# Focused Web Crawling in the Acquisition of Comparable Corpora*

Tuomas Talvensaari     Ari Pirkola     Kalervo Järvelin
Martti Juhola     Jorma Laurikkala

February 25, 2008

### Abstract

CLIR resources, such as dictionaries and parallel corpora, are scarce for special domains. Obtaining comparable corpora automatically for such domains could be an answer to this problem. The Web, with its vast volumes of data, offers a natural source for this. We experimented with focused crawling as a means to acquire comparable corpora in the genomics domain. The acquired corpora were used to statistically translate domain-specific words. The same words were also translated using a high-quality, but non-genomics-related parallel corpus, which fared considerably worse. We also evaluated our system with standard IR experiments, combining statistical translation using the Web corpora with dictionary-based translation. The results showed improvement over pure dictionary-based translation. Therefore, mining the Web for comparable corpora seems promising.

Keywords: cross-language information retrieval, focused crawling, comparable corpora

## 1   Introduction

In Cross-Language Information Retrieval (CLIR), the aim is to find documents that are written in a language different 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.

---

*This is a preprint of an article accepted for publication in Information Retrieval. Please refer to the final version.

1

In dictionary-based translation, the source language query keys are replaced by their target language counterparts in a dictionary. This seems straightforward, but multiple translation alternatives may introduce ambiguity into the resulting target language query. In addition, dictionaries are limited in scope, often missing crucial query vocabulary, such as proper nouns and domain-specific terminology. Naturally, such shortcomings can severely impair query performance (Pirkola et al., 2001).

MT systems aim to produce readable, grammatically correct translations. However, queries are often just lists of words, and using sophisticated natural language processing techniques on them would seem out of proportion. Similarly to dictionary-based translation, domain-specific vocabulary is often missing from MT systems. On the other hand, MT systems often have ways to "guess" (e.g. by weighting) the most probable translation candidate, which decreases translation ambiguity.

Fuzzy cognate matching is the least resource-laden of the mentioned techniques. Proper nouns and technical terms often vary only slightly between languages, and rather simple techniques, such as $n$-gram matching, can be used to "translate" such words. Also, transformation rules can be learned and used to capture stereotypical variation between languages. The English-German word pair *construction-konstruktion* is a typical example of such variation. However, fuzzy matching is usually inadequate when used alone (save maybe for languages that are closely related, such as Swedish and Norwegian). More often it is used as a complementary technique (Pirkola et al., 2006).

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 (Sheridan and Ballerini, 1996). The aligned documents can be, e.g., accounts of the same news event written independently in different countries.

Naturally, the most reliable translation knowledge is obtained from large parallel corpora, such as the Canadian Hansard corpus (Gale and Church, 1991) or the JRC-Acquis corpus of EU legislation (Steinberger et al., 2006). However, such collections are relatively rare, and they are often not available for particular domains. Moreover, CLIR resources in general are scarce for special domains. For this reason, the acquisition and use of comparable corpora in domain specific CLIR is an appealing idea.

The Web, with its vast volumes of data in almost any domain and language, is a natural resource for corpus-based CLIR. In this paper, we experiment with focused Web crawling as a means to build domain-specific comparable corpora. To our knowledge, such experiments have not previously been published. Focused crawling refers to the acquisition of material specific to a given subject from the Web, taking advantage of its hyperlink structure (Chakrabarti et al., 1999). The domain of choice for our experiments is genomics, a fast-growing field with a fast-growing vocabulary. Cross-lingual resources for such a domain

would have to keep up with the pace of the field, and we think our method has potential in this respect as well.

We aim to show that:

- it is possible to mine comparable texts in predefined domains and languages

- the gathered texts can be aligned and the alignments can be employed as a similarity thesaurus

- it is possible to derive good quality translations from the alignments

- a domain-specific comparable Web corpus provides better translations than a general-purpose parallel corpus, even if the latter is of much higher alignment quality

- the system can achieve competitive CLIR performance when used together with other resources

In Section 3 we introduce our focused crawler that was used to gather genomics-specific text in English, Spanish, and German. In Section 4, a brief overview of some of the tools used in the study, is presented. The crawled text was aligned at paragraph level – Spanish and German paragraphs were aligned with the English ones. This procedure is explained in Section 5. The alignments were employed to extract statistical translation knowledge. This was done with our Comparable Corpus Translation program, Cocot (Talvensaari et al., 2007), which is introduced in Section 6. We performed IR experiments based on the genomics track of the 2004 TREC conference (Hersh, 2005). Several translation system setups and translation approaches were used in the tests, which are described in Section 7.1. Section 7.2 gives a more in-depth analysis of the quality of the translations provided by Cocot. We translated individual domain-specific words from the source languages (Spanish and German) into the target language (English), and compared the quality of the translations to that achieved with the help of the JRC-Acquis (Steinberger et al., 2006) parallel corpus. Section 8 provides a discussion on how well the fore-mentioned aims were achieved.

## 2   Previous work

Most of the research on the automatic creation of comparable corpora has involved established research corpora, such as the collections of TREC and CLEF conferences (Sheridan and Ballerini, 1996; Braschler and Schäuble, 1998). Surprisingly few studies exist on the acquisition of such corpora from the Web. Cheng et al. (2004) note that

> Comparable corpora are far easier to obtain; however, how to automatically gather appropriate comparable corpora from the Web is still a challenging task.

Utsuro et al. (2002) come closest to this. They collected Japanese and English news articles and aligned those having matching dates. Unfortunately, date-based alignment is not generally applicable, because the assumption that articles published on the same day report on the same events does not hold in the Web in general. Hassan et al. (2007) derived named entity (i.e. proper noun) translations from comparable and parallel corpora. They used a sophisticated method for aligning comparable texts that was based on word clustering. They did not, however, address the problem of acquiring the comparable corpora. Steinberger et al. (2005) use thesauri and named entities to cluster news documents cross-lingually. However, they do not apply the clusters to CLIR, but to browsing and information extraction.

*Parallel* corpora, on the other hand, have been mined from the Web. Multilingual news services and web sites of multinational corporations are only some examples of parallel content (Nie et al., 1999; Resnik, 1999; Yang and Li, 2004). However, the more specific the domain of interest, the harder it is to find parallel pages in the Web.

The present work is an extension of our previous work and it differs from the previous one (Talvensaari et al., 2007) in these respects:

- The comparable corpora are mined from the Web. Previously, we used readily available IR test corpora. This is a profound extension that greatly improves the portability of our method.

- We concentrate on a specific domain. Additional CLIR resources are especially needed for domain-specific vocabulary.

- The alignments are made on paragraph-level, not document-to-document.

- Unlike in the news domain, the alignments could not be made based on the dates of the documents.

- We compare our system's CLIR performance to more other CLIR approaches than previously. Hence, the experiment setup is more competitive.

## 3   Focused crawling

The process of focused crawling can be summarized as follows (Cho et al., 1998; Bra et al., 1994; Chakrabarti et al., 1999). At the start of the process a set of seed URLs are inserted into a queue. One by one, the URL at the head of the queue is removed, and the web page pointed to by the URL is retrieved. The page is processed in some manner (e.g. it can be scanned to build an index for a search engine) and the out-links of the page are extracted and inserted into the queue. The queue can be prioritized, for example, based on how well the anchor text of the links matches against a *driver query* that consists of words of the wanted domain. The driver query is meant to steer the crawler to pages
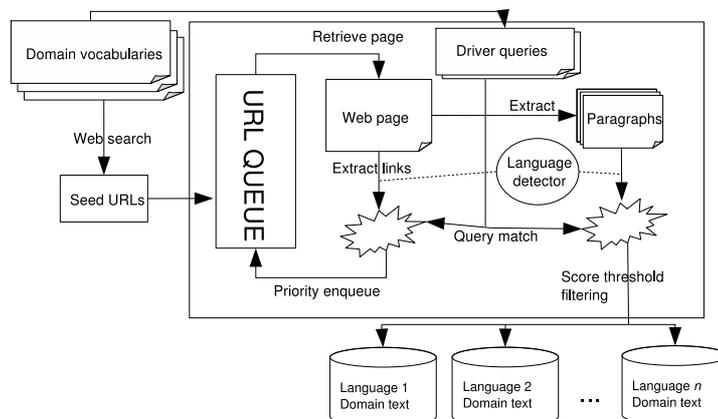
Figure 1: The crawling process

whose content is relevant to the domain in question. Crawling continues until the URL queue is empty or the process is interrupted.

In our experiments, the crawler collected domain-specific text in some pre-defined languages to be used in statistical translation. This brings forth special requirements for the implementation of the crawler:

1. In statistical translation, it is essential that the words to be translated appear in their natural contexts; random lists of words cannot be used as sources of translation knowledge. However, web pages often contain lots of noise from the domain's point of view – e.g. links to out-of-domain pages or personal contact information. For this reason, our crawler extracts text paragraphs from the retrieved pages – entire pages are not used. Also, we made the alignments at the paragraph level, because this level of granularity seems appropriate for statistical translation. A lengthy web page can handle various topics, whereas a paragraph is most often a concise expression of a single idea.

2. Unlike in many focused crawling applications, we are not that interested in the "popularity" or "importance" of the encountered pages; we only need some specific words appearing in their contexts. Consequently, our crawler need not to use well-known measures such as PageRank to evaluate the pages.

Next, we describe the functioning of our crawler in detail. The crawler was coded in Perl. Figure 1 provides an outline of the crawler.

## 3.1 Acquiring the seed URLs

Prior to the actual crawling phase, we semi-automatically collected domain-specific vocabulary in all of the three languages. The vocabularies play a central

5

role in the crawling process – they are used in acquiring the seed URLs and as driver queries to filter domain-specific content in the actual crawling process, as seen in Figure 1.

The vocabularies were acquired with simple Google searches, such as, for English, `(biotechnology OR genomics) AND (vocabulary OR lexicon)`. The word lists were extracted from the found pages, and, for each language, we constructed a combined word list that had the words sorted according to decreasing frequency in the lists. This phase involved manual work, since the lists had to be extracted from the non-uniformly coded web pages. It should be stressed that the lists were collected independently for each language, and it took roughly one working day from a non-domain-expert (the first author) to collect them.

We constructed Boolean search phrases from the acquired word lists, to search for the seed URLs. Each query consisted of two parts connected with the AND operator: a constant *context facet* and a varying set of domain words that were chosen based on their frequency. The context facet provided the correct context for the query; it contained the ten most frequent vocabulary words. (In the English queries, the context facet was `(gene OR dna OR pcr OR mutation OR karyotype OR genotype OR genome OR translocation OR translation OR transcription)`.) The second set also contained ten words which were connected with the OR operator. In the first query it consisted of words from the ranks 11 to 20 in the frequency-sorted domain lexicon. For the second query, the set included words from the ranks 21 to 30, and so on. A total of 50 queries were constructed and run for each language.

The aim of this procedure was to use a wide array of domain words as search keys, in order to find lots of prospective seed URLs – the order in which they were picked was not important. Shorter queries were also needed due to technical limitations. There is a limit to effective query length in Google (2006): query keys in excess of 32 key did not affect the results at all, they were presumably ignored by the search engine.

After the queries were executed with Google, the retrieved URLs (no more than 1000 for each query) were scored. The more times a URL appeared in the results, and the higher it ranked, the more points it scored. After this, the scores were combined host-wise, that is, each host's score was the sum of its indivual pages' URLs. The hosts were sorted according to their score. For the highest-scoring hosts (a few dozen per language), the page that had the highest individual score was chosen as a seed URL. Note that the root of a host (e.g. `http://en.wikipedia/`) could not be automatically chosen as a seed URL, since the domain-specific content might be only a small proportion of the overall content of a host.

The described approach ensured that there was at most one seed URL for each host. It is probable, and certainly preferable from a user's point of view, that other domain-specific pages are reachable through the link structure of the site.

Table 1: Sizes of the acquired corpora

| Language | Size (MB) | Words ($\cdot 10^6$) | Paragraphs |
|----------|-----------|----------------------|------------|
| English | 154 | 21.5 | 149,500 |
| Spanish | 25 | 3.5 | 30,800 |
| German | 73 | 8.8 | 84,200 |

## 3.2 The crawl

For each web page encountered, its text paragraphs were extracted (see Figure 1). This was done by examining the intended lay-out of the page (as laid out by Perl's `HTML::FormatText` module), not the HTML mark-up. A text segment that spanned more than one line and contained three or more sentences was considered a paragraph. Sentences, on the other hand, were defined as character sequences that start with an upper-case letter and end with one of the punctuation symbols. An array of miscellaneous heuristics was applied to prune exceptions to this simple rule. For example, in *George W. Bush*, *George W.* was not considered a sentence.

The language of each paragraph was detected with a simple $n$-gram-based algorithm (Cavnar and Trenkle, 1994). If the paragraph was in one of the sought-for languages, it was matched against the driver query of the particular language. The queries consisted of about 300 domain words acquired earlier. If the match score exceeded a threshold, the paragraph was saved to disk. The threshold was decided by manually sampling brief test crawls with varying thresholds.

The score was simply the proportion of domain words (words in the driver query) versus the total number of words in the paragraph. A more sophisticated *tf.idf* score could also be used, but this would require collection-wide statistics. They could be incorporated from some readily available corpus, and accumulated as the crawl would progress.

The out-links of the page were extracted and scored. The score of a link $l$, which is located on page $P$ is calculated as follows:

$$score(l) = w_a \cdot \rho(a(l)) + w_p \cdot \rho(P) + w_h \cdot \rho(host(P)),$$

where $a(l)$ is the anchor text of link $l$, $host(P)$ is the set of pages visited so far that have the same host as $P$, and $\rho(x)$ is the proportion of domain words in a text segment $x$. The weights $w_a$, $w_p$ and $w_h$ add up to 1, and after some experimentation we ended up choosing $w_p = w_h = 0.45$, $w_a = 0.1$. The link URLs were priority-enqueued, based on the scores.

The encountered URLs were kept in memory, so that every URL was visited only once. Paragraphs were also tracked, because often the same paragraph came up on different pages and different URLs pointed to pages with the same content. Table 1 depicts the sizes of the corpora acquired with the described approach.

7

# 4   Tools employed

In the present research we make frequent use of the following tools: the Utaclir query translator, the FITE-TRT cognate translator, the Babelfish MT system, the Cocot query translator, and the Lemur search engine. These are briefly described below.

The Utaclir query translator (Keskustalo et al., 2002) is a dictionary based query-generator. It employs stop word elimination, lemmatization and stemming, compound splitting, and off-the-shelf translation dictionaries. It produces synonym structured queries (the Pirkola method (Pirkola, 1998)), where each source word is translated into a synonym set `#syn(...)` and these are combined by a probabilistic sum operator `#sum(...)`. The sizes of the dictionaries used in this study were 29,000 and 35,000 source word entries for German-English and Spanish-English, respectively.

The FITE-TRT cognate translator is based on transliteration rules automatically mined from bilingual word lists (Pirkola et al., 2006). In addition, it uses large frequency lists for the source and target languages of translation in order to resolve between candidate translations generated.

The BabelFish MT system (`babelfish.altavista.com`) is used in the CLIR experiments (see Section 7).

The Lemur search engine (Lemur homepage) is based on language modelling. It supports various modes of operation, including structured queries of the kind Utaclir produces. In the present study, Lemur was used in the InQuery mode (Allan et al., 1996).

The Cocot Comparable Corpus Translation Program is introduced in Section 6.

# 5   Paragraph alignment

In the following section, the word *document* actually refers to the paragraphs extracted from the Web pages in the crawling phase. *Document* is used instead of *paragraph*, because generally the aligned entities are not necessarily paragraphs; they can be of any chosen granularity.

Let $d_S \in C_S$ and $d_T \in C_T$ be documents in the source and target collections, respectively. We aim 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, we aim to map each source document to a set of target documents whose similarity with the source document exceeds some threshold $\theta$. Each set $D$ is called a *hyper document*. It is not realistic to expect that we could find a satisfying counterpart for every source language document. Thus, we expect that $|A| < |C_S|$.

The alignment method resembles the one by Talvensaari et al. (2007). First, queries were formed from each source document. Second, the queries were translated into the target language (English) with Utaclir. Words that were not in Utaclir's dictionary were transmuted by FITE-TRT. Third, the translated queries were run against the English paragraphs with the Lemur IR toolkit.

Table 2: Alignment statistics

| Languages | $|A|$ | Avg. $|D|$ | Unique target paragraphs | Source words | Target words |
|---|---|---|---|---|---|
| Spa-Eng | 16,073 | 6.7 | 21,664 | 1,100,000 | 1,700,000 |
| Ger-Eng | 30,087 | 5.6 | 30,049 | 3,800,000 | 3,200,000 |

(See Section 4 for description of the mentioned tools). The Lemur 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 $D$. The threshold was chosen among a few predefined threshold levels. For each level, a test alignment was created, and the alignments were used to translate a test vocabulary with Cocot (see Section 6). The level that brought the best translation quality was chosen.

Table 2 depicts the statistics of the alignments created for the Spanish-English, and the German-English comparable corpora. First, the number of source paragraphs for which at least one alignment pair was found, is shown. Average $|D|$ is the average number of target paragraphs aligned per source document, while the fourth column depicts the number of target documents that appear in at least one hyper document. The last two columns show the number of words in source and unique target documents, respectively.

# 6 Cocot – employing the alignments

Cocot, a Comparable Corpus Translation program (Talvensaari et al., 2007), 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:

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

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 $D_k$ (see Section 5) in which a word $t_i$ appears, the weight is

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

9

Table 3: Example Cocot translations

| Rank | alelo | | alergia | | alogénico | |
|------|-------|------|--------------|-----|-----------|-----|
| 1 | allele | 14.0 | allergy | 4.1 | tcr | 9.2 |
| 2 | dominant | 10.6 | allergic | 2.5 | allogenic | 6.0 |
| 3 | recessive | 10.6 | allergen | 2.0 | mhc | 5.4 |
| 4 | gene | 9.8 | ragweed | 1.9 | apc | 5.4 |
| 5 | heterozygous | 9.6 | non-allergic | 1.6 | lfa-1 | 5.0 |

where $rank_{jk}$ is the rank of the document $d_j$ in the hyper document $D_k$, i.e. the rank calculated by Lemur in the alignment phase. The lower the rank, the less similar the target document is to the source document, according to Lemur. 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:

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

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. (Singhal et al., 1996). The *slope* value is a parameter of this scheme (we used $slope = 0.2$). 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 3 shows Cocot ranks for three genomics-related Spanish 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 $WCV = 4, \theta = 4.0$ mean that for the word *alelo*, the four highest ranking words would be returned, whereas, for *alergia*, only the first word would be used as the translation.

## 7 Test runs and results

To evaluate our system, we experimented with the test topics of the genomics track of the 2004 TREC conference (Hersh, 2005). The test collection was a subset of the MEDLINE database of about 4.6 million documents. There were 50 English topics, which were translated into Spanish and German by knowledgeable speakers. Figure 2 presents an example topic in English and German. Only the *title* and *need* parts of the topics were used.

The experiments consisted of two distinct set-ups:

```
<TOPIC>
<ID>2</ID>
<TITLE>Generating transgenic mice</TITLE>
<NEED>Find protocols for generating
transgenic mice.</NEED>
<CONTEXT>Determine protocols to generate
transgenic mice having a single
copy of the gene of interest at a
specific location.</CONTEXT>
</TOPIC>

<TOPIC>
<TITLE>Erzeugung von transgenen Mäusen
</TITLE>
<NEED>Finde Protokolle für die Erzeugung
von transgenen Mäusen.</NEED>
</TOPIC>
```

Figure 2: Example topic in English and German.

1. Standard IR experiments with the topics. Queries were constructed from the translated topics, which were then translated back to English with various CLIR systems. Cocot, with the Web corpus alignments, was combined with dictionary-based translation, and the performance of the combination was compared to that of other CLIR systems. These experiments are reported on in Section 7.1.

2. Translating individual domain-specific words extracted from the Spanish and German topics. To gain more detailed analysis on the performance of the Web corpus Cocot, we compared its performance in translating individual domain words with the performance of Cocot using the JRC-Acquis corpus. The experiments and their results are presented in Section 7.2.

## 7.1 Retrieval experiments

In the experiments, the Spanish and German topics were translated into English with several different translation approaches. Test runs were performed with the translated queries and the original English topics, which provided the monolingual baseline. The first 20 topics were used as training topics to decide Cocot's translation parameters. The values $WCV = 3, \theta = 4.0$ were chosen for both language pairs. The acquired comparable Web corpus was used as Cocot's translation corpus. For the last 30 topics used in the evaluation runs, there were 6,594 relevant documents in the recall base.

Besides the monolingual baseline, we applied six query construction strategies that used different translation methods:

1. Utaclir alone (UC for short).

2. Utaclir and Cocot (UC-CC). Utaclir with Cocot to translate out-of-vocabulary (OOV) words, that is, words not found in Utaclir's dictionary. Since Cocot's strength lies in its ability to translate domain vocabulary, and provide expansion keys (see Section 7.2), we paired Cocot with a resource that could handle the more general vocabulary. Comparing the

UC-CC approach with UC provided us evidence of the improvement Cocot brings to pure dictionary-based translation. The queries were formed in the following way. Utaclir produces a synonym set of translations, enclosed with the #sum operator: #sum( #syn($T_1$) #syn($T_2$) ... #syn($T_n$) ), where $T_i$ is the set of dictionary translations to some source word. It also produces a set of untranslated words $w_1, w_2 \ldots w_n$. The UC-CC approach produces the query #sum( #syn($T_1$) #syn($T_2$) ... #syn($T_n$) #syn($C1$) #syn($C2$) ... #syn($C_n$) ), where $C_i$ is the Cocot's translation set (its size determined by parameters WCV, $\theta$) of the word $w_i$.

3. FITE-TRT alone. (FI)

4. Utaclir-FITE-TRT (UC-FI). Utaclir used with the FITE-TRT technique to translate OOV words.

5. Utaclir-FITE-TRT-Cocot (UC-FI-CC). The above combination concatenated with the Cocot translations of Utaclir's OOV words.

6. Babel Fish (BA).

Tablee 4 shows the results for the runs in mean average precision and precision after 10 documents; Figures 3(a) and 3(b) depict the recall-precision curves for the Spanish-English and German-English runs, respectively. The Friedman test (Siegel and Castellan, 1988) showed significant ($p < 0.05$) difference in the performance for both language pairs. In the pairwise comparisons, the monolingual baseline was most often significantly better than the other methods, as expected. Significant differences between the different translation approaches are depicted in Table 4.

All translation methods perform clearly better in the Spanish-English runs than in the German-English runs. This is probably due to the large number of Latin-based cognates between Spanish and English, and perhaps also to the morphological complexity of German when compared to Spanish. In the Spanish runs, UC-FI performed exceptionally well, nearly matching the monolingual baseline. Cocot seems to bring some improvement into dictionary-based translation, and in the German runs, the improvement was statistically significant. In the German runs, all of the CLIR approaches, save for UC, appear to perform quite alike. The combination UC-FI-CC has a slight edge over the rest of the approaches.

It is noteworthy that in some cases, an added resource actually decreases the performance level. This is true in the Spanish runs with UC-FI-CC, and with UC-FI in the German runs. This could perhaps be fixed by weighting the components.

Figures 4 and 5 present a query-by-query analysis of the results for the Spanish and German runs, respectively. The median of the average precision of each query represents the zero level in the histograms. The median was calculated only among the translation approaches, the monolingual baseline was not included.
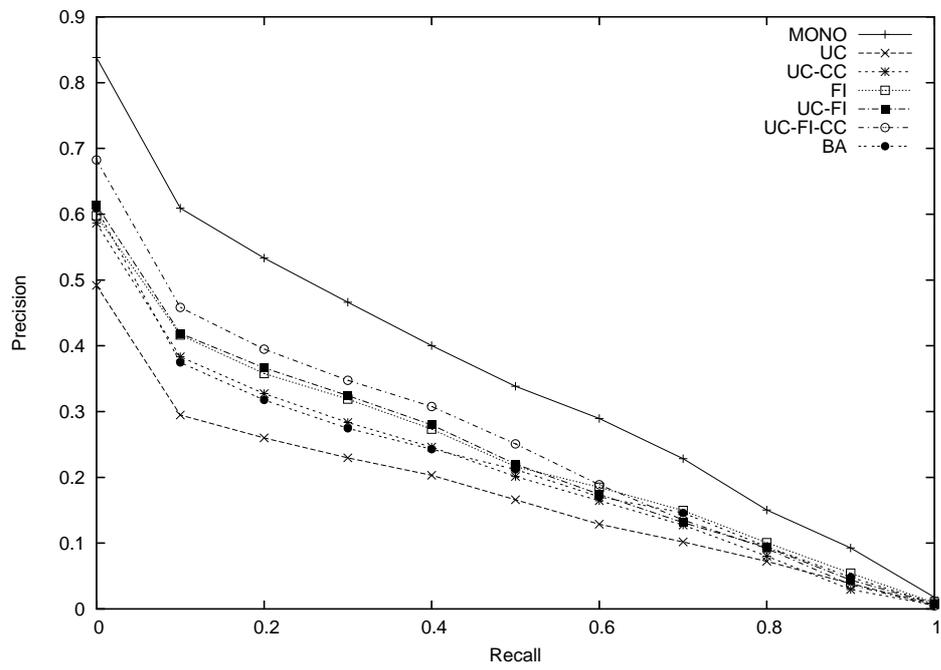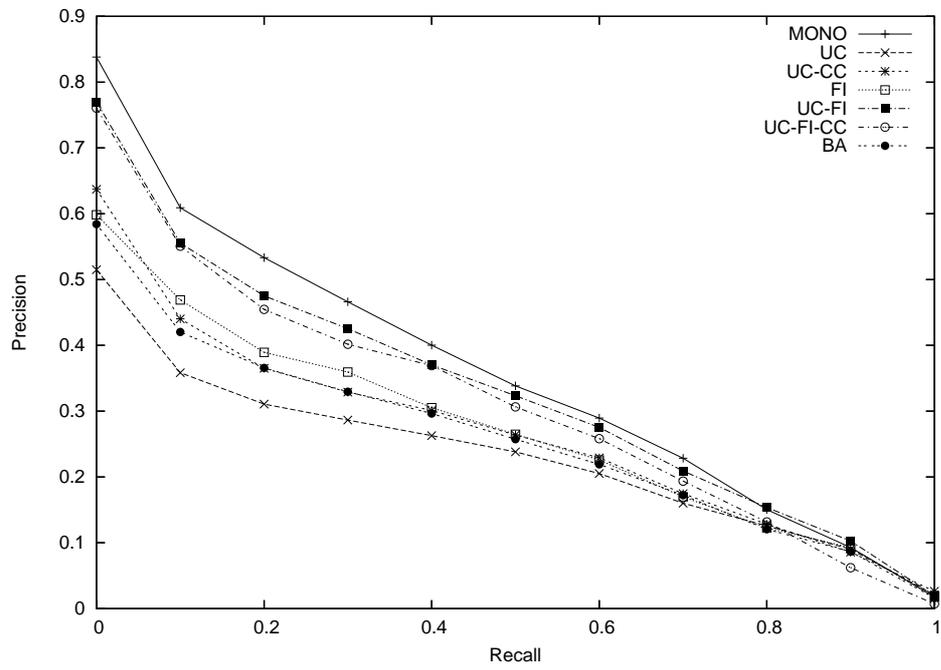
(a) Spanish-English



(b) German-English

Figure 3: Recall-precision curves for the retrieval experiments

13

Table 4: Mean average precision and precision after 10 retrieved documents for monolingual baseline and 4 translation approaches, $N = 30$. '> XX' indicates statistically significant ($p < 0.05$) difference over method XX.

| | Spa-Eng | | Ger-Eng | |
|---|---|---|---|---|
| | **MAP** | **P@10** | **MAP** | **P@10** |
| **MONO** | 0.34 | 0.64 | 0.34 | 0.64 |
| **UC** | 0.22 | 0.39 | 0.17 | 0.33 |
| **UC-CC** | 0.26 | 0.48 | 0.20 (> UC) | 0.42 |
| **FI** | 0.26 | 0.44 | 0.23 (> UC) | 0.41 |
| **UC-FI** | 0.32 (> UC, FI) | 0.58 (> BA, FI, UC) | 0.22 | 0.41 |
| **UC-FI-CC** | 0.30 | 0.57 (> BA, FI, UC) | 0.25 (> UC) | 0.48 (> UC) |
| **BA** | 0.25 (> UC) | 0.43 | 0.21 (> UC) | 0.42 |

In the Spanish runs, FI and BA are wildly uneven, while UC is consistently below the median. UC-CC performs quite near the median all the way, which confirms the improvement that Cocot brings to Utaclir. In the German runs, BA and UC are much like in the Spanish runs. The other approaches seem to perform quite steadily above the median in all of the queries. In general, combining different resources seems to bring consistency to the performance: in the combined approaches there are very few queries that drop significantly below the median.

## 7.2   Word translation tests

The word translation tests are meant to test whether the following two assumptions hold:

1. In addition to correct translations, Cocot (and similarity thesauri in general) gives words related to the correct translations which are often good expansion keys in IR.

2. Cocot, using the comparable Web corpus, can translate genomics-specific vocabulary better than Cocot that uses a high-quality parallel corpus with more general vocabulary.

As the parallel corpus, we used the JRC-Acquis corpus (Steinberger et al., 2006), which consists of official EU documents in all official EU languages. Its size for the languages used in this study is depicted in Table 5. The version of the corpus used was 2.2. The alignments in the corpus were mostly on paragraph level. They were created by Steinberger et al. (2006) with an algorithm that was based on the famous algorithm by Gale and Church (1991).

To test the performance of Cocot with different translation corpora, we needed a measure for the "goodness" of the words returned by Cocot in relation to the queries they are part of. Since a good word may be expected to appear more often in the relevant documents than in the rest of the documents, we devised a simple measure of goodness by using document frequencies.
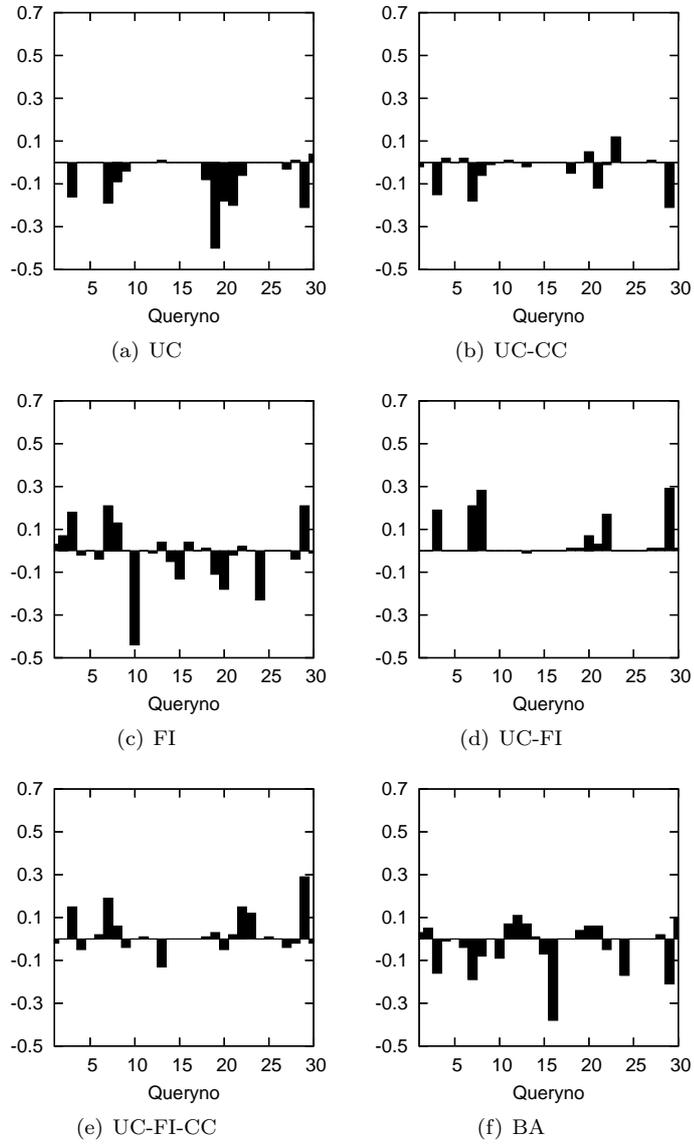
Figure 4: Difference to median average precision query-by-query, Spanish queries

Table 5: Size of the JRC-Acquis parallel corpus, version 2.2

| Language | Words |
|----------|-----------|
| English  | 7,547,154 |
| Spanish  | 8,006,579 |
| German   | 6,481,949 |

(a) UC

(b) UC-CC
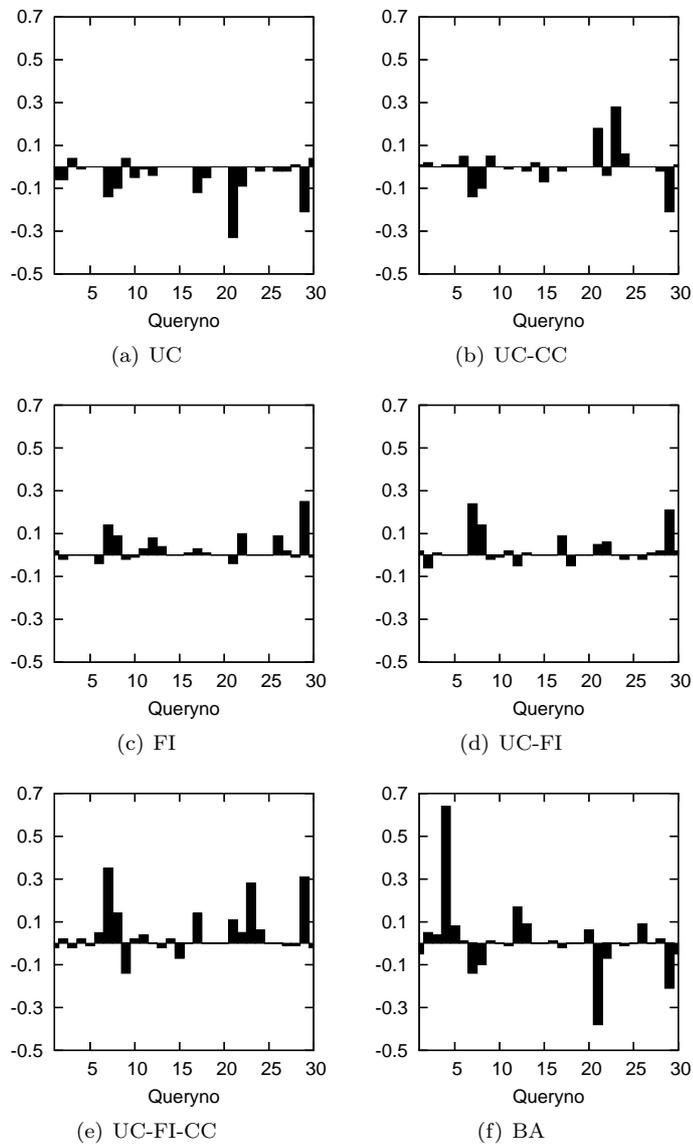
(c) FI

(d) UC-FI

(e) UC-FI-CC

(f) BA

Figure 5: Difference to median average precision query-by-query, German queries

16

**Definition 1.** The *relative document frequency* of a word $w$ in a document set $D$ is $rdf(w, D) = df(w, D)/|D|$, where $df(w, D)$ is the document frequency of $w$ in $D$.

**Definition 2.** Let $C$ and $R_q$ ($\subseteq C$) be the target test collection set and the set of relevant documents for a query $q$, respectively. Furthermore, let $t$ be a Cocot translation to a word in a source language query $q$. The *goodness* of the key $t$ in the query $q$ is $g(t, q) = rdf(t, R_q) - rdf(t, C \setminus R_q)$. In other words, the goodness is the difference in the relative document frequency of a key among documents relevant to the query ($R_q$) and among the rest of the documents ($C \setminus R_q$). The larger the positive difference, the better the key, and vise versa. The measure has the range $[-1, 1]$.

**Definition 3.** A translation key $t$ is good, if $g(t, q) \geq 0.2$ and the one-tailed binomial test (Siegel and Castellan, 1988) indicates that relative document frequency was significantly ($p < 0.05$) higher in $R_q$ than in $C \setminus R_q$.

We extracted domain-specific vocabulary from the Spanish and German topics and translated them with Cocot, using the two translation corpora. To gain more reliable results, we only considered topics for whom $|R_q| \geq 5$.

For example, the word *transgenen* appears in the German example topic 2 in Figure 2. For this topic, $|R_2| = 101$, i.e. there are 101 relevant documents for the topic. Since there are 4,591,008 documents in the entire collection, $|C \setminus R_2| = 4591008 - 101 = 4590907$. When the Web corpus is used as the translation corpus, and WCV = 3 is applied, Cocot gives the translation set $\{plant, transgene, transgenic\}$ for the word *transgenen*. The word *plant* appears in 105,292 documents in the collection, of which only one is in $R_2$. Accordingly, for the word *plant* in topic 2, $g(plant, 2) = 1/101 - 105291/4590907 = -0.01$ ($p = 0.90$). According to this measure, the other translation set words are much better, because $g(transgene, 2) = 0.30$ ($p \approx 0$) and $g(transgenic, 2) = 0.97$ ($p \approx 0$).

Table 6 shows the results for the word translation tests. In the tests, we used parameters WCV = 5, $\theta = 0$ for Cocot.

The number of OOV words (i.e. words not found in the source language documents) is a crucial statistic; for example, in the German JRC runs, 93 out of 148 unique source language words could not be translated at all. When the Web corpus is used, this number reduces to 31. The German Web corpus seems to work slightly better than the Spanish one, perhaps due to its larger size. The JRC corpus, on the other hand, seems to perform evenly for both languages, as indicated by the percentage of good translations and average goodness of the translations. All of the measures indicate that the Web corpus outperforms the JRC corpus in translating genomics-related words. Note also that for the Web corpus in both languages, the number of good translations exceeds the number of source language words, which shows that Cocot also provides related expansion keys, besides correct translations. Therefore, both of the above-stated

Table 6: Results of the word translation tests. Cocot returned 5 words per source language word, unless the word was OOV.

| | Spanish | | German | |
|---|---|---|---|---|
| Queries | 46 | | 46 | |
| Source words | 187 | | 200 | |
| Unique source words | 132 | | 148 | |
| | **Web** | **JRC** | **Web** | **JRC** |
| OOV | 30 | 70 | 31 | 93 |
| Translations | 775 | 520 | 785 | 410 |
| Good translations | 234 | 106 | 307 | 83 |
| Good translations % | 30 | 20 | 39 | 20 |
| Avg. goodness | 0.20 | 0.13 | 0.24 | 0.13 |

| | German | | | | Spanish | | | |
|---|---|---|---|---|---|---|---|---|
| | JRC | | Web | | JRC | | Web | |
| | MAP | P@10 | MAP | P@10 | MAP | P@10 | MAP | P@10 |
| Source | 0.17 | 0.24 | 0.17 | 0.24 | 0.18 | 0.24 | 0.18 | 0.24 |
| Good | 0.20 | 0.32 | 0.22 | 0.36 | 0.24 | 0.37 | 0.22 | 0.32 |
| Bad | 0.16 | 0.22 | 0.16 | 0.22 | 0.17 | 0.23 | 0.17 | 0.25 |

Table 7: CLIR results based on the goodness analysis

assumptions seem to hold.

### 7.2.1 Validating the translation goodness analysis

In the previous section it was shown that our genomics-specific corpora could produce translations that appear frequently in relevant documents in the recall base. However, it is not self evident that the "good" translations (as determined by the above analysis) would also result in good CLIR performance. The validity of the goodness analysis was tested by making IR experiments based on the tests.

For both of the language pairs, we first made a baseline run where the source language queries were run without translation. Then, for both language pairs and translation corpora, we formed "good" and "bad" queries. In the good queries, only the translations that were judged as good by the above analysis were used as translations. In the bad queries, only the translations that were not judged as good were used. In both query sets, the output of Cocot was concatenated with the baseline queries, so that the untranslated source language words were also part of both good and bad queries. All 50 topics of the TREC genomics track were used in these experiments. Table 7 presents the results of the runs in mean average precision and precision after 10 documents. The good queries always clearly outperform the bad ones. Also, the good queries always outperform the source language queries. These observations prove that translation goodness and CLIR performance correlate strongly.

It should be noted that the results of Tab. 7 should not be compared to

the "official" IR results of Tab. 4, or used to compare the performance of the translation corpora. This is because prior knowledge of the relevant documents was used in forming the good and bad queries.

# 8   Discussion

Our aim was to devise a novel technique for multilingual focused crawling for the creation of comparable corpora, and to use the acquired corpora in statistical query translation. We built a focused crawler that applied language detection, driver queries, and URL prioritizing to collect domain specific text in predefined languages. We managed to collect considerable amounts of text in the genomics domain in Spanish, German, and English. Alignments were made between the collections at the paragraph level; we mapped source language texts with similar counterparts in the target language collection. The alignments were employed in statistical translation with the Cocot translation system.

In the word translation experiments, we showed that Cocot was capable of providing good quality translations of domain vocabulary, as well as contributing semantically connected, but morphologically or taxonomically unrelated, expansion keys. Such query expansion is a feature missing from the other CLIR approaches. We were also able to prove that a high-quality parallel corpus with more general vocabulary was unable to provide equally good translation knowledge for domain-specific vocabulary. This indicates that resorting to noisier comparable corpora that can be created (semi-)automatically is necessary when dealing with special domains. In the standard IR experiments Cocot, using the Web corpus as the translation corpus, clearly improved dictionary-based translation.

However, compared to other translation approaches the results are more varied. Especially in the Spanish runs, the UC-CC approach was inferior to the approach where FITE-TRT was used in OOV translation (although the difference was statistically insignificant). The excellent performance of FITE-TRT in the Spanish runs could be explained by the large number of Latin-based cognates between Spanish and English. This kind of cross-lingual variation is where FITE-TRT is at its best. The greater similarity between Spanish and English may also be the reason behind the better overall results in the Spanish-English runs than in the German-English ones. Cocot's poorer relative performance in the Spanish runs is probably also due to the smaller size of the corpus. In the German runs, UC-CC fared as well as UC-FI and machine translation. It might be that in general, Cocot would be of better use with language pairs that have relatively low number of cognates.

We also note that there are various aspects of the present Cocot configuration that could be improved: the alignment process, the choice of Cocot's translation parameters, more complex query structuring (e.g. weighting strategies) could be employed, and combination with other resources could be made more effective.

Our method should naturally be easily portable to other domains, and indeed we believe it is. Gathering domain-specific vocabulary for the crawling phase

is not a trivial task, but it can be done quickly even by a non-specialist, as in our experiments. The vocabulary gathering could further be automated by automatically extracting domain-specific vocabulary from user-specified seed pages. In the alignment phase a smallish general purpose dictionary suffices. In addition, some kind of cognate matching can be used, such as FITE-TRT which was used in the present study.

It could be argued that the genomics domain is a relatively easy domain for CLIR, because much of the central vocabulary – protein names etc. – is the same across languages. In other technical domains, where more than cognate matching is needed, Cocot could perhaps bring greater improvement over other methods. This, though, remains to be shown – readily available test environments for special domains are rare.

While the contribution of the automatically acquired comparable corpora was not dramatic, the results were at least encouraging. This research contributed the first method for the acquisition of such corpora. It should also be noted that the techniques presented here are not the only way to employ the acquired corpora. For instance, they could also be used to prune translation alternatives in dictionary-based translation. A further potential application of the cross-language word associations Cocot provides on the basis of comparable corpora is cross-lingual document similarity calculation (e.g. for cross-lingual document retrieval through query by example document; cross-lingual plagiarism detection, etc.). Even in its present design, however, the method resulted in good translations of OOV words and provided useful expansion keys that improve CLIR effectiveness. Therefore, the method seems competitive and promising, especially for rapidly developing special domains.

## Acknowledgments

# References

Allan, J., Callan, J. P., Croft, W. B., Ballesteros, L., Broglio, J., Xu, J., & Shu, H. (1996). Inquery at TREC-5. In: TREC-5: The Fifth Text Retrieval Conference (pp. 119–132). National Institute of Standards and Technology.

Bra, P. D., Houben, G.-J., Kornatzky, Y., & Post, R. (1994). Information retrieval in distributed hypertexts. In *Proceedings of the 4th RIAO Conference* (pp. 481–491).

Braschler, M. & Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *ECDL '98: Proceedings of the Sec-*

ond *European Conference on Research and Advanced Technology for Digital Libraries* (pp. 183–197). London: Springer-Verlag.

Cavnar, W. B. & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (pp. 161–175).

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). 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). New York: Elsevier North-Holland, Inc.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., & Chien, L.-F. (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 146–153). New York: ACM Press.

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. In *WWW7: Proceedings of the seventh international conference on World Wide Web* (pp. 161–172). Amsterdam: Elsevier Science Publishers B. V.

Gale, W. A. & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *ACL '91: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (pp. 177–184). Morristown, NJ: Association for Computational Linguistics.

Hassan, A., Fahmy, H., & Hassan, H. (2007). Improving named entity translation by exploiting comparable and parallel corpora. In *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop.*

Hersh, W. R. (2005). Report on the TREC 2004 genomics track. *SIGIR Forum*, 39(1), 21–24.

Keskustalo, H., Hedlund, T., & Airio, E. (2002). 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). New York: ACM Press.

Lemur homepage. The Lemur toolkit homepage. `http://www.lemurproject.org/`. Accessed 22 February 2008.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 74–81). New York: ACM Press.

Pirkola, A. (1998). 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). New York: ACM Press.

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr.*, 4(3-4), 209–230.

Pirkola, A., Toivonen, J., Keskustalo, H., & Järvelin, K. (2006). FITE-TRT: a high quality translation technique for OOV words. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing* (pp. 1043–1049). New York: ACM Press.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 527–534). Morristown, NJ: Association for Computational Linguistics.

Sheridan, P. & Ballerini, J. P. (1996). 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). New York: ACM Press.

Siegel, S. & Castellan, N. J. Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.

Singhal, A., Buckley, C., & Mitra, M. (1996). 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). New York: ACM Press.

Steinberger, R., Pouliquen, B., & Ignat, C. (2005). Navigating multilingual news collections using automatically extracted information. *Journal of Computing and Information Technology*, 13(4), 257–264.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., & Varga, D. (2006). 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.*

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., & Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1), 4.

Utsuro, T., Horiuchi, T., Chiba, Y., & Hamamoto, T. (2002). Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users* (pp. 165–176). London: Springer-Verlag.

Yang, C. C. & Li, K. W. (2004). Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Inf. Process. Manage.*, 40(6), 939–955.