Combining Query Translation Techniques to Improve Cross-Language Information Retrieval

Benjamin Herbert, György Szarvas*, and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Computer Science Department,
Technische Universität Darmstadt,
Hochschulstr. 10, D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract. In this paper we address the combination of query translation approaches for cross-language information retrieval (CLIR). We translate queries with Google Translate and extend them with new translations obtained by mapping noun phrases in the query to concepts in the target language using Wikipedia. For two CLIR collections, we show that the proposed model provides meaningful translations that improve the strong baseline CLIR model based on a top performing SMT system.

1 Introduction

Multilingual information search becomes increasingly important due to the growing amount of online information available in non-English languages and the rise of multilingual document collections. Query translation for CLIR became the most widely used technique to access documents in a different language from the query. As CLIR is less accurate than monolingual IR, the combination of query translation techniques is a promising way to approximate monolingual accuracy. Despite the importance of the task, previous combination approaches showed limited success. Combination of statistical machine translation (SMT), machine readable dictionary (MRD) based models or similarity thesauri (ST) proved to be difficult [1] due to the difference in the accuracy of individual models (SMT tends to be superior); the aggregation of translation errors; or the topic drift caused by integrating multiple translations in a single query. Studies that report successful combination of different models require substantial extra computation and resources (syntactic analysis and NP translation patterns [2]).

For query translation, one can i) use an online translation service; ii) train an SMT system using parallel corpora; iii) employ MRDs to translate query words; or iv) make use of large scale multilingual knowledge sources like Wikipedia for cross-lingual mapping. CLIR based on information in Wikipedia [5] can reach 60-70% of monolingual accuracy, while using Google Translate is reported to reach 90% of the accuracy of monolingual search [4]. Other approaches usually perform in between, e.g. [3]. In this study, we develop translation methods that are simple and accurate to be good candidates for extending a high performance SMT system in a realistic search scenario.

* On leave from Research Group on AI of the Hungarian Academy of Sciences.

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 712-715, 2011.

2 Translation Models

Google Translate. As a baseline CLIR model, we use query translation by Google Translate. Due to robustness across domains and strong performance in translating Named Entities, using Google Translate for CLIR achieved the best results in the recent CLIR evaluation at CLEF 2008 [4].

Wikipedia based concept mapping. Wikipedia provides a natural source of multilingual information, with redirects and cross-language links between articles in different languages. E.g., the phrase *German school system* maps to the German concept *'Bildungssystem in Deutschland'*. Wikipedia, being an encyclopedia, typically uses formal terminology which is less likely to be ambiguous and detrimental to retrieval performance than lay terms. To exploit this, we mine all redirect and cross-language links to build a translation table which maps concepts to their target language equivalent. This offers a reliable, but incomplete source of translation information (e.g. adjectives are seldomly contained in Wikipedia). To map queries to Wikipedia concepts (titles), we first try to map the whole query, and then gradually proceed with mapping shorter word sequences. Thus, the query *German Spelling Reform* is mapped as a single phrase, while *Nuclear Transport in Germany* is mapped to *'Nuclear transport'* and *'Germany'*.

3 Experimental Setup

Document collections. We used two CLIR collections introduced in the CLEF *Domain Specific (DS)* and *Ad Hoc (AH)* tracks. They consist of 151,319 German social science and 294,339 newspaper articles and were used in CLEF between 2003-2008 and 2001-2003. We used two times 75 queries for DS (2003-5 and 2006-8) and 100 + 60 queries for AH (2001-2 and 2003). We used the 2-3 words long title field as the query. For more details, see the CLEF website.

Retrieval model. For retrieval and query expansion via pseudo relevance feedback (PRF), we used Terrier's Okapi BM25 model and Bo1 term weighting method, with their default parameters. We tokenized the queries and documents, removed stopwords and used stemming (with SnowBall). Since German is a compounding language and decompounding can add further, less specific terms to enrich a query (e.g. *Milchkonsum (Milk consumption)* can be split to *Milch* and *Konsum*), we used a compound splitter for German (BananaSplit package).

Combination of alternative translations. To improve Google based CLIR, we add the phrases obtained from the Wikipedia-based concept mapping to the query. However, alternative terms for noun phrases can cause topic drift. For the query *Maternity Leave in Europe*, using two translations for *Maternity Leave* can cause documents that contain both to be ranked higher than those containing only one and the term *Europe*. To avoid this, we downweight further translations.

4 Experimental Results

Our results are summarized in Table 1. The CLIR models, using the concept mapping with Wikipedia, Google Translate and their combination are presented

for all four collection parts used, together with a monolingual retrieval run using German queries as reference. We provide the mean average precision (MAP) scores and the relative accuracies to the monolingual run, with just tokenization and stemming used (BASE), with compound splitting (CSPLIT) and with query expansion based on PRF (CS+QE). For the combined *WP + Google* runs, the weighting parameter used in queries was set to the optimal value on the other query set. That is, we used the DS 2003-2005 queries to determine the weight value used for the DS 2006-2008 query set, etc. For the BASE, CSPLIT and CS+QE configurations, the weight values were 0.3, 0.4, 0.2 for DS 2003-5; 0.4, 0.3, 0.2 for DS 2006-8; 0.5, 0.3, 0.6 for AH 2001-2; 0.4, 0.4, 0.2 for AH 2003, respectively.

Table 1. MAP values on different collections. Significant improvements (paired t-test, $p < 0.05$) over the Google based models are marked with † for the *WP + Google* model.

Collection	Method	BASE		CSPLIT		CS+QE	
		MAP	% Monolingual	MAP	% Monolingual	MAP	% Monolingual
DS 2003-5	Wikipedia	0.2397	69.20	0.2501	62.74	0.2739	63.65
	Google Trans.	0.3304	95.38	0.3543	88.89	0.3844	89.33
	WP + Google	0.3562†	102.83	0.3753†	94.15	0.4034†	93.75
	Monolingual	0.3464	–	0.3986	–	0.4303	–
DS 2006-8	Wikipedia	0.1742	59.72	0.1740	52.41	0.1888	54.98
	Google Trans.	0.2878	98.66	0.3081	92.80	0.3293	95.89
	WP + Google	0.3204†	109.84	0.3194	96.20	0.3424†	99.71
	Monolingual	0.2917	–	0.3320	–	0.3434	–
AH 2001-2	Wikipedia	0.2193	82.04	0.2237	80.44	0.2611	79.58
	Google Trans.	0.2879	107.71	0.2886	103.78	0.3342	101.86
	WP + Google	0.2984†	111.63	0.2960†	106.44	0.3417	104.15
	Monolingual	0.2673	–	0.2781	–	0.3281	–
AH 2003	Wikipedia	0.1878	62.98	0.1902	55.84	0.2216	53.58
	Google Trans.	0.3166	106.17	0.3328	97.71	0.3904	94.39
	WP + Google	0.3339†	111.97	0.3487†	102.38	0.3937	95.19
	Monolingual	0.2982	–	0.3406	–	0.4136	–

As can be seen in Table 1, the individual CLIR models perform similar to the results reported in previous works: the Wikipedia model achieves 50-80% of the monolingual result, while Google Translate performs around 90% of the monolingual run. The Wikipedia based concept mapping performs slightly worse than the more complex WP model by [5] but we use just title fields.

As regards the combination of the Wikipedia based and the Google translations, we see consistent improvements over the CLIR models using a single translation. In particular, the combination improves over the results obtained using Google Translate which was argued to be a very strong CLIR model [4], and performs very close to the monolingual result. Moreover, these improvements are statistically significant (except for the AH collection when using QE and DS 2006-8 for CSPLIT). The positive effect of alternative translations is observed regardless of using compound splitting or not, which indicates that Wikipedia provided genuinely different translation terms. These improvements are in part complementary to query expansion, which indicates that the additional phrases are not always contained in the top retrieved documents and recovered by QE.

5 Conclusion and Future Work

In this study, we introduced a simple CLIR model using Wikipedia, mapping concepts in one language to their equivalents in another language based on the redirect and cross-language links in multilingual Wikipedia versions. This simple WP-based model performs similar to previous results obtained by Wikipedia-based CLIR (60-70% accuracy of monolingual retrieval). We also showed that the Wikipedia translations are accurate and often quite different from the translations of an SMT service, and are capable of improving the accuracy of an SMT-based CLIR model. In particular, to our knowledge we are first to show consistent (and in most settings significant) improvements over the CLIR based on Google Translate, which is reported to perform very well for CLIR.

There are many ways to improve our results. In particular, in the current study we used a single global term weighting parameter for Wikipedia translations. However, the benefit of WP translations is highly correlated with the coverage of the concept mapping for the given query (the higher portion of the query is mapped, the more beneficial it is), and with the average length of the mappings (concepts corresponding to longer phrases are more beneficial). This calls for a query-dependent weighting scheme, which we just started to develop. Another promising future work is to improve the coverage of our concept mapping by exploiting further information in Wikipedia (e.g. anchor texts [6]), or by adding a complementary resource to map verbs and adjectives.

Acknowledgements

This work was supported by the German Ministry of Education and Research under grant 'Semantics- and Emotion-Based Conversation Management in Customer Support (SIGMUND)', No. 01ISO8042D, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program (grant No. I/82806).

References

1. Braschler, M.: Combination approaches for multilingual text retrieval. Information Retrieval 7(1-2), 183–204 (2004)
2. Gao, J., Nie, J.Y., Xun, E., Zhang, J., Zhou, M., Huang, C.: Improving query translation for cross-language information retrieval using statistical models. In: Proceedings of SIGIR, pp. 96–104. ACM, New York (2001)
3. Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K.: Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. Information Retrieval 7(1-2), 99–119 (2001)
4. Kürsten, J., Wilhelm, T., Eibl, M.: The Xtrieval framework at CLEF 2008: domain-specific track. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 215–218. Springer, Heidelberg (2009)
5. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, R.B., Hiemstra, D., Jong de, F.M.G.: WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia. In: Proceedings of CLEF, pp. 58–65 (2009)
6. Roth, B., Klakow, D.: Cross-language retrieval using link-based language models. In: Proceedings of SIGIR, pp. 773–774. ACM, New York (2010)