evaluated whether translation between text and concepts could be improved by combining translation models. Despite the simplicity of the term-by-term translation models, the results showed that a combination of translation models could improve retrieval effectiveness when combined with a word-based representation.

The results also demonstrated the added value of the concept representations MeSH and an extended version of UMLS (UMLS++). MeSH turned out to be primarily a recall enhancing device, especially useful for citation retrieval. For a MeSH-based representation to be effective, however, many (also indirectly related) terms were required to represent the information need. The UMLS++ representation can be used to precisely represent information needs and can be used as a precision enhancing device.

We conclude that the proposed cross-lingual framework offers a transparent view on the integration of a concept-based representation for monolingual biomedical IR. Based on the promising results with relatively simple translation and retrieval models, we have high expectations for more sophisticated translation and retrieval models.

Table 7: Retrieval effectiveness after reweighting.

(a) Using MeSH concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	**0.3889**	0.2796
word + KNN	**0.3868**△	0.2429∧	0.3736	**0.2916**
M1 reweigh	0.3686	0.2428△	0.3400▽	0.2228▼
PTT reweigh	0.3608	**0.2436**△	0.3171▽	0.2147▼
PMI reweigh	0.3697	0.2425△	0.3463▽	0.2243▼
THES reweigh	0.3699	0.2399△	0.3509▽	0.2273▼

(b) Using UMLS++ concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3929**△	0.2285	0.4048	0.2981
M1 reweigh	0.3821	0.2300	0.4217	0.3045
PTT reweigh	0.3833△	0.2320	0.4215	0.3025
PMI reweigh	0.3835△	0.2341	0.4237	0.3051
STATTHES reweigh	0.3824△	**0.2345**	0.4226	0.3029
THES reweigh	0.3807△	0.2332	**0.4262**	**0.3070**

Table 8: Retrieval effectiveness after structuring.

(a) Using MeSH concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	**0.3889**	0.2796
word + KNN	**0.3868**△	0.2429∧	0.3736	**0.2916**
M1 struct	0.3733	0.2475△	0.3594	0.2306▽
PTT struct	0.3787	**0.2521**△	0.3546	0.2250▼
PMI struct	0.3762	0.2477△	0.3572	0.2288▽
THES struct	0.3739	0.2414△	0.3378▽	0.2267▽

(b) Using UMLS++ concepts

	2004	2005	2006	2007
baseline	0.3576	0.2219	0.3889	0.2796
word + KNN	**0.3929**△	0.2285	0.4048	**0.2981**
M1 struct	0.3790	0.2362△	0.4237	0.2964
PTT struct	0.3847	**0.2406**	0.4239	0.2976
PMI struct	0.3861	0.2311	0.4244	0.2949
STATTHES struct	0.3857	0.2358	0.4223	0.2930
THES struct	0.3770△	0.2371	**0.4266**	0.2972

8. ACKNOWLEDGMENTS

We thank Martijn Schuemie (Erasmus MC, Rotterdam) for annotating the document collections with Peregrine. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). This research was supported by the Netherlands Organization for Scientific Research (NWO, under project number 612-066-513).

References

[1] J. Bai and J.-Y. Nie. Adapting information retrieval to query contexts. *Inform. Process. Manag.*, 44(6):1901–1922, 2008.

[2] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *SIGIR '98*, pages 64–71, 1998.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR '99*, pages 222–229, 1999.

[4] G.-W. Bian and H.-H. Chen. Integrating query translation and document translation in a cross-language information retrieval system. In *Proceedings of AMTA '98*, pages 250–265, 1998.

[5] M. Boughanem, C. Chrisment, and N. Nassr. Investigation on disambiguation in clir: Aligned corpus and bi-directional translation-based strategies. In *CLEF '01*, pages 158–168, 2002.

[6] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June 1993.

[7] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16 (1):22–29, 1990.

[8] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the Second Text REtrieval Conference, TREC-2*, pages 243–252, 1993.

[9] J. Gao, J.-Y. Nie, and M. Zhou. Statistical query translation models for cross-language information retrieval. *ACM TALIP*, 5(4):323–359, 2006.

[10] W. Hersh, R. Bhupatiraju, L. Ross, A. Cohen, D. Kraemer, and P. Johnson. TREC 2004 genomics track overview. In *Proceedings of TREC 2004*, 2004.

[11] W. R. Hersh, D. H. Hickam, R. B. Haynes, and K. A. McKibbon. A performance and failure analysis of saphire with a medline test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60, 1994.

[12] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR '04*, pages 178–185, 2004.

[13] W. Kraaij. *Variations on Language Modeling for Information Retrieval*. PhD thesis, University of Twente, June 2004.

[14] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37 (6):512–526, 2004.

[15] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR '02*, pages 175–182, 2002.

[16] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, USA., 1999.

[17] N. Moreau. Best practices in language resources for multilingual information access. TrebleCLEF: Evaluation, Best Practices & Collaboration for Multilingual Information Access, Oct 2009.

[18] S.-H. Na, I.-S. Kang, and J.-H. Lee. Parsimonious translation models for information retrieval. *Inform. Process. Manag.*, 43(1):121–145, 2007.

[19] M. Schuemie, R. Jelier, and J. Kors. Peregrine: lightweight gene name normalization by dictionary lookup. In *Second BioCreative Workshop*, pages 131–133, Madrid, 2007.

[20] P. Srinivasan. Retrieval feedback in medline. *JAMIA*, 3(2): 157–167, 1996.

[21] N. Stokes, Y. Li, L. Cavedon, and J. Zobel. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17–50, 2009.

[22] D. Trieschnigg. *Proof of Concept: Concept-based Biomedical Information Retrieval*. PhD thesis, University of Twente, Enschede, 2010.

[23] D. Trieschnigg, W. Kraaij, and F. de Jong. The influence of basic tokenization on biomedical document retrieval. In *SIGIR '07*, pages 803–804, 2007.

[24] D. Trieschnigg, P. Pezik, V. Lee, W. Kraaij, F. de Jong, and D. Rebholz-Schuhmann. MeSH Up: Effective MeSH Text Classification and Improved Document Retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[25] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.