Iterative Translation Disambiguation for Cross-Language Information Retrieval*

ABSTRACT

Finding a proper distribution of translation probabilities is one of the most important factors impacting the effectiveness of a cross-language information retrieval system. In this paper we present a new approach that computes translation probabilities for a given query by using only a bilingual dictionary and a monolingual corpus in the target language. The algorithm combines term association measures with an iterative machine learning approach based on expectation maximization. Our approach considers only pairs of translation candidates and is therefore less sensitive to data-sparseness issues than approaches using higher n-grams. The learned translation probabilities are used as query term weights and integrated into a vector-space retrieval system. Results for English-German cross-lingual retrieval show substantial improvements over a baseline using dictionary lookup without term weighting.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Cross-Language Retrieval, Query Formulation

Keywords

Term Weighting, Term Co-Occurrence measures, Translation Disambiguation

1. INTRODUCTION

Web-accessible documents are becoming increasingly available in languages other than English. Users that are able to read documents in more than one language can benefit from cross-language retrieval, where the information need is expressed in a user's native language (the source language) and a ranked list of documents is returned in another language (the target language).

*(Produces the permission block, and copyright information). For use with SIG-ALTERNATE.CLS. Supported by ACM.

WOODSTOCK '97 El Paso, Texas USA

The two most straightforward techniques for cross-language retrieval are: (1) Automatically translate all documents in the collection into the source language and then apply monolingual retrieval on the translated document collection; and (2) Automatically translate the user-posed query into the target language and then apply monolingual retrieval with the translated query on the original document collection in the target language.

The second approach, query translation, is by far the most common cross-lingual retrieval approach. However, Chen and Gey [2] showed that translating the entire document collection outperforms query translation and also that a combination of query translation and document translation can lead to further improvements in retrieval effectiveness. Nevertheless, most approaches use only query translation because document translation is very time consuming and requires re-indexing of the entire collection each time the automatic translation system is modified to produce a new target collection.

Query translation requires access to some form of translation dictionary. Three approaches may be used to produce the translations:

1. Application of a machine translation system to translate the entire query into the target language.

2. Use of a dictionary to produce a number of target-language translations for words or phrases in the source language.

3. Use of a parallel corpus to estimate the probabilities that word w in the source language translates into word w' in the target language.

Each approach has advantages and disadvantages. The machine-translation approach is likely to produce the best translation results. On the other hand, this approach assumes that the input is a syntactically well-formed unit, i.e., a declarative sentence (such as *I am looking for documents that discuss the Cuba crisis and President Kennedy's role*), or a question (such as *What was President Kennedy's role in the Cuba crisis?*). Machine translation systems rely, to some degree, on the linguistic context to decide what the most likely translation should be. The problem is that most users don't pose their query as a well-formed sentence, but as a list of keywords (e.g., *President Kennedy, Cuba crisis*), disregarding any word order or other syntactic constraints.

The other shortcoming of many machine translation systems—especially commercial systems—is that they only return the most likely translation, where 'most likely' is defined in terms of the internal algorithm of the translation system. But this does not mean that there are no other equally good—or, by some other objective standard, maybe even better—alternative translations.

A simple alternative to a full-fledged machine translation is a dictionary lookup approach that uses a bilingual machine-readable

dictionary. Dictionary lookup will not provide a single best translation, but a set of possible translations for the query terms. The advantage of this approach is that it does not require syntactic well-formedness, i.e., it can translate a simple unordered list of keywords. The disadvantage is that the translations are normally not prioritized, as most dictionaries do not include information about how likely it is that a word w translates into w' versus w''. Even in cases where such information is available, these preferences are ranked, not quantitatively expressed, and they typically refer to global preferences disregarding any contextual information from the topic.

The third approach is to use a parallel corpus to estimate probabilities of translations between source- and target-language words. A parallel corpus is a collection of source-language documents coupled with a target-language (human) translation of each document. Sentences are typically aligned to indicate the correspondence between the original sentence and the translated sentence. This approach merges the advantages of the machine-translation framework with those of the dictionary-lookup approach in that the parallel corpus allows one to compute translation probabilities based on the frequency of co-occurrences between a source-language word and a target-language word in the parallel corpus. Most successful statistical machine-translation systems exploit parallel corpora (e.g., [18, 25]).

On the other hand, there are several drawbacks to the parallel-corpus approach. First, although parallel corpora are available for many of the European languages as well as for Arabic and Chinese, there are many languages for which there are still no parallel corpora large enough to estimate translation probabilities. Second, most of the parallel corpora belong to a rather specific domain, such as the Europarl corpus,[1] which contains the proceedings of the European parliament in 11 languages for the years 1996–2003. This introduces a bias toward the domain of the parallel corpus and makes the learned translation probabilities less reliable for other domains. A third disadvantage is that the translation probabilities induced from parallel corpora are typically based on single-word mappings, although recent template-based statistical methods facilitate the acquisition of phrase translations [18].

In this paper, we propose an approach that does not require a parallel corpus to induce translation probabilities. We use a more sophisticated approach to exploiting context to compute translation probabilities. Instead of using n-grams, we use co-occurrence statistics between all translation candidates for the sentence in question. In order to compute translation distributions for the foreign words in a sentence, our approach only requires two resources: a machine-readable dictionary (without any rankings or frequency statistics) and a monolingual corpus in the target language. The approach is novel in that uses an iterative expectation-maximization based algorithm for computing translation probabilities from these resources for any given information need. It also allows us to use a minimal cross-lingual component (a bilingual dictionary, not a bilingual corpus) which addresses the issue of scarce parallel resources for certain languages. In addition, the approach is much easier to adapt to new domains as it does not require a parallel corpus for the domain at hand, but only a monolingual domain-specific corpus, which is much easier to obtain.

This paper is organized as follows: The next section describes some of the problems that arise in determining correct translations. Section 3 describes our novel approach for computing translation probabilities. Section 4 provides the details of the experimental setting and the evaluation itself. Section 5 provides an overview of approaches that are related to our approach. In Section 6, we draw some conclusions and give an outlook on future research.

[1]The Europarl corpus is freely available from `http://www.isi.edu/~koehn/europarl/`.

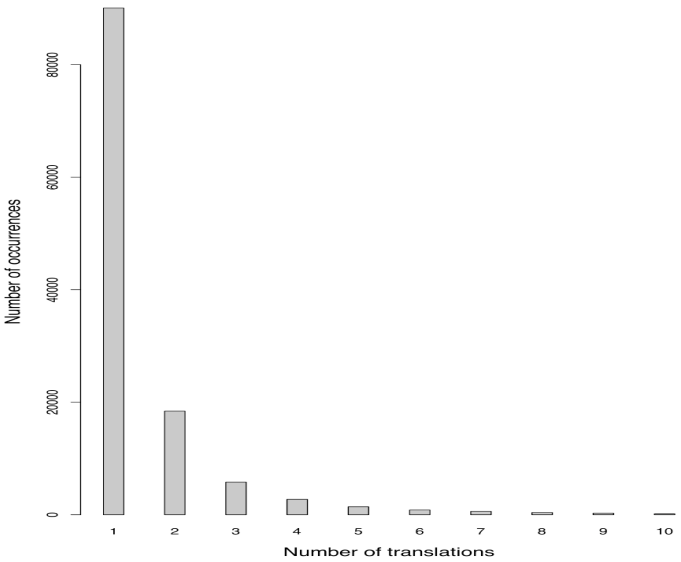

Figure 1: Frequency distribution of number of translation in the dictionary.

2. TRANSLATION SELECTION

Many words or phrases in one language can be translated into another language in a number of ways. For instance, the English word *penalty* can be translated as *Elfmeter* (as in soccer) or *Strafe* (as in punishment), and the choice of the translation depends on the context in which *penalty* occurs. Translation ambiguity is very common. Figure 1 shows the frequency distribution of dictionary entries with their corresponding number of translations in the DING dictionary, an English-German machine readable dictionary.[2]

One way to address the word-choice problem is to apply word-sense disambiguation on the source-language sentence and then use only those translation candidates that are associated with the appropriate word sense. Unfortunately, word-sense disambiguation is a non-trivial enterprise and for most languages the appropriate resources, e.g., ontologies like WordNet [9], do not exist. Also sense-annotated corpora that are used to train a word-sense disambiguation system are rare in foreign languages, and the process of building them is very laborious.

Our alternative approach to modeling context for the problem of word selection is to use co-occurrences between terms. For instance, the simultaneous occurrence of the terms w_1 and w_2 count as a co-occurrence if they appear within a certain window, where a window can be a particular number of words, a sentence, a paragraph, or a document. Co-occurrences are more flexible than linear n-grams as they do not put any constraints on adjacency or word order. Because of the higher degree of flexibility, it makes sense to base lexical selection on co-occurrence rather than on n-grams. For instance, given three source language terms s_1, s_2, and s_3, where s_1 can translate into $t_{1,1}, \ldots t_{1,3}$, s_2 can translate into $t_{2,1}$ and $t_{2,2}$, and s_3

[2]`http://www-user.tu-chemnitz.de/~fri/ding/`

can translate into $t_{3,1}$, one compares all possible triples and selects the pair of terms that co-occur most frequently as the most likely translation of s_1 and s_2:

$$freq(t_{1,1}, t_{2,1}, t_{3,1}) = n_1$$
$$freq(t_{1,2}, t_{2,1}, t_{3,1}) = n_2$$
$$freq(t_{1,3}, t_{2,1}, t_{3,1}) = n_3$$
$$freq(t_{1,1}, t_{2,2}, t_{3,1}) = n_4$$
$$freq(t_{1,2}, t_{2,2}, t_{3,1}) = n_5$$
$$freq(t_{1,3}, t_{2,2}, t_{3,1}) = n_6$$

Using term co-occurrence for selecting the target-language translations for two source-language terms is simple. Although the expansion of this approach to three, four, or more source terms may seem trivial, computing co-occurrence statistics for a larger number of terms induces a data-sparseness issue, similar to the situation for higher n-grams in language modeling. To overcome the data-sparseness problem we could use very large corpora for counting co-occurrence frequencies or we could apply smoothing techniques. Although large corpora for English are available (e.g., the Gigaword corpus[3]), it is doubtful that these provide frequency counts for a sufficient number of co-occurrences of four or more terms. An obvious alternative is to use Internet search engines to compute frequencies for larger sizes of co-occurring terms. Researchers have shown that the Internet can be used to address the problem of data-sparseness for bi-grams [13] but it is unclear to what extent this approach will resolve the issue of data-sparseness for higher n-gram models.

The other approach for tackling data sparseness is smoothing. Several smoothing techniques have been developed and many of them are successfully applied in language modeling [3]. The problem is that smoothing techniques are generally evaluated with respect to bi- or tri-gram models. It is unclear to what extent these techniques scale up successfully to models using a larger context such as four or five terms.

3. ITERATIVE DISAMBIGUATION

In the previous section noted that using co-occurrence frequencies for all possible translations is not only computationally expensive, but also suffers from the problem of data sparseness. In this section we propose an algorithm that overcomes this problem by using co-occurrences only between pairs of possible translations.

Assume there are three source terms in a source language sentence, s_1, s_2, and s_3, and each of these has a number of translations in the target language. For instance, s_1 can be translated as $t_{1,1}$, $t_{1,2}$, or $t_{1,3}$, s_2 can be translated as $t_{2,1}$, $t_{2,2}$, and s_3 can be translated as $t_{3,1}$. As mentioned above, one way to select the appropriate translations for s_1 and s_2 is to use the co-occurrence frequencies of all possible translations and then select the one with the highest co-occurrence count. In the case of two terms this might be feasible, but if the source sentence contains three or more terms, one runs into the problem of data-sparseness. That is, it is likely that most co-occurrence counts are zero, thus rendering such an approach useless for selecting a possible translation.

Instead of looking at the co-occurrences between the possible translations of all source terms at the same time, we propose to examine pairs of terms in order to gather partial evidence for the likelihood of a translation in a given context.

When looking at individual pairs only, disambiguation is done locally. Consider Figure 2, where each link between two translation candidates indicates that we consider the co-occurrence frequency of that pair of translation candidates. Here, the link strength between two translations is computed in terms of some co-occurrence based association measure. Note that there are no links between translation candidates for the same source term.

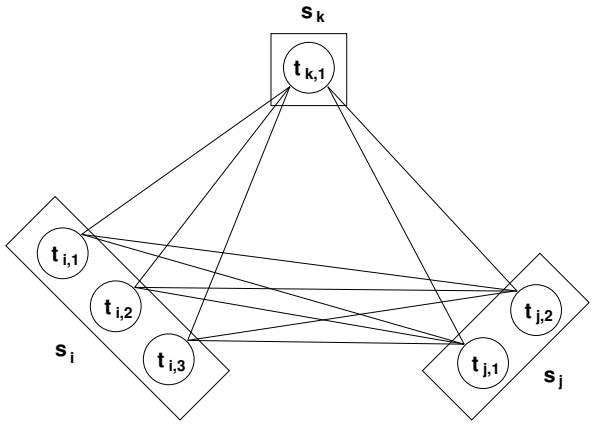

Figure 2: Co-occurrence network.

Assume that $t_{i,1}$ occurs more frequently with $t_{j,1}$ than any other pair of candidates between a translation for s_i and s_j. In this local context, $t_{i,1}$ would be the preferred translation for s_i, and $t_{j,1}$ for s_j. On the other hand, assume that $t_{i,1}$ and $t_{j,1}$ do not co-occur with $t_{k,1}$ at all, but $t_{i,2}$ and $t_{j,2}$ do. Given the co-occurrence information from other local contexts, the question is, which should be preferred: (1) $t_{i,1}$ and $t_{j,1}$; or (2) $t_{i,2}$ and $t_{j,2}$?

One solution is simply to add up all the co-occurrence frequencies and then, for each source word, take the translation candidate with the highest overall co-occurrence frequency. The disadvantage of this approach is that the probability that a target word is a translation for a certain source word is not taken into account when the link strength between two candidates is computed in the first place.

Our approach combines the link-strength computation with the prior probability of a translation given the other words in the query by computing them iteratively, in a fashion similar to the Expectation Maximization (EM) algorithm [6].

Initially, all possible translations of a source term are considered equally likely, where we associate with each translation candidate a confidence weight $w_T(\cdot|\cdot)$ that it is indeed the appropriate translation. That is, $w_T(t_{i,1}|s_i) = \ldots = w_T(t_{i,n}|s_i) = 1/n$, for all n possible translations of s_i. For all terms in the source sentence, the sets of translations are initialized this way. Next, each translation candidate $t_{i,l}$ is linked to each translation candidate $t_{j,m}$ where $i \neq j$, i.e., different translations of the same source term are not linked to each other. This situation is depicted in Figure 2 for a source sentence with three terms, where s_i has three translations, s_j has two, and s_k has one.

In addition to the translation weights, the links between translation candidates are also weighted. The link weight $w_L(t_{i,l}, t_{j,m})$ between two translations is computed by some measure of association strength based on the co-occurrence frequency of $t_{i,l}$ and $t_{j,m}$. Below we provide the definitions of some association measures.

The first step towards recomputing the term weights is to initialize the weights. Since we use a bilingual dictionary that does not contain any information about the translation probability of a certain source word or phrase, the weights are initialized making the weakest possible assumption, i.e., using a uniform distribution.

[3]Gigaword is distributed by the Linguistic data Consortium: `http://ldc.upenn.edu`.

Assuming that t is a translation candidate for s_i, i.e. $t \in tr(s_i)$, initialization is defined as:

Initialization Step:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|} \quad (1)$$

After all translation candidates have been initialized appropriately, each term weight is recomputed based on two different inputs: the weights of the terms that link to the term; and the respective link weight. This formulated as follows:

Iteration Step:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} w_L(t,t') \cdot w_T(t'|s_i) \quad (2)$$

where $inlink(t)$ is the set of translation candidates that are linked to t. For instance, in Figure 2, $inlink(t_{i,2}) = \{t_{j,1}, t_{j,2}, t_{k,1}\}$.

After each term weight has be re-computed, term weights are normalized so that all weights associated with translation candidates of the same source word sum up to 1:

Normalization Step:

$$w_L^n(t|s_i) = \frac{w_L^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w_L^n(t_{i,m}|s_i)} \quad (3)$$

Following this, steps (2) and (3) are repeated. The iteration stops if the changes in term weights become smaller than some predefined threshold θ. More formally, let w_T^n be the vector of all term weights in the network for iteration n and let $|V|_1$ be the L^1 norm of a vector V:

$$|V|_1 = \sum_k |V_k| \quad (4)$$

where V_k is the kth element in the vector, and $|V_k|$ is the absolute value of V_k. Then, the iteration stops if $|w_T^n - w_T^{n-1}|_1 < \theta$.

Note that the algorithm described above can also be considered a modification of the PageRank algorithm [19], allowing for nodes in the network to be clustered.

There are a number of ways to compute the association strength between two terms. We focus here on three alternatives: *Pointwise mutual information*, *Dice coefficient*, and *Log Likelihood Ratio*. The point-wise mutual information between to terms t and t' is defined as follows [4]:

$$MI(t,t') = \log_2 \frac{p(t,t')}{p(t) \cdot p(t')} \quad (5)$$

where $p(t,t')$ is the probability that the terms t and t' occur in the same document. Thus, as this value gets larger, the joint probability of t and t' occurring together is increasingly larger than the combined probability of them occurring together individually. The probabilities in (5) are estimated by counting the number of (co-)occurrences and dividing this result by the number of text windows that were used in the whole corpus.

Measuring association strength in terms of mutual information has some shortcomings. For instance, given two pairs of terms (t_1, t_1') and (t_2, t_2'), where both pairs of terms always co-occur with each other, but the pair (t_1, t_1') is less frequent than the pair (t_2, t_2'), the pair (t_1, t_1') will have a higher mutual information value than the pair (t_2, t_2'). Although this is certainly counter-intuitive, mutual information is widely used for measuring association strength (see, e.g., [16]), and we therefore include it as one possibility for computing the link weight.

Source: English	Target: German	Iteration		
		0	1	2
europe	europa	1.0000	1.0000	1.0000
trade	branche	0.0833	0.2072	0.2040
trade	handel	0.0833	0.1312	0.1376
trade	gewerbe	0.0833	0.1059	0.1018
trade	geschaeft	0.0833	0.1001	0.1016
trade	abschluss	0.0833	0.0930	0.0998
trade	handeln	0.0833	0.0774	0.0760
trade	beruf	0.0833	0.0624	0.0572
trade	trade	0.0833	0.0476	0.0478
trade	handwerk	0.0833	0.0474	0.0466
trade	eintauschen	0.0833	0.0430	0.0430
trade	handel treiben	0.0833	0.0425	0.0424
trade	schachern mit etw.	0.0833	0.0416	0.0416
union	union	0.2000	0.4678	0.4554
union	gewerkschaft	0.2000	0.1748	0.1893
union	vereinigung	0.2000	0.1271	0.1264
union	verbindung	0.2000	0.1220	0.1214
union	verein	0.2000	0.1081	0.1072
trade union	gewerkschaftlich	0.5000	0.5415	0.4664
trade union	gewerkschaft	0.5000	0.4584	0.5335

Table 1: The first two iterations of the term re-weighting algorithm.

Alternatively, the link weight can be computed using the Dice coefficient, which is defined as follows:

$$DC(t,t') = \frac{2 \cdot freq(t,t')}{freq(t) + freq(t')} \quad (6)$$

where $freq(t,t')$ is the number of times t and t' co-occur. One advantage of the Dice coefficient is that its value ranges between 0 and 1 (where 1 is perfect co-occurrence), whereas mutual information has no upper bound.

The last measure of association strength we consider here is the Log Likelihood ratio, which compares two hypotheses:

$$H_1\text{:}\ \ p(w^2|w^1) = p = p(w^2|\neg p^1)$$

$$H_2\text{:}\ \ p(w^2|w^1) = p_1 \neq p_2 = p(w^2|\neg p^1)$$

Hypothesis H_1 states that the probability of both w^2 and w^1 occurring together is the same as the probability that w^2 occurs without w^1. In other words, H_1 formalizes independence between w^2 and w^1. H_2 states that the two probabilities are not the same and hence w^2 and w^1 do not occur independent of each other.

The log likelihood is then defined as [7]:

$$\begin{aligned} \log\lambda &= \log\frac{L(H_1)}{L(H_2)} \\ &= \log L(c_{1,2}, c_1, p) + \log L(c_2 - c_{1,2}, p) \\ &\quad -\log L(c_{1,2}, c_1, p_1) - \log L(c_2 - c_{1,2}, N - c_1, p_2) \end{aligned} \quad (7)$$

where p, p_1, and p_2 are defined as in H_1 and H_2 above, c_1 is the frequency or word w_1, c_2 is the frequency of word w^2, $c_{1,2}$ is the frequency of both words occurring together, N is the number of tokens in the corpus, and $L(k,n,x) = x^k(1-x)^{n-k}$

Computing the translation probabilities iteratively can help resolve some of the translation ambiguities properly. Consider Table 1. The source topic is *Trade Unions in Europe*. During the first iteration, *gewerkschaftlich* is the preferred translation for *Trade Union*, but during the second iteration, *gewerkschaft*, which is the proper translation, has a higher translation probability.

Documents source	Years	Size	No. documents
Frankfurter Rundschau	1994	320 MB	139,715
Der Spiegel	1994–1995	63 MB	13,979
SDA German	1994	144 MB	71,677
SDA German	1995	141 MB	69,438
Total:		668 MB	294,809

Table 2: Documents used for the English-German bilingual task at CLEF 2003.

```
<num> C141 </num>
<EN-title> Letter Bomb for Kiesbauer
<DE-title> Briefbombe für Kiesbauer
<num> C142 </num>
<EN-title> Christo wraps German Reichstag
<DE-title> Christo verhüllt den Deutschen Reichstag
<num> C164 </num>
<EN-title> European Drug Sentences
<DE-title> Europäische Strafurteile zu Drogen
<num> C182 </num>
<EN-title> 50th Anniversary of Normandy Landings
<DE-title> 50.  Jahrestag der Landung in der Normandie
<num> C195 </num>
<EN-title> Strikes by Italian Flight Assistants
<DE-title> Streik italienischer Flugbegleiter
```

Table 3: Example topics (title field only) used for the English-German bilingual task at CLEF 2003.

4. EXPERIMENT

This section describes our evaluation of the term re-weighting algorithm described above. We present the set-up of our experiment (test data, morphological processing, retrieval model, statistical significance tests) and our experimental results.

4.1 Experimental Set-Up

Below we describe our experimental set-up, specifically: the test data used; the morphological processing required for the original source-language topics and the target-language queries; the model underlying our retrieval system (a standard vector space model); and our choice of statistical significance tests.

4.1.1 Test Data

The document collection used in our experiments consists of the CLEF 2003 English to German bilingual data. Specific documents are listed in Table 2.

The document collection contains 60 topics, four of which were removed by the CLEF organizers, as no relevant documents were found in the collection, leaving us with 56 topics. Each topic has a title, a description, and a narrative field. For our experiments, we used only the title field to formulate the queries. Although it is common practice in CLEF and TREC to use both the title and description to formulate the queries, title-only queries are more realistic, as most queries posed by actual users tend to be short (i.e., two to four terms). Table 3 shows some of the topics that were used, in combination with the corresponding topic in the target language (as formulated by one of the CLEF assessors).

4.1.2 Morphological Normalization

Morphological normalization is required for the original source-language topics as well as for the translated (target-language) queries and documents.

```
Source language title field:
<EN-title> Wimbledon Lady Winners
Morphologically normalized source title:
Wimbledon lady winner
Weighted query:
w(1,wimbledon), w(1,dame), w(0.548,sieger),
w(0.452,gewinner)
Splitted query:
w(1,wimbledon) w(1,wimbl) w(1,imble) w(1,mbled)
w(1,bledo) w(1,ledon) w(1,dame) w(0.548,sieger)
w(0.548,siege) w(0.548,ieger) w(0.452,gewinner)
w(0.452,gewin) w(0.452,ewinn) w(0.452,winne)
w(0.452,inner)
```

Figure 3: Intermediate results of the query formulation process.

First, source-language words (English, in our case) from the original topic are normalized to match entries in the bilingual dictionary. The words in the topic may bear morphological inflection such as tense information for verbs and plural information for nouns. Since the dictionary only contains base forms, the words in the topics must be mapped to their respective base forms as well. Here, we used TreeTagger [23], which is a part-of-speech tagger that also provides the lemma (or base form) for each word. This form of morphological normalization is less aggressive than a rule-based stemmer, such as Porter's stemmer [21].

Since the target language is German—a morphologically more complex language than English—additional normalization steps are required. The translation candidates need not be mapped to their base forms because they are taken from the bilingual dictionary which contains only the base forms. On the other hand, compounds are very frequent in German and it has been shown that de-compounding can improve retrieval effectiveness substantially [17].

Instead of de-compounding, we use character n-grams, an approach that yields almost the same retrieval performance as de-compounding. Specifically, it has been shown that using 5-grams leads to the best performance [11]. Thus, we split all tokens in documents and translated queries into 5-grams, without crossing word boundaries, for all mono-lingual and cross-lingual runs.

For the runs involving term weights, we must decide how to assign weights to 5-gram substrings. In our experiments, we simply gave the substrings the same weight as the original term. In addition to *n*-gram splitting, all words in the target language (including the translations) were mapped to lower case. Figure 3 shows the intermediate results of the query formulation process for one of the topics in the CLEF 2003 test set.

4.1.3 Retrieval Model

The model underlying our retrieval system is the standard vector space model. All our mono- and bi-lingual runs were based on the Lnu.ltc weighting scheme [1]. That is, to compute the similarity between between a query (q) and a document (d):

$$sim(q,d) = \sum_{i \in q \cap d} \frac{\frac{1+\log(\text{freq}_{i,d})}{1+\log(\text{avg}_{j \in d}\text{freq}_{j,d})} \cdot \frac{\text{freq}_{i,q}}{\max_{j \in q}\text{freq}_{j,q}} \cdot \log\left(\frac{N}{n_i}\right)}{((1-sl)\cdot pv + sl \cdot \text{uw}_d) \cdot \sqrt{\sum_{i \in q}\left(\frac{\text{freq}_{i,q}}{\max_{j \in q}\text{freq}_{j,q}} \cdot \log\left(\frac{N}{n_i}\right)\right)^2}} \quad (8)$$

In our experiments, we used a slope (sl) of 0.1. The pivot (pv) was set to the average number of unique words per document. The

parameter uw_d refers to the number of unique words in document d.

Neither the mono-lingual run nor the dictionary baseline include term weights—both used the document similarity measure in (8).

In order to integrate term weights, the formula in (8) was modified. Weighted document similarity is defined in (9):

$$sim_w(q,d) = \sum_{i \in q \cap d} \frac{w(i) \cdot \frac{1+\log(\text{freq}_{i,d})}{1+\log(\text{avg}_{j \in d}\text{freq}_{j,d})} \cdot \frac{\text{freq}_{i,q}}{\max_{j \in q}\text{freq}_{j,q}} \cdot \log\left(\frac{N}{n_i}\right)}{((1-sl)\cdot pv + sl \cdot \text{uw}_d) \cdot \sqrt{\sum_{i \in q}\left(w(i) \cdot \frac{\text{freq}_{i,q}}{\max_{j \in q}\text{freq}_{j,q}} \cdot \log\left(\frac{N}{n_i}\right)\right)^2}} \quad (9)$$

where the weight of term i is computed as described in Section 3.

4.1.4 Statistical Significance

There are many techniques for drawing statistical inferences. The paired t-test is probably the best-known technique (see, e.g., [14]). Many of the inference techniques make certain assumptions about the data to which they are applied. The most common assumption, which also underlies the paired t-test, is that the data is taken from a population which is normally distributed. In the setting of retrieval this means that for a number of queries, the differences between two methods are normally distributed. Whether this assumption holds for text retrieval has been the subject of debate in retrieval evaluation [24].

To determine whether the observed differences between two retrieval approaches are statistically significant and not just caused by chance, we used the bootstrap method, a powerful non-parametric inference test [8]. The method was previously applied to retrieval evaluation [22, 26]. The basic idea of the bootstrap is to simulate the underlying distribution by randomly drawing (with replacement) a large number of samples of size N from the original sample of N observations. These new samples are called *bootstrap samples*; we set the number of bootstrap samples to 2,000 as using the standard size of 1,000 has been shown to be a less reliable approach to inducing a normal distribution [5].

The mean and the standard error of the bootstrap samples allow computation of a confidence interval for different levels of confidence (typically 0.95 and higher). We compare two retrieval methods a and b by one-tailed significance testing. If the left limit of the confidence interval is greater than zero, we reject the null hypothesis, stating that method b is not better than a, and conclude that the improvement of b over a is statistically significant, for a given confidence level. Analogously, if the right limit of the confidence interval is less than zero, we conclude that method b performs significantly worse than a.

In the results reported below, we indicate improvements at a confidence level of 90% with "$^{\triangle}$" and at a confidence level of 95% with "$^{\blacktriangle}$". Analogously, decreases in performance at a confidence level of 90% are marked with "$^{\triangledown}$" and at a confidence level of 95% with "$^{\blacktriangledown}$". No mark-up is used if neither an increase nor a decrease in performance is significant at either of the 90% or 95% confidence levels.

4.2 Experimental Results

For evaluating the effectiveness of our term re-weighting algorithm, we compared five runs against each other. First, we determined the performance ceiling by using manually translated topics provided by the CLEF organizers. As a baseline, we used the English-German bilingual dictionary without any weights assigned to the translation, i.e. all translation candidates were considered equally likely. All cross-language runs, i.e. all runs except the mono-lingual run, used the DING English-German dictionary.

Run	MAP	Rel. impr.	Perc. Mono-ling.
Mono-lingual	0.3171	–	–
Unweighted baseline	0.1708	–	53.9%
Mutual Inf. weighted	0.1972	+15.5%	62.2%
Dice weighted	0.1994	+16.7%	62.9%
Log-likel. weighted	0.2013	+17.6%$^{\triangle}$	63.5%

Table 4: Experimental results for the different association measures.

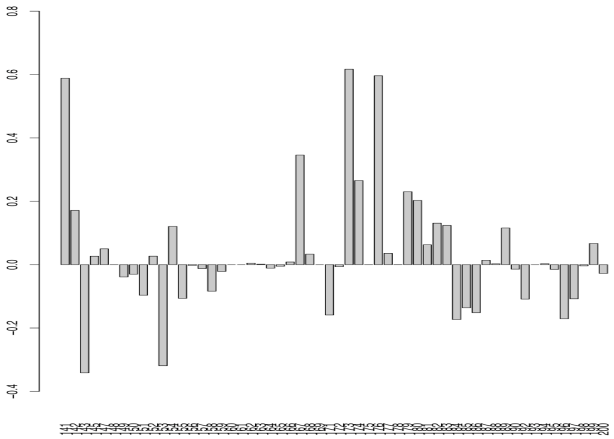

Figure 4: Absolute differences in average precision between the baseline and Log Likelihood weighted run for the individual queries.

As described in Section 3, we used three different association measures for our term re-weighting algorithm: Mutual Information, Dice Coefficient, and Log Likelihood Ratio. Table 4 lists the mean average precision (MAP), relative improvements over the baseline, and percentage of the German mono-lingual effectiveness for all five runs. The frequencies of term occurrences and co-occurrences were computed on the German corpus of the CLEF 2003 document collection.

The results in Table 4 show that retrieval using term re-weighting outperforms the baseline. On the other hand, the improvement was weakly statistically significant for the run using log likelihood ratio as the association measure. Figure 4 shows the absolute differences in average precision between Log Likelihood re-weighting and the baseline for individual topics.

Despite that the mean average precision of Log Likelihood re-weighting is substantially higher than the baseline, individual average precision decreases for a number of queries. One explanation for some of the decreases is the treatment of unknown words during translation. Unknown words are treated as if they were proper names, and therefore the original word from the source language is included in the target language query. Although this fall-back strategy works well in many cases, there are also cases where it harms retrieval performance. For instance, in the English source topic in (10)

```
Women's Conference Beijing                    (10)
```

the word *Women* is falsely considered a proper noun by the part-of-speech tagger and, therefore, not mapped to its base form *woman*. The plural form *women* is not in the bilingual dictionary, although the singular form *woman* is. Because this word is assumed to be a proper noun during query translation, the original source word *women* is used as translation.

Although faulty translations of this type affect both the baseline system and the run using term weights, the latter is affected more severely. The reason is that the conditional translation weight is 1 (i.e., the source-word *women* only has one translation candidate) which induces a weight of 1 for *women*, whereas all the other terms in the query receive lower weights (the other source words had more than one translation candidate).[4] As a consequence the term *woman* dominates the document similarity, and most top-ranked documents contain *women* as the only matching term.

To deal with out-of-vocabulary terms properly, additional machinery (outside of the scope of this paper) is necessary since they strongly bias retrieval results.

5. RELATED WORK

Cross-language retrieval has a long history in the broader field of information retrieval. In addition, techniques for word-sense disambiguation and phrase-based machine translation have recently evolved to the point where they are ripe for investigation in the context of cross-language retrieval, although others have not yet used these to the extent that we have.

For example, Pirkola's [20] approach does not consider disambiguation during query formulation at all. Pirkola uses structured queries to cluster together all translations of a word or phrase in the source topic. Disambiguation takes place implicitly during retrieval. The underlying assumption is that top-ranked documents contain at least one translation for the majority of the clusters and, since all the translations occur in the same document, they are likely to be appropriate translations. Although this assumption makes sense and is appealing in its simplicity, Pirkola's method is sensitive to skewed translations for retrieval systems using inverted document frequencies for term weighting. Given some source word (or phrase) w_i, if one translation $w_{i,j}$ has a very high inverted document frequency score, it can bias document similarity toward this translation, and therefore cause the top-ranked documents to contain only few translations of other source words from the topic. This bias reduces the effect that co-occurrence with translations of other source words has on selecting an appropriate translation.

The work by Jang et al. [12] is closely related to ours in that they also use a word-association measure, mutual information in their case, to re-compute translation probabilities for cross-language retrieval. Their approach differs from ours in two ways: First, their system only considers mutual information between consecutive terms in the query. For example, for a query of the form w_1, w_2, w_3, they only consider mutual information scores of the form $MI(w_{1,i}, w_{2,j})$ and $MI(w_{2,k}, w_{3,l})$, where $w_{n,i}$ is a translation candidate of w_n. The sensitivity to word order seems somewhat problematic in the context of keyword-based query formulation which should allow for free-word order. Second, they do not compute the translation probabilities in an iterative fashion. Thus, their approach does not benefit from the power of multiple iterations, as in our approach, where disambiguated information from a previous iteration induces more accurate decisions in the current iteration.

Gao et al. [10] use a decaying mutual-information score in combination with syntactic dependency relations. The decay factor is based on the average distance between two words in the target language. In our model, we did not consider distances between words, but simply counted the number of times two words occur in the same document. Integrating a distance factor might be beneficial to our approach. The dependency model used in [10] requires the topics to bear some form of syntactic structure, e.g., verb-argument or noun-modifier relations. Unfortunately, simple keyword-based topics (such as the title field in the CLEF topics) are not in this form, but are typically just simple lists of noun phrases. For that reason we did not try to carry out any deeper linguistic analysis between the words in the topic.

Maeda et al. [15] compare a number of co-occurrence statistics with respect to their usefulness for improving retrieval effectiveness. As in our own approach, they consider all pairs of possible translations of words in the query—not just co-occurrences of consecutive words. On the other hand, Maeda et al. use co-occurrence information to select translations of words from the topic for query formulation, instead of re-weighting them. If the association strength between two translation candidates, measured by co-occurrence statistics (e.g., mutual information) is greater than some pre-defined threshold, both translation candidates are included in the query; otherwise they are excluded. This approach has two potential shortcomings: First, it requires a proper estimation of the threshold on some development data set of topics. Second, it does not result in a probability distribution over the possible translations of a word in the source topic. By contrast, our approach allows for a more fine-grained estimation of the usefulness of a particular translation in the context of the given topic.

6. CONCLUSIONS

This paper has introduced a new algorithm for computing topic-dependent translation probabilities for cross-language information retrieval. These translation probabilities were integrated as term weights into a vector-space retrieval system.

Experimental results for English to German cross-language retrieval showed that our approach improves retrieval effectiveness significantly compared to a baseline using bilingual dictionary lookup. For estimating translation probabilities we experimented with different term association measures: Mutual Information, Dice Coefficient, and Log Likelihood Ratio. The experimental results show that Log Likelihood Ratio has the strongest positive impact on retrieval effectiveness, although the differences in performance between the three measures are relatively small. An important advantage of our approach is that it only requires a bilingual dictionary and a monolingual corpus in the target language to compute the translation probability distributions for a given topic.

An issue that remains open at this point is the computation of query terms that are not covered by the bilingual dictionary. In our approach, we have set the query term weight to be the conditional translation probability, which causes translations of unknown words to bias the document similarity computation. We plan to investigate ways of addressing this problem.

7. REFERENCES

[1] C. Buckley, A. Singhal, and M. Mitra. New retrieval approaches using SMART: TREC 4. In D. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 25–48. NIST Special Publication 500-236, 1995.

[2] A. Chen and F. C. Gey. Combining query translation and document translation in cross-language retrieval. In *Proceedings of the 4th Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, 2003.

[4] Moreover, *Beijing* also had two translations, viz. *Peking* and *Beijing*.

[3] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th conference on Association for Computational Linguistics*, pages 310–318, 1996.

[4] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[5] A. Davison and D. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[7] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[8] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[9] C. Fellbaum, editor. *WordNet: An Electronical Lexical Database*. MIT Press, 1998.

[10] J. Gao, J.-Y. Nie, H. He, W. Chen, and M. Zhou. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependency relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 183–190, 2002.

[11] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Information Retrieval*, 7:33–52, 2004.

[12] M.-G. Jang, S. H. Myaeng, and S. Y. Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pages 223–229, 1999.

[13] F. Keller and M. Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

[14] L. Kitchens. *Exploring Statistics: A Modern Introduction to Data Analysis and Inference*. Brooks/Cole Publishing Company, 2nd edition, 1998.

[15] A. Maeda, F. Sadat, M. Yoshikawa, and S. Uemura. Query term disambiguation for web cross-language information retrieval using a search engine. In *IRAL '00: Proceedings of the 5th International Workshop on on Information Retrieval with Asian Languages*, pages 25–32. ACM Press, 2000.

[16] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. M.I.T. Press, 1999.

[17] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, LNCS 2406, pages 262–277. Springer Verlag, 2002.

[18] F.-J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguitics*, 30(4):417–449, 2004.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library, 1999.

[20] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language retrieval. In B. Crof, A. Moffat, C. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, 1998.

[21] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[22] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33(4):495–512, 1997.

[23] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[24] C. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

[25] A. Venugopal, S. Vogel, and A. Waibel. Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 319–326, 2003.

[26] J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20(4):270–284, 1994.