Transitive probabilistic CLIR models

Wessel Kraaij
TNO TPD
Stieltjesweg 1
Delft
kraaijw@acm.org

Franciska de Jong
Universiteit Twente, dept. of Computer Science
P.O. Box 217
Enschede
fdejong@ewi.utwente.nl

Abstract

Transitive translation could be a useful technique to enlarge the number of supported language pairs for a cross-language information retrieval (CLIR) system in a cost-effective manner. The paper describes several setups for transitive translation based on probabilistic translation models. The transitive CLIR models were evaluated on the CLEF test collection and yielded a retrieval effectiveness up to 83% of monolingual performance, which is significantly better than a baseline using the synonym operator.

1 Introduction

Combination approaches have been well studied in Cross-Language Information Retrieval (CLIR). Especially the combination of different lexical translation resources to remedy gaps in the resources for certain translation pairs has been shown to be quite effective and supports the generally adopted view that lexical coverage is one of the most determining factors for CLIR effectiveness. In this paper we propose a different way to combine translation resources, namely by stacking translation resources in order to bridge languages for which no direct translation resources are available. Such a cascaded configuration of translation resources is usually referred to as *transitive* or *pivoted* translation. Earlier work on transitive translation is based on a combination of machine readable dictionaries with Dutch as a pivot language for bilingual retrieval experiments [Kraaij & Hiemstra, 1998], a combination of statistical MT steps [Franz et al., 1999], or a transitive dictionary-based method using structured queries [Ballesteros, 2000, Lehtokangas & Airio, 2002]. A common characteristic of all these approaches is that translation and retrieval are seen as a separate step, whereas we propose a more tighter integrated (probabilistic) model. In this paper, we will address two research questions:

1. How do the proposed configurations for transitive probabilistic CLIR perform in comparison with transitive probabilistic CLIR based on structured queries using the synonym operator?

2. Are there interactions between the CLIR models and the particular translation resources used. We will compare translation resources built from parallel Web corpora with machine readable dictionaries.

Our aim is to design a low-cost CLIR system based on freely available resources covering as many languages as possible. These requirements affect not only the selection of translation lexicons, but also the system architecture. When the requirement to cover as many languages as possible is really taken seriously, the only feasible architecture is one based on an interlingua or pivot language. A system based on direct translation between N languages would require translation resources for $N(N-1)$ directional language pairs, whereas a system based on an interlingua has linear scalability: $2(N-1)$ directional language pairs. Since currently English is the language for which the most translation resources are available, either as source or target language, it is the obvious candidate to fill the role of interlingua. Of course there are other languages that could fulfil an interlingua role, e.g. French, Spanish, Russian. When resources for multiple pivot languages are available they can be combined in order to enhance lexical coverage [Gollins & Sanderson, 2001, Ballesteros & Sanderson, 2003].

As high quality translation dictionaries for English are not always easily available, we propose to build statistical translation dictionaries based on parallel corpora mined from the Web. In previous work, we have shown that this approach can lead to very effective CLIR performance [Kraaij & Spitters, 2003].

The paper is structured as follows: In Section 2 we describe two alternative CLIR models based on statistical word-by-word translation. In Section 3 we sketch the procedure to construct the required translation resources. In Section 4 we present the three different models for pivoted CLIR, and the evaluation of the models, based on the methods and test collections made available via TREC[1] and CLEF[2], is discussed in Section 5. The paper ends with some preliminary conclusions.

2 Probabilistic CLIR Models

CLIR architectures can conceptually be distinguished on the basis of the elements that are submitted to translation: the queries, the document collection, or both. In [Kraaij et al., 2003] we compared several models for CLIR. We showed that probabilistic CLIR models based on mapping a language model of the query onto a language model of the document (a form of query translation), or mapping a language model of the document onto a query language model (a form of document translation), significantly outperformed CLIR models that do not use translation probabilities. An example of the latter approach is the popular structured query model that is based on INQUERY's *synonym* operator [Pirkola, 1998, Ballesteros & Croft, 1998]. These performance figures gave rise to two follow-ups research actions: (i) a comparison between different configurations of generative transitive CLIR models, and (ii) and exploration of the possible interactions between CLIR models and translation resources. Therefore we will take probabilistic models as a starting point for our experiments with transitive translation. In addition to performance improvement, an advantage of these models is that translation is not seen as a separate process, but is tightly integrated into the retrieval model. As a consequence our models can be argued to cover not just the IR part but the entire CLIR-task, and thereby contribute to the theoretical framework for CLIR.

There are strong arguments to integrate translation more tightly into the retrieval model rather than just combining MT and monolingual IR. The goal of CLIR is to find relevant documents across a language barrier. Instead of using one translation which is optimized for readability and well-formedness, we propose to improve recall by using as many alternative correct translations. A successful CLIR system should be able to outperform a monolingual system, since the multiple translations can be used for query expansion. However, simply substituting each query term by all its translations will not work, because then words with many translations will dominate the result [Kraaij et al., 2003]. Several methods have been proposed to overcome this problem [Pirkola, 1998, Hiemstra & de Jong, 1999], but as said, we follow the tradition of IR models based on language modelling [Ponte & Croft, 1998, Hiemstra, 2001] and the combination with statistical translation [Berger & Lafferty, 1999].

Our preferred formulation of an document ranking function based on language models (LM) is the *cross-entropy reduction* (CER). This measure is defined as the difference between two cross-entropies: the cross-entropy as measured between the query model and the document model and the cross-entropy

[1] TREC: Text REtrieval Conference. Cf. http://trec.nist.gov/

[2] CLEF: Cross-Language Evaluation Forum. Cf. http://clef.iei.pi.cnr.it/

as measured between the query model and a background language model: [Kraaij, 2004]:

$$CER(Q; C, D) = H(Q, C) - H(Q, D) = \sum_{i=1}^{n} P(\tau_i|Q) \log \frac{P(\tau_i|D_k)}{P(\tau_i|C)} \tag{1}$$

where $P(\tau_i|Q)$ is the unigram language model estimated for the query (representing the user's view of relevant documents), $P(\tau_i|D_k)$ is the language model representing the document and $P(\tau_i|C)$ models the background language.

Cross-entropy is a measure of average surprise; so the better a document model 'fits' a query distribution, the lower its cross-entropy will be. If the document model fits the query better than the background model, this will result in a positive ranking score $CER(Q; C, D)$, the document model with the best fit (the lowest cross-entropy) will yield the highest score. The cross-entropy between the query model and the background model acts as a normalizing constant (depending on the query), which is important when document ranking score distributions have to be compatible across queries [Kraaij & Spitters, 2003]. All models are estimated by maximum likelihood procedures, but since the document models are sparse and the query (model) contains several non-content terms, we apply linear interpolation with the background model as a smoothing function:

$$P(\tau_i|D_k) = (1 - \lambda)P_{ml}(\tau_i|D_k) + \lambda P_{ml}(\tau_i|C) \tag{2}$$

where P_{ml} is estimated by relative frequencies. λ is taken as a constant (0.3), optimized on a previous test collection. The choice of λ has proved to be not very critical. Of course, there is very little data to accurately estimate a language model of the user's information need. The query terms are the only information. However, techniques exist to use the document collection in order to expand this model [Lavrenko & Croft, 2003].

In the following subsections, we will describe two ways to extend this monolingual IR model with translation: before measuring the cross-entropy one can either map the query language model onto the document language, or one can map the document language model onto the query language. In terminology inspired by the MT domain, query words will also be referred to as source language elements (s_j), while document language will also be referred to as target language (t_i). For reasons of clarity, we will use s and t subscripts for Q, D and C in order to show whether they are stated in the source of target language.

2.1 Estimating the query model in the target language

Instead of translating a query before estimating a query model (e.g. by using an MT system), we propose to directly estimate the query model in the document language. This form of query translation can be described also as: estimating the source language model in the target language. This can be achieved by decomposing the problem into two components that are easier to estimate:

$$P(t_i|Q_s) = \sum_{j}^{S} P(s_j, t_i|Q_s) = \sum_{j}^{S} P(t_i|s_j, Q_s)P(s_j|Q_s) \approx \sum_{j}^{S} P(t_i|s_j)P(s_j|Q_s) \tag{3}$$

where S is the size of the source language vocabulary. Thus, $P(t_i|Q_s)$ can be approximated by combining the translation model $P(t_i|s_j)$, which we can estimate e.g. on a parallel corpus, and the familiar $P(s_j|Q_s)$, which can be estimated using relative frequencies.

This simplified model, from which we have dropped the dependency of $P(t_i|s_j)$ on Q, can be interpreted as a way of mapping the probability distribution function in the source language event space $P(s_j|Q_s)$ (which can easily be estimated using maximum likelihood procedure) onto the event space of the target language vocabulary. Since this probabilistic mapping function involves a summation over all possible translations, mapping the source language model can be implemented as the matrix product of a vector representing the query probability distribution over source language terms with the translation matrix $P(t_i|s_j)$. The result is a probability distribution function over the target language vocabulary.

Now we can substitute the query model $P(\tau_i|Q)$ in the smoothed version of formula (1) by the target language query model in (3), yielding CLIR-model QT (Query Translation):

$$\text{QT:} \quad CER(Q_s; C_t, D_t) = \sum_{i=1}^{n} \sum_{j=1}^{S} P(t_i|s_j) P(s_j|Q_s) \log \frac{(1-\lambda)P(t_i|D_t) + \lambda P(t_i|C_t)}{P(t_i|C_t)} \tag{4}$$

2.2 Estimating the document model in the source language

Another way to embed translation into the IR model is to estimate the document model in the query language, or in other terms: to estimate the target language model in the source language:

$$P(s_i|D_t) = \sum_{j}^{T} P(s_i, t_j|D_t) = \sum_{j}^{T} P(s_i|t_j, D_t) P(t_j|D_t) \approx \sum_{j}^{T} P(s_i|t_j) P(t_j|D_t) \tag{5}$$

where T is the size of the target language vocabulary. Obviously, we need a translation model in the reverse direction for this approach. Now we can substitute (5) for $P(\tau_i|D)$ in the smoothed version of formula (1). After a similar substitution operation for $P(\tau_i|C)$, we arrive at

$$\text{DT:} \quad CER(Q_s; C_t, D_t) = \sum_{i=1}^{n} P(s_i|Q_s) \log \frac{\sum_{j=1}^{T} P(s_i|t_j)((1-\lambda)P(t_j|D_t) + \lambda P(t_j|C_t))}{\sum_{j=1}^{T} P(s_i|t_j) P(t_j|C_t)} \tag{6}$$

So, though this model has been often described as a model for query Translation (e.g. [Hiemstra, 2001]), we would rather view it as a CLIR model based on a simple form of document translation (using a word-by-word approach), which on the basis of document terms generates a query. However, contrary to other document translation approaches like [Oard, 1998] and [Franz et al., 1999], only those terms in the document are translated that do lead to a match with query terms. It is therefore a more efficient and more scalable approach.

Note that both the QT and DT models are based on context-insensitive translation, since translation is added to the IR model after the term independence assumption has been made. Recently, a more complex CLIR model based on relaxed assumptions - context-sensitive translation but term-independence based IR - has been proposed in [Federico & Bertoldi, 2002]. In experiments on the CLEF test collections, the aforementioned model also proved to be more effective than a probabilistic CLIR model based on word-by-word translation. However, it has the disadvantage of reducing efficiency, due to a Viterbi search procedure.

3 Translation resources

As said, the generation of well-formed target language expressions is not an issue in the context of CLIR. In our probabilistic framework translation can thus be performed on a word-by-word basis. As a consequence the role of translation resources is to translate between words. Therefore dictionaries will suffice. In this section we will describe some procedures to generate translation models and dictionaries on the basis of freely available resources. These will be compared with costly high quality machine readable dictionaries.

3.1 Building parallel web corpora by mining parallel pages

Parallel corpora seem an ideal resource for the construction of translation models, since we can benefit from proven word alignment techniques, which have been developed for statistical MT. Translation models can be derived from parallel corpora by first aligning the sentences in source and target language text and subsequently aligning words using statistical algorithms that maximize a probabilistic criterion. A probabilistic translation dictionary can subsequently be derived from the word-aligned texts. A serious drawback of resorting to parallel texts as a translation resource is that it is difficult to acquire large parallel corpora for many language pairs. For many language pairs, large parallel corpora are not available, or access is restricted. This problem can partially be overcome by using the web as a resource of parallel pages [Resnik, 1998, Nie et al., 1999]. Many non-English web sites offer English translations of their pages, which can form the basis for the construction of parallel corpora with English as one of the languages. Another potential drawback of using parallel corpora is that they introduce a domain dependency into the estimation of translation probabilities. This might be a problem when e.g. using a parallel corpus trained on legal documents in the medical domain. A similar problem is present for hand-crafted term-bank translations for specialized domains.

We have developed several parallel corpora based on parallel web pages for the CLEF 2001 evaluation in close cooperation with the RALI laboratory of the Université de Montréal. The PTMiner tool [Nie et al., 1999] was used to find web pages that have a high probability to be translations of each other. One of the key elements in the mining algorithm is to look at regularities in file names and to search for specific anchor texts, such as ”English version”.

The mining process was run for four language pairs and resulted in one large and three modestly sized parallel corpora. Table 1 lists sizes of the corpora.

language	# cleaned pairs
EN-IT	4768
EN-DE	5743
EN-NL	2907
EN-FR	18807

Table 1: Size of the corpora expressed as number of parallel pairs of web pages; ISO 639 language codes

3.2 Building translation models using parallel web pages

The parallel web corpora were used to construct simple statistical translation models (IBM model 1) [Nie et al., 1999]. The construction of the translation models is documented in [Kraaij et al., 2003]. Here the major aspects will be summarized.

Format conversion In this first step, the textual data are extracted from the Web-pages. Of the HTML markup tags, only paragraph markers and sentence boundary information is retained, since these markers are important for the sentence alignment process.

Sentence alignment After a pair of web pages has been converted in neatly structured documents consisting of paragraphs consisting of sentences, the document pair is aligned. This alignment produces so-called *couples* i.e. minimal-size pairs of text segments from both documents. The couples usually consist of two sentences, but sometimes a sentence cannot be aligned, or is aligned to more than one sentence. The alignment procedure we used was based on [Simard et al., 1992]

Tokenization, Lemmatization and Stopwords Since the final goal of our procedure is a word-alignment, sentences have to be tokenized first. The end goal is to use the translation models in an IR context, so it seems natural to have both the translation models and the IR system operate on the same type of data. We therefore lemmatize the sentences and remove stopwords. Since we did not have access to full morphological analysis for Italian, we used a simple, freely-distributed stemmer from the Open Muscat project[3]. For French and English, we lemmatized each word-form by lookup in a morphological dictionary using its POS-label (assigned by a HMM-based POS-tagger [Foster, 1991]) as a constraint. As a final step, stopwords were removed.

[3] Currently distributed by OMSEEK: `http://cvs.sourceforge.net/cgi-bin/viewcvs.cgi/omseek/om/languages/`

Word Alignment Following common practice, only one-to-one aligned sentence pairs were used for the word alignment process. A simple statistical translation model, IBM's *Model 1*, was trained on the pre-processed aligned sentences. This model disregards word order (which is ignored in most IR systems) and is relatively easy to train. As a by-product, the training procedure for Model 1 yields the conditional probability distribution $P(s|t)$, which we need for our CLIR model. The following table provides some statistics on the processed corpora.

	EN-FR	EN-IT
# 1-1 alignments	1018K	196K
# tokens	6.7M/7.1M	1.2M/1.3M
# unique stems	200K/173K	102K/87K
# unique stems (P $>$ 0.1)	81K/73K	42K/39K

Table 2: Sentence-aligned corpora

Pruning the model The $P(s|t)$ distribution is estimated on the corpus of aligned sentences, using the Expectation-Maximisation (EM) algorithm. As in any other corpus-based approach to learning properties of natural language data, sparseness poses a real problem. A complex model requires a large dataset in order to estimate parameters in a reliable way. IBM's *M*odel 1 is not a very complex model, but it contains many parameters, since $P(s|t)$ covers the cross-product of source and target language vocabularies. Since the aligned corpora are not extremely large, translation parameters for which there is not much training data (rare English and French words) cannot be reliably estimated. We noticed from preliminary experiments, that the retrieval effectiveness of a CLIR system based on a probabilistic model can be improved by deleting parameters (translation relations) for which indications exist that they are less reliable [Kraaij et al., 2003]. The pruning method that we used for the translation models that are used for the experiments with transitive translation are based on taking the 100K best model parameters (IBM1 gains) and to delete the parameters that contribute the least to the quality of the model. One way to measure quality is the normalized log-likelihood of a target language test corpus given a source language test corpus. The individual contribution of each parameter (translation probability) can be rated by computing the aforementioned log-likelihood based on the full translation model in comparison with the log-likelihood of the translation model where the parameter is set to zero. The log-likelihood ratio for a reliable parameter will be high, indicating that pruning such a parameter would seriously hurt the performance of the model [Foster, 2000]. Pruning the model is than a matter of ordering, thresholding and re-normalization.

Since translation models trained on parallel corpora will not have a complete coverage of names, we applied one back-off rule in the translation model: if a word is not found it's translation is the identical form, in the hope that the target language translation is in fact a cognate. Fuzzy matching strategies might even help improve recall.

3.3 Building translation models using MRD's

As a contrast to Web-based translation models, which use freely available data, we also generated probabilistic translation dictionaries using the VLIS lexical database owned by the publishing company Van Dale Lexicografie BV, which contains manually generated, high quality translation relations for several European languages, with Dutch as a pivot language. (See [Kraaij & Pohlmann, 2001] for more general information about VLIS.)

We have prepared several translation models based on the information in the VLIS database. All models are based on using just the simple lemmas. The basic idea is to look up all possible translations for a certain lemma. Both the search term and the translations[4] are normalized to minimize lookup

[4] We often used the translation relations in reverse direction.

problems; part-of-speech (POS) information for both search terms and lexical entries is available.

Before translation, queries are pre-processed in a series of steps in order to normalize them to a lemma format:

Tokenizing The query string is separated into individual words and punctuation characters.

Part-of-speech tagging Word forms are annotated with a part-of-speech tag using the Xelda toolkit developed by Xerox Research Centre in Grenoble for tagging and lemmatisation.

Lemmatisation Inflected word forms are lemmatized (replaced with their base form).

Stopword removal Non-content words like articles, auxiliaries etc, are removed.

More often than not, a translation consists of more than one word. It can be a phrase, or a list of alternatives, but also often some context is given in parentheses. A clean-up procedure has been defined based on a couple of heuristics: removing context in parentheses, removing punctuation and stopwords, lemmatizing the remaining words, treating each as a separate translation.

Estimation based on counts of translations Due to polysemy, but also due to fine grained sense distinctions, which are important for translators, multiple senses are available for the majority of the lemmas, each again possibly with several translations. Since the VLIS lexical database does not contain any frequency information about translation relations, we can only approximate $P(t|s)$ in a crude way. Some lemmas have identical translations for different senses. The Dutch lemma *bank*, for example, translates to **bank** in English in five different senses: ”institution”, ”building”, ”sand bank”, ”hard layer of earth” and ”dark cloud formation”. Other translations are **bench**, **couch**, **pew**, etc.

```
VLIS-query(English translations of bank(NL))
```

bank (institution), **bank** (building), **bank** (sand bank), **bank** (hard layer of earth), **bank** (dark cloud formation), **bench** (seat), **couch** (seat), **pew** (seat)

It is easy to compute the forward translation probability $P(t_j|s_i)$ for this (simplified) example: $P(\text{bench}|\text{bank}) = 1/8$. In a more formal way:

$$P(t_j|s_i) = \frac{c(s_i, t_j)}{\sum_j c(s_i, t_j)} \tag{7}$$

Here, $c(s_i, t_j)$ is the number of times the translation relation (s_i, t_j) is found in the lexical database.

The computation of the reverse translation probability $P(s_i|t_j)$ is slightly more elaborate. First, we select all lemmas in the target language that translate to the query term in the source language. We subsequently translate the target language lemmas to the source language and count the number of times that the target lemma translates to the literal query term.

```
VLIS-query:  Dutch translations of English translation of bank/NL
```

bank (English) → *bank (2x), oever, reserve, rij* etc.
pew (English) → *(kerk)bank, stoel*
couch (English) → *bank, sponde, (hazen)leger,* etc.

The probability that **bank** (E) translates to *bank* (NL) is twice as high as the probability that **bank** (E) translates to *oever*. The estimation of $P(s_i|t_j)$ on the VLIS database can be formalized as:

$$P(s_i|t_j) = \frac{c(s_i, t_j)}{\sum_i c(s_i, t_j)} \tag{8}$$

So far we have discussed translating from and into Dutch, the pivot language in the VLIS database.

For transitive translation via Dutch as a pivot language (e.g. French into Italian), we investigated two estimation methods. The first estimation method disregards the fact that Dutch is used as a pivot language and is based on (7) and (8). The second estimation procedure explicitly models the individual translations steps, to and from the interlingua:

$$P(t_j|s_i) \approx \sum_k P(t_j|d_k)P(d_k|s_i) = \sum_k \frac{c(d_k,t_j)}{\sum_j c(d_k,t_j)} \frac{c(s_i,d_k)}{\sum_k c(s_i,d_k)} \tag{9}$$

$$P(s_i|t_j) \approx \sum_k P(s_i|d_k)P(d_k|t_j) = \sum_k \frac{c(s_i,d_k)}{\sum_i c(s_i,d_k)} \frac{c(d_k,t_j)}{\sum_k c(d_k,t_j)} \tag{10}$$

where d_k represents a term from the Dutch interlingua. We hypothesized that this more detailed estimation procedure would improve retrieval performance. We will give a symbolic example to show the difference between the direct and transitive estimation procedure. Suppose the French word $f1$ has two Dutch translations d_1 and d_2. Now d_1 has one English translation e_1 and e_2 has two English translations e_2 and e_3. The direct translation probability estimates for translating F1 into English are $P(e_1|f_1) = P(e_2|f_1) = P(e_3|f_1) = 1/3$. The transitive estimates are: $P(e_1|f_1) = \sum_i P(e_1|d_i)P(d_i|f_1) = 1/2$, and in a similar fashion: $P(e_1|f_2) = P(e_1|f_3) = 1/4$.

Surprisingly, the experiments with the simpler approach (direct estimation: (7) and (8)) yielded better results than (9) and (10), we therefore did not pursue the transitive probability estimates further. We hypothesize that the performance decrease is due to the fact that VLIS contains roughly twice as many concepts as lemmas. This means that in a transitive estimation procedure, the probability mass is spread equally over each sense. Now if some of the word senses are actually just sense variations -sense differences bear upon varying levels of granularity-, then the transitive estimation procedure will assign most probability mass to related word senses, which might down-weight clearcut word senses. The direct estimation procedure suffers less from this problem.

4 Transitive translation

An important advantage of CLIR based on parallel web corpora is that it will lead to resources for many more language pairs than covered by commercial MT systems. For most supported pairs English will be one of the two languages, since it is the dominant language in international business and science. Therefore we hypothesized already that English could be used as a pivot language to maximize the number of different language pairs for which CLIR resources are available. In the following section we will report some preliminary experiments that were carried out with transitive approaches to CLIR based on parallel Web corpora and MRD's: FR-EN-IT and IT-EN-FR. The different approaches are illustrated in Figure 1.

We evaluated three different ways to to implement such a transitive approach, the first two alternatives use the convolution operation to combine two language models:

1. Model (4) - the QT model - based on a transitive estimate of $P(t|Q_s)$ involving a pivot language:

$$P(t_i|Q_s) \approx \sum_k^I \sum_i^S P(t_j|v_k)P(v_k|s_i)P(s_i|Q_s) \tag{11}$$

   Here, v_k is a term in the pivot language and I is the vocabulary size of the interlingua.

2. Model (6) - the DT model - based on a transitive estimate of $P(t|D_t)$ involving a pivot language:

$$P(s_i|D_t) \approx \sum_k^I \sum_j^T P(s_i|v_k)P(v_k|t_j)P(t_j|D_t) \tag{12}$$

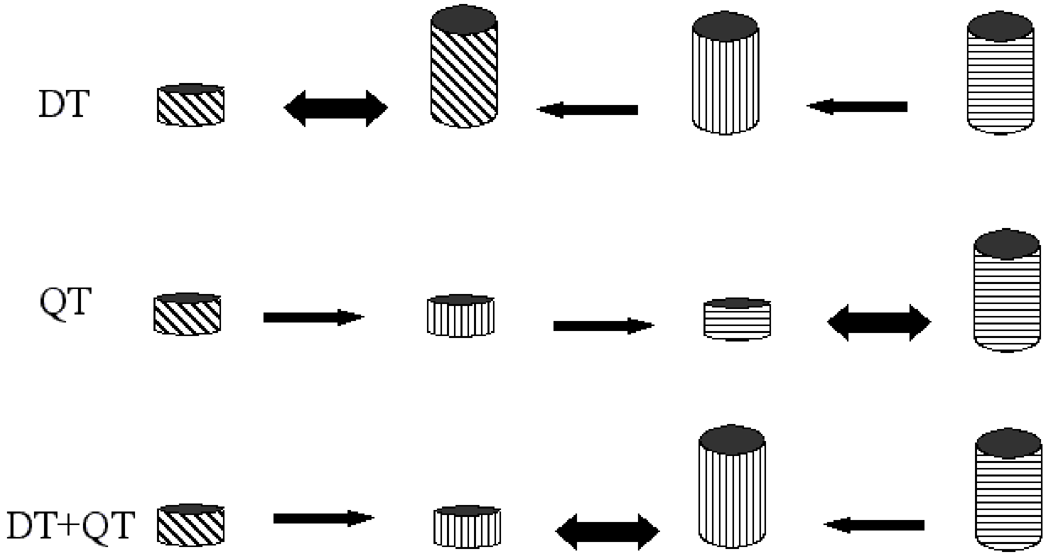

Figure 1: Schematic view of the three different ways to use a pivot language for CLIR. The small cylinders represent query models. The large cylinders represent document models. The source, pivot and target language are each represented by a different line pattern. The double arrow represents the matching operation between query and document model. The single arrow represents a translation step.

3. A variant where both the query and the documents are translated and matching takes thus place in the pivot language:

DT+QT:

$$CER(Q_s;C_t,D_t)=\sum_{k=1}^{I}\sum_{i=1}^{S}P(v_k|s_i)P(s_i|Q_s)\log\frac{\sum_{j=1}^{T}P(v_k|t_j)((1-\lambda)P(t_j|D_t)+\lambda P(t_j|C_t))}{\sum_{j=1}^{T}P(v_k|t_j)P(t_j|C_t)} \tag{13}$$

5 Evaluation

We have evaluated the models on the combined topic collections of CLEF-2000, 2001 and 2002, using the French and Italian corpus of CLEF 2000 (cf. [Kraaij et al., 2003]). We provide two baselines: a monolingual run (MONO) and a run based on a transitive version of the structured query model using the synonym operator (SYN). For further comparison, we also provide the results for bilingual runs, with the queries stated in the pivot language (English). This helps to judge the relative quality of each translation step.

Table 3 presents the results of the experiment using the Web-based translation models. Inspection of 11-point recall-precision plots indicates that the performance differences are consistent across all recall levels, e.g. the plots are parallel. With a performance ranging between 68% and 79% with respect to the monolingual baseline, the results of all three methods are at least at a comparable level as those reported in [Franz et al., 2000] and [Lehtokangas & Airio, 2002] and are significantly better than the SYN baseline.

All the LM-based methods are significantly better than the SYN baseline at the 0.01 level. It is perhaps not surprising anymore that the LM-based methods perform better than the SYN baseline, since the SYN based model cannot leverage the probabilities of the translation alternatives. All translation alternatives are equally probable in this approach and many translation alternatives amounts thus

	IT-EN-FR		FR-EN-IT	
monolingual baseline	0.4233		0.4542	
bilingual baselines:				
QT (EN→{FR|IT})	0.3878	(-8%)	0.3519	(-23%)
DT (EN←{FR|IT})	0.3909	(-8%)	0.3728	(-18%)
transitive runs:				
SYN(Pirkola/Ballesteros)	0.1469	(-65%)	0.2549	(-44%)
QT (target match)	0.2924	(-31%)	0.3287	(-28%)
DT (source match)	0.3149	(-26%)	0.3598	(-21%)
QT+ DT (pivot match)	0.2866	(-32%)	0.3361	(-26%)
% missed qt	14.5		17	
% missed dt	16		20	
% missed qt+dt	11		11	
# translations qt	9.6		9.4	
# translations dt	84.0		123	
# translations qt+dt	55.0		97	

Table 3: Results of transitive CLIR runs based on a combination of 100K models (mean average precision)

to high ambiguity. Recent work by Ballesteros confirms this weakness of the SYN based approach [Ballesteros & Sanderson, 2003]. The weakness of the SYN based method can be compensated by a probabilistically motivated version of weighted structured queries [Darwish & Oard, 2003], but the resulting model is less transparent than our cross-entropy based approach where translation is a part of the model.

The QT, DT, and QT+ DT methods have a slightly different performance, but differences are not consistent across both language pairs. We performed sign tests and did not find any significant differences between the QT, DT and QT+DT methods for both IT-EN-FR and FR-EN-IT. We think that the differences between the LM-based CLIR variants are due to lexical mismatches between the constituting models. For a CLIR run on the Italian test collection using French queries and the QT model with English as an interlingua, the first translation model maps the French query model onto an English query model, whereas the second translation model maps the English query model onto an Italian query model. Not all terms that can be translated by one model, have a non-zero translation probability in the other model. An important reason for this imbalance is the fact that the models are trained on different parallel corpora of different sizes. Since the translation models themselves are not symmetric, this will result in differences between methods. A comparison of the number of missed translations with the runs based on just a single translation (the bilingual runs in Table 3) shows that this is a serious effect. The EN-IT statistical translation dictionary is substantially smaller than the EN-FR translation dictionary (about 35K vs. 50K entries). This explains why mean average precision is hurt much more by going from EN-FR to IT-EN-FR than going from EN-IT to FR-EN-IT.

The data seem to suggest a positive correlation between the number of translations and the mean average precision (with the exception of QT+DT for IT-EN-FR). Indeed, this seems plausible, since more translation relations would help to provide a more robust mapping of a language model from one language to another. However, previous work on bilingual CLIR [Kraaij et al., 2003] using the same test collection does not show this correlation. In the bilingual experiments, translation effectiveness seemed dependent on the relative verbosity of the languages involved. Translation from the more verbose language to the less verbose language (e.g. French → English) was more effective. Moreover, the experiment with pivoted translation using symmetric data from MRD's as reported below does not suggest a correlation.

We repeated the experiment with translation models based on VLIS although we used the "direct" estimation method of (7) and (8) instead of the convolution approach (9) and (10). Results are presented in Table 4. The performance of the VLIS-based runs on Italian and French documents does not differ dramatically from the results based on English queries presented in [Kraaij et al., 2003]. This can hardly come as a surprise, since all the runs use the interlingual Dutch representation as a pivot language and thus do not differ in a principal way. Again there is no clear sign that either of the models QT, DT or QT+ DT is clearly superior over the other. This time, we can directly compare QT and DT since the number of translations is exactly the same. Sign tests show that there is no significant difference between QT, DT and QT+DT for the IT-NL-FR runs, but QT and DT are significantly better than SYN. For the FR-NL-IT runs, the DT run is significantly better (at the 0.05 level) than the other methods. Since the order of CLIR models based on retrieval effectiveness is different for the IT-NL-FR runs, we will not draw any strong conclusions. There seems to be a strong interaction between the translation resource, the query set, the document collection and retrieval performance. Further research is needed to explore the nature of this interaction, e.g. by performing a query-by-query analysis.

	IT-NL-FR		FR-NL-IT	
monolingual baselines	0.4233		0.4542	
transitive runs:				
SYN	0.3421	(-19%)	0.3171	(-30%)
QT	0.3542	(-16%)	0.3171	(-30%)
DT	0.3468	(-18%)	0.3391	(-25%)
QT+ DT	0.3473	(-18%)	0.3080	(-32%)
% missed qt	6.2		10	
% missed dt	6.2		10	
% missed qt+dt	6.2		10	
# translations qt	3.4		6.4	
# translations dt	3.4		6.4	
# translations qt+dt	4.6		10.6	

Table 4: Transitive translation based on different VLISbased models (performance difference with EN-FR and EN-IT respectively in brackets)

6 Conclusions

We have evaluated three different variants of a CLIR model involving a pivot language, where matching occurred in the source, pivot or target language. The alternative configurations have roughly equivalent effectiveness. For Web translation models, all probabilistic models outperform the SYN baseline (the Pirkola/Ballesteros approach); the SYN-based baseline cannot handle many translations. We have found that the most important factor for retrieval effectiveness is the lexical coverage of the complete translation chain, which is determined by the weakest translation resource. We did not find evidence for a clear superior model after analysing these preliminary experiments. Both Web-based and dictionary-based transitive CLIR methods yielded performances between 68-83% of a monolingual setting, depending on query language and translation resource, which is quite acceptable if direct translation resources are lacking. All in all, Web-based translation resources should be considered as competitive with high quality MRD resources, and transitive translation as a viable approach to CLIR.

Acknowledgments

We would like to thank the RALI lab of Université de Montréal for making available several tools to mine parallel Web pages and especially Michel Simard for his help with training the translation models.

References

[Ballesteros & Croft, 1998] Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In [Croft et al., 1998], pp. 64–71.

[Ballesteros & Sanderson, 2003] Ballesteros, L., & Sanderson, M. (2003). Addressing the lack of direct translation resources for cross-language retrieval. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*, pp. 147–152. acm.

[Ballesteros, 2000] Ballesteros, L. A. (2000). Cross-language retrieval via transitive translation. In Croft, W. B., editor, *Advances in Information Retrieval*. Kluwer Academic Publishers.

[Berger & Lafferty, 1999] Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In [Hearst et al., 1999], pp. 222–229.

[Croft et al., 1998] Croft, W., Moffat, A., van Rijsbergen, C., Wilkinson, R., & Zobel, J., editors (1998). *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM Press.

[Darwish & Oard, 2003] Darwish, K., & Oard, D. W. (2003). Probabilistic structured query methods. In Callan, J., Cormaxk, G., Clarke, C., Hawking, D., & Smeaton, A., editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pp. 338–343. ACM Press.

[Federico & Bertoldi, 2002] Federico, M., & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In Beaulieu, M., Baeza-Yates, R., Myaeng, S. H., & Järvelin, K., editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pp. 167–174. ACM Press.

[Foster, 2000] Foster, G. (2000). A maximum entropy / minimum divergence translation model. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[Foster, 1991] Foster, G. F. (1991). Statistical Lexical Disambiguation. Msc thesis, McGill University, School of Computer Science.

[Franz et al., 1999] Franz, M., McCarley, J., & Roukos, S. (1999). Ad hoc and multilingual information retrieval at IBM. In Voorhees, E. M., & Harman, D. K., editors, *The Seventh Text REtrieval Conference (TREC-7)*, volume 7. National Institute of Standards and Technology, NIST. NIST Special Publication 500-242.

[Franz et al., 2000] Franz, M., McCarley, J. S., & Ward, R. T. (2000). Ad hoc, cross-language and spoken document retrieval at IBM. In Voorhees, E. M., & Harman, D. K., editors, *The Eigth Text REtrieval Conference (TREC-8)*, volume 8. National Institute of Standards and Technology, NIST. NIST Special Publication 500-246.

[Gollins & Sanderson, 2001] Gollins, T., & Sanderson, M. (2001). Improving cross language retrieval with triangulated translation. In Croft, W., Harper, D., D.H.Kraft, & Zobel, J., editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pp. 90–95. ACM Press.

[Hearst et al., 1999] Hearst, M., Gey, F., & Tong, R., editors (1999). *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM Press.

[Hiemstra, 2001] Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente.

[Hiemstra & de Jong, 1999] Hiemstra, D., & de Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In *European Conference on Digital Libraries*, pp. 274–293.

[Kraaij, 2004] Kraaij, W. (2004). *Variations on language modeling for information retrieval*. PhD thesis, University of Twente. forthcoming.

[Kraaij & Hiemstra, 1998] Kraaij, W., & Hiemstra, D. (1998). Cross language retrieval with the twenty-one system. In Voorhees, E. M., & Harman, D. K., editors, *The Sixth Text REtrieval Conference (TREC-6)*, volume 6. National Institute of Standards and Technology, NIST. NIST Special Publication 500-240.

[Kraaij et al., 2003] Kraaij, W., Nie, J.-Y., & Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3),381–419.

[Kraaij & Pohlmann, 2001] Kraaij, W., & Pohlmann, R. (2001). Different approaches to cross language information retrieval. In Daelemans, W., Sima'an, K., Veenstra, J., & Zavrel, J., editors, *Computational Linguistics in the Netherlands 2000*, number 37 in Language and Computers: Studies in Practical Linguistics, pp. 97–111, Amsterdam. Rodopi.

[Kraaij & Spitters, 2003] Kraaij, W., & Spitters, M. (2003). Language models for topic tracking. In Croft, B., & Lafferty, J., editors, *Language Models for Information Retrieval*. Kluwer Academic Publishers.

[Lavrenko & Croft, 2003] Lavrenko, V., & Croft, B. (2003). Relevance models in information retrieval. In Croft, B., & Lafferty, J., editors, *Language Models for Information Retrieval*, pp. 11–56. Kluwer Academic Publishers.

[Lehtokangas & Airio, 2002] Lehtokangas, R., & Airio, E. (2002). Translation via a pivot language challenges direct translation in CLIR. In *Proceedings of the SIGIR 2002 Workshop: Cross-Language Information Retrieval: A Research Roadmap*.

[Nie et al., 1999] Nie, J., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts in the web. In [Hearst et al., 1999], pp. 74–81.

[Oard, 1998] Oard, D. (1998). A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of AMTA 1998*, pp. 472–483.

[Pirkola, 1998] Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In [Croft et al., 1998], pp. 55–63.

[Ponte & Croft, 1998] Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In [Croft et al., 1998], pp. 275–281.

[Resnik, 1998] Resnik, P. (1998). Parallel stands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of AMTA*, number 1529 in Lecture Notes in Artificial Intelligence, pp. 72–82.

[Simard et al., 1992] Simard, M., Foster, G., & Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodogical Issues in Machine translation (TMI92)*.