Effects of Aligned Corpus Quality and Size in Corpus-based CLIR

Tuomas Talvensaari

Department of Computer Sciences, University of Tampere
Kanslerinrinne 1, FIN-33014, University of Tampere, Finland
tuomas.talvensaari@uta.fi

Abstract. Aligned corpora are often-used resources in CLIR systems. The three qualities of translation corpora that most dramatically affect the performance of a corpus-based CLIR system are: (1) topical nearness to the translated queries, (2) the quality of the alignments, and (3) the size of the corpus. In this paper, the effects of these factors are studied and evaluated. Topics of two different domains (news and genomics) are translated with corpora of varying alignment quality, ranging from a clean parallel corpus to noisier comparable corpora. Also, the sizes of the corpora are varied. The results show that of the three qualities, topical nearness is the most crucial factor, outweighing both other factors. This indicates that noisy comparable corpora should be used as complimentary resources, when parallel corpora are not available for the domain in question.

1 Introduction

In Cross-Language Information Retrieval (CLIR), the aim is to find documents that are written in a different language from the query. Consequently, besides the usual information retrieval (IR) issues, in CLIR one has to address the problem of crossing the language barrier. Usually, the query is translated from the *source language* into the *target language*, i.e., the language of the documents, after which a normal monolingual retrieval process can take place. The query translation approaches can be categorized according to the linguistic resources employed. The main approaches use either machine-readable dictionaries, machine translation (MT) systems, fuzzy cognate matching, multilingual corpora, or a combination of these resources [1, 2].

In approaches based on multilingual corpora, the translation knowledge is extracted statistically from the corpora used. These methods can further be categorized based on the relatedness of the corpora. A *parallel corpus* consists of document pairs that are more or less exact translations of each other. In a *comparable corpus*, the document pairs are not exact translations but have similar vocabulary [3]. The aligned documents can be, e.g., accounts of the same news event written independently in different countries. Therefore, the *alignment quality* of the corpus at hand can vary significantly – a parallel corpus and a noisy comparable corpus represent the extremes of this characteristic.

Besides alignment quality, the *topical nearness* between the corpus and the translated queries is also a significant factor. For example, a parallel corpus consisting of sports news is not likely to provide dependable translation knowledge if the queries concern quantum physics. Many of the query words would be out-of-vocabulary (OOV) for the system.

Naturally, the *size* of the corpus is also an important factor. The more aligned documents we have, the more reliable the translation knowledge is. Therefore, a large parallel corpus with good coverage of domain vocabulary would be ideal for CLIR. The availability of such corpora is often a problem, however. This is especially true with rarer languages and special domains. Consequently, noisier but more easily available comparable corpora may have to be used.

In this paper, the aim is to examine the effects of the above three factors to the performance of a CLIR system. This is done by applying translation corpora of varying alignment quality and size to retrieval topics of different domains. To our knowledge, studies where all the three factors are simultaneously experimented with, do not exist previously. The results show that although the alignment quality and size are important factors, it is essential that the corpus covers the vocabulary of the domain in question.

The rest of this paper is organized as follows. In Sec. 2, we take a look at related work done previously. Section 3 introduces the corpora used in the experiments, and also the methods that are used to utilize them. Section 4 describes the test runs and results, and Sec. 5 provides a brief conclusion.

2 Previous Work

Franz et al. [4] varied the size of various parallel corpora and found that the performance of a CLIR system based on parallel corpora is inversely proportional to the query OOV rate of the system. This would mean that the OOV rate would be a handy and easily calculated measure of system performance. Of the three qualities discussed in Sec. 1, topical nearness is the one closest related to the OOV rate. However, when comparable corpora are concerned, things are not that straightforward. It is not enough for the words just to appear in the corpus – they also have to appear frequently and the document alignments have to be of good quality.

Zhu and Wang [5] degraded a rule-based MT system by decreasing the size of the rule base and the dictionary. They found that removing dictionary entries (i.e., increasing the OOV rate) impaired the performance of the system more dramatically than removing rules. This seems to be in line with [4].

Xu and Weischedel [6] studied CLIR performance as a function of parallel corpus and dictionary size. They found that a large dictionary can compensate for a small parallel corpus and vice versa, and that the combination of these resources always performs better than either of them alone. As the size of the corpus was increased, the performance also improved, up to a point. Again, though, it remains unclear whether the same would be true when a comparable corpus is used.

3 Data and Methods

Two topic sets of different domains and languages were used in the experiments. The Swedish topics were provided by the CLEF consortium [7] and cover news events from the mid-90's. The German topics were the German translations of a TREC genomics track. The topics themselves are more closely introduced in Sec. 4. For both of the topic sets, two translation corpora were applied. Figure 1 presents the corpora in relation to their alignment quality and topical nearness to the topics.

All of the corpora and the queries were preprocessed by removing stopwords and using a lemmatizer program to normalize the words to their base form. Next, the translation corpora are introduced in more detail.

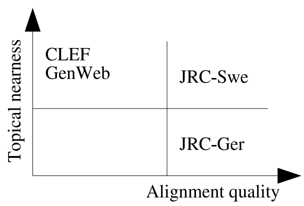

Fig. 1. Alignment quality and topical nearness of the used translation corpora

3.1 The JRC-Acquis Parallel Corpus

For both of the language pairs and topic sets, the JRC-Acquis parallel corpus [8] was applied. The corpus consists of official EU documents in all official EU languages. The alignments in the corpus were made on paragraph or sentence level, and the aligned documents were exact translations of each other – as one would expect from a parallel corpus. With respect to the German genomics topics, JRC-Acquis is topically distant (see Fig. 1), while the distance to the Swedish news topics is smaller. The ultimate measure for the topical distance is the corpus' OOV rate with respect to the queries, which is discussed in Sec. 4.

3.2 The CLEF Swedish-English Comparable Corpus

The CLEF comparable corpus consists of Swedish news-wire reports by the news agency TT from 1994-95, aligned with news articles by the Los Angeles Times from 1994. Both of the collections are provided by CLEF [7]. The CLEF corpus is topically very near the topics (Fig. 1), since it covers news events from the same period as the topics do. The alignments were created by Talvensaari et al. [9] in the following way.

Let $d_S \in C_S$ and $d_T \in C_T$ be documents in the source and target collections, respectively. The aim was to produce a set of alignments $A = \{\langle d_S, D \rangle \mid D \neq \emptyset\}$, where $D = \{d_T | sim(d_S, d_T) > \theta\}$, in other words, to map each source document to a set of target documents whose similarity with the source document exceeds some threshold θ. Each set D is called a *hyper document*. It is not realistic to expect that a satisfying counterpart for every source language document could be found. Thus, $|A| < |C_S|$.

To find the similar counterparts in the target collection, queries were first formed from each source document. Second, the queries were translated into the target language (English) with a smallish dictionary. Third, the translated queries were run against the L.A. Times corpus with the InQuery retrieval system [10]. The InQuery score was used as an indication of the similarity between the source document and the target documents. For each source document, at most 20 target documents whose similarity exceeded a score threshold were chosen into the set D.

3.3 The Genomics Web Corpus

The genomics Web corpus (GenWeb) consists of paragraphs extracted from German and English genomics-related Web pages. They were acquired by Talvensaari et al. [11] by means of focused Web crawling, which refers to the acquisition of Web pages that cover a specific topic [12]. The alignments between the German and English paragraphs were made in the same manner as for the CLEF corpus. The GenWeb corpus is topically quite near the topics that is is used to translate (see Fig. 1).

Table 1 presents the translation corpora in more detail. The alignments in the JRC corpora are all 1-to-1, that is, each source document is aligned to exactly one target document.

Table 1. The translation corpora in detail

Corpus	Source language	$\|A\|$	Avg. $\|D\|$	Unique target documents	Avg. source document length (words)	Avg. target document length (words)
GenWeb	German	39,143	6.5	39,190	114	139
CLEF	Swedish	12,579	4.3	7,732	183	421
JRC-Ger	German	282,417	1	282,417	21	23
JRC-Swe	Swedish	277,735	1	277,735	21	24

3.4 COCOT: Employing the Corpora

Cocot, a Comparable Corpus Translation program [9], uses the aligned corpus as a *similarity thesaurus*, which implies calculating similarity scores between a source language word and the words in the target documents. The similarity thesaurus' similarity score can be calculated by using traditional IR weighting approaches, reversing the roles of documents and words. A source language word is thought of as the query, and target language words are retrieved as the answer.

For a document d_j, in which a word t_i appears, the Cocot system calculates the weight w_{ij} as follows:

$$w_{ij} = \begin{cases} 0 & \text{if } tf_{ij} = 0 \\ \left(0.5 + 0.5 \cdot \frac{tf_{ij}}{Maxtf_j}\right) \cdot \ln\left(\frac{NT}{dl_j}\right) & \text{otherwise} \end{cases} , \quad (1)$$

where tf_{ij} is the frequency of word t_i in document d_j, $Maxtf_j$ the largest term frequency in document d_j, dl_j the number of unique words in the document. NT can be the number of unique words in the collection, or its approximation.

For a hyper document D_k in which a word t_i appears, the weight is

$$W_{ik} = \sum_{d_j \in D_k} \frac{w_{ij}}{\ln(rank_{jk} + 1)} , \quad (2)$$

where $rank_{jk}$ is the rank of the document d_j in the hyper document D_k, i.e. the rank calculated by InQuery in the alignment phase. The lower the rank, the less similar the target document is to the source document, according to InQuery. Thus, the lower rank documents can be trusted less as a source of translation knowledge. This is echoed in the equation above.

Finally we can calculate Cocot's similarity score between a word s_i appearing in the source documents, and a word t_j appearing in the target hyper documents:

$$sim(s_i, t_j) = \frac{\sum_{\langle d_k, D_k \rangle \in A} w_{ik} \cdot W_{jk}}{\|\mathbf{s_i}\| \cdot \left((1 - slope) + slope \cdot \frac{\|\mathbf{t_j}\|}{avg\_trg\_vlength}\right)} , \quad (3)$$

where A is the set of alignments, $\mathbf{s_i}$ and $\mathbf{t_j}$ are the feature vectors representing s_i and t_j, and $avg\_trg\_vlength$ the average length of the target word vectors. The formula employs the pivoted vector length normalization scheme, introduced by Singhal et al. [13]. The *slope* value is a parameter of this scheme, and its range is $[0, 1]$. The scheme was applied because standard cosine normalization favors words with short feature vectors, i.e. rare words.

When the above score is calculated between a source language word and every word appearing in the target documents, we get a rank of the target words. Table 2 shows Cocot ranks for three genomics-related German words. Score thresholding and word cut-off values (WCV) can be used as translation parameters to define Cocot's query translation behavior. For example, the parameters $WCV = 4, \theta = 6.0$ mean that for the word *dna*, the four highest ranking words would be returned, whereas, for *reparatur*, only the first word would be used as the translation. For the word *oxidativer*, no words would be returned, and the word would be effectively out-of-vocabulary.

4 Experiments

Two different topic sets and target collections were used in the experiments. Table 3 provides an overview of them. The TREC genomics collection [14] is a

Table 2. Example Cocot translations

Rank	dna		reparatur		oxidativer	
1	dna	13.8	repair	8.8	oxidative	2.3
2	sequence	8.1	damage	5.3	ros	2.3
3	base	7.4	dsb	5.2	superoxide	2.1
4	strand	7.2	excision	4.9	dismutase	1.9
5	rna	6.7	dsbs	4.4	peroxide	1.7

subset of the MEDLINE database of about 4.6 million documents. This collection was used in the German-English experiments. The 50 topics were translated into German by a molecular biologist, who is also a native speaker of German. The CLEF collection used in this study consists of newspaper articles by the Los Angeles Times, published in 1994. This collection was used in the Swedish-English experiments. The collection is the same as the target collection in the Swedish-English CLEF comparable corpus. The 7,732 documents that were part of the comparable corpus (see Tab. 1) were removed from the test database. The 70 topics were provided by CLEF in Swedish and English. The topic sets were further divided into two subsets, one for training the parameters of COCOT (i.e., θ, WCV and the slope value of Eq. 3), and one for the actual tests. InQuery [10] retrieval system was used in the experiments.

Table 3. Test collections

Collection	Source	Documents	Training topics	Test topics
TREC Genomics	MEDLINE	4,591,008	20	30
CLEF English	L.A. Times	113,005	30	40

4.1 Tests on Corpus Quality and Topical Nearness

For both of the languages pairs, COCOT was used to translate the queries with two different translation corpora, the JRC parallel corpus and a comparable corpus (GenWeb for the German queries, CLEF for the Swedish ones). Besides COCOT, we also applied Utaclir [15], a dictionary-based query translation program, in the experiments. In the German-English experiments, Utaclir used a German-English dictionary of 29,000 entries, while the Swedish-English dictionary had 36,000 entries.

Utaclir was applied to provide a baseline CLIR performance. Utaclir was also used in collaboration with COCOT. In these experiments, the concatenated output of the two programs was used as the target language query. This represents a more realistic approach – in a working CLIR system different translation approaches are likely to be used in collaboration (see, e.g., [16]). It should be noted, however, that the aim in these experiments was not to build a fully functional CLIR system, but to experiment with the qualities of aligned corpora. Consequently, performance-enhancing techniques such as pre-translation query expansion were not used.

The *title* and *need* parts of the TREC topics were used in the experiments – the longer *context* field was ignored. Table 4 presents an example topic in English; the same topic in German after stopword removal and word form normalization; and the translations provided by Utaclir and COCOT that uses the GenWeb comparable corpus. Both programs enclose multiple translation alternatives inside InQuery's syn-operator. This causes InQuery to treat the enclosed words as synonyms. This kind of query structuring reduces the translation ambiguity brought by multiple translation alternatives [17]. Initial experiments showed that this approach performed better than approaches based on weighting the query words according to their COCOT scores. OOV words were left unchanged in all of the approaches.

Table 4. From English topic to German query to English query.

English topic	DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress.
German query	dna reparatur oxidativer stress hang zusammen zusammenhang dna weg reparatur reparaturweg oxidativem stress
Utaclir	#sum(dna repair oxidativer stress hang #syn(context coherence) dna #syn(lane path road course way track channel walk) repair oxidativem stress)
COCOT ($slope = 0.6, \theta = 2.0$, WCV= 5)	#sum(#syn(dna sequence base strand rna) #syn(repair damage dsb excision dsbs) #syn(oxidative ros superoxide) stress hang zusammen zusammenhang #syn(dna sequence base strand rna) #syn(cell protein gene pathway stem) #syn(repair damage dsb excision dsbs) #syn(ner nhej dsb repair dsbs) #syn(antioxidant oxidative p4502e1 roi peroxide) stress)

Tables 5 and 6 present the results of the Swedish-English and German-English runs respectively. In addition to mean average precision and precision after 10 documents, the OOV rate and the type of translation resource is provided for each CLIR approach. Figures 2 and 3 depict the recall-precision curves of the runs.

In the German-English runs, the combined GenWeb-Utaclir approach performs significantly better than approaches based on the JRC parallel corpus. This happens because the GenWeb corpus is topically closer to the genomics topics than JRC. The OOV rate of GenWeb is surprisingly high, though, considering that GenWeb consists of genomics-related text. However, the GenWeb OOV words are mostly general, non-topical words. This is proven by the low OOV rate of GenWeb-Utaclir. In the combined approach, Utaclir, with its general-purpose dictionary, can cover much of GenWeb's OOV words. Therefore, GenWeb seems

Table 5. Results for the German-English runs, '> X' means the translation method is significantly better ($p < 0.05$) than method X according to the Friedman test (UC=Utaclir)

CLIR approach	MAP	P@10 docs	OOV %	Translation quality
JRC-Ger	0.087	0.210	35.0	Parallel
GenWeb	0.137	0.297	29.1	Comparable
UC	0.170	0.270	56.6	Dictionary
JRC-UC	0.136	0.303	29.3	Combined
GenWeb-UC	0.225 (> JRC-Ger,JRC-UC)	0.407	17.3	Combined

Table 6. Results for the Swedish-English runs, '> X' means the translation method is significantly better ($p < 0.05$) than method X according to the Friedman test (UC=Utaclir)

CLIR approach	MAP	P@10 docs	OOV %	Translation quality
JRC-Swe	0.247	0.245 (> UC)	27.6	Parallel
CLEF	0.204	0.190	15.2	Comparable
UC	0.186	0.168	24.2	Dictionary
JRC-UC	0.294 (> CLEF, UC)	0.288 (> CLEF, UC)	4.6	Combined
CLEF-UC	0.282 (> CLEF, JRC-Swe, UC)	0.293 (> CLEF,UC)	3.2	Combined

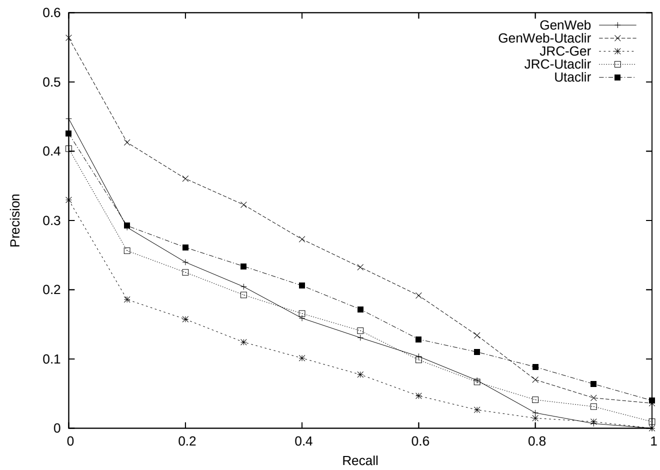

Fig. 2. Precision at 11 recall points for the German-English runs

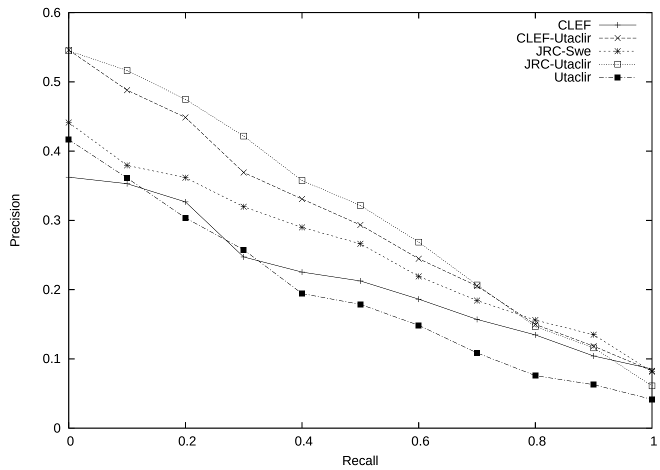

Fig. 3. Precision at 11 recall points for the Swedish-English runs

to fulfill its purpose to a certain extent. The role of a noisy comparable corpus is usually to complement other, more general, resources. Therefore, the relatively low performance of GenWeb alone is not alarming.

In the Swedish-English runs, the CLEF comparable corpus is arguably topically even closer to the topics than GenWeb to the genomics topics. The topics are news events from 1994-95, and the CLEF corpus consists of news articles from the same period. The vocabulary of the news domain, however, is much more general than genomics vocabulary. Consequently, the JRC corpus fares much better than in the German runs. When JRC is used as the sole translation corpus, it quite clearly outperforms CLEF. The combined approaches perform evenly, and they both achieve very low OOV rates.

Table 7 presents a closer look at the performance of individual queries. For each CLIR approach, the average precision of each query was compared to the median performance of the query. The table depicts the number of queries for whom average precision was significantly greater or smaller (absolute difference of 5 %) than the median, for each approach. This analysis was adopted instead of a more complete query-by-query analysis because of its compactness.

The analysis echoes the results presented earlier. It is notable that in the Swedish runs, the approach JRC-Utaclir has more queries that perform significantly below median than CLEF-Utaclir (5 against 1), although JRC-Utaclir had higher MAP. This indicates that JRC-Utaclir performed very well on some individual queries, while CLEF-Utaclir was better overall. Also the stability of GenWeb-Utaclir is notable – not one its queries perfomed significantly worse than the median.

Table 7. The number of queries that perform significantly better or worse (absolute difference of 5% in average precision) than the (language-specific) median of each query.

	German(30 queries)				
	JRC-Ger	GenWeb	Utaclir	JRC-Utaclir	GenWeb-Utaclir
Better	2	7	6	4	11
Worse	7	6	3	2	0
	Swedish(40 queries)				
	JRC-Swe	CLEF	Utaclir	JRC-Utaclir	CLEF-Utaclir
Better	7	5	4	13	13
Worse	9	12	13	5	1

4.2 Tests on Corpus Size

The effect of corpus size on the performance level of COCOT was tested by increasing the sizes of the translation corpora step-by-step from 500,000 source document words onwards, until they reached their full sizes (see Tab. 1) . The shrinked corpora were created by removing alignments from the sets A (see Sec. 3.2) randomly. On each size level, COCOT was first trained with the same set of training topics as in Sec. 4. Then, the actual test queries were run with the learned parameters. Figure 4 presents the mean average precision of the runs plotted against the increasing corpus size.

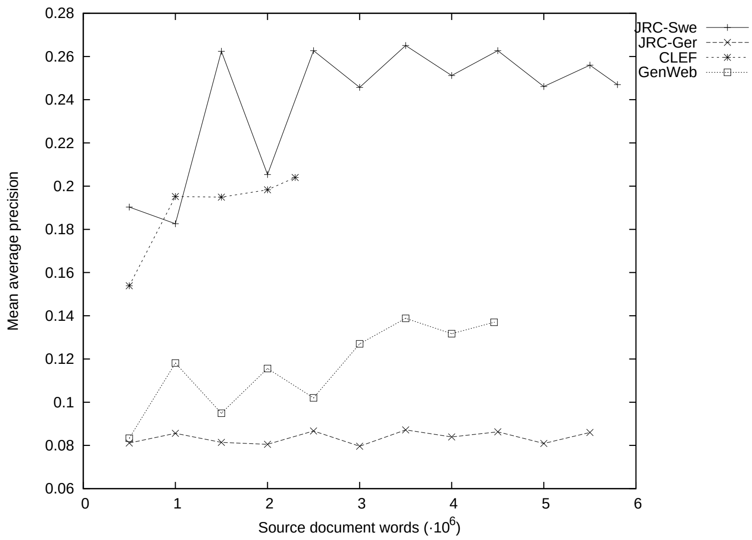

Fig. 4. MAP plotted against corpus size for four translation corpora.

For each corpus, save for JRC-Ger, there seems to be a significant difference in performance between the smallest level and the full size. However, from about 1M words onwards, the corpora seem to reach a performance level comparable to that of the full corpus quite quickly. This is somewhat surprising and puzzling. The JRC-Ger corpus performs badly on all size levels. This is an example of topical distance weighing more than alignment quality or corpus size. The fluctuations in performance (especially for JRC-Swe) are perhaps due to unoptimal COCOT parameters on some of the levels.

5 Conclusion

The performance of CLIR systems based on aligned corpora are affected by three qualities of the corpora: 1. The topical nearness of the corpus to the translated queries. 2. The alignment quality of the corpus – a parallel corpus is better than a noisy comparable corpus. 3. The size of the corpus. Based on the experiments discussed in this paper, the topical nearness seems to be the most crucial factor. The JRC-Ger corpus was of high alignment quality and sufficiently large – yet it performed badly, when it was used to translate queries from the genomics domain. This indicates that noisier, but easily available comparable corpora should be used for special domain vocabulary, if parallel corpora are not available. Comparable corpora are most effective as a complimentary resource.

The effect of alignment quality was also shown quite clearly. The JRC-Swe corpus was topically not as close to the topics as the CLEF corpus. However, JRC-Swe outperformed CLEF because it is a high-quality parallel corpus. The effect of corpus size was not as clear, though. The results seem to indicate that after a certain threshold in corpus size, the performance does not increase significantly.

Acknowledgements. This research was partly funded by the Tampere Graduate School in Information Science and Engineering (TISE)

References

1. Oard, D.W., Diekema, A.R.: Cross-Language Information Retrieval. Annual Review of Information Science and Technology (ARIST) 33, 223–256 (1998)
2. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Inf. Retr. 4, 209–230 (2001)
3. Sheridan, P., Ballerini, J.P.: Experiments in Multilingual Information Retrieval Using the SPIDER System. In: SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 58–65. ACM Press, New York (1996)
4. Franz, M., McCarley, J.S., Ward, T., Zhu, W.J.: Quantifying the Utility of Parallel Corpora. In: SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 398–399. ACM Press, New York (2001)

5. Zhu, J., Wang, H.: The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval. In: ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp. 593–600. Association for Computational Linguistics, Morristown, NJ. (2006)
6. Xu, J., Weischedel, R.: Empirical Studies on the Impact of Lexical Resources on CLIR Performance. Inf. Process. Manage. 41, 475–487 (2005)
7. Peters, C.: What Happened in CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M.(eds.) CLEF 2006. LNCS, vol. 4730, pp. 1–10. Springer, Heidelberg (2007)
8. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, Erjavec, Tufiş, D., and Varga, D.: The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In: LREC'2006:Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 2142-2147.European Language Resources Association, Paris (2006)
9. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., Keskustalo, H.: Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. ACM Trans. Inf. Syst. 25 4 (2007)
10. Allan, J., Callan, J.P., Croft, W.B., Ballesteros, L., Broglio, J., Xu, J., Shu, H.: Inquery at TREC-5. In: TREC-5: The Fifth Text Retrieval Conference, pp. 119–132. National Institute of Standards and Technology (1996)
11. Talvensaari, T., Järvelin, K., Pirkola, A., , Juhola, M., Laurikkala, J.: Focused Web Crawling in Acquisition of Comparable Corpora. (2007) submitted to Information Retrieval.
12. Chakrabarti, S., van den Berg, M., Dom, B.: Focused Crawling: a New Approach to Topic-specific Web Resource Discovery. In: WWW '99: Proceeding of the Eighth International Conference on World Wide Web, pp. 1623–1640. Elsevier North-Holland, Inc. (1999)
13. Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In: SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–29. ACM Press, New York (1996)
14. Hersh, W.R.: Report on the TREC 2004 Genomics Track. SIGIR Forum. 39, 21–24 (2005)
15. Keskustalo, H., Hedlund, T., Airio, E.: Utaclir : General Query Translation Framework for Several Language Pairs. In: SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 448–448. ACM Press, New York (2002)
16. McNamee, P., Mayfield, J.: Comparing Cross-language Query Expansion Techniques by Degrading Translation Resources. In: SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159–166. ACM Press, New York (2002)
17. Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-based Cross-language Information Retrieval. In: SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55–63. ACM Press, New York (1998)