Statistical Machine Translation and Cross-Language IR: QMUL at CLEF 2006

Christof Monz
Department of Computer Science
Queen Mary, University of London
London E1 4NS, UK
Email: christof@dcs.qmul.ac.uk
Web: http://www.dcs.qmul.ac.uk/~christof

Abstract

In this year's CLEF submissions we focus on using a state-of-the-art statistical machine translation approach for ad-hoc cross-language retrieval. Our machine translation approach is phrase-based as opposed to statistical word-based approaches that have been previously used for query translation in cross-language IR. The phrase translation probabilities were estimated by using the Europarl corpus. For query formulation, we also use the n-best lists of translation candidates to assign weights to query terms. Our results show that a statistical phrase-based approach is a competitive alternative to commercial, rule-based machine translation approaches in the context of cross-language IR.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Information Retrieval, Performance, Experimentation

Keywords

Cross-Language Information Retrieval, Statistical Machine Translations

1 Introduction

The two basic approaches to cross-language retrieval are: (1) Automatically translate all documents in the collection into the source language and then apply monolingual retrieval on the translated document collection; and (2) Automatically translate the user-posed query into the target language and then apply monolingual retrieval with the translated query on the original document collection in the target language.

Query translation requires access to some form of translation dictionary. Three approaches may be used to produce the translations:

1. Application of a machine translation system to translate the entire query into the target language.
2. Use of a dictionary to produce a number of target-language translations for words or phrases in the source language.

3. Use of a parallel corpus to estimate the probabilities that word w in the source language translates into word w' in the target language.

The main shortcoming of most commercial machine translation systems is that they only return the most likely translation, where 'most likely' is defined in terms of the internal algorithm of the translation system. But this does not mean that there are no other equally good—or, by some other objective standard, maybe even better—alternative translations.

On the other hand, there are several drawbacks to the parallel-corpus approach as well. First, although parallel corpora are available for many of the European languages, there are many languages for which there are still no parallel corpora large enough to estimate translation probabilities. Second, most of the parallel corpora belong to a rather specific domain, such as the Europarl corpus,[1] which contains the proceedings of the European parliament in 11 languages for the years 1996–2003. This introduces a bias toward the domain of the parallel corpus and makes the learned translation probabilities less reliable for other domains. A third disadvantage is that the translation probabilities induced from parallel corpora are typically based on single-word mappings, although recent template-based statistical methods facilitate the acquisition of phrase translations [12].

Here, we propose an approach that uses a state-of-the-art phrase based statistical machine translation system. Since this system is trained on the aforementioned Europarl corpus, we wanted to investigate whether such a translation approach results in a well-performing cross-language IR system despite the difference in genre between the corpus that is used to train the translation model and the retrieval corpus.

This paper is organized as follows: The next section provides some background on statistical machine translation and its application to cross-language IR. Section 3 describes the experimental set-up and results of our runs submitted to CLEF 2006. Finally we draw some conclusions in Section 4.

2 Statistical Machine Translation

This section provides some background on statistical machine translation and how it is applied to cross-language retrieval.

2.1 Phrase-Based Translation

Word-based statistical machine translation approaches have been used in several cross-language IR systems, see, e.g. [3, 7, 8]. These approaches learn word translations probabilities from a parallel corpus, or bi-text, which is a set of sentence pairs, where both sentences in a pair are translations of each other. There are a number of approaches computing these translation probabilities, such as the IBM models [1], Hidden Markov Models [18] and other co-occurrence measures. In practice most cross-language IR systems use IBM model 1, the most the simple of the IBM models.

In the area of machine translation word-based translation approaches more recently have been replaced by phrase-bases approaches, see, e.g. [12], which generate translations of substantially better quality. Phrase-based models use translation-probabilities for sequences of consecutive words—although not necessarily linguistic phrases—instead of individual words. The advantage is that more contextual information is captured this way, and at the same time this approach models some of the local word re-orderings that can take place when translating from one language into another. Word re-ordering does not appear to be a prime concern in cross-language retrieval, where most systems use a bag-of-words approach, and fluency of the translation can be disregarded. On the other hand, aiming to generate fluent output can also have a positive impact on choosing a correct phrase translation. In general, the probability of the English sentence e being a translation of the foreign sentence f is computed by applying Bayes' rule:

$$p(e|f) \quad = \quad \frac{p(f|e) \cdot p(e)}{p(f)} \tag{1}$$

[1] The Europarl corpus is freely available from `http://www.isi.edu/~koehn/europarl/`.

Since we are only interested in finding the most likely English translation for a given foreign sentence, $p(f)$ can be disregarded:

$$\text{argmax}_e\, p(e|f) \quad = \quad \text{argmax}_e\, p(f|e) \cdot p(e) \tag{2}$$

Here, the likelihood, $p(e|f)$, models the faithfulness of translation, i.e. to what extent it covers the semantic content of the foreign sentence, and the prior, $p(e)$, models the fluency of the generated translation.

During decoding, which is the process of finding the English sentence maximizing $p(e|f)$ according to equation (2), English translation candidates are generated incrementally by applying matching phrase translations and considering the probability of the fluency of each translation candidate. A number of decoding approaches are used in statistical machine translation, and here we follow the multi-stack beam search design used in [5].

In multi-stack beam search decoding partially generated translation candidates are grouped in stacks, where candidates are put into the same stack if they have translated the same number of words from the foreign sentence. Since comprehensive search is computationally prohibitive, and in fact NP-complete, see [4], heuristics have to be used to restrict the search space. See [5] for more details on multi-stack decoding.

2.2 Phrase Extraction

Although most alignment approaches, such as the IBM models 1–5 [1], are restricted to word-level alignments, it is possible to extract phrase translations by post-processing the word-level alignments, and there are a number of approaches to phrase extraction, see, e.g., [11, 17]. For our experiments we use the approach described in [6], which is also part of the distribution of the Pharaoh Statistical Machine Translation system.[2]

In this approach, first the intersection of the alignment links of both alignment directions, i.e. e-to-f and f-to-e, is taken. This initial alignment is then expanded by adding links that are adjacent to existing links, which is done iteratively until no further alignment links can be added.

During phrase extraction, groups of adjacent words are extracted on the foreign and the English side. Two phrases are considered translations of each other if the words in the phrase pair are only aligned to each other and not to words outside the phrase pair. The conditional phrase translation probabilities are simply computed by marginalizing the joint distribution of the phrase pair:

$$p(\bar{f}|\bar{e}) = \frac{\text{freq}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{freq}(\bar{f},\bar{e})}$$

where $\bar{f}$ is a phrase in the foreign language, and $\bar{e}$ is an English phrase. For more details see [5].

2.3 Language Modeling

When estimating the probability of a translation candidate, it is impossible to use maximum likelihood estimates on the entire string, by using the chain rule

$$p(e) \quad = \quad p(e_1) \cdot p(e_2|e_1) \cdot p(e_3|e_1,e_2) \cdot \ldots \cdot p(e_n|e_1,\ldots,e_{n-1}) \tag{3}$$

where e is a string in target language of length n, as one would run into serious data sparseness issues. Instead most language modeling approaches use a limited horizon of two words, called tri-gram models:

$$p(e) \quad \approx \quad p(e_1) \cdot p(e_2|e_1) \cdot p(e_3|e_1,e_2) \cdot \ldots \cdot p(e_n|e_{n-2},e_{n-1}) \tag{4}$$

Despite the limited context that is taken into account they still form a good compromise between prediction quality, robustness and computational tractability.

[2] Pharaoh is freely available from `http://www.isi.edu/licensed-sw/pharaoh/`.

2.4 Parameter Estimation

During decoding the different probabilities are combined by using a log-linear model. In addition to the translation and language model probabilities our model contains parameters that regulate, for example, word re-ordering, length of the translation, and number of phrase matches. Manually estimating the appropriate weights for all parameters is nearly impossible as the number of combinations is simply too large. Hence some form of optimization is required. Here we use the minimum error rate training (MERT) approach introduced by Franz Och [10].[3] Although MERT is not guaranteed to find a global optimum, its parameter estimations generally lead to substantial improvements in translation quality.

2.5 Translating Retrieval Queries

As mentioned above, full-fledged machine translation systems expect their input to be well-formed linguistic units, normally sentences. Statistical machine translation systems are no exception, but it should be pointed out that this restriction does not hold in principle, but is mainly due to the way statistical MT systems are trained.

When translating the queries, no stop word removal is applied. The only pre-processing that is applied is tokenization and case folding, as the entries in the phrase table are case-folded as well. The main reason for keeping stop words is that we want to keep the input to the MT system fluent.

One of the major disadvantages of using commercial machine translation systems is that they tend to return only the most likely translation, despite the fact that there are many more translations that might be almost as good and formulate the translation slightly different. In addition, if the machine translation system is mainly rule-based, scoring different translation candidates can be rather difficult as it is hard to consistently assign weights to human-generated translation rules.

On the other hand it is very easy to generate n-best list for statistical machine translation systems, as all translation and language model probabilities are estimated by using training data.

3 CLEF 2006 Experiments

This year, we participated in the English-to-French and English-to-Portuguese cross-language retrieval tasks. The following three subsections describe the configurations and data used for our machine translation component and the information retrieval system, and report the official CLEF 2006 results for those runs.

3.1 Experimental Set-Up: Machine Translation

For both translation directions we used the Europarl parallel corpus to estimate the phrase translation probabilities. Table 1 summarizes some collection statistics of the data used to build the machine translation systems for our cross-language retrieval experiments.

	English-French	**English-Portuguese**
no. of sentence pairs	645,518	568,446
no. of English words	14.4M	13.2M
no. of foreign words	16.3M	13.5M
no. of distinct phrase pairs	28.0M	29.3M
no. of distinct phrase pairs (CLEF 2006)	0.6M	0.6M
no. of words in language model corpus	87.5M	126.4M
no. of distinct 1-grams	512.9K	509.7K
no. of distinct 2-grams	5.8M	10.3M
no. of distinct 3-grams	5.8M	10.3M

Table 1: Overview of the data used in the statistical MT system.

[3] We are grateful to Philipp Koehn who provided us with a re-implementation of Och's optimization procedure.

The first five rows in Table 1 refer to parallel corpora from Europarl that were used to estimate phrase translation probabilities. The resulting phrase translation table is rather big and hard to keep in memory, and it was filtered to retain only phrases that match word sequences occurring in the CLEF 2006 topic set (see row 5).

In order to build the French and the Portuguese language models, we used the French and Portuguese side of the parallel corpus in combination with the documents in the CLEF retrieval corpus: ATS 1994–1995 and Le Monde 1994–1995 for French, and Fohla 1994–1995 and Publico 1994–1995 for Portuguese, see Table 1 for the respective sizes. We used the SRI Language Modeling toolkit [16] to train the tri-gram model, using modified Kneser-Ney smoothing [2]. The distribution of n-grams is shown in the last three rows of Table 1.

The parallel corpora were tokenized and case-folded. We used the simple rule-based tokenization scripts for all three languages involved. Diacritics were not removed.

For query translation we used the title and description fields only, and each field was translated separately. Note that in Section 2 we assumed that we translate from a foreign language into English, as most MT approaches focus on translation into English. For our experiments we translate from English into French and Portuguese, respectively. Hence English is referred to as the foreign language in the discussion below.

Parameter estimation for the translation models used CLEF test sets from previous years. Most of the CLEF topics are available in several languages where topics have been created by a human translator. For English-to-French translation we used the topics 210 from CLEF 2001–2004, and for English-to-Portuguese translation we used the 100 topics from 2002 and 2004.

3.2 Experimental Set-Up: Information Retrieval

The submitted runs used FLEXIR, a home-grown information retrieval system [9]. The main goal underlying FLEXIR's design is to facilitate flexible experimentation with a wide variety of retrieval components and techniques. FLEXIR is implemented in Perl; as it is built around the standard UNIX pipeline architecture, and supports many types of preprocessing, scoring, indexing, and retrieval tools. The retrieval model underlying FLEXIR is the standard vector space model. The Okapi BM25 weighting scheme [13] is used to compute the similarity between a query and a document. The BM25 parameters were set to $k_1 = 1.5$, $k_3 = 7$, and $b = 0.7$ as suggested in [15].

The text collection files were pre-processed using the same French and Portuguese tokenizers that were used to pre-process the parallel corpus. Stop words were removed from the translated queries, using the stop word list provided by the University of Neuchatel.[4]

Stemming and n-gram splitting was used for the French run, where we used the SNOWBALL stemmer, which was provided by the University of Neuchatel as well. After stemming, n-gram splitting was applied with $n = 4$. Note that n-gram splitting, was only applied to individual tokens and did not cross word boundaries.

The Portuguese run used only ngram splitting and no stemming. Here $n = 5$ and n-grams did cross word boundaries. For both French and Portuguese, these settings were determined by using the CLEF 2005 test collection as a development set.

Blind relevance feedback was applied to expand the original query with related terms. Term weights were computed by using the standard Rocchio method [14], where we considered the top 10 documents to be relevant and the bottom 500 documents to be non-relevant. We allowed at most 20 terms to be added to the original query.

As we mentioned above, both topic fields, i.e., the title and the description, were translated separately. For retrieval both translations were merged into a single query. For our official submissions we also used the n-best lists generated by our statistical machine translation system. The n-best lists are used to assign weights to terms in the query:

$$weight_f(t) \quad = \quad \frac{\sum_{i=1}^{n} occurs\_in(t, s_i) \cdot p(s_i|f)}{\sum_{i=1}^{n} \sum_{t \in s_i} p(s_i|f)} \qquad (5)$$

[4] The stop word lists are available at `http://www.unine.ch/info/clef/`.

English-to-French	
QMUL06e2f10b	0.3396
English-to-Portuguese	
QMUL06e2p10b	0.3526

Table 2: Official CLEF 2006 results

where s_i is the i-th translation candidate for f and f is the original foreign language topic (note: here, English is the foreign language as we translate *from* English *into* French and Portuguese). We experimented with n-best lists of different length, and setting $n = 10$ yielded the best results for the CLEF 2005 test set. The term weight in (5) is computed for each occurrence of a term in a query.

3.3 CLEF 2006 Experimental Results

We submitted two official runs to CLEF 2006, one for the English-to-French cross-lingual task and one for the English-to-Portuguese cross-lingual task. The MAP scores for both runs are shown in Table 2. Although using n-best machine translation output for term re-weighting did lead to slight improvements for the CLEF 2005 development set, the improvements were still rather small. In part this could be due to the small variation in the n-best lists. The differences between the translation candidates tend to be minute and are often causes by different derivations of almost identical translations. For example, Table 3 shows the top ten translations for topic 301.

Rank	Translation
	English: what brands are marketed by nestle around the world ?
1	les marques sont commercialisees par nestle dans le monde ?
2	trans=ce que sont marques commercialisees par nestle dans le monde ?
3	les marques sont commercialisees par nestle dans le monde - ?
4	ce que marques sont commercialisees par nestle dans le monde ?
5	ce que sont marques commercialisees par nestle dans le monde - ?
6	ce que marques sont commercialisees par nestle dans le monde - ?
7	les marques sont commercialisees par nestle dans le monde ?
8	ce qu ’ marques sont commercialisees par nestle dans le monde ?
9	ce qui marques sont commercialisees par nestle dans le monde ?
10	les marques sont commercialisees par nestle dans le monde entier ?

Table 3: Top ten translations for topic 301

As one can see, the translation candidates are almost identical to each other, and there is little variation in terms of lexical translation choices. Using longer n-best lists did not change this fact. It seems that more elaborate selection strategies are needed to filter out translation candidates that are merely due to word-order or phrase-matching differences.

4 Conclusions

With our submissions we tried investigate to what extent statistical machine translation is effective in the context of cross-language information retrieval. In particular with respect to machine translation approaches that translate the entire topic as opposed to word-translation models like the simpler IBM models. One of the main advantages of statistical machine translation is the ease with which n-best translation output can be generated, but unfortunately we were only able to get minor improvements from this. More elaborate ways of exploiting n-best lists remains a topic that we want to pursue further in future research.

Acknowledgments

This work has been supported, in part, by the Nuffield Foundation, Grant No. NAL/32720.

References

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] Stanley Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, 1998.

[3] Martin Franz, J. Scott McCarley, Todd Ward, and Wei-Jing Zhu. Quantifying the utility of parallel corpora. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 398–399, 2001.

[4] Kevin Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615, 1999.

[5] Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Robert E. Frederking and Kathryn B. Taylor, editors, *Proceedings of the 6th Conference of the Association for Machine Translations in the Americas (AMTA 2004)*, pages 115–124, 2004.

[6] Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models, user manual. Technical report, USC Information Science Institute, 2004.

[7] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3):381–419, 2003.

[8] Paul McNamee. Exploring new languages with HAIRCUT at CLEF 2005. In *Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, 2005.

[9] Christof Monz, Jaap Kamps, and Maarten de Rijke. The University of Amsterdam at CLEF 2002. In *Working Notes for the CLEF 2002 Workshop*, pages 73–84, 2002.

[10] Franz-Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, 2003.

[11] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52, 2003.

[12] Franz-Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[13] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 109–126. NIST Special Publication 500-225, 1994.

[14] Joseph J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System — Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.

[15] Jacques Savoy and Pierre-Yves Berger. Report on CLEF-2005 evaluation campaign: Monolingual, bilingual, and GIRT information retrieval. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J.F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, 2005.

[16] Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, 2002.

[17] Ashish Venugopal, Stephan Vogel, and Alex Waibel. Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 319–326, 2003.

[18] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational Linguistics (COLING '96)*, pages 836–841, 1996.