Combining Evidence for Cross-Language Information Retrieval

Jaap Kamps, Christof Monz, and Maarten de Rijke

Language & Inference Technology Group, ILLC, U. of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
{kamps,christof,mdr}@science.uva.nl

Abstract. This paper describes the official runs of our team for CLEF 2002. We took part in the monolingual tasks for each of the seven non-English languages for which CLEF provides document collections (Dutch, Finnish, French, German, Italian, Spanish, and Swedish). We also conducted our first experiments for the bilingual task (English to Dutch, and English to German), and took part in the GIRT and Amaryllis tasks. Finally, we experimented with the combination of runs.

1 Introduction

In this year's CLEF evaluation exercise we participated in four tasks. We took part in the monolingual tasks for each of the seven non-English languages for which CLEF provides document collections (Dutch, Finnish, French, German, Italian, Spanish, and Swedish). We also conducted our first experiments for the bilingual task (English to Dutch, and English to German), and took part in the GIRT and Amaryllis tasks.

Our participation in the monolingual task was motivated by a number of aims. First, we wanted to refine and improve our morphological normalization tools for the languages for which we took part in CLEF 2001: Dutch, German, and Italian. Furthermore, during the 2001 evaluation exercise we found that compound splitting significantly improved retrieval effectiveness for Dutch and German [14]. However, building tools such as compound splitters is not only highly language dependent but also resource intensive. And for some languages lexical resources are hard to obtain. For this reason we also wanted to develop, and experiment with, 'zero knowledge' language independent morphological normalization tools. As an aside, the availability of different kinds of runs (such as linguistically motivated vs. zero-knowledge runs) made it possible to experiment with combinations of runs, a method which has been shown to lead to improvements in retrieval effectiveness over the underlying base runs; see e.g., [8, 1, 12, 13].

This year was the first time we participated in the bilingual task. Therefore our experiments were rather modest, the main purpose being to establish a reasonable base line for English-to-Dutch and English-to-German retrieval. We used a simple dictionary-based approach to query translation, where all possible translations of a phrase or word are considered and no attempts to disambiguate the query were made.

One of the main goals of our participation in the GIRT and Amaryllis tasks was to experiment with the keywords used in the collections. Many domain-specific collections, such as the scientific collections of GIRT and Amaryllis, contain keywords. Our strategy for CLEF 2002 was to compute the similarity of keywords based on their occurrence in the collection, and explore whether the resulting keyword space can be used to improve retrieval effectiveness.

The paper is organized as follows. In Section 2 we describe the FlexIR system as well as the approaches used for each of the tasks in which participated. Section 3 describes our official runs for CLEF 2002, and in Section 4 we discuss the results we have obtained. Finally, in Section 5 we offer some conclusions regarding our document retrieval efforts.

2 System Description

All submitted runs used FlexIR, an information retrieval system developed by the second author. The main goal underlying FlexIR's design is to facilitate flexible experimentation with a wide variety of retrieval components and techniques. FlexIR is implemented in Perl; it is built around the standard UNIX pipeline architecture, and supports many types of preprocessing, scoring, indexing, and retrieval tools, which proved to be a major asset for the wide variety of tasks in which we took part this year.

2.1 Approach

The retrieval model underlying FlexIR is the standard vector space model. All our official mono- and bilingual runs for CLEF 2002 used the Lnu.ltc weighting scheme [2] to compute the similarity between a query and a document. For the experiments on which we report in this note, we fixed *slope* at either 0.1 or 0.2; the pivot was set to the average number of unique words per document.

2.2 Morphological Normalization

Previous retrieval experiments [9] in English have not demonstrated that morphological normalization such as rule-based stemming [17] or lexical stemming [11] consistently yields significant improvements. As to the effect of stemming on retrieval performance for languages that are morphologically richer than English, such as Dutch, German, or Italian, in our experiments for CLEF 2001 we consistently found that morphological normalization does improve retrieval effectiveness [14].

Stemming/Lemmatizing. For this year's monolingual experiments the aim was to improve our existing morphological analysis for languages that we had dealt with before (i.e, Dutch, German, and Italian), and to extend it to languages that we had not dealt with before (i.e., Finnish, French, Spanish, and Swedish). Where available we tried to use a lexical-based stemmer, or lemmatizer: for French, German, and Italian we used lemmatizers that are part of TreeTagger [19]. For Dutch we used a Porter stemmer developed within the Uplift project [21]; for Spanish we also used a version of Porter's stemmer [3]. We did not have access to (linguistically informed) morphological normalization tools for Finnish or Swedish.

For the GIRT and Amaryllis task, we used TreeTagger for processing the main text. The keywords, i.e., GIRT's controlled-terms and Amaryllis' controlled vocabulary, were indexed as given, indexing the keywords or keyword-phrases as a single token.

Compound splitting. For Dutch and German, we applied a compound splitter to analyze complex words, such as, *Autobahnraststätte* (English: highway restaurant), *Menschenrechte* (English: human rights), *Friedensvertrag* (English: peace agreement), etc. In addition to these noun-noun compounds, there are several other forms of compounding, including verb-noun (e.g., German: *Tankstelle*, English: gas station), verb-verb (e.g., German: *spazierengehen*, English: taking a walk), noun-adjective (e.g., German: *arbeitslos*, English: unemployed), adjective-verb (e.g., German: *sicherstellen*, English: to secure); etc., see [6] for a more detailed overview. In last year's participation we focused on noun-noun compound splitting, but this year we tried to cover the other forms for German as well. This resulted in a much larger compound dictionary for German. Whereas last year's dictionary contained 108,489 entries, it grew up to 772,667 for this year's participation. An entry in the compound dictionary consists of a complex word and its parts, where each part is lemmatized. See [14] for further details on the actual process of compound splitting.

For retrieval purposes, each document in the collection is analyzed and if a compound is identified, both the compound and all of its parts are added to the document. Compounds occurring in a query are analyzed in a similar way: the parts are simply added to the query. Since we expand both the documents and the queries with compound parts, there is no need for compound formation [16].

Ngrams. To obtain a zero-knowledge language independent approach to morphological normalization, we implemented an ngram-based method in addition to our linguistically informed methods.

	Dutch	*Finnish*	*French*	*German*	*Italian*	*Spanish*	*Swedish*
Avg. word length	5.4	7.3	4.8	5.8	5.1	5.1	5.4
Ngram length	5	6	4	5	5	5	5

Table 1. Average word length and ngram length used for the ngram base runs.

For each of the seven non-English languages in the monolingual task we determined the average word length, and set the ngram-length to be the largest integer less than the average word length, except for Finnish, where we set the ngram-length to be 6, while the average word length is 7.3; see Table 1 for the details. For each word we stored both the word itself and all possible ngrams that can be obtained from it without crossing word boundaries. For instance, the Dutch version of Topic 108 contains the phrase *maatschappelijke gevolgen* (English: societal consequences); using ngrams of length 5, this becomes:

maatschappelijke maats aatsc atsch tscha schap chapp happe appel ppeli pelij elijk lijke gevolgen gevol evolg volge olgen

2.3 Blind Feedback

Blind feedback was applied to expand the original query with related terms. Term weights were recomputed by using the standard Rocchio method [18], where we considered the top 10 documents to be relevant and the bottom 500 documents to be non-relevant. We allowed at most 20 terms to be added to the original query. For Dutch and German, the added words are also decompounded, and the complex words and their parts are added to the query.

The text runs for the GIRT and Amaryllis tasks used blind feedback, while it was switched off for the keyword runs. To aid comparison with the monolingual runs, the same feedback settings were used. There is a remarkable difference in the effect of feedback: virtually no words are added for the GIRT and Amaryllis tasks.

2.4 Combined Runs

In addition to our morphological interests we also wanted to experiment with combinations of (what we believed to be) different kinds of runs in an attempt to determine their impact on retrieval effectiveness. More specifically, for each of the languages for which we had access to language specific morphological normalization tools (i.e., stemmers or lemmatizers), we created a base run using those tools. In addition, we used ngrams in the manner described above to create a second base run. We then combined these two base runs in the following manner. First, we normalized the retrieval status values (RSVs), since different runs may have radically different RSVs. For each run we re-ranked these values in $[0.5, 1.0]$ using:

$$RSV_i' = 0.5 + 0.5 \cdot \frac{RSV_i - min_i}{max_i - min_i}$$

and assigned all documents not occurring in the top 1000, the value 0.5; this is a variation of the Min_Max_Norm considered in [13].[1] Next, we assigned new weights to the documents using a linear interpolation factor λ representing the relative weight of a run:

$$RSV_{new} = \lambda \cdot RSV_1 + (1 - \lambda) \cdot RSV_2.$$

For $\lambda = 0.5$ this is similar to the simple (but effective) summation function used by Fox and Shaw [8], and later by Belkin et al. [1] and Lee [12, 13]. The interpolation factors λ were obtained from experiments on the CLEF 2000 and 2001 data sets (where available).

For the GIRT and Amaryllis task, we created alternative base runs based on the usage of the keywords in the collection, and combined these with the text-based runs.

[1] We also conducted pre-submission experiments with a product combination rule, for which our normalization yielded better results than the standard normalization of [13]. For the combination method used, the $[0.5, 1]$ normalization is identical to the standard $[0, 1]$ normalization.

Run	*Language*	*Type*	*Factor*
UAmsC02DuDuNGiMO	Dutch	Ngram/Morphological	0.71
UAmsC02DuDuNGram	Dutch	Ngram	–
UAmsC02FiFiNGram	Finnish	Ngram	–
UAmsC02FrFrNGiMO	French	Ngram/Morphological	0.60
UAmsC02GeGeLC2F	German	Morphological	–
UAmsC02GeGeNGiMO	German	Ngram/Morphological	0.285
UAmsC02GeGeNGram	German	Ngram	–
UAmsC02ItItNGiMO	Italian	Ngram/Morphological	0.25
UAmsC02SpSpNGiSt	Spanish	Ngram/Morphological	0.70
UAmsC02SwSwNGram	Swedish	Ngram	–

Table 2. Overview of the monolingual runs submitted. For combined runs column 3 gives the base runs that were combined, and column 4 gives the interpolation factor λ.

3 Runs

We submitted a total of 27 runs: 10 for the monolingual task, 7 for the bilingual task, and 5 each for the GIRT and Amaryllis tasks. Below we discuss our runs in some detail.

3.1 Monolingual Runs

All our monolingual runs used the title and description fields of the topics. Table 2 provides an overview of the runs that we submitted for the monolingual task. The third column in Table 2 indicates the type of run:

- *Morphological* — topic and document words are lemmatized and compounds are split (Dutch, German), using the morphological tools described in Section 2.
- *Ngram* — both topic and document words are ngram-ed, using the settings discussed in Section 2.
- *Combined* — two base runs are combined, an ngram run and a morphological run, using the interpolation factor λ given in the fourth column.

Both topics and documents were stopped. First of all, for each language we used a stop phrase list containing phrases such as 'Find documents that discuss ...'; stop phrases were automatically removed from the topics. We then stopped both topics and documents using the same stop word list. We determined the 400 most frequent words, then removed from this list content words that we felt might be important despite their high frequency. For instance, in most of the document collections terms such as 'Europe' and 'dollar' occur with high frequency. We did not use a stop ngram list, but in our *ngram* runs we first used a stop *word* list, and then ngram-ed the topics and documents. For the ngram runs we did not replace diacritic letters by their non-diacritic counterparts, for the morphological runs we did.

3.2 The Bilingual Task

We submitted a total of 7 bilingual runs, using English as the topic language, and Dutch and German as document languages.

Run	*Topics*	*Documents*	*Type*	*Factor*
UAmsC02EnDuMorph	English	Dutch	Morphological	–
UAmsC02EnDuNGiMO	English	Dutch	Ngram/Morphological	0.71
UAmsC02EnDuNGram	English	Dutch	Ngram	–
UAmsC02EnGeLC2F	English	German	Morphological 1	–
UAmsC02EnGeMOiMO	English	German	Morphological/Morphological 2	0.50
UAmsC02EnGeNGiMO	English	German	Ngram/Morphological 1	0.285
UAmsC02EnGeNGram	English	German	Ngram	–

Table 3. Overview of the bilingual runs submitted.

For the bilingual runs, we followed a dictionary-based approach. The translations of the words and phrases of the topic are simply added to the query in an unstructured way; see [15] for a more elaborated way of query formulation. The original queries are translated to Dutch using the Ergane dictionary [7], and to German using the Ding dictionary [5], version 1.1. The Ergane dictionary contains 15,103 English head words and 45,068 translation pairs in total. The Ding dictionary contains 103,041 English head words and 145,255 translation pairs in total.

Since the Ergane dictionary is rather small, we used a pattern-based approach to extend the translation dictionary with additional translation pairs. Table 4 shows some of the patterns. Notice that the vast majority of the words that match one or more of these patterns are words that are derived from Latin. If an English word was not in the Ergane dictionary each matching pattern was applied and all translations were added to the query. Of course, this rather ad-hoc approach to translation is far from perfect. For instance, *privatization* will be translated as *privatisering* (correct), by applying pattern (5), and *privatisatie* (incorrect), by applying pattern (6). Although this is unacceptable for machine translation applications, those erroneous translations have virtually no im-

Patterns	*Example Translation Pairs*	
	English	*Dutch*
(1) `s/acy$/atie/`	democracy	democratie
(2) `s/ency$/entie/`	urgency	urgentie
(3) `s/ency$/ens/`	tendency	tendens
(4) `s/([aeiou])ssion$/$1ssie/`	commission	commissie
(5) `s/zation$/sering/`	privatization	privatisering
(6) `s/zation$/satie/`	realization	realisatie
(7) `s/ation$/atie/`	relation	relatie
(8) `s/ical$/isch/`	medical	medisch
(9) `s/ical$/iek/`	identical	identiek
(10) `s/idal$/idaal/`	suicidal	suicidaal
(11) `s/ic$/iek/`	specific	specifiek
(12) `s/([gmr])y$/$1ie/`	industry	industrie
(13) `s/ty$/teit/`	university	universiteit
(14) `s/ism$/isme/`	realism	realisme

Table 4. Patterns to extend the English-Dutch dictionary.

pact on retrieval effectiveness, because allmost all of them are non-existing words that do not occur in the inverted index anyway.

Just like our Dutch and German monolingual runs, we prepared morphological and ngram-based runs, and combined these in order to improve effectiveness; see Table 3 for the details.

3.3 The GIRT and Amaryllis Tasks

As pointed out in Section 1, our strategy for the GIRT and Amaryllis tasks in CLEF 2002 was to compute the similarity of keywords based on their occurrence in the collection, and investigate whether the resulting keyword space can be used to improve retrieval effectiveness. We assumed that keywords that are frequently assigned to the same documents, will have similar meaning. We determined the number of occurrence of keywords and of co-occurrences of pairs of keywords used in the collection, and used these to define a distance metric. Specifically, we used the Jaccard similarity coefficient on the log of (co)occurrences, and used 1 minus the Jaccard score as a distance metric [10]. For creating manageble size vectors for each of the keywords, we reduced the matrix using metric multi-dimensional scaling techniques [4]. For all calculations we used the best approximation of the distance matrix on 10 dimensions. This resulted in a 10-dimensional vector for each of the 6745 keywords occurring in the GIRT collection. The Amaryllis collection uses a much richer set of 125360 keywords, which we reduced by selecting the most frequent ones; this resulted in vectors for 10274 keywords occurring $\geq$ 25 times in the collection. For our official CLEF runs we experimented with these keywords spaces for two specific purposes: keyword recovery and document re-ranking.

We used the following strategy for determining vectors for the documents and for the topics: we took the top 10 documents from a base run (not using the keywords). For each of these documents we collected the keywords, and determined a document vector by taking the mean of the keyword vectors. Next, we determined a vector for the topic by taking the weighted mean of the vectors for the top 10 documents. For document re-ranking, we simply re-ranked the documents retrieved in the base run by the distance between the document and topic vectors. For keyword recovery, we considered the keywords used in the top 10 document, and selected the ten keywords that are closest to the topics vector. Table 5(a) shows the keywords recovered for GIRT topic 51.

For the Amaryllis task, we can compare the provided topic keywords in the narrative field (shown in Table 5(b)), with the topic keywords resulting from our automatic keyword recovery (shown in Table 5(c)). The recovered keywords are subsequently used in a keyword-only run.

For the GIRT task, we submitted three monolingual runs and two bilingual (English to German) runs. All our GIRT runs use the title and description fields of the topics. The morphological base run mimics the settings of our monolingual morphological base run for German. Based on the top 10 documents from the base run, we use the keyword space for recovering keywords for the topics as discussed above. The topic vector based on the top 10 documents of the base run is also used for re-ranking the documents retrieved in our base run. Experimentation on topics of CLEF 2000 and CLEF 2001 revealed that the keyword and re-rank runs perform worse than the base text run, yet a

Selbstbewußtsein	*Concentration et toxicité des polluants*	*Qualité air*
familiale Sozialisation	*Mécanisme de formation des polluants*	*Moteur diesel*
Junge	*Réduction de la pollution*	*Trafic routier urbain*
Adoleszenz	*Choix du carburant*	*Autobus*
Subkultur	*Réglage de la combustion*	*Azote oxyde*
Erziehungsstil	*Traitement des gaz d'échappement*	*Exposition professionnelle*
soziale Isolation	*Législation et réglementation*	*Véhicule à moteur*
Marginalität		*Carburant diesel*
Bewußtseinsbildung		*Inventaire source pollution*
Pubertät		*Carburant remplacement*
(a) GIRT topic 51 (recovered)	(b) Amaryllis topic 1 (monolingual, given)	(c) Amaryllis topic 1 (bilingual, recovered)

Table 5. Keywords for the GIRT and Amaryllis tasks.

combination of the base run with either a keyword or a re-rank run helps to improve the performance. Our runs for the bilingual GIRT task (English topics) used the translation method of the German bilingual task (using the *Ding* dictionary) for translation of the title and description fields. For the rest, the bilingual runs mimic the monolingual runs. We made a base morphological run, and recovered keywords for a keyword-only run and a document re-ranking; see Table 6 for the details.

Run	*Task*	*Topics*	*Documents*	*Type*	*Factor*
UAmsC02GeGiTT	GIRT	German	German	Morphological	–
UAmsC02GeGiTTiKW	GIRT	German	German	Morphological/Keyword	0.70
UAmsC02GeGiTTiRR	GIRT	German	German	Morphological/Re-rank	0.60
UAmsC02EnGiTTiKW	GIRT	English	German	Morphological/Keyword	0.70
UAmsC02EnGiTTiRR	GIRT	English	German	Morphological/Re-rank	0.60
UAmsC02FrAmTT	Amaryllis	French	French	Morphological	–
UAmsC02FrAmKW	Amaryllis	French	French	Keyword	–
UAmsC02FrAmTTiKW	Amaryllis	French	French	Morphological/Keyword	0.70
UAmsC02EnAmTTiKW	Amaryllis	English	French	Morphological/Keyword	0.70
UAmsC02EnAmTTiRR	Amaryllis	English	French	Morphological/Re-rank	0.60

Table 6. Overview of the runs submitted for the GIRT and Amaryllis tasks.

For the Amaryllis task, we submitted three monolingual runs and two bilingual (English to French) runs. Our morphological base run uses the same settings as the monolingual French run. For the keyword-only run, keywords were taken from the narrative fields of the topics. For the bilingual Amaryllis task, we used Systran [20] to translate the title and description fields of the English topics. We did not use the provided English keywords, nor the special dictionary provided. We made a morphological base run (similar to the monolingual task), and collected the keywords from the top 10 documents, which were then used for determining a document re-ranking and for keyword recovery; again, see Table 6 for the details.

4 Results

This section summarizes the results of our CLEF 2002 submissions.

4.1 Monolingual Results

Table 7 contains our non-interpolated average precision scores for all languages. In addition to the scores for our submitted runs, the table also lists the scores for the base runs that were used to generate the combined runs.

	Dutch	*Finnish*	*French*	*German*	*Italian*	*Spanish*	*Swedish*
Morphological	*0.3673*	–	*0.4063*	0.4476	*0.4285*	*0.4370*	–
Ngram	0.4542	**0.3034**	*0.4481*	0.4177	*0.3672*	*0.4512*	**0.4187**
Combined Ngrm./Mrph.	**0.4598**	–	**0.4535**	**0.4802**	**0.4407**	**0.4734**	–
	(+1.2%)		(+1.2%)	(+7.3%)	(+2.8%)	(+4.9%)	

Table 7. Overview of non-interpolated average precision scores for all submitted monolingual runs and for the underlying base runs. Best scores are in boldface; base runs that were not submitted are in italics. The figures in brackets indicate the improvement of the combined run over the best underlying base run.

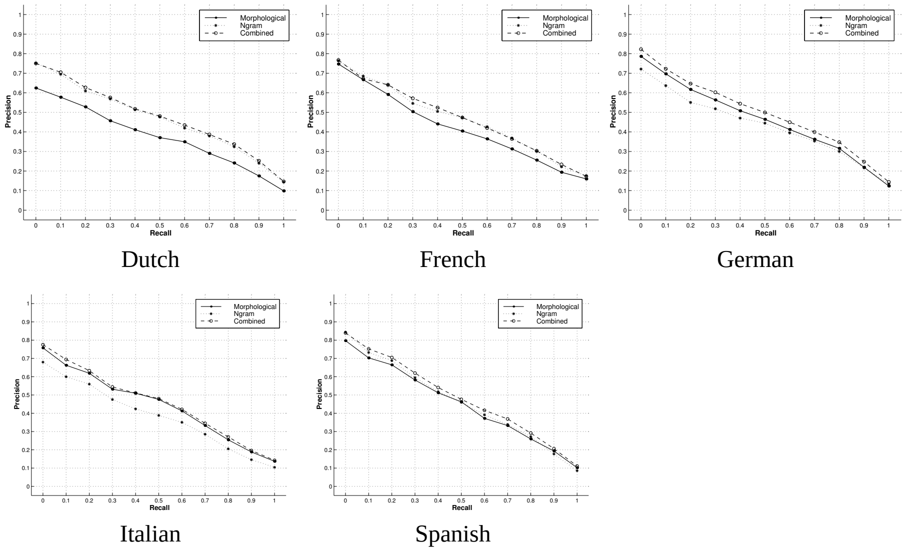
We were somewhat surprised by the low scores of our morphological run for Dutch (0.3673) and of the ngram run for Italian (0.3672). The former is probably due to the fact that we used a reasonably crude stemmer, instead of a proper lemmatizer; the latter may be due to the fact that we did not replace diacritic characters by the corresponding non-diacritic letters.

Observe that for all languages for which we submitted combined runs, the combined run outperforms the underlying base runs; in some cases the differences do not seem to be significant, but in others they do. Figure 1 displays the interpolated precision-recall curves for all languages for which we submitted combined runs. The superior performance of the combined runs can again be observed here. Several authors have proposed the following rationale for combining (high quality) runs: one should maximize the overlap of relevant documents between base runs, while minimizing the overlap of non-relevant documents. Lee [12] proposed the following coefficients $R_{overlap}$ and $N_{overlap}$ for determining the overlap between two runs run_1 and run_2:

$$R_{overlap} = \frac{R_{common} \times 2}{R_1 + R_2} \qquad N_{overlap} = \frac{N_{common} \times 2}{N_1 + N_2},$$

where R_{common} (N_{common}) is the number of common relevant (non-relevant) documents, and R_i (N_i) is the number of relevant (non-relevant) documents in run_i. (A document is relevant if, and only if, it receives relevance score equal to 1 in the qrels provided by CLEF.) Table 8 shows the overlap coefficients for the base runs used to produce combined runs.

A few comments are in order. First, for French and Spanish the base runs are of similar (high) quality, but because the $N_{overlap}$ coefficient is high, the combinations do



Dutch French German

Italian Spanish

Fig. 1. 11pt interpolated average precision for all combined monolingual runs, and the underlying base runs.

not improve all that much. Furthermore, we conjecture that the reason for the limited gains of the combined runs over the best base runs for Dutch and Italian is due to the somewhat low quality of one of the base runs for these languages. Finally, the significant improvement obtained by combining the two German base runs may be explained as follows: both base runs are high quality runs, their $R_{overlap}$ coefficient is high, and their $N_{overlap}$ is fairly low — under these circumstances, Lee's rationale predicts that the combined run is of high quality.

	Dutch	*French*	*German*	*Italian*	*Spanish*
$R_{overlap}$	0.9359	0.9606	0.9207	0.9021	0.9172
$N_{overlap}$	0.4112	0.5187	0.4180	0.4510	0.5264

Table 8. Degree of overlap among relevant and non-relevant documents for the base runs used to form the combined ngram/morphological runs for the monolingual task. The coefficients are computed over all topics).

4.2 Bilingual Results

After we had received our results from CLEF, it emerged that one of the base runs submitted for the English to German task (UAmsC02EnGeLC2F) was not the correct one. As a consequence, the combinations in which this base run was used were also incorrect (UAmsC02EnGeNGiMO and UAmsC02EnGeMOiMO). The results and figures below have been obtained with the *correct* version of UAmsC02EnGeLC2F, using the qrels provided by CLEF.

To begin, Table 9 shows our non-interpolated average precision scores for both bilingual sub tasks: English to Dutch and English to German. For English to Dutch, we submitted one morphological run, where both stemming and compound splitting were applied. For English to German, we created two morphological runs, one with a large decompounding lexicon (*Morphological 1*), and one with last year's settings, i.e., a smaller decompounding lexicon (*Morphological 2*). For both target languages we also submitted a single n-gram run. In addition, we combined the n-gram run with the morphological run for both languages, and for German we also combined the two morphological runs.

	English to Dutch	*English to German*
Morphological 1	0.2576	0.3363
Morphological 2	–	*0.3094*
Ngram	0.2807	0.2614
Combined Ngram/Morp. 1	**0.2933** (+4.5%)	**0.3514** (+4.5%)
Combined Morph. 1/Morph. 2	–	0.3451 (+2.6%)

Table 9. Overview of non-interpolated average precision scores for all correct bilingual runs. Best scores are in boldface. The figures in brackets indicate the improvement of the combined run over the best underlying base run.

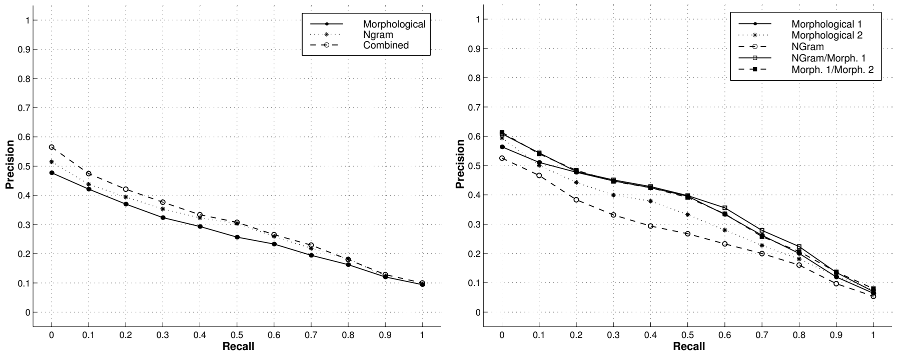
Table 10 shows the decrease in effectiveness compared to the best monolingual run for the respective target language.

	Dutch	*German*
Best monolingual	0.4598	0.4802
Best bilingual	0.2933 (−36.2%)	0.3514 (−26.8%)

Table 10. Decrease in effectiveness for bilingual runs.

If we consider the difference in retrieval effectiveness between monolingual and bilingual, we can observe a significant difference between Dutch and German. It is very likely that this is due to the difference in size between the translation dictionaries that were used to formulate the target queries: the Dutch translation dictionary contained 15,103 head words plus translation rules, whereas the German dictionary contained 103,041 head words; see Section 3.

As with the monolingual runs, we also analyzed the overlap coefficients for base runs that were combined; see Table 11. The gains in effectiveness of the combination over the best base runs is consistent with the coefficients, with comparable gains for the ngram/morphological combinations for Dutch and German; note that both have a fairly low $N_{overlap}$ coefficient. The two (German) morphological runs share many non-



Bilingual English to Dutch

Bilingual English to German

Fig. 2. 11pt interpolated average precision for all correct bilingual runs.

relevant documents, and as a consequence the combination of these two runs is less effective than the combination of the ngram run with the morphological 1 run.

	English to Dutch	*English to German*	*English to German*
	Ngram/Morphological	*Ngram/Morphological 1*	*Morphological 1/Morphological 2*
$R_{overlap}$	0.7737	0.7898	0.9338
$N_{overlap}$	0.2516	0.3588	0.5853

Table 11. Degree of overlap among relevant and non-relevant documents for the base runs used to form the combined bilingual runs. The coefficients are computed over all topics).

4.3 Results for the GIRT and Amaryllis Tasks

Table 12 contains our non-interpolated average precision scores for the GIRT and Amaryllis tasks. In addition to the scores for our submitted runs, the table also lists the scores for the base runs that were used to generate the combined runs.

The results for the GIRT tasks are outright disappointing. Our morphological base run fails to life up to the performance of the corresponding monolingual German runs (average precision 0.1639 for GIRT versus 0.4476 for German). On our pre-submission experiments on the GIRT topics of CLEF 2000 and CLEF 2001, we also noticed a drop in performance, but far less dramatic than for the CLEF 2002 run (average precision around 0.31 for both runs versus 0.1639 this year). Still, the combination of the morphological run with either the keyword run or re-rank run improves retrieval effectiveness. For the English to German GIRT task, only the combination of the morphological and re-rank base runs improves compared to the base runs; this may be due to the extremely low precision at 10 of the bilingual base run (0.1417).

Our runs for Amaryllis are more in line with the results for the monolingual French task (average precision 0.2681 for the base run versus 0.4063 for French). The keyword-

	GIRT (mono)	*GIRT (bi)*	*Amaryllis (mono)*	*Amaryllis (bi)*
Morphological	0.1639	*0.0666*	0.2681	*0.2325*
Keyword	*0.0349*	*0.0210*	0.2684	*0.0890*
Re-rank	*0.1015*	*0.0405*	–	*0.1029*
Combined Mrph./KW	0.1687 (+2.9%)	0.0620 (−6.9%)	**0.3401** (+26.7%)	**0.2660** (+14.4%)
Combined Mrph./RR	**0.1906** (+16.3%)	**0.0704** (+5.7%)	–	0.2537 (+9.1%)

Table 12. Overview of non-interpolated average precision scores for all submitted GIRT and Amaryllis runs, and for the underlying base suns. Best scores are in boldface; base runs that were not submitted are in italics. The figures in brackets indicate the improvement of the combined run over the best underlying base run.

only run using the provided keywords even out-performs the morphological base run. The combination of the two runs leads to an impressive improvement in retrieval effectiveness (+26.7%). The English to French Amaryllis task performs fairly well compared to the monolingual Amaryllis task. The combination runs of the morphological base run with the recovered keywords, and of the morphological base run with the re-ranking show significant improvement.

Figure 3 contains precision-recall plots for the GIRT and Amaryllis tasks. In addition to the scores for our submitted runs, the figure also plots the scores for the base runs that were used to generate the combined runs.

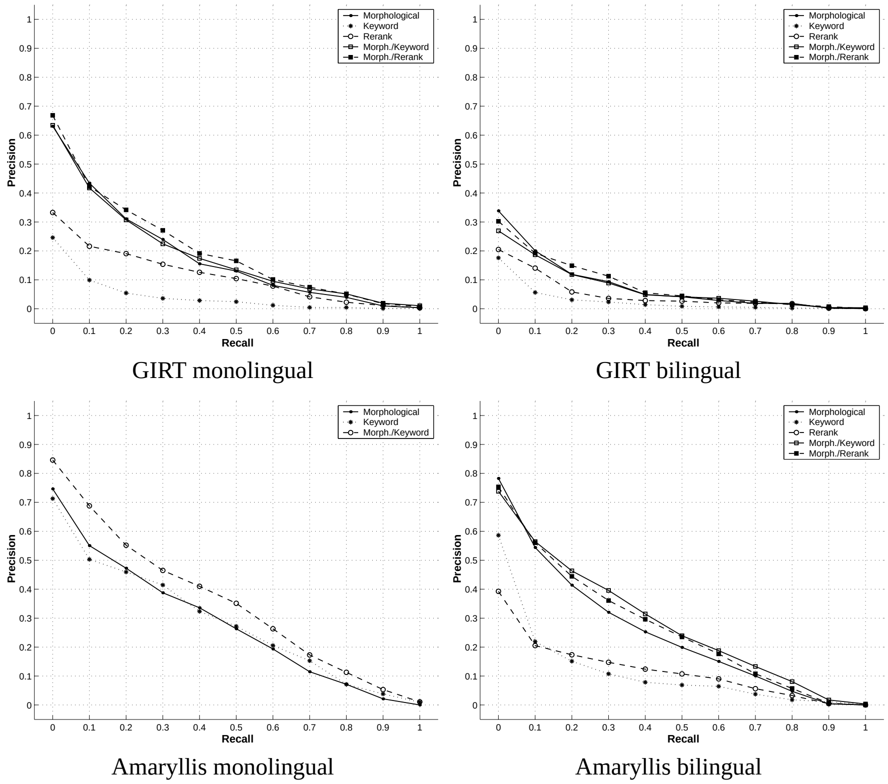
	GIRT (mono)	*GIRT (bi)*	*Amaryllis (mono)*	*Amaryllis (bi)*
$R_{overlap}$	0.4493	0.2984	0.6586	0.6506
$N_{overlap}$	0.1031	0.0756	0.1236	0.1301

Table 13. Degree of overlap among relevant and non-relevant documents for the base runs used to form the combined morphological/keyword runs for the GIRT and Amaryllis tasks. The coefficients are computed over all topics).

As with the other tasks, we analyzed the overlap coefficients for base runs that were combined; see Table 13. As expected, gains in effectiveness are due to a high $R_{overlap}$ coefficient combined with a relatively low $N_{overlap}$ coefficient. It is interesting to note that the coefficients for the combined monolingual Amaryllis runs (using the provided keywords) are similar to those of the bilingual runs (using the recovered keywords). This may provide a partial explanation of why the combination of a base run with a much lower quality run can still improve retrieval effectiveness.

5 Conclusions

The experiments on which report in this not indicate a number of things. First, morphological normalization does improve retrieval effectiveness significantly, especially for languages such as Dutch and German, that have a more complex morphology than English. We also showed that ngram-based retrieval can be a viable option in the absence of linguistic resources to support deep morphological normalization. Further-



GIRT monolingual

GIRT bilingual

Amaryllis monolingual

Amaryllis bilingual

Fig. 3. 11pt interpolated average precision for all submitted GIRT and Amaryllis runs, and the underlying base runs.

more, combining runs provides a method that can consistently improve base runs, even high quality base runs; moreover, the interpolation factors required for the best gain in performance seem to be fairly robust across topics sets. Finally, our results for the bilingual task indicate that simple word/phrase translation, where all possible translations are used to formulate the target query in an unstructured way, leads to a significant decrease in effectiveness, when compared to the respective monolingual runs. Therefore, we plan to investigate more restrictive ways of formulating target queries.

Acknowledgments

We want to thank Willem van Hage and Vera Hollink for their technical support, and Maarten Marx for useful discussions.

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grant # 400-20-036). Christof Monz was supported by the Physical Sciences Council with financial support from the Netherlands Organization for Scientific Research (NWO), project 612-13-001. Maarten de Rijke was supported by grants from the Netherlands Organization for Scientific Research (NWO), under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, and 612.000.207.

References

1. N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw. Combining evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448, 1995.
2. 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.
3. CLEF resources at the University of Neuchâtel. `http://www.unine.ch/info/clef`.
4. T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, London UK, 1994.
5. Ding: A dictionary lookup program. `http://ding.tu-chemnitz.de`.
6. G. Drosdowski, editor. *Duden: Grammatik der deutschen Gegenwartssprache*. Dudenverlag, fourth edition, 1984.
7. Ergane: a free multi-lingual dictionary programme. `http://download.travlang.com/Ergane/frames-en.html`.
8. E.A. Fox and J.A. Shaw. Combination of multiple searches. In *Proceedings TREC-2*, pages 243–252, 1994.
9. W. Frakes. Stemming algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Strcutures & Algorithms*, pages 131–160. Prentice Hall, 1992.
10. J. C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 1986.
11. D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42:7–15, 1991.
12. J.H. Lee. Analyses of multiple evidence combination. In *Proceedings SIGIR'97*, pages 267–276, 1997.
13. J.H. Lee. Combining multiple evidence from different relevant feedback methods. In *Database Systems for Advanced Applications*, pages 421–430, 1997.
14. 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 CLEF 2001*, LNCS 2406, pages 262–277. Springer Verlag, 2002.
15. A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4(3–4):209–230, 2001.
16. R. Pohlmann and W. Kraaij. Improving the precision of a text retrieval system with compound analysis. In J. Landsbergen, J. Odijk, K. van Deemter, and G. Veldhuijzen van Zanten, editors, *Proceedings of the 7th Computational Linguistics in the Netherlands Meeting (CLIN 1996)*, pages 115–129, 1996.
17. M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
18. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System — Experiments in Automatic Document Processing*. Prentice Hall, 1971.
19. H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.
20. Systran Online Translator. `http://www.systransoft.com/`.
21. UPLIFT: Utrecht project: Linguistic information for free text retrieval. `http://www-uilots.let.uu.nl/˜uplift/`.