

Translation Corpus Source and Size in Bilingual Retrieval

Paul McNamee and James Mayfield

Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD 21218, USA
{paul.mcnamee, james.mayfield}@jhuapl.edu

Charles Nicholas

Dept. of Computer Science and Electrical Engineering
UMBC
Baltimore, MD 21250, USA
nicholas@umbc.edu

Abstract

This paper explores corpus-based bilingual retrieval where the translation corpora used vary by source and size. We find that the quality of translation alignments and the domain of the bitext are important. In some settings these factors are more critical than corpus size. We also show that judicious choice of tokenization can reduce the amount of bitext required to obtain good bilingual retrieval performance.

1 Introduction

Large parallel corpora are an increasingly available commodity. Such texts are the fuel of statistical machine translation systems and are used in applications such as cross-language information retrieval (CLIR). Several beliefs are commonly held regarding the relationship between parallel text *quality* and *size* for CLIR. It is thought that larger texts should be better, because the problems of data sparseness and untranslatable terms are reduced. Similarly, parallel text from a domain more closely related to a document collection should lead to better bilingual retrieval performance, again because better lexical translations are available.

We compared four sources of parallel text using CLEF document collections in eight languages (Braschler and Peters, 2004). English topic sets from 2000 to 2007 were used. Corpus-based translation of query terms was performed and documents were ranked using a statistical language model approach to retrieval (Ponte and Croft, 1998). Experiments were conducted using unlemmatized words and character 5-grams. No use was made of pre-translation query expansion or automated relevance feedback.

2 Translation Corpora

Information about the four parallel texts used in our experiments is provided in Table 1. We restricted our focus to Dutch (NL), English (EN), Finnish (FI), French (FR), German (DE), Italian (IT), Portuguese (PT), Spanish (ES), and Swedish (SV). These languages are covered by each parallel corpus.

2.1 Bible

The *bible* corpus is based on the 66 books in the Old and New Testaments. Alignments at the verse level were used; there are 31 103 verses in the English text.

2.2 JRC-Acquis v3

This parallel text is based on EU laws comprising the *Acquis Communautaire* and translations are available in 22 languages. The English portion of the *acquis* data includes 1.2 million aligned passages containing over 32 million words, which is approximately 40 times larger than the Biblical text. Alignments were provided with the corpus and were produced by the *Vanilla* algorithm.¹ The alignments are at roughly the sentence level, but only 85% correspond to a single sentence in both languages.

2.3 Europarl v3

The Europarl corpus was assembled to support experiments in statistical machine translation (Koehn, 2005). The documents consist of transcribed dialogue from the official proceedings of the European Parliament. We used the precomputed alignments that are provided with the corpus, and which are based on the algorithm by Gale and Church (1991). The alignments are believed to be of high quality.

¹ Available from <http://nl.ijs.si/telri/vanilla/>

Name	Words	Wrds/doc	Alignments	Genre	Source
<i>bible</i>	785k	25.3	Near Perfect	Religious	http://unbound.biola.edu/
<i>acquis</i>	32M	26.3	Good	EU law (1958 to 2006)	http://wt.jrc.it/lt/acquis/
<i>europarl</i>	33M	25.5	Very Good	Parliamentary oration (1996 to 2006)	http://www.statmt.org/europarl/
<i>ojeu</i>	84M	34.5	Fair	Governmental affairs (1998 to 2004)	Derived from documents at http://europea.eu.int/

Table 1: Parallel texts used in experiments.

2.4 Official Journal of the EU

The Official Journal of the European Union covers a wide range of topics such as agriculture, trade, and foreign relations. We constructed this parallel corpus by downloading documents dating from January 1998 through April 2004 and converting the texts from Adobe’s Portable Document Format (PDF) to ISO-8859-1 encoded text using *pdftotext*. The documents were segmented into pages and into paragraphs consisting of a small number of sentences (typically 1 to 3); however this process was complicated by the fact that many documents have outline or tabular formatting. Alignments were produced using Church’s *char_align* software (1993).

Due to complexities of decoding the PDF, some of the accented characters were not extracted properly, but this is a problem mostly for the earlier material in the collection. In total about 85 million words of text per language was obtained, which is over twice the size of either the *acquis* or *europarl* collections.

3 Translation

Using the pairwise-aligned corpora described above, parallel indexes for each corpus were created using words and 5-grams. Query translation was accomplished as follows. For each query term s , source language documents from the aligned collection that contain s are identified. If no document contains this term, then it is left untranslated. Each target language term t appearing in the corresponding documents is scored:

$$Score(t) = (F_l(t) - F_c(t)) \times IDF(t)^{1.25} \quad (1)$$

where F_l and F_c are relative document frequencies based on local subset of documents and the whole corpus. $IDF(t)$ is the inverse document frequency, or $\log_2(\frac{N}{df(t)})$. The candidate translation with the highest score replaced the original query term and

the transformed query vector is used for retrieval against the target language collection.

This is a straightforward approach to query translation. More sophisticated methods have been proposed, including bidirectional translation (Wang and Oard, 2006) and use of more than one translation candidate per query term (Pirkola et al., 2003).

Subword translation, the direct translation of character n -grams, offers several advantages over translating words (McNamee and Mayfield, 2005). N -grams provide morphological normalization, translations of multiword expressions are suggested by translation of word-spanning n -grams, and out-of-vocabulary (OOV) words can be partly translated with n -gram fragments. Additionally, there are few OOV n -grams, at least for $n = 4$ and $n = 5$.

4 Experimental Results

We describe two experiments. The first examines the efficacy of the different translation resources and the second measures the relationship between corpus size and retrieval effectiveness. English was the sole source language.

4.1 Translation Resources

First the relationship between translation source and bilingual retrieval effectiveness is studied. Table 2 reports mean average precision when word-based tokenization and translation was performed for each of the target collections. For comparison the corresponding performance using topics in the target language (*mono*) is also given. As expected, the smallest bitext, *bible*, performs the worst. Averaged across the eight languages only 39% relative effectiveness is seen compared to monolingual performance. Reports advocating the use of religious texts for general purpose CLIR may have been overly optimistic (Chew et al., 2006). Both *acquis* and *europarl* are roughly 40 times larger in size than *bible*

Target	<i>mono</i>	<i>bible</i>	<i>acquis</i>	<i>europarl</i>	<i>ojeu</i>
DE	0.3303	0.1338	0.1802	0.2427	0.1937
ES	0.4396	0.1454	0.2583	0.3509	0.2786
FI	0.3406	0.1288	0.1286	0.2135	0.1636
FR	0.3638	0.1651	0.2508	0.2942	0.2600
IT	0.3749	0.1080	0.2365	0.2913	0.2405
NL	0.3813	0.1502	0.2474	0.2974	0.2484
PT	0.3162	0.1432	0.2009	0.2365	0.2157
SV	0.3387	0.1509	0.2111	0.2447	0.1861
Average	0.3607	0.1407	0.2142	0.2714	0.2233
		39.0%	59.4%	75.3%	61.9%

Table 2: Mean average precision for word-based translation of English topics using different corpora.

Target	<i>mono</i>	<i>bible</i>	<i>acquis</i>	<i>europarl</i>	<i>ojeu</i>
DE	0.4201	0.1921	0.2952	0.3519	0.3169
ES	0.4609	0.2295	0.3661	0.4294	0.3837
FI	0.5078	0.1886	0.3552	0.3744	0.3743
FR	0.3930	0.2203	0.3013	0.3523	0.3334
IT	0.3997	0.2110	0.2920	0.3395	0.3160
NL	0.4243	0.2132	0.3060	0.3603	0.3276
PT	0.3524	0.1892	0.2544	0.2931	0.2769
SV	0.4271	0.1653	0.3016	0.3203	0.2998
Average	0.4232	0.2012	0.3090	0.3527	0.3286
		47.5%	73.0%	83.3%	77.6%

Table 3: Mean average precision using 5-gram translations of English topics using different corpora.

and both do significantly better; however *europarl* is clearly superior and achieves 75% of monolingual effectiveness. Though nearly twice the size, *ojeu* fails to outperform *europarl* and just barely beats *acquis*. Likely reasons for this include difficulties properly converting the *ojeu* data to text, problematic alignments, and the substantially greater length of the aligned passages.

The same observations can be seen from Table 3 where 5-grams were used for tokenization and translation instead of words. The level of performance with 5-grams is higher and these improvements are statistically significant with $p < 0.01$ (t -test).² Averaged across the eight languages gains from 30% to 47% were seen using 5-grams, depending on the resource. As a translation resource *europarl* still outperforms the other sources in each of the eight languages and the relative ordering of $\{europarl, ojeu, acquis, bible\}$ is the same in both cases.

²Except in four cases: *mono*: In ES & IT $p < 0.05$; *bible*: 5-grams were not significantly different than words in FI & SV

4.2 Size of Parallel Text

To investigate how corpus size effects bilingual retrieval we subsampled *europarl* and used these smaller subcorpora for translation. The entire corpus is 33 million words in size, and samples of 1%, 2%, 5%, 10%, 20%, 40%, 60%, and 80% were made based on counting documents, which for *europarl* is equivalent to counting sentences. Samples were taken by processing the data in chronological order.

In Figure 1 (a-d) the effect of using larger parallel corpora is plotted for four languages. Mean average precision is on the vertical axes, and for visual effect the chart for each language pair uses the same scale. The general shape of the curves is to rise quickly as increasing subsets from 1% to 10% are used and to flatten as size increases further. Curves for the other four languages (not shown) are quite similar. The deceleration of improvement with increasing corpus size can be explained by Heap’s Law. Similar results have been obtained in the few studies that have sought to quantify bilingual retrieval performance as a function of translation resource size (Xu and Weischedel, 2000; Demner-Fushman and Oard, 2003). In the higher complexity languages such as German and Finnish, n-grams appear to be gaining a slight improvement even when the entire corpus is used; vocabulary size is greater in those languages.

The data for the 0% condition were based on cognate matches for words and ‘cognate n-grams’ that require no translation. The figure reveals that even very small amounts of parallel text quickly improve performance. The 2% condition is roughly the size of *bible*, but is higher performing, likely due to a better domain match.³ Using a subsample of only 5% of available data from the highest performing translation resource, *europarl*, 5-grams outperformed plain words using any amount of bitext.

5 Conclusion

We examined issues in corpus-based bilingual retrieval, including the importance of parallel corpus selection and size, and the relative effectiveness of alternative tokenization methods. Size is not the only important factor in corpus-based bilingual re-

³For example, the Biblical text does not contain the words *nuclear* or *energy* and thus is greatly disadvantaged for a topic about nuclear power.

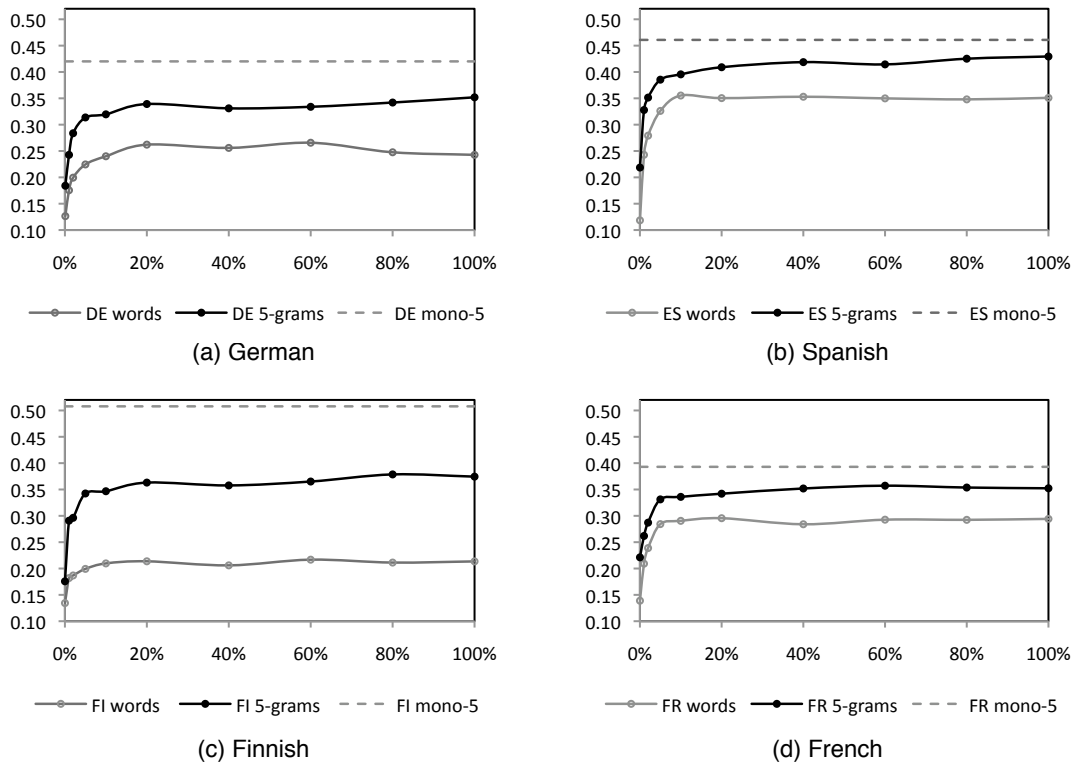


Figure 1: Performance improvement with corpus growth.

trieval, the quality of alignments, compatibility in genre, and choice of tokenization are also important.

We found that character 5-gram tokenization outperforms words when used both for translation and document indexing. Large relative improvements (over 30%) were observed with 5-grams, and when only limited parallel data is available for translation, n-grams are markedly more effective than words.

Future work could address some limitations of the present study by using bidirectional translation models, considering other language families and source languages other than English, and applying query expansion techniques.

References

- Martin Braschler and Carol Peters. 2004. Cross-language evaluation forum: Objectives, results, achievements. *Inf. Retr.*, 7(1-2):7–31.
- P. A. Chew, S. J. Verzi, T. L. Bauer, and J. T. McClain. 2006. Evaluation of the Bible as a resource for cross-language information retrieval. In *Workshop on Multilingual Language Resources and Interoperability*, pages 68–74.
- Kenneth Ward Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings ACL*, pages 1–8.
- Dina Demner-Fushman and Douglas W. Oard. 2003. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *HICSS*, pages 108–117.
- William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings ACL*, pages 177–184.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.
- Paul McNamee and James Mayfield. 2005. Translating pieces of words. In *ACM SIGIR*, pages 643–644.
- Ari Pirkola, Deniz Puolamäki, and Kalervo Järvelin. 2003. Applying query structuring in cross-language retrieval. *Inf. Process. Manage.*, 39(3):391–402.
- Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *ACM SIGIR*, pages 275–281.
- Jianqiang Wang and Douglas W. Oard. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In *ACM SIGIR*, pages 202–209.
- Jinxi Xu and Ralph Weischedel. 2000. Cross-lingual information retrieval using hidden Markov models. In *EMNLP*, pages 85–103.