This is a preprint of the publication intended for rapid communication of the results:
Lehtokangas, R. & Keskustalo, H. & Järvelin, K. (2008). Experiments with Transitive Dictionary Translation and Pseudo-Relevance Feedback Using Graded Relevance Assessments. Journal of the American Society for Information Science and Technology (JASIST) 59(3), 476-488.
Please refer to the final published version when citing this work.

Experiments with Transitive Dictionary Translation and Pseudo-Relevance Feedback Using Graded Relevance Assessments

Raija Lehtokangas[a], Heikki Keskustalo, and Kalervo Järvelin

Department of Information Studies, FIN-33014 University of Tampere, Finland

Raija.Lehtokangas@uta.fi, Heikki.Keskustalo@uta.fi, Kalervo.Jarvelin@uta.fi

tel. +358 3 3551 8840, +358 3 3551 8972, +358 3 3551 6953

Abstract. Research on cross-language information retrieval (CLIR) has typically been restricted to settings using binary relevance assessments. In this paper, we present evaluation results for transitive dictionary-based CLIR using graded relevance assessments in a best match retrieval environment. A text database containing newspaper articles and a related set of 35 search topics were used in the tests. Source language topics (in English, German, and Swedish) were automatically translated into the target language (Finnish) via an intermediate (or pivot) language. Effectiveness of the transitively translated queries was compared to that of the directly translated and monolingual Finnish queries. Pseudo-relevance feedback (PRF) was also used to expand the original transitive target queries. CLIR performance was evaluated on three relevance thresholds: stringent, regular, and liberal. The transitive translations performed well achieving, on the average, 85-93% of the direct translation performance, and 66-72% of monolingual performance. Moreover, PRF was successful in raising the performance of transitive translation routes originally performing worse.

[a] Corresponding author

Introduction

A lot of CLIR research has been carried out during the last years, e.g., at TREC[2], CLEF[3], and NTCIR[4]. The research is, however, mainly based on binary relevance assessments. So, there is not sufficient knowledge on how CLIR methods treat documents of various relevance levels. In this paper, we concentrate on this aspect in CLIR performance evaluation. At NTCIR, empirical results with graded relevance assessments have been presented (see, e.g., Zhou et al., 2004; Fujii & Ishikawa, 2004; Kwok, Dinstl, & Choi, 2004), but these results have not been interpreted from the point of view we have in this paper.

Using binary relevance assessments (documents are either relevant or non-relevant) ignores the fact that documents are to different degrees relevant with respect to search requests, thus considering a marginally relevant document as valuable as a highly relevant one. This is a real problem since a majority of documents relevant in a database may be only marginally relevant (Sormunen, 2002). Normally, searchers prefer documents with a higher degree of relevance. In the present information overload it is more vital than ever to be able to pick the best documents. So, degrees of relevance should be taken into account when evaluating IR systems and methods, and systems and methods able to retrieve the most valuable documents should be credited for this.

Evaluation of IR methods and systems by various relevance levels has recently become possible for two reasons. First, evaluation methods for handling graded relevance data have been developed (Järvelin & Kekäläinen, 2000; Kekäläinen & Järvelin, 2002). Secondly, test collections exist that

[2] TREC Homepage. Available: http://trec.nist.gov/
[3] CLEF Homepage. Available: http://clef.iei.pi.cnr.it
[4] NTCIR Homepage. Available: http://research.nii.ac.jp/ntcir/index-en.html

provide graded relevance assessments (Sormunen, 2000; Sormunen, 2002; Kishida et al., 2004; Lee et al., 2002; Voorhees, 2001).

In dictionary-based CLIR queries expressed in one language (*source language*) are translated into the language of the test collection (*target language*) by means of an electronic dictionary, usually accepting all the translation equivalents the dictionary gives for a word. Even though electronic dictionaries are becoming more available it may still be difficult to find a suitable dictionary for a given pair of languages. This may be true for common European languages, not to speak of more rare languages. In a situation like this, translation through an intermediate language (*pivot language*) may, however, be possible: first from the source language into the pivot language, then from the pivot language into the target language. This is called *transitive translation*. Besides sometimes being a necessity because of lacking resources, translation through a pivot language may also be advantageous even if bilingual dictionaries between (at least some of) the languages exist. Namely, if one has to perform translations between a large number of languages, i.e., when there are many source and target languages at the same time, the number of the individual translations needed can be effectively reduced using a suitable pivot language between the source and target languages. Good candidates for pivot languages are languages popular in bilingual dictionaries, such as English.

There are a few studies on transitive translation. For the most part, transitive translation studies have used dictionary-based translation. It was used by, e.g., Ballesteros (2000), Gollins and Sanderson (2001), Ballesteros and Sanderson (2003), and Lehtokangas, Airio, and Järvelin (2004). Kwok, Dinstl, and Choi (2004) used machine translation in translating queries. Kraaij and de Jong (2004), for their part, had a different approach in their study using methods based on language models. Transitive translation has often been studied in combination with other techniques. These

have been used to combat problems connected with transitive translation. Generally, one of the basic problems with dictionary translation is ambiguity (Ballesteros, 2000; Pirkola et al., 2001). Natural language words often have more than one sense. When a word is translated, most often all the senses are taken into the translated query even though not all of them are relevant. The problem is accentuated for transitive translation because of the additional translation phase needed. Gollins and Sanderson (2001) tried to solve the problem of ambiguity through triangulation, i.e., by using several translation routes. They used several pivot languages and merged the translation results from the different routes. This had a favorable effect. The overall performance of the study was low, mainly due to poor translation resources. Ballesteros (2000) reduced the ambiguity of transitive translation by query structuring and various expansion techniques. In a later study, Ballesteros and Sanderson (2003) experimented with transitive translation and triangulation, with and without synonym operator based query structuring. In most cases, triangulation had a positive effect, and this was true even when query structuring was used simultaneously. Lehtokangas, Airio, and Järvelin (2004) achieved reasonable transitive performance in comparison to direct performance by only using synonym operator based query structuring. In that study triangulation proved useful only for unstructured queries.

This paper presents novel CLIR results based on graded relevance assessments when translation of source language topics into the target language is performed via a pivot language. Our main research question is how well dictionary-based transitive CLIR is able to find documents relevant to different degrees, in particular highly relevant documents. A four-point relevance scale is used in the tests: documents in the test database are highly, fairly or marginally relevant, or non-relevant. CLIR performance is evaluated by precision and recall at three relevance thresholds: 1) *stringent* (only highly relevant documents accepted as relevant) 2) *regular* (both highly and fairly relevant documents accepted), 3) *liberal* (highly, fairly and marginally relevant documents accepted). These

three relevance thresholds were already used in our previous study (Lehtokangas, Keskustalo, & Järvelin, 2006). In that study they were used to evaluate target queries translated directly from the source language. Reasonable CLIR performance was achieved for the *liberal* and *regular* relevance threshold. For the *stringent* threshold equally high performance could not be achieved.

Moreover, we experiment with expansion of translated target queries. Query expansion (QE) means query reformulation by changing search keys (or their weights) to make the query better match relevant documents. QE has been studied extensively because selection of good search keys is difficult but crucial for good results (Efthimiadis, 1996; Kekäläinen, 1999). QE may be based on external, collection independent knowledge structures (such as thesauri), collection-dependent knowledge structures (e.g., word co-occurrence statistics) or search results. Relevance feedback (RF) is a method based on search results. In interactive RF the searcher examines retrieved documents and gives the IR system feedback at the level of (ir)relevant documents or at the level of candidate search keys extracted from top ranking documents. In *pseudo RF* (PRF) the IR system assumes the top ranking documents to consist of relevant documents and automatically, without user interaction, extracts QE keys by statistical means. In the present paper we examine PRF in enhancing queries, based on results of initial dictionary-based CLIR queries. We are particularly interested in whether PRF is capable of reducing query ambiguity due to dictionary translation and thereby enhancing the retrieval of highly relevant documents. We employ the RATF formula by Pirkola, Leppänen, and Järvelin (2002b) in the extraction of candidate QE keys from the top ranking initial results.

We evaluate CLIR performance in a laboratory setting, using a best match retrieval system (InQuery) and a test database consisting of Finnish newspaper articles. Search topics in English, German and Swedish are translated into the target language Finnish via pivot languages English or

Swedish by an automated process using morphological analyzers, machine-readable dictionaries and stopword lists (Hedlund et al., 2001). *n*-Gram techniques are applied to words that are untranslatable by the dictionaries, and target queries are structured by using the synonym operator of InQuery.

Test Design

Training and Test Collections

Our test database TUTK consists of 53,893 Finnish newspaper articles from three newspapers (Sormunen, 2000; Kekäläinen & Järvelin, 2002). Finnish is a highly inflectional language and rich in compounds (words written together as one unit), therefore a lemmatizer was used in index building. Words recognized by the lemmatizer were turned into their lemmas in the index, and in addition to this, compounds were split. Words not recognized by the lemmatizer were put into the index as such (thus typically in inflected forms). The resulting index contains about 241,000 unique recognized words (or compound components) as lemmas and about 118,000 unique unrecognized word forms. There are 35 test topics, each expressing a search request in 1-4 sentences. The themes of the topics are distributed as follows: person (5 topics), organization (12), geographical place (10), general theme (8). The topics are originally expressed in Finnish, but have been translated by professional translators into English, German, and Swedish.

Graded Relevance Assessments

A recall base for the 35 TUTK topics has been collected by extensive pooling. With respect to the 35 topics, altogether 17,338 documents have been evaluated by human assessors using a four-point relevance scale. Four relevance judges were employed, and the relevance of 20 topics was assessed by two persons, and the remaining 15 topics by one person. (Sormunen, 2000; Järvelin & Kekäläinen, 2000).

A four-point scale was used in the relevance assessments. Relevance level 0 is used to denote non-relevant documents not about the subject of the topic. Relevance level 1 denotes marginally relevant documents - documents referring to the topic but not giving more information than the topic itself. Relevance level 2 is used to denote fairly relevant documents - documents that contain some new facts with regard to the topic. Finally, relevance level 3 is used to denote highly relevant documents - documents that contain valuable information with regard to the topic. (Sormunen, 2000)

The relevance assessors agreed in 73% of the parallel assessments. In 21% of the cases the difference was one point. In the remaining 6% of the cases the difference was two or three points. Disagreements in judgments were resolved in the following way: if the difference was one point, the assessment was selected from each judge in turn. If the difference was two or three points, the researcher made the final decision about the relevance level. (Sormunen, 2000; Järvelin & Kekäläinen, 2000)

As a result of the relevance assessments, 444 documents are considered highly relevant (relevance level 3), 829 documents fairly relevant (level 2), and 993 documents marginally relevant (level 1). Thus, the recall base contains 2,266 documents evaluated as relevant for the 35 topics. The rest of

the database is considered to contain only non-relevant documents with respect to the topics (relevance level 0).

Resources Used

The retrieval system used in the experiments was InQuery (v. 3.1), a probabilistic retrieval system provided by the Center for Intelligent Information Retrieval at the University of Massachusetts (Broglio, Callan, & Croft, 1994). InQuery queries are either natural language queries (e.g., English sentences) or structured queries. Structured queries are constructed by using, e.g., the operator *syn*, which treats all of its arguments as instances of one search key. All operators are preceded by the hash sign *#*, and the arguments are delimited by parentheses, e.g., *#syn(ship vessel boat)*. If no operator is given, the operator *sum* is used as default. This treats all of its arguments as having an equal influence on the result.

Large machine-readable dictionaries, provided by Kielikone plc., Finland, were used for the translations Swedish/English/German-Finnish, Swedish-English and English-Swedish (number of entries in the dictionaries being 55,000, 110,000, 60,000 and 60,000, respectively). For the translations German-English and German-Swedish, bilingual wordlists were compiled from the dictionaries Oxford Duden German Dictionary (260,000 entries) and Norstedts Tyska Ordbok (127,000 entries) by Norstedts Ordbok AB, Sweden, respectively. For normalizing source and target language words, morphological analyzers provided by Lingsoft plc., Finland, were used. The stoplists in the respective languages had the following number of words: English 402, Finnish 737, German 637, Swedish 658.

Monolingual Queries

Monolingual queries used as a baseline of the study were automatically constructed from topics by lemmatizing their words and forming InQuery synonym sets (*#syn*).[5] If a word was not recognized by the lemmatizer, approximate string matching was applied to find the most similar strings from the target index. We used skip-grams (Pirkola et al., 2002a) for selecting the two best matching strings. Finally, stop words were removed.

As an example, after processing the Finnish topic *"OPEC:n öljyn hintaa ja tuotantomääriä koskevat päätökset" (The decisions of OPEC concerning oil prices and production levels)* the following baseline query (in InQuery syntax) was formed:

#sum(#syn(opec) #syn(n) #syn(öljy) #syn(hinta) #syn(tuotantomääri) #syn(tuotantomäärä) #syn(päätös))

In the example above, the words *OPEC*, *n* (genitive suffix), *öljyn* (inflected word form referring to *oil*), *hintaa* (inflected word form referring to *price*), *tuotantomääriä* (inflected compound word referring to *production volume*) and *päätökset* (inflected form referring to *decision*) are lemmatized successfully. (Note that the word *tuotantomääriä* generates two lemmatized word forms, *tuotantomääri* and *tuotantomäärä*.) The remaining query words are recognized as stopwords (*ja* meaning *and*, *koskevat* - inflected form referring to *related*) and are removed from the query.

Transitive Queries

[5] #syn clauses are, of course, not needed for unary arguments. This is due to using the same UTACLIR process for both monolingual and CLIR queries.

The transitive queries were automatically constructed by applying direct translation twice, i.e., by translating the topics in English, German and Swedish into target language Finnish via English or Swedish. When a source language word was untranslatable, it was not processed in the pivot phase at all, but fuzzy-matched against the target database index. This matching was based on the s-grams (a variation of n-grams) and two best matches were added to the query (Pirkola et al., 2002a). The n-gram based methods are effective for finding spelling variants between several European languages sharing words written differently but having the same origin, e.g., technical terms derived from latin or Greek, or proper names (Keskustalo et al., 2003). Yet, using n-gram based matching for all non-identified and non-translatable words also adds noise to the query (Hedlund et al., 2002).

In transitive translation, it is natural that the number of keywords increases every time a translation is performed. Table 1 presents the keyword increases in this study (stopwords were excluded from the figures at each stage). When translating the topics into the pivot languages, the number of words was increased by a factor of 3 to 4 in all the translation routes. In translating from the pivot languages into the target language, the picture was more differentiated: for the routes Swedish-English-Finnish and German-English-Finnish the growth factor was approximately 14, for the routes English-Swedish-Finnish and German-Swedish-Finnish remarkably smaller: 4. On the whole, the number of words in the target queries varied a lot from topic to topic, the maximum number of words being in one translation route 62 times as large as the minimum number (Table 2). When pivot language queries were translated into the target language, increases in the number of words varied likewise. For example, in one language route the maximum growth factor among the 35 topics was 22, the minimum 4 (Table 3).

In performing translations, not only increases the number of words but also widens the semantic area covered by the words. In our study, pivot queries could be relatively reasonable with respect to the subject area of topics, whereas target queries often were prone to cover the whole gamut of life. For example, when translating the Swedish topic "*OPEC:s beslut om priset och produktionsmängderna för olja" (The decisions of OPEC concerning oil prices and production levels)*, the resulting pivot language (English) query had the following 13 words: *opec, decision resolution, price, prize, production, work, quantity, number, crowd, oil, cook, oil* (the last two words originating from the phrase 'cooking oil'). By target query words (205), instead, meanings more surprising for this context were conveyed, e.g., *elf, zander, radar, treasure, to bulldoze, to bewitch, to cultivate.*

Query Expansion Based on Pseudo-Relevance Feedback

In our previous study (Lehtokangas, Keskustalo, & Järvelin, 2006), target language queries translated directly from source language were expanded by pseudo-relevance feedback (PRF). In selecting expansion terms we used the RATF formula (Pirkola, Leppänen, & Järvelin, 2002b).

Our PRF scenario was characterized by three variables:

1) number of documents: how many top documents are used in the expansion process (the values to be used were: 10, 20 or 30)
2) number of terms per document: how many keys are selected by the RATF formula to represent each document (the values to be used were: 20, 50 or 100)
3) number of expansion terms: how many keys are used to expand the original queries (the values to be used were: 10 or 30).

The expansion process is described in more detail in Lehtokangas, Keskustalo, and Järvelin (2006). In that study a CLEF collection and two related sets of topics were used for training purposes. Experiments were made to find the best combinations for the three variables above. According to the tests, the best value combinations were: 10_50_30 and 20_100_30 (top 10/20 documents, 50/100 RATF keys, 30 query expansion keys). These combinations were used in the final query expansion runs of the previous study, and the same combinations were used in the present study, too.

Findings

Unexpanded transitive runs

In our experiments, four transitive translation routes were used: Swedish-English-Finnish, English-Swedish-Finnish, German-English-Finnish, and German-Swedish-Finnish. The effectiveness of the transitively translated queries was compared to that of the directly translated on three relevance thresholds. Also, comparisons to the monolingual baseline were made. These results are presented in Table 4. Comparison to direct translation shows the possible loss of effectiveness if resources for direct translations are not available or are not developed. Comparison to monolingual shows the overall effectiveness.

The transitive runs performed well compared to the monolingual and direct baselines at all relevance thresholds, achieving, on the average, 66-72% of the monolingual and 85-93% of the direct performance. Compared to the directly translated queries the transitive queries performed

best on the stringent threshold: the transitive runs achieved, on the average, 93% of the baseline direct performance (differences between the translation routes were, however, large, the transitive performance ranging from 68 to 112% of the direct baseline). On the regular and liberal thresholds, the transitive runs achieved an average performance of 85% of the direct baseline. When compared to the monolingual baseline, for all the transitive routes but one, the best performance was achieved on the liberal threshold, the worst on the stringent. As an example on the transitive performance at the stringent threshold, precision-recall curves for one route (German-Swedish-Finnish) are presented in Figure 1, also giving the direct and monolingual baselines. On all the 11 recall points, the transitive run outperforms the corresponding direct run.

Expanded transitive runs

We also tried improving the results of the transitive runs by query expansion using pseudo-relevance feedback. As in our previous study (Lehtokangas, Keskustalo, & Järvelin, 2006), PRF combinations 10_50_30 and 20_100_30 (see Section *Query Expansion Based on Pseudo-Relevance Feedback* above) were selected to be used in the runs. Also here, the combination 10_50_30 gave better results, so these results are presented in Table 5 comparing the PRF expanded transitive queries to the directly translated unexpanded queries and to the unexpanded monolingual baseline. We also formed monolingual PRF queries by using the RATF based query expansion with combination 10_50_30 (10 top documents, 50 terms per document, and 30 final expansion terms were selected). The set of expansion terms were added unweighted as in case of transitive expansion and the compound words were split to make the component words searchable (Finnish is rich in compounds). A Finnish stop list was used during the process to prevent adding nonsense words as expansion terms to the queries.

For all the translation routes and relevance thresholds, query expansion had a positive effect. For all the translation routes except for the route having Swedish as source language, the expanded transitive queries outperformed the original direct queries. Also the monolingual PRF queries outperformed the monolingual baseline when liberal and regular relevance thresholds were used (on the average, by +7.4 % units and by +5.6 % units, respectively). Yet, the PRF expanded queries were slightly inferior to the unexpanded queries at the stringent threshold (by -0.6 % units).

The expansion was specially favorable for the transitive runs originally not succeeding well in comparison to the direct runs, i.e., Swedish-English-Finnish, and German-English-Finnish. The effectiveness of these improved at its best by 43% and 39%, respectively, when compared to the original unexpanded transitive runs.

When having English or German as source language, the expanded transitive runs outperformed the original direct runs. This happened at all relevance thresholds, and the effectiveness of the expanded transitive runs ranged from 103 to 120% of the corresponding direct runs. Also the runs having Swedish as source language did well, the effectiveness of the expanded transitive runs ranging from 90 to 97% of the corresponding direct runs.

Above, query expansion was used as a means of boosting transitive runs, often performing worse than direct runs. This is why the transitive runs were there compared to the unexpanded direct runs. This, indeed, proved to be a viable approach since the biggest expansion improvements were achieved in the transitive runs originally performing worse. Figure 2 presents the gains achieved for one transitive route, Swedish-English-Finnish, showing both the original and the expanded transitive run, and the direct and monolingual baseline. The expanded transitive run competes well

with the baseline direct run, outperforming it at the later recall points. In Table 5 also figures for the expanded transitive and expanded direct runs are given, thus presenting a more fair comparison between the two. Using expanded direct runs instead of unexpanded runs makes the comparison less favorable for the expanded transitive runs.

Statistical Tests

Wilcoxon signed ranks 2-tailed test was used to test the statistical significance of differences between the runs. Primarily, we used the significance level 0.01 in the tests. Transitive runs were compared to the direct or monolingual baseline. Moreover, transitive runs expanded by PRF (see Section *Expanded transitive runs*) were compared to the monolingual or direct baseline and to the unexpanded transitive runs. On the whole, quite few statistically significant differences were found. Among the transitive runs there was only one run (Swedish-English-Finnish) showing statistically significant differences with respect to the corresponding direct run. There were two transitive runs (Swedish-English-Finnish and German-English-Finnish) with significant differences to the monolingual baseline on the stringent and regular thresholds, whereas on the liberal threshold there were three (in addition to the ones above also English-Swedish-Finnish). Interestingly, when expanded transitive runs were used in the tests, differences to the direct or monolingual baselines were no longer significant.

Contrary to the cases above, significant differences were desirable between the transitive runs using original queries and transitive runs using PRF expanded queries. Significant differences were indeed found between these runs on the regular threshold for two translation routes (German-English/Swedish-Finnish) and on the liberal threshold for all the four translation routes.

Using the significance level 0.05, more significant differences were found: on the stringent threshold between the runs German-Swedish-Finnish and monolingual baseline, German-English-Finnish expanded and German-English-Finnish, German-English-Finnish expanded and monolingual; on the regular threshold between English-Swedish-Finnish and monolingual, and Swedish-English-Finnish expanded and Swedish-Finnish expanded; on the liberal threshold between Swedish-English-Finnish expanded and Swedish-Finnish expanded. In Tables 4 and 5 differences significant at the level 0.05 are indicated by * and at the level 0.01 by **.

Discussion

In this CLIR study, we used target queries translated transitively, i.e., through an intermediate language, from the source language into the target language. We wanted to find out how well these queries retrieve documents at three relevance thresholds: 1) *stringent* (only highly relevant documents accepted as relevant) 2) *regular* (both highly and fairly relevant documents accepted), 3) *liberal* (highly, fairly and marginally relevant documents accepted).

Compared to the monolingual and direct baselines, the transitive runs performed well, achieving, on the average, 66-72% of the monolingual and 85-93% of the direct performance. Interestingly, the transitive queries performed best on the stringent threshold when compared to the direct performance. When comparing to the monolingual performance, the best performance was observed on the liberal threshold.

We also experimented with expanding the transitive target queries by pseudo-relevance feedback. For all the translation routes and relevance thresholds, query expansion had a positive effect, improving the performance at its best by 43% in relation to the unexpanded run. As we have a special focus on retrieving highly relevant documents we were interested if the performance of the stringent threshold in relation to the more relaxed thresholds could be improved by using query expansion. As in our previous study where directly translated target queries were expanded (Lehtokangas, Keskustalo, & Järvelin, 2006), in the present study, too, the performance of the stringent threshold in relation to the other thresholds could not be raised by this method. Interestingly, the performance of the transitive translation routes originally performing worse (i.e., routes having English as pivot language) was easily, by using pseudo-relevance feedback, raised to the level of the originally better performing routes (i.e., routes having Swedish as pivot language).

Table 6 compiles central research that has up till now been done on transitive translation. For each study the following facts are given: mean average precision (MAP) values for the transitive retrieval results; how these values compare to the corresponding monolingual and/or direct baselines in the study; the performance levels of the monolingual and/or direct runs, if they are reported in the study. The performance of transitive translation has varied a lot, ranging from 0.4 to 40.2 MAP, the average value being 21.0. The transitive performance in comparison to the monolingual or direct baselines has ranged from 2 to 84% of monolingual, and 8 to 130% of direct performance, being, on the average, 63% of monolingual and 89% of direct performance. On the other hand, the performance levels of monolingual or direct runs have varied remarkably from study to study, being rather low in some of them. Transitive runs in the present study reached MAP values in the range of 24.0 to 29.9 on the liberal threshold, which is 64 to 80% of monolingual and 72 to 99% of direct baseline. These results compare well with what has been achieved in earlier transitive research. Up till now graded relevance assessments have only been used in few transitive translation studies. In

Kwok, Dinstl, and Choi (2004) and Kwok, Choi, and Dinstl (2005). Chinese queries were translated in both studies into Korean via English. Two relevance thresholds were used, *rigid* corresponding our regular and *relax* corresponding our liberal threshold. MAPs were roughly on the same level in Kwok, Dinstl, and Choi (2004), in Kwok, Choi, and Dinstl (2005) and in the present study (see Table 6) but the present study shows higher figures when MAPs are compared to the monolingual baseline, especially on the liberal/relax threshold. Kwok, Dinstl, and Choi (2004) employed pre-translation query expansion to the transitive runs, while Kwok, Choi, and Dinstl (2005) did not use pre-translation expansion. The latter study also discusses pivot translation stage issues. In our study, figures for *unexpanded* transitive runs are used in the comparison above.

Typical of transitive translation, word number increases in the present study are substantial. Increases were still rather moderate when translating from the source into the pivot language but were sometimes enormous when translating from the pivot into the target language. In the latter translation, differences between the translation routes in the increase of word numbers were noticeable. Routes having English as pivot language showed larger word number increases than routes having Swedish as pivot language. This, of course, goes down to the dictionaries used in the pivot-target translation, the English-Finnish dictionary obviously giving more translation equivalents for a word than the Swedish-Finnish dictionary. It was our intention to study if the word number increases were due to appearance of new word senses or due to increase of translation equivalents inside already existing senses. Unfortunately, the dictionaries that we used did not give the necessary information about word senses.

There was an interrelation between word number increases in the pivot-target translation and effectiveness values in that where there were large word number increases, i.e., routes having English as pivot language, effectiveness was hurt, and vice versa. This was seen on the stringent

threshold in particular. As the fact how much word number is increased during the translations seemed to have an effect on the transitive CLIR performance, we did a further experiment on this. For each translation route and relevance threshold, topics were divided into two groups according to word number growth factor (calculated as follows: number of words in the target query divided by the number of words in the pivot query), one group consisting of topics where growth factors were under the median value, the other group consisting of all the remaining topics. Mean average precision values were calculated for the two groups. We hypothesized that the group having smaller growth factors would have higher MAPs, and vice versa. Indeed, averaged at each relevance threshold, MAPs were lower in the under-median group but the differences between the two groups were smaller than expected. The following values of MAPs averaged over the translation routes were obtained for the two groups on the three relevance thresholds:

- *Stringent threshold*: growth under Median: MAP 20.0; growth as large as Median or over: MAP 18.0
- *Regular threshold*: growth under Median: MAP 27.3; growth as large as Median or over: MAP 24.4
- *Liberal threshold*: growth under Median: MAP 28.3; growth as large as Median or over: MAP 25.6.

When translation routes were observed individually, in three out of 12 MAP was lower in the under-median group. This suggests that other factors besides word number increase may be decisive when it comes to the performance of an individual topic, e.g., whether the treatment of some central word in the topic (e.g., proper name) has succeeded or not.

To get a better insight into issues affecting effectiveness in transitive translation, correlations between effectiveness and various word number measures were calculated. In these calculations, correlations were based either on measures for individual topics or on measures for a set of 35 topics. Correspondingly, average precision or mean average precision values were used to measure effectiveness. We used the following measures to express word numbers or their changes during different phases of the translation process: word number in topic, pivot query or target query, word number growth factor in translating from topic to pivot query, from pivot query to target query or from topic to target query. Based on individual topics, also correlations of effectiveness between transitive, direct and monolingual runs were calculated. All these correlations are presented below.

Correlations between average precision values and various word number measures based on individual topics are presented in Table 7. In most cases (61 out of 72) there was a negative correlation between average precision and word number measures. This of course was expected. Correlations varied and were rather low, 71% of the absolute values not exceeding 0.20 and the highest of them being 0.39. All this indicates that word number variation has an influence on effectiveness but on the level of individual topics the interrelation of these two is not as straightforward as when topics are studied as a whole. When the total word number measures for the topics and MAP values for the transitive runs were used instead, high negative correlations were found between MAPs and, respectively, word numbers in target queries, word number growth factors in translating from pivot to target queries and from topics to target queries (Table 8). Correlations were especially high between MAPs and word number growth factors in translating from pivot to target queries (-0.95, -0.94 and -0.98 on the respective relevance thresholds). This confirms what was earlier in this section observed about the interrelation between word number increases in the pivot-target translation and effectiveness values. Positive correlations were found between MAPs and the rest of the word number measures (word numbers in topics, word numbers

in pivot queries, and growth factors in translating from topic to pivot queries). In these word measures differences between the translation routes were rather small (see Table 1), and correlations with MAPs could be coincidental, e.g., a translation route with a verbose source language (English) for some reason performing better than a translation route with a less verbose source language (Swedish).

Correlations were also calculated between transitive, direct and monolingual effectiveness, using average precision values of individual topics (Table 9). Correlations between the runs were high, ranging between transitive and direct runs from 0.67 to 0.83, between transitive and monolingual from 0.41 to 0.78, and between direct and monolingual from 0.51 to 0.80. The high correlations between the runs involving one, two or three languages indicate that there are other factors decisive to the results than only the ones directly related to the translation process. In our previous research (Lehtokangas, Keskustalo, & Järvelin, 2006), some reasons for mismatches between topics and documents came up. Outside the translation process, the wording of topics was a typical source of mismatch. The topic might, e.g., be on a wrong level of generality, the wording of it being too general or specific with respect to relevant documents. Also, something essential might be missing in the topic or be expressed by a word not in a right form. These were problems found in direct translation but the same phenomena would undoubtedly be found in transitive translation too.

Between transitive and monolingual runs, correlation was also calculated separately for two groups, one group containing topics with under-median performance in the monolingual run and the other group containing all the remaining topics. For all but one translation route (Swedish-English-Finnish) on all the relevance thresholds correlations were considerably higher in the latter group, i.e., among the topics performing well in the monolingual case (e.g., for English-Swedish-Finnish vs. monolingual on the stringent threshold, correlations in the two groups were 0.05 and 0.81,

respectively). When wordings of the topics in the two groups were studied it was found out that there were far more proper names (denoting persons, organizations, or geographical places) in the group of topics that performs well in the monolingual case. One reason for the higher correlations in this group might be that the performance of these topics is to a great degree determined by the proper names (normally untranslated), and to a much lesser degree by the outcome of the translation process. In the other group, instead, the situation is more open, leaving the performance more dependent on the outcome of the translation process, which, for one, may be successful or unsuccessful.

Conclusion

In this paper, dictionary-based transitive CLIR was tested in a best match retrieval environment, using graded relevance assessments. A four-point relevance scale was used in the test database, which consists of Finnish newspaper articles. Topics in English, German and Swedish were translated into the target language Finnish via pivot languages English and Swedish by an automated process using morphological analyzers, machine-readable dictionaries, stopword lists, *n*-gramming of untranslatable words, and structured queries. The effectiveness of the translated queries was evaluated on three relevance thresholds: *stringent* (accepting only highly relevant documents), *regular* (accepting highly and fairly relevant documents), and *liberal* (accepting highly, fairly and marginally relevant documents).

Compared to the monolingual and direct baselines, the transitive runs performed well, achieving, on the average, 66-72% of the monolingual and 85-93% of the direct performance. Thus translation through a pivot language can be a viable approach in a situation where there are not resources for

direct translation, especially because the transitive runs of this study performed best on the stringent threshold compared to the corresponding direct runs.

Query expansion based on pseudo-relevance feedback was applied to the transitive runs. This proved favorable on the whole but especially for runs originally not performing well. By using query expansion, the performance of the originally poorly performing runs could be raised to the level of the better performing transitive runs. On the whole, the expanded transitive runs achieved 70-94% of the monolingual baseline and 78-119% of the expanded direct translation baselines.

Typical of transitive translation, number of words in queries was heavily increased during the translations. There seemed to be an interrelation between word number increases and respective performance figures but the relation was not straightforward. It is obvious that other factors besides word number growth play a part there. Length of the original source queries is also a major variable. Kwok (2001) discovered that in *direct* CLIR lengthening the source queries with free text may enhance the precision of the translated queries.

The major conclusions of this study are:

- Retrieving highly relevant documents causes problems. This is seen in absolute figures and in relation to the monolingual baseline.
- Query expansion based on pseudo-relevance feedback is favorable for poorly performing transitive runs raising their performance to the level of originally better performing transitive runs.

- There is not one single factor deciding the performance level of transitive translation but multiple factors are interacting. Among the factors are the words produced by the translation process (their quality and quantity) and the wording of topics in each source language.

Acknowledgements

The InQuery search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Atro Voutilainen and Juha Heikkilä.
FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1993.
GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft plc.
SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright (c) 1998 Fred Karlsson and Lingsoft plc.
TWOL-R (Run-time Two-Level Program): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1992.
MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone plc., Finland.

This research was funded by the Academy of Finland, under Project numbers 204978 and 1209960.

References

Ballesteros, L.A. (2000). Cross language retrieval via transitive translation. In W.B. Croft (Ed.), Advances in information retrieval: Recent research from the CIIR (pp. 203-234). Boston: Kluwer Academic Publishers.

Ballesteros, L., & Sanderson, M. (2003). Addressing the Lack of Direct Translation Resources for Cross-Language Retrieval. In D. Kraft, O. Frieder, J. Hammer, S. Qureshi, & L. Seligman (Eds.), Proceedings of the 12th International Conference on Information and Knowledge Management CIKM 2003 (pp. 147-152). New York, NY: ACM Press.

Broglio, J., Callan, J., & Croft, W.B. (1994). INQUERY system overview. In Proceedings of the TIPSTER text program (Phase I). San Francisco, CA: Morgan Kaufmann Publishers.

Efthimiadis, E.N. (1996). Query expansion. In M.E. Williams (Ed.), Annual Review of Information Science and Technology ARIST 31 (pp. 121-187). Medford, NJ: Information Today, Inc.

Fujii, A., & Ishikawa, T. (2004). Cross-Language IR at University of Tsukuba: Automatic Transliteration for Japanese, English, and Korean. In Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html

Gollins, T., & Sanderson, M. (2001). Improving cross language retrieval with triangulated translation. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 90-95). New York, NY: ACM Press.

Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., & Järvelin, K. (2002). UTACLIR@CLEF 2001 – Effects of compound splitting and n-gram techniques.. In C. Peters (Ed.), Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Lecture Notes in Computer Science, 2406 (pp. 118-136) Berlin: Springer.

Hedlund, T., Keskustalo, H., Pirkola, A., Sepponen, M., & Järvelin, K. (2001). Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. In C. Peters (Ed.), Cross-language information retrieval and evaluation. Workshop of Cross-Language Evaluation Forum CLEF 2000. Lecture Notes in Computer Science, 2069 (pp. 210-223). Berlin: Springer.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 41-48). New York, NY: ACM Press.

Kekäläinen, J. (1999). The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Tampere, Finland: University of Tampere, Department of Information Studies. Ph.D. Thesis. Acta Universitatis Tamperensis 678. Available: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf

Kekäläinen, J., & Järvelin, K. (2002). Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology, 53(13), 1120-1129.

Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., Järvelin, K. (2003) Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants: SPIRE 2003, pp. 252-265

Kishida, K., Chen, K., Lee, S., Kuriyama, K., Kando, N., Chen, H.H., Myaeng, S.H., & Eguchi, K. (2004). Overview of CLIR Task at the Fourth NTCIR Workshop. In Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html

Kraaij, W., & de Jong, F. (2004). Transitive probabilistic CLIR models. In Proceedings of RIAO 2004, Recherche d'Information Assistée par Ordinateur (pp. 69-81). Paris: Le Centre de Hautes Etudes Internationales d'Informatique Documentaire (C.I.D). Available: http://www.riao.org/sites/RIAO-2004/Proceedings-2004/papers.html

Kwok, K.L., Dinstl, N., & Choi, S. (2004). NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS. In Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html

Kwok, K.L., Choi, S., & Dinstl, N. (2005). Rich Results From Poor Resources: NTCIR-4 Monolingual and Cross-lingual Retrieval of Korean Texts Using Chinese and English. ACM Transactions on Asian Language Information Processing. Vol. 4, No. 2, June 2005, pp. 135-158. Available at http://delivery.acm.org

Lee, S., Myaeng S.H., Kim, H., Seo, J.H., Lee, B., & Cho, S. (2002). Characteristics of the Korean Test Collection for CLIR in NTCIR-3. In Working Notes of NTCIR-3, Tokyo, October 8-10, 2002. Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html

Lehtokangas, R., Airio, E., & Järvelin, K. (2004). Transitive dictionary translation challenges direct dictionary translation in CLIR. Information Processing and Management, 40(6), 973-988.

Lehtokangas, R., Keskustalo, H., & Järvelin, K. (2006). Experiments with Dictionary-Based CLIR Using Graded Relevance Assessments: Improving Effectiveness by Pseudo-Relevance Feedback. To appear in Information Retrieval 9(4), September 2006.

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Information Retrieval, 4 (3/4), 209-230. Available: http://www.info.uta.fi/tutkimus/fire/archive/dictionary_based.pdf

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.-P., & Järvelin, K. (2002a). Targeted *s*-Gram Matching: a Novel *n*-Gram Matching Technique for Cross- and Monolingual Word Form Variants. Information Research, 7(2). Available: http://informationr.net/ir/7-2/infres72.html

Pirkola, A., Leppänen, E., & Järvelin, K. (2002b). The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. Information Research, 7(2). Available: http://informationr.net/ir/7-2/infres72.html

Sormunen, E. (2000). A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Tampere, Finland: University of Tampere, Department of Information Studies. Ph.D. Thesis. Available: http://acta.uta.fi/pdf/951-44-4732-8.pdf

Sormunen, E. (2002). Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 320-330). New York, NY: ACM Press.

Voorhees, E. (2001). Evaluation by Highly Relevant Documents. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 74-82). New York, NY: ACM Press.

Zhou, Y., Qin, J., Chau, M., & Chen, H. (2004). Experiments on Chinese-English Cross-language Retrieval at NTCIR-4. In Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html

TABLE 1. Total number of words in topics (n = 35) and in pivot and target queries; increase of words by translation phases.

Language route	Word no in topics	Word no in pivot queries	Word no in target queries	GF: topics to pivot queries	GF: pivot to target queries	GF: topic to target queries
Swe-Eng-Fin	426	1298	18499	3.0	14.3	43.4
Eng-Swe-Fin	510	1962	8437	3.8	4.3	16.5
Ger-Eng-Fin	437	1818	24577	4.2	13.5	56.2
Ger-Swe-Fin	437	1750	7635	4.0	4.4	17.5

Note. GF = Word number growth factor.

TABLE 2. Average, minimum, maximum and median number of words in target queries among the topics (n = 35).

Language route	Average no of target words	Minimum no of target words	Maximum no of target words	Median no of target words
Swe-Eng-Fin	528	53	1517	379
Eng-Swe-Fin	241	49	575	219
Ger-Eng-Fin	702	28	1748	669
Ger-Swe-Fin	218	26	1097	199

TABLE 3. Average, minimum, maximum and median increase in the number of words among the topics (n = 35) in translating from pivot to target language.

Language route	Average growth factor	Minimum growth factor	Maximum growth factor	Median growth factor
Swe-Eng-Fin	13.9	3.8	21.7	13.5
Eng-Swe-Fin	4.2	2.9	7.3	4.1
Ger-Eng-Fin	13.0	3.5	18.4	13.5
Ger-Swe-Fin	4.3	2.9	6.6	4.2

TABLE 4. Mean average precisions (MAP) of the transitive target queries (n = 35), compared to the monolingual and direct baseline (differences significant at the level 0.05 are indicated by * and at the level 0.01 by **).

Stringent	MAP	Of monolingual performance (%)	Of direct performance (%)
Swe-Eng-Fin	14.0	49**	68**
Eng-Swe-Fin	23.8	84	106
Ger-Eng-Fin	16.1	57**	87
Ger-Swe-Fin	20.8	73*	112
Mean	18.7	66	93
Regular	**MAP**	**Of monolingual performance (%)**	**Of direct performance (%)**
Swe-Eng-Fin	21.8	59**	68**
Eng-Swe-Fin	27.5	75	88
Ger-Eng-Fin	24.3	66**	83
Ger-Swe-Fin	29.3	79	100
Mean	25.7	70	85
Liberal	**MAP**	**Of monolingual performance (%)**	**Of direct performance (%)**
Swe-Eng-Fin	24.0	64**	72**
Eng-Swe-Fin	28.6	76**	87
Ger-Eng-Fin	25.1	67**	83
Ger-Swe-Fin	29.9	80	99
Mean	26.9	72	85

TABLE 5. Mean average precisions (MAP) of the PRF expanded transitive target queries (n = 35) compared to the PRF expanded monolingual, plain monolingual, direct, and PRF expanded direct baselines (differences significant at the level 0.05 are indicated by * and at the level 0.01 by **).

Stringent	MAP	Of expanded monolingual performance (%)	Of monolingual performance (%)	Of direct performance (%)	Of expanded direct performance (%)
Swe-Eng-Fin	20.0	72*	70	97	87
Eng-Swe-Fin	24.8	89	87	110	93
Ger-Eng-Fin	22.2	80	78*	120	113
Ger-Swe-Fin	23.3	84	82	126	119
Mean	22.6	81	79	113	103
Regular	MAP	Of expanded monolingual performance (%)	Of monolingual performance (%)	Of direct performance (%)	Of expanded direct performance (%)
Swe-Eng-Fin	28.7	68**	78	90	78*
Eng-Swe-Fin	32.3	76*	88	103	87
Ger-Eng-Fin	33.6	79	91	115	104
Ger-Swe-Fin	34.5	81	94	118	107
Mean	32.3	76	88	107	94
Liberal	MAP	Of expanded monolingual performance (%)	Of monolingual performance (%)	Of direct performance (%)	Of expanded direct performance (%)
Swe-Eng-Fin	30.9	69**	82	92	81*
Eng-Swe-Fin	34.0	76*	90	104	87
Ger-Eng-Fin	34.8	77	93	115	101
Ger-Swe-Fin	35.4	79	94	117	103
Mean	33.8	75	90	107	93

TABLE 6. Summary of research results on transitive translation.

	Transitive performance (MAP)	Of monoling performance (%)	Of direct performance (%)	Monolingual performance (MAP)	Direct performance (MAP)
Ballesteros (2000)	12.3	44	61	27.7	20.1
Gollins & Sanderson (2001)	0.4-1.1	2-4	8-19	28.9	5.5
Ballesteros & Sanderson (2003)	10.0-17.0	-	73-130	-	10.0-17.0
Lehtokangas, Airio & Järvelin (2004)	17.5-40.2	48-84	74-108	36.1-48.0	23.5-40.5
Kraaij & de Jong (2004)	28.7-36.0	68-84	-	42.3-45.4	-
Kwok, Dinstl & Choi (2004)[1]	22.6-29.6 ** 24.0-30.8 *	61-64 ** 60-63 *	-	35.5-48.5 ** 38.3-51.6 *	-
Kwok, Choi & Dinstl (2005)[2]	12.5-29.7 ** 12.6-31.5 *	? ** ? *	-	35.5-46.6 ** 38.3-50.2 *	-
Present study	14.0-23.8 *** 21.8-29.3 ** 24.0-29.9 *	49-84 *** 59-79 ** 64-80 *	68-112 *** 68-100 ** 72-99 *	28.4 *** 36.9 ** 37.6 *	18.5-22.5 *** 29.2-31.9 ** 30.3-33.4 *

*** = stringent, ** = regular/rigid, * = liberal/relax relevance threshold

[1,2] PRF as a default.

TABLE 7. Correlations between average precisions and word number measures of individual topics.

Stringent	Average precision of topic vs.					
	Word no in topic	Word no in pivot query	Word no in target query	GF: topic to pivot	GF: pivot to target	GF: topic to target
Swe-Eng-Fin	-0.07	0.00	0.09	0.07	0.13	0.15
Eng-Swe-Fin	-0.23	-0.24	-0.24	-0.17	-0.06	-0.18
Ger-Eng-Fin	-0.22	-0.39	-0.37	-0.35	-0.18	-0.31
Ger-Swe-Fin	-0.18	-0.14	-0.13	-0.07	-0.09	-0.07
Regular	**Average precision of topic vs.**					
	Word no in topic	Word no in pivot query	Word no in target query	GF: topic to pivot	GF: pivot to target	GF: topic to target
Swe-Eng-Fin	0.16	-0.04	-0.01	-0.16	-0.11	-0.10
Eng-Swe-Fin	-0.17	-0.17	-0.13	-0.09	0.13	-0.02
Ger-Eng-Fin	-0.09	-0.23	-0.27	-0.26	-0.27	-0.28
Ger-Swe-Fin	-0.12	-0.17	-0.15	-0.18	-0.11	-0.16
Liberal	**Average precision of topic vs.**					
	Word no in topic	Word no in pivot query	Word no in target query	GF: topic to pivot	GF: pivot to target	GF: topic to target
Swe-Eng-Fin	0.16	-0.05	0.02	-0.18	-0.02	-0.09
Eng-Swe-Fin	-0.14	-0.15	-0.10	-0.07	0.18	0.03
Ger-Eng-Fin	-0.07	-0.22	-0.26	-0.27	-0.32	-0.30
Ger-Swe-Fin	-0.09	-0.20	-0.18	-0.26	-0.16	-0.25

Note. GF = Word number growth factor.

TABLE 8. Correlations between mean average precisions (MAP) of transitive runs and total word number measures for the topics (n=35).

	MAP vs.					
Relevance threshold	Word no in topics	Word no in pivot queries	Word no in target queries	GF: topics to pivot queries	GF: pivot to target queries	GF: topic to target queries
Stringent	0.83	0.79	-0.83	0.45	-0.95	-0.86
Regular	0.45	0.72	-0.80	0.62	-0.94	-0.81
Liberal	0.48	0.64	-0.88	0.50	-0.98	-0.89

Note. GF = Word number growth factor.

TABLE 9. Correlation of mean average precision between transitive and direct, transitive and monolingual, and direct and monolingual runs on three relevance thresholds.

Stringent

Source language	Transitive vs. Direct	Transitive vs. Monolingual	Direct vs. Monolingual
Swedish	0.75	0.41	0.62
English	0.76	0.74	0.69
German [a]	0.67	0.48	0.58
German [b]	0.73	0.78	0.58
All routes	0.66	0.61	0.62

Regular

Source Language	Transitive vs. Direct	Transitive vs. Monolingual	Direct vs. Monolingual
Swedish	0.75	0.43	0.51
English	0.80	0.70	0.77
German [a]	0.72	0.64	0.63
German [b]	0.82	0.68	0.63
All routes	0.76	0.61	0.64

Liberal

Source language	Transitive vs. Direct	Transitive vs. Monolingual	Direct vs. Monolingual
Swedish	0.83	0.54	0.55
English	0.81	0.71	0.80
German [a]	0.72	0.66	0.55
German [b]	0.80	0.63	0.55
All routes	0.78	0.63	0.61

[a] English as pivot language, [b] Swedish as pivot language.

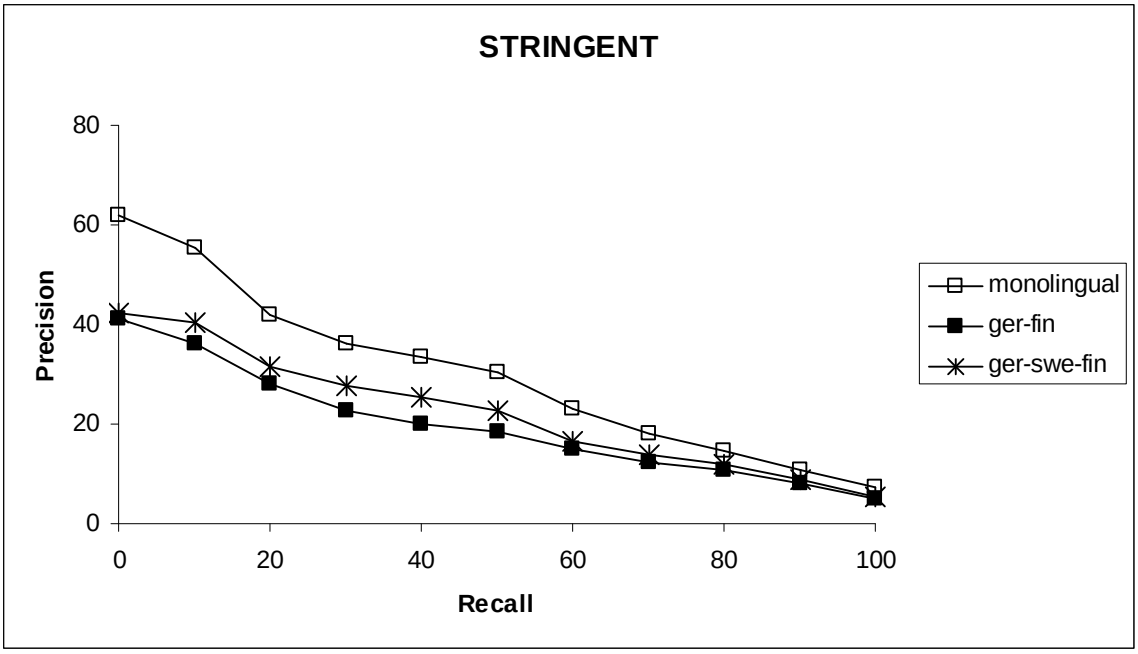

FIG. 1. Precision-recall curves for directly and transitively translated German topics on stringent relevance threshold.

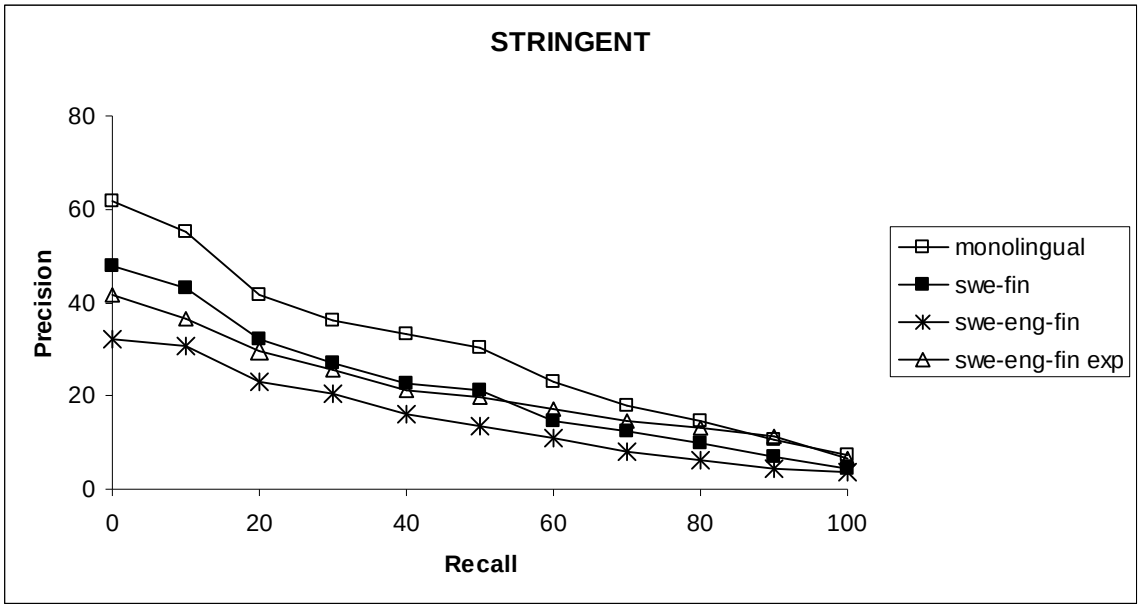

FIG. 2. Precision-recall curves for original and expanded transitive run on stringent relevance threshold.