TNO at CLEF-2001: Comparing Translation Resources

Wessel Kraaij

TNO-TPD
P.O. Box 155, 2600 AD Delft
The Netherlands
kraaij@tpd.tno.nl

Abstract

This paper describes the official runs of TNO TPD for CLEF-2001. We participated in the monolingual, bilingual and multilingual tasks. The main contribution of this paper is a systematic comparison of three types of translation resources for bilingual retrieval based on query translation. We compared several techniques based on machine readable dictionaries, statistical dictionaries generated from parallel corpora with a baseline of the Babelfish MT service, which is available on the web. The study showed that the topic set is too small to draw reliable conclusions. All three methods have the potential to reach about 90% of the monolingual baseline performance, but the effectiveness is not consistent across language pairs and topic collections. Because each of the individual methods are quite sensitive to missing translations, we tested a combination approach, which yielded consistent improvements up to 98% of the monolingual baseline.

1 Introduction

Research on Cross Language Information Retrieval has been on the agenda of TNO TPD since 1997. TNO TPD participated in the CLIR tracks of TREC and CLEF, usually together with the University of Twente under the flag of the "Twenty-One" project, a EU project focusing on cross language dissemination of information. University of Twente and TNO still cooperate in various research domains in the DRUID project. For CLEF 2001, we did not change our basic approach to CLIR, but experimented with different translation resources and the best way to integrate them. Therefore we will restrict ourselves to a rather concise description of the basic retrieval model and concentrate on this year's experiments.

2 The CLIR model

The basic approach for our CLIR experiments is query translation. The major advantage of query translation is its scalability. Some groups have shown however that document translation can yield competitive results, especially in combination with query translation[3],[1]. At this point we have chosen to refine our approach instead of testing all kinds of combination strategies, which could suffer from collection dependency.

All runs were carried out with an information retrieval system based on a simple unigram language model[4]. The basic idea is that documents can be represented by simple statistical language models. Now, if a query is more probable given a language model based on document d_1, than given e.g. a language model based on document d_2, then we hypothesise that the document d_1 is more relevant to the query than document d_2. Thus the probability of generating a certain query given a document-based language model can serve as a score to rank documents with respect to relevance.

$$P(T_1, T_2, \cdots, T_n|D_k)P(D_k) \;=\; P(D_k)\prod_{i=1}^{n}(1-\lambda_i)P(T_i) + \lambda_i P(T_i|D_k) \qquad (1)$$

Formula 1 shows the basic idea of this approach to information retrieval, where the document-based language model is interpolated with a background language model to compensate for sparseness. In the formula, T_i is a random variable for the query term on position i in the query ($1 \leq i \leq n$, where n is the query length), which sample space is the set $\{t^{(0)}, t^{(1)}, \cdots, t^{(m)}\}$ of all terms in the collection. The probability measure $P(T_i)$ defines the probability of drawing a term at random from the collection, $P(T_i|D_k)$ defines the probability of drawing a term at random from the document; and λ_i defines the importance of each query term. The optimal λ (0.3) was found by tuning on earlier CLIR collections. The a-priori probability of relevance $P(D_k)$ is usually taken to be a linear function of the document length, modelling the empirical fact that longer documents have a higher probability of relevance.

The retrieval model has been extended for the CLIR task, by integrating a statistical translation step into the model [5]. The CLIR extension is presented in the following formula:

$$\begin{aligned} &P(D_k, S_1, S_2, \cdots, S_n) \ = \\ &P(D_k) \prod_{i=1}^{n} \sum_{j=1}^{m} P(S_i|T_i{=}t^{(j)})((1-\lambda_i)P(T_i{=}t^{(j)}) + \lambda_i P(T_i{=}t^{(j)}|D_k)) \end{aligned} \tag{2}$$

Here S_i refers to terms in the source (query) language and T_i refers to terms in the target (document) language, $P(S_i|T_i = t^{(j)})$ represents the probability of translating a term from the target language $t^{(j)}$ in a source language term S_i. Note that the notions of source and target language are a bit confusing here, because the CLIR retrieval model contains a translation component, which translates target language terms in source language terms.

An informal paraphrase of the extension is: the relevance of a document in a target language with respect to a query in a different source language can be modelled by the probability that the document generates the query. We know that several words T_j in the target language can be translated into the query term S_i, we also assume for the moment that we know their respective translation probabilities. The calculation of the probability involves an extra step: the probability of generating a certain query term is the sum of the probabilities that a document in the target language generates a word which in turn is translated to the query term. These probabilities are a product of the probability $P(T_j)$ as in Formula 1 with the translation probability $P(S_i|T_j)$. We refer to [7] and [5] for a technical description of the model. A crucial aspect of the approach is that the model treats alternative translations as a probabilistic disjunction resulting in highly structured queries.

3 Integrating prior knowledge

In previous years (CLEF2000) we have seen that adding a so-called document prior conditioned on the document length results in a significant improvement of retrieval performance especially for short queries. The document prior was simply integrated into the model as a multiplication, assuming independence.

The approach has some problems:

1. The formalisation is ad-hoc.
2. It is hard to normalise scores.
3. Its effect on short (title) queries is much larger than on longer queries.

Especially problem (2) is important for CLIR, because for the multilingual task, we want to normalise scores. Document scores in the original model are linearly related to the query length. Now suppose, we translate the query into different languages using Systran. It could very well be the case that a translation into a compounding language results in a shorter query, thus the translated queries will not have a homogeneous length. An easy strategy is to divide the document scores by the query length. But this division would also affect the "document prior". We propose to model the a-priori probability of relevance conditioned on the document length and the normalised generative probability of the query based on the document unigram model as separate models.

$$RSV_{final} = \mu \log(P(D_k = rel|len(D_k) = d)) + (1 - \mu) \log(P(T_1, T_2, \cdots, T_n|D_k))/n \tag{3}$$

The scores of these models are subsequently interpolated via linear interpolation, in formula (3) the final retrieval status value (RSV) is composed of two components, the first addend is the a-priori probability of relevance of a document conditioned on its length, the second component is the unigram component.

In studies on other collections we found that the optimal value for the interpolation parameter is relatively stable across collections, but is dependent on the query characteristics. The optimal value of the interpolation parameter is inversely related to the query length. This can be explained as follows: for short queries, the probability of relevance is not dependent on the document length because of the document normalisation which comes with the maximum likelihood estimates for $P(T_i|D_k)$. However, there is empirical evidence that longer documents have a higher probability to be relevant [15], because longer documents often contain more information. Robertson calls this the "scope hypothesis" [14]. Therefore it is beneficial to have a rather high value of μ for title queries, which give the hybrid retrieval model a slight bias to longer documents. For longer queries, the generative model will have an increasing bias for longer documents, because of the "soft" coordination effect, thus it is important to use a small (or even zero) value of μ. For our CLEF experiments, we used formula (3) with $\mu = 0.2$, which was found to give optimal results on title+description queries from the CLEF2000 topic collection.

4 Lemmatisation

The major part of the experiments was based on an indexing procedure which involves morphological lemmatisation. The lemmatiser is based on the Xelda morphological toolkit[1], developed by XRCE Grenoble, which supports a wide range of European languages. The lemmatiser removes inflectional suffixes, performs case normalisation (proper nouns and German nouns keep their capitalisation) and can segment compounds. This year, we tested whether keeping the capitalisation would improve precision (cf. Section 6.1). For German and Dutch, the basic procedure for handling compounds was to add both the full compound (inflection removed) and the recognised compound segments to the set of index terms (at indexing and retrieval time). This approach has proved to work well [9], but is a very ad-hoc plugin to the probabilistic retrieval framework that we apply for term weighting [4]. Several runs were done based on a Porter like stemming procedure [13]. The Dutch version of Porter [8] cannot split compounds but removes also some derivational suffixes, which is sometimes beneficial [9],[6].

5 Translation resources

While the development of a theoretical framework for CLIR is extremely important, a more practical but not less important issue is the acquisition or development of translation resources. CLEF 2000 showed that there are several classes or resources can be used for successful CLIR.

1. High quality machine readable dictionaries are available for the major European languages[5].
2. Commercial MT systems with different levels of quality can be used at no or relatively small costs. CLEF 2000 showed successful use of the Babelfish translation service based on Systran, the Powertranslator system from L&H and several other systems[10],[1].
3. Some groups exploited parallel or comparable corpora for CLIR. Parallel corpora can be used to train probabilistic bilingual translation dictionaries[12]. Comparable corpora can be used to generate similarity thesauri[1].

In this paper we will compare these three types of resources in a quantitative and qualitative way. For the machine readable dictionaries we used the VLIS lexical database of lexicon publisher Van Dale. For a description of the structure and contents of the VLIS database we refer to our CLEF 2000 paper [5]. We estimated the "reverse" translation probabilities $P(S_i|T_j)$ (the translation probability of a term in the source language given a term in the target language) using some very sparse pseudo frequency information extracted from the lexical database (cf. [5]).

[1] cf. `http://www.xrce.xerox.com/ats/xelda/overview.html`

We contrasted query translation using VLIS with translation based on the Babelfish MT service and word by word translation based on dictionaries derived from a collection of parallel web documents.

For CLEF 2000 we had already developed three parallel corpora based on web pages in close cooperation with RALI, Université de Montréal. We used the PTMiner tool [12] to locate web pages which have a high probability to be translations of each other. The mining process consisted of the following steps:

1. Query a web search engine for web pages with a hyperlink anchor text "English version" and respective variants.
2. (For each web site) Query a web search engine for all web pages on a particular site.
3. (For each web site) Try to find pairs of path names that match certain patterns, e.g.: `/department/tt/english/home.html` and `/department/tt/italian.html`.
4. (For each pair) download web pages, perform a language check using a probabilistic language classifier, remove pages which are not positively identified as being written in a particular language.

The mining process was repeated for three language pairs (EN-IT, EN-DE and EN-NL) and resulted in three modestly sized parallel corpora. RALI had already mined a parallel web corpus for English-French. Table 1 lists some quantitative data about these corpora.

language	nr of web sites	nr of candidate pages	nr of candidate pairs	retrieved + cleaned pairs
EN-IT	3651	1053649	23447	8504
EN-DE	3817	1828906	33577	10200
EN-NL	3004	1170082	24738	2907
EN-FR	n.a.	n.a.	n.a.	18807

Table 1: Intermediate sizes during corpus construction

The parallel corpora were used to train simple translation models (IBM model 1, [2]) for both translation directions. For CLEF 2000 we did some preliminary runs with encouraging results. This year we decided to do some more extensive tests with the translation models for two language pairs: EN-IT and EN-FR. We chose these language pairs because the translation models are trained on lemmatised/stemmed corpora. The IR indexes are also based on lemmatised documents, thus both lemmatisation/ stemming schemes have to be sufficiently equal. For the pairs EN-IT and EN-FR we were able to synchronise the lemmatisation/stemming schemes of the translation models and the IR indexes. The translation models were built at RALI and used a RALI lemmatiser for English and French and the Porter stemmer from MUSCAT[2] for Italian. All IR indexes were built using the Xelda lemmatisation tools. For our bilingual EN-IT runs, we built an index using the Porter stemmer for Italian. We tested several alternatives to include the translation probabilities into the model on the CLEF 2000 collection. Strangely enough, the forward probabilities $P(T_j|S_i)$ proved to be more effective than reverse probabilities, single most probable translation and equal translation probabilities. Pruning the translation model proved to be a key issue. We will elaborate on these aspects in a later publication. For CLEF 2001, we decided to use the best performing estimates i.e. the forward probabilities, neglecting the fact that the theoretical model calls for reverse translation probabilities.

6 Experiments

Although our focus this year was on the bilingual task, we also participated in the monolingual and multilingual tasks. All reported runs used only the title and description fields for the automatic construction of queries.

[2] `http://open.muscat.com`

6.1 Monolingual results

Our main goal for the monolingual runs was to improve the pool and to provide a baseline for bilingual experiments. We also did a minor experiment with case sensitivity and fuzzy query term expansion.

run tag	language	m.a.p.	judged@100	description
tnodd1	DE	0.3946	81	baseline
tnodd2	DE	0.3945	81	fuzzy lookup
tnodd3	DE	0.4111	87	fuzzy expansion
tnoee1	EN	0.5144	75	baseline
tnoee2	EN	0.5130	75	fuzzy lookup
tnoee3	EN	0.5289	75	fuzzy expand
tnoff1	FR	0.4877	90	baseline
tnoff2	FR	0.4883	90	fuzzy lookup
tnoff3	FR	0.4796	90	fuzzy expansion
tnoff4	FR	0.4904	91	case normalisation
tnoii1	IT	0.4411	87	baseline
tnoii2	IT	0.4449	87	fuzzy lookup
tnoii3	IT	0.4534	88	fuzzy expansion
tnoii4	IT	0.4508	88	case normalisation
tnonn1	NL	0.3795	84	baseline
tnonn1p	NL	0.3643	74	Porter stemmer (case insensitive)
tnonn2	NL	0.3720	84	fuzzy lookup
tnonn3	NL	0.3917	86	fuzzy expansion
tnonn5	NL	0.3071	72	no stemming
tnonn6	NL	0.3977	84	Xelda, case insensitive
tnoss1	ES	0.5181	88	baseline
tnoss2	ES	0.5182	88	fuzzy lookup
tnoss3	ES	0.5234	90	fuzzy expansion

Table 2: Monolingual results (italic=unofficial run)

Table 2 shows the results of our monolingual runs. We added some unofficial runs to complete the picture. Since case insensitivity had deteriorated the performance of some topics in previous years (when a proper name and a normal noun are homonyms e.g. Turkey and turkey) we decided to test the proper name recognition function of the Xelda morphological analyser. We did not apply any case normalisation after the Xelda lemmatisation step. Xelda can often correctly disambiguate homonyms, but when the context is not sufficient, our indexer chose the proper noun reading.

As in previous years, we also tested the effectiveness of a fuzzy lookup scheme. If a query term does not match with a term in the index vocabulary, a fuzzy lookup procedure (based on n-grams) substitutes the nearest matching term. This function is practical for misspellings in queries. The same fuzzy matching procedure can also be used to expand the query with near matches of all query terms. This function can be practical for queries that contain diacritics or such as accents, or spelling variants like the German ß/ss. The CLEF 2001 monolingual runs showed some modest improvements for the fuzzy expand option. We suspect however, that this is mostly due to undoing the case sensitivity, which is a side-effect of the expansion operation. For most languages, fuzzy conflation increases average precision by about 1% (absolute scale), the only exception being French. We will look at two languages (French and Dutch) in some more detail.

For French, mean average precision is hurt by the expansion operation. Looking closer at runs tnoff1 and tnoff3 shows that the main differences occur at topic 70 an 75. Topic 70 is about *Kim Il Sung* while Le Monde almost invariably spells his name as *Kim Il-sung*. The tokeniser converts dashes into spaces, a strategy which is necessary here, but it is clear also that case normalisation is necessary for proper matches. Omitting case normalisation (tnoff1) is 0.14 % less effective than doing case normalisation (tnoff3). Topic 75 is an odd topic because it only contains one relevant document; run tnoff3 has the relevant document on

the first position, whereas run tnoff1 retrieves the document in third position. The key point here is that the topic mentions *tuerie*. The baseline run retrieves the single relevant document in first position, the fuzzy expansion run conflates *tuerie* with *tuer* which has the effect that two other marginally relevant documents have an increased score and the single relevant document has rank 3, decreasing the map for this topic with 0.666. We did a control run (tnoff4) which was similar to the baseline (no fuzzy expansion) but which did apply case normalisation. The run performed marginally better at 0.4904, confirming our conjecture that our sub-optimal proper noun recognition module effectively hurts average precision.

For Italian, we also have conclusive evidence that case sensitivity deteriorated retrieval performance slightly instead of an expected improvement. There are marked differences only for a small number of topics. E.g. topic 88 talks about *Spongiforme Encefalopatia*. The Italian Xelda lexicon does not contain these words, and thus the capitalisation is left in place, which hurts performance here, because in most documents these terms are used in lowercase. A possible strategy could be to lowercase unknown words, but this information is not available for all languages.

For Dutch the fuzzy expansion run increases mean average precision w.r.t the baseline run. The increase is mainly due to topic 52 (0.0165 versus 0.5000). Here, the effect is merely due to case normalisation: Xelda sometimes sometimes assigns the lemma *Chinese* and sometimes *chinese*, which causes a great loss of recall in the baseline (case sensitive) run. The control run (tnonn6) confirms again that case sensitivity hurts average precision in the current set-up.

Our conclusion is that in principle it is good to do proper noun recognition (which we implemented by capitalisation) because sometimes a proper noun is homonymous with another word (e.g. in English: (Kim Il) *Sung* and *sung*). However, if the proper noun recognition is not perfect, e.g. it cannot deal with spelling variation (Kim Il-sung versus Kim Il Sung) it might hurt recall more than it improves precision. In practice, an imperfect proper name recognition module will yield inconclusive results which depend on the specific topic set. Our conjecture is that a proper name recognition module in combination with a proper name normalisation module (which conflates spelling variants of proper nouns) could improve retrieval effectiveness in a convincing way. The fuzzy matching technique as a means to find spelling variants other than case alternatives can help to improve retrieval performance as a fall-back option (i.e. when a query term is not found in the indexing vocabulary). Expanding every query term to find spelling variants is probably only useful when it is constrained to special word classes:

1. Languages with accentuated words: sometimes accents are left out or diacritics like the umlaut are mapped to a different form: $\ddot{u} \rightarrow ue$.

2. spelling variants of proper nouns: transliteration of non-Latin script names into a Latin script is usually done in different ways.

A principled approach would be to recognise these word classes (which are language dependent!) and build tailored normalisation/conflation modules for each of them. Our non-principled fuzzy conflation procedure does the job, but often also hurts precision because it is not constrained to these word classes.

The second experiment was a comparison between Xelda inflectional lemmatisation and a Dutch variant of the Porter stemming algorithm [9]. We did an unofficial case insensitive run based on Xelda lemmatisation (tnonn6) in order to do a clean comparison. The Xelda based run performs noticeably better than the Porter based run (+0.03) Big performance gains by the lemmatisation based approach can be seen in topic 55 (+0.55), where it is crucial to split the compound *alpeninitiatief* and in topic 68 (+0.50), where dictionary based approach correctly removes the inflectional suffix of the query term *synagogen* whereas the Dutch Porter fails to conflate *synagogen* and *synagoge*. This confirms our earlier conclusions that splitting of compound terms is essential [9]. The Xelda compound splitter is based on a lexicon of segmented compounds. This approach is not optimal, since compounding is a highly productive process (e.g. the terms *gekkekoeienziekte*, *schatzoekactiviteiten*, *schatzoekactiviteiten* from the CLEF 2001 topic collection). Unfortunately, we lacked the time to test existing better compound splitting algorithms, like the one we used for our experiments with the UPLIFT collection [16].

The 4th column of Table 2 shows the percentage of documents of the top 100 which is judged. In comparison with last year, this percentage has increased slightly, possibly due to the increased pool depth (60 instead of 50) and/or the increased number of participants. Only the English pool is of a slightly lower quality, questioning again why monolingual English runs are excluded from the pool.

6.2 Bilingual results

Table 3 shows the results of our official and unofficial (italic) bilingual runs, complemented with monolingual baselines.

FR-EN: Babelfish versus corpus A striking result is that the web corpus runs perform at the same level as the Systran based Babelfish service. Again we looked at some topics with marked differences in average precision. First, the topics where the web corpus run performs better: in topic 47 (+0.55), Systran lacks the translation of Tchétchénie (Chechnya); topic 58 (+0.46), Systran translates *mort* and *mourir* with *died* and *die*, whereas the web corpus has the additional concepts of *death* and *dead*; topic 82 (about IRA attacks, +0.4) Systran translates *l'IRA* erroneously by *WILL GO*, the corpus based translation brings in the related term *bomb* as a translation of attack. Secondly, the topics where Systran performs much better: topic 65 (-0.39) the corpus translations of *trésor* are *treasury* and *board*, which would be a fine phrase translation. In this context however, *trésor* does not have the financial meaning and because our system does not recognise phrases, *treasury* and *board* are used as separate query terms, which has the effect that the much more frequent term *board*, brings in a lot of irrelevant documents. Topic 75 (-0.98) suffers from a wrong interpretation of the word *sept*, which is translated to *sept (September)* and *7*, the latter term is discarded by the indexer. The month abbreviation retrieves a lot of irrelevant documents, resulting in a low position of the single relevant document; in topic 80 (about hunger strikes) *faim* is translated both by *hunger* and *death*. *Death* might be a related term in some cases, but it also retrieves documents about strikes and death, hurting precision; topic 89 talks about an *agent immobilier*, Systran produces the correct translation *real estate agent*, but the corpus based translation has *officer* and *agent* as additional translation. Here, the phrase translation of Systran is clearly superior.

run tag	language pair	m.a.p.	% of baseline	description
tnoee1	EN-EN	0.5144	100	baseline
tnoee3	EN-EN	0.5289	103	fuzzy expand
tnofe1	FR-EN	0.4637	90	RALI parallel web corpus, forward probabilities, fuzzy expansion
tnofe1a	FR-EN	0.4320	84	RALI parallel web corpus, forward probabilities, no fuzzy expansion
tnofe2	FR-EN	**0.4735**	92	Babelfish MT, fuzzy expansion
tnofe3	FR-EN	0.3711	73	VLIS MRD, inverse probabilities, fuzzy expansion
tnonn1	NL-NL	0.3795	100	baseline
tnoen1	EN-NL	0.3336	87	VLIS MRD, inverse probabilities, fuzzy expansion
tnoff1	FR-FR	0.4877	100	baseline
tnoef3	EN-FR	**0.4051**	83	VLIS MRD, inverse probabilities, fuzzy expansion
tnoef4	EN-FR	0.4039	82	Babelfish MT, fuzzy expansion
tnoef5	EN-FR	0.3642	76	RALI parallel web corpus, forward probabilities, fuzzy expansion
tnoii1	IT-IT	0.4411	100	baseline
tnoei3	EN-IT	**0.3549**	80	VLIS MRD, inverse probabilities, fuzzy expansion
tnoei4	EN-IT	0.2824	64	Babelfish MT, fuzzy expansion
tnoei5	EN-IT	0.3137	70	RALI parallel web corpus, forward probabilities

Table 3: Bilingual results CLEF 2001

FR-EN: Babelfish versus VLIS We also looked at some topics that revealed marked differences between the Systran run and the VLIS run. Topic 58 is a clear example where VLIS gives the best results (+0.44) , it correctly translates the key term *euthanasie* by *euthanasia* instead of the non standard translation *euthanasy* by Systran. In most cases however, Systran gives better results, some examples: topic 79 (-1.00), here VLIS fails to translate *Ulysse* into *Ulysses*, the word by word translation strategy also fails for *sonde spatiale*, VLIS translates sonde into *sampler;sound;probe;catheter;gauge;plumb;sink;auger* and spatiale into *spatial;dimensional*. Probably the fact that the query terms *Ulysses* and *space* are missing is more detrimental then the fact that VLIS generates some irrelevant translations for *sonde*, since the correct translation (*probe*) is found. In topic 62 (-0.50) both *Japon* is not found in VLIS and the multi-word unit *tremblement de terre* is not recognised as the French translation of *earthquake*. In topic 66 (-0.50) the seminal proper noun *Lettonie* is not found in VLIS, but is successfully translated by Systran. The proper nouns are probably not found in VLIS because in French, country names are usually denoted in combination with a determiner *La France*, *Le Quèbec*,..., our lexical lookup routine was not aware of this fact. In topic 80 (-0.65) the seminal query term *faim* is translated to *appetite;lust* instead of hunger (Systran). In topic 83 (-0.40), VLIS translates *enchère* by *raise;bid*, whereas Systran gives the contextual better translation *auction*. Finally, the low effectiveness of the VLIS based translation for topic 86 (-0.50) is due to a combination of factors, the dictionary based translation is rather fertile (e.g *usage* is translated in *currency;commonness;use;custom;habit;way;practice;usage;word*) and also the fuzzy expansion process is active, to correct for the case sensitivity of the index, which brings in more unwanted terms. Summarising, the Systran based Babelfish service outperforms the VLIS based run, because i) VLIS lacks translations of some proper nouns, ii) the word by word based translation fails for some topics (we currently have not accessed the phrasal translations in VLIS) and iii) VLIS has no method for sense disambiguation. Babelfish most probably uses phrase translations as a form of contextual sense disambiguation: the translation in isolation of *enchères* is *bidding*, *ventes aux enchères* gives *auction sales* and *ventes enchères* gives *sales biddings*.

EN-IT and EN-FR We do not want to base our judgement of the effectiveness of translation resources on one language pair and one topicset. Therefore we included two other languages pairs: EN-IT and EN-FR. For these language pairs, the VLIS based runs is clearly superior, which is not trivial, since these translations use Dutch as a pivot language. The EN-IT web based run performs surprisingly good (better than Systran), given its relatively small size and the fact that the corpus is hardly cleaned.

CLEF 2000 topic collection For an even better perspective, Table 4 shows the results for the same language pairs based on the CLEF 2000 topics.

method	language pair	m.a.p.	% of baseline
mono	FR-FR	0.4529	100
web corpus	EN-FR	**0.3680**	82
Babelfish	EN-FR	0.3321	73
VLIS	EN-FR	0.2773	62
mono	EN-EN	0.4164	100
web corpus	FR-EN	0.3995	95
Babelfish	FR-EN	**0.4007**	95
VLIS	FR-EN	0.2971	71
mono	IT-IT	0.4808	100
web corpus	EN-IT	**0.3771**	79
Babelfish	EN-IT	0.3564	75
VLIS	EN-IT	0.3266	69

Table 4: Bilingual results CLEF 2000

When we make a comparison of CLEF 2000 and CLEF 2001 results, we hardly see any consistent results, this confirms experiences we had with the various CLIR tracks at TREC6/7/8. The bilingual re-

sults depend strongly on lexical coverage. When a resource misses a few important concepts in the topic collection, its performance is seriously affected. In other words the mean average precision of a run is proportional to the lexical coverage. Unfortunately a set of 50 topics proves to be too small to measure the retrieval performance of a system based on a particular translation resource and its inherent lexical coverage in a reliabel way. We could do a few things to remedy this problem:

- Devise a special test for lexical coverage.
- Remove topics from the collection, for which one of the methods has serious lexical gaps. This might very well introduce a bias, but has the advantage that we can concentrate on some interesting research questions:
  - How well deal the different methods with the translation of phrases?
  - Is query term disambiguation really necessary?
  - Can we exploit synonym translations?

Combination of translation resources If we are merely interested in a strategy yielding "the best" result, it is fairly obvious that a combination of lexical resources could help to remedy the gaps of the individual translation resources. Some groups experimented with this idea, but with mixed results[1],[11]. We took a very straightforward approach and simply concatenated the (structured) translations of the different methods, which indeed improved upon the results of the individual runs. Results are presented in Table 5. This simple approach proved consistently effective: every combination of runs is more effective than the individual composing runs. The fact that combining a good and a bad translation resource does not degrade the performance is another indication that (at least for t+d queries) it is much more important to have at least one good translation and that the retrieval model is fairly robust against "noise" translations.

method	language pair	m.a.p.	% of baseline
mono	EN-EN	0.5144	100
web corpus	EN-FR	0.4637	90
Babelfish	EN-FR	0.4735	92
VLIS	EN-FR	0.3711	73
corpus&Babelfish	EN-FR	0.4895	96
corpus&VLIS	EN-FR	0.4672	92
VLIS&Babelfish	EN-FR	0.4783	94
VLIS&Babelfish&corpus	EN-FR	0.5032	98

Table 5: Combination runs CLEF 2001

6.3 Multilingual results

Our strategy for the multilingual task was identical to previous years. First we ran retrieval runs for each of the sub-collections, involving a translation step for most of these runs (the only exception is the EN-EN run). Subsequently we merged results using a naive merging strategy based on raw scores. The main difference with last year was that the VLIS lexical database was extended with translations from Dutch to Italian. Spanish did not pose any problem, because it was supported both by VLIS and Xelda. This gave us the opportunity to compare two set-ups: one based on the Babelfish service and one based on the VLIS lexical database. We did not have time to run a multilingual experiment based on the parallel web corpora.

Table 6 presents the results of our official multilingual runs. We can make several observations, the VLIS based run performs a little bit better than the Babelfish based run based on English topics. This comes as a surprise after the detailed comparison of Babelfish and VLIS on the bilingual FR-EN task. The only explanation could be that the translation quality of the various language pairs offered by Babelfish and/or Systran is not homogeneous. We will assess this shortly. Secondly, the VLIS run based on Dutch topics performs better than the VLIS run based on English topics. This is not so obvious, since the NL-X

run tag	language pair	mean average precision	description
tnoex3	EN-X	0.2634	VLIS
tnoex4	EN-X	0.2413	Babelfish
tnonx3	NL-X	0.2513	VLIS

Table 6: Multilingual results CLEF 2001

run does not contain a monolingual EN-EN run. Apparently the advantage to use the direct translations instead of translating from English via Dutch outweighs the disadvantage of replacing a monolingual (EN-EN) by a bilingual (NL-EN) run.

We have computed the mean average precision of the partial runs which make up the multilingual runs, results are shown in table 7. The number of topics on which the mean average precision is based is shown between brackets. We discovered that something went wrong with the Babelfish translations for some of the topics from English into Spanish (hence the 47 instead of 49).

run tag	languages	English	French	German	Italian	Spanish
tnoex3	EN-X	0.5289(47)	0.4051(49)	0.3184(49)	0.3549(47)	0.3990(49)
tnoex4	EN-X	0.5289(47)	0.4039(49)	0.2827(49)	0.2824(47)	0.3910(47)/**fix:**0.4135(49)
tnonx3	NL-X	0.4196(47)	0.4189(49)	0.3419(49)	0.3359(47)	0.3422(49)
mono run		0.5289	0.4877	0.3946	0.4411	0.5181

Table 7: mean average precision of intermediate runs

Rerunning the Babelfish EN-X multilingual run yielded a mean average precision of 0.2465, which does not really change the picture.

When we compare the Babelfish bilingual EN-X translation runs with the official FR-EN run (90%), we see that these runs compare less favourable with respect to the corresponding monolingual run: EN-FR:86%, EN-DE: 72%, EN-IT: 64%, EN-ES: 80%. Indeed, the translation quality of Babelfish is not homogeneous across languages for this set of topics.

7 Conclusions

At CLEF 2001, we concentrated on monolingual and bilingual experiments. Our hypothesis was that proper noun recognition could improve precision, because proper nouns are sometimes homonymous with other words. Our implementation of a proper noun aware indexing strategy turned out to hurt average precision. This is probably caused by a sub-optimal way to deal with the sometimes sill ambiguous output of the Xelda lemmatiser. We also experimented with a fuzzy query expansion method, in other to deal with spelling variation of especially proper nouns. A control experiment showed that the effectiveness of this algorithm is largely due to the conflation of capitalised/uncapitalised forms. We suggest a class based expansion scheme instead. Further we compared two different lemmatisation schemes for Dutch: the morphological lemmatisation (which includes the decomposition of compounds) proved to be markedly more effective than the Dutch variant of the Porter suffix stripper. For the bilingual task, we compared three different translation resources: a bilingual MRD (VLIS), a statistical dictionary based on a parallel web corpus and the Babelfish MT service. For the translation pair French-English, the web based and the MT based run reach a quite impressive level of 90 and 92% of the monolingual EN-EN run. The VLIS based run reached a level of 73%, which is due to several factors: deficiencies in the lexical lookup of proper names, lack of phrase handling and translation via a pivot language. For the translation pair English-Dutch, the VLIS based run scored better at 87% of the monolingual baseline, but failure analysis showed that phrase translation could improve results substantially. We think that the good results of the Babelfish based runs are mostly due to its ability to translate phrases. We consider the competitive results of the runs based on a web corpus based dictionary as a breakthrough in CLIR, because parallel web corpora

for EN-* language pairs are relatively easy to acquire. We hope to improve upon these results by training more complex models which allow for phrase translations. We also looked at several other bilingual tasks and did a comparison with CLEF 2000 topics. Our conclusion is that the topic sets are too small to really compare techniques to integrate translation resources into the retrieval model. Retrieval performance is proportional too lexical coverage. The set of 50 topics is too small to estimate lexical coverage, thus results are highly dependent on the particular topic set. We tested a combination approach, which merely concatenates the translated queries and proved to be consistently effective. Finally, in the multilingual task, our best result was achieved by a run based on the Dutch topicset and the VLIS lexical database.

Acknowledgements

We thank XRCE Grenoble for making the Xelda morphological toolkit available to us. We thank Van Dale Data to make an extended version of VLIS available which includes translations into Italian. Furthermore we would like to thank Michel Simard (RALI, Université de Montréal) for helping with the construction of aligned corpora and building translation models. We also thank George Foster and Jian-Yun Nie (also RALI) for general discussions about the application of statistical translation models for CLIR.

References

[1] Martin Braschler and Peter Schäuble. A language-independent approach to european text retrieval. In Carol Peters, editor, *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*, 2001, to appear.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

[3] M. Franz, J.S. McCarley, and S. Roukos. Ad hoc and multilingual information retrieval at IBM. In Ellen Voorhees and Donna Harman, editors, *The Seventh Text REtrieval Conference (TREC-7)*. National Institute for Standards and Technology, 1999. Special Publication 500-242.

[4] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In Christos Nicolaou and Constantine Stephanides, editors, *Research and Advanced Technology for Digital Libraries - Second European Conference, ECDL'98, Proceedings*, number 1513 in Lecture Notes in Computer Science, pages 569–?? Springer Verlag, September 1998.

[5] Djoerd Hiemstra, Wessel Kraaij, Renée Pohlmann, and Thijs Westerveld. Twenty-one at clef-2000: Translation resources, merging strategies and relevance feedback. In Carol Peters, editor, *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*, 2001, to appear.

[6] David Hull. Stemming algorithms – a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 1996.

[7] W. Kraaij, R. Pohlmann, and D. Hiemstra. Twenty-one at TREC-8: using language technology for information retrieval. In *The Eighth Text Retrieval Conference (TREC-8)*. National Institute for Standards and Technology, 2000.

[8] Wessel Kraaij and Renée Pohlmann. Porter's stemming algorithm for Dutch. In L.G.M. Noordman and W.A.M. de Vroomen, editors, *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie*, pages 167–180, 1994.

[9] Wessel Kraaij and Renée Pohlmann. Viewing stemming as recall enhancement. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR96)*, pages 40–48, 1996.

[10] Paul McNamee and James Mayfield. A language-independent approach to european text retrieval. In Carol Peters, editor, *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*, 2001, to appear.

[11] Jian-Yun Nie, Michel Simard, and George Foster. Using parallel web pages for multi-lingual ir. In Carol Peters, editor, *Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*, 2001, to appear.

[12] J.Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts an d automatic mining of parallel texts in the web. In *Proceedings of the 22nd ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR99)*, pages 74–81, 1999.

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

[14] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[15] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, 1996.

[16] T. G. Vosse. *The Word Connection*. PhD thesis, Rijksuniversiteit Leiden, Neslia Paniculata Uitgeverij, Enschede, 1994.