Adapting SMT Query Translation Reranker to New Languages in Cross-Lingual Information Retrieval

Shadi Saleh and Pavel Pecina
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
{saleh,pecina}@ufal.mff.cuni.cz

ABSTRACT

We investigate adaptation of a supervised machine learning model for reranking of query translations to new languages in the context of cross-lingual information retrieval. The model is trained to rerank multiple translations produced by a statistical machine translation system and optimize retrieval quality. The model features do not depend on the source language and thus allow the model to be trained on query translations coming from multiple languages.

In this paper, we explore how this affects the final retrieval quality. The experiments are conducted on medical-domain test collection in English and multilingual queries (in Czech, German, French) from the CLEF eHealth Lab series 2013–2015. We adapt our method to allow reranking of query translations for four new languages (Spanish, Hungarian, Polish, Swedish). The baseline approach, where a single model is trained for each source language on query translations from that language, is compared with a model co-trained on translations from the three original languages.

1. INTRODUCTION

The growth of the Internet and its content in the world increases its diversity in terms of natural languages. Early study showed that the English language is not predominant on the internet anymore. This gave rise for the Cross-Lingual Information Retrieval (CLIR) task, in which users can pose queries in a language that is different from the language of documents. The big challenge in this task is to have multilingual resources for different languages, since such resources are expensive to prepare. In this work, we tackle the problem of lacking enough assessment information, which is required to develop machine learning-based systems to support more languages.

We present a machine learning approach that is based on source-language-independent features. This enables us to enlarge the training set by combining the data coming from different languages and at the end, improve the CLIR performance.

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

2. RELATED WORK

In the CLIR task, both queries and collection should be in the same language in order to conduct the retrieval when using models that are based on term-matching approach. Translating queries by statistical machine translation (SMT) systems proved to be the state-of-art approach in the CLIR task [16, 2, 10, 5]. Researchers recently have started to look inside SMT systems rather than using them as black boxes to translate the queries. Nikoulina et al. [8] used a machine learning model to rerank the *n-best-list* translations that are provided by their SMT system. For training, they used parallel queries obtained from several various CLEF tracks from 2000–2008.A similar approach was presented by Ture et al. [15] in which they converted the problem of reranking into classification problem. While Sokolov et al. [12] developed different approach that optimized the SMT system to immediately return the hypothesis that gives the better retrieval performance. In our previous work [11], we used discriminative features taken from the SMT system, alternative translations and external resources (e.g. Wikipedia and UMLS Metathesaurus) to build machine learning reranker that predicts the best hypothesis for better CLIR performance. We build on this work and explore the use of source-language-independent features to adapt the reranker to new languages, namely: Spanish, Hungarian, Polish, and Swedish.

3. SYSTEM DESCRIPTION

3.1 Data

We use collection taken from CLEF 2015 eHealth Task 2: User-Centred Health Information Retrieval. The collection is indexed using Terrier, an open source information retrieval system[1]. The queries come from CLEF 2013–2015 eHealth information retrieval Tasks [4, 3, 13] and are split into 100 queries for training and 66 queries for testing. The English queries were constructed by medical experts and then translated into Czech, French, and German also by medical experts, native speakers of those languages. To extend the task, we asked medical experts to translate these queries also into Spanish, Hungarian, Polish and Swedish. However, a complete relevance assessment (of top 10 documents in all the experiments) is available only for the three original languages. Results for the new languages are not completely assessed. The ratio of unjudged documents among the top 10 retrieved documents is around 25%.

[1] http://terrier.org/

Table 1: Cross-validation of language-specific models on the training set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90
Oracle	52.10	25.86	86.50	51.80	24.55	79.40	47.40	22.16	77.50	50.10	23.23	78.70
Baseline	41.90	22.02	78.80	40.10	20.67	74.40	**37.30**	**19.55**	72.80	39.60	20.07	73.80
SMT	*44.30*	22.65	89.10	40.40	20.39	87.30	34.90	17.51	79.00	38.90	19.00	87.80
SMT+Rank	*43.00*	22.46	87.80	40.50	20.63	88.30	35.70	18.18	81.40	38.80	19.83	86.90
ALL	***45.30***	**23.91**	90.30	**41.30**	**21.59**	89.40	35.50	18.38	82.10	**39.80**	**20.07**	87.00

Table 2: Final evaluation results of language-specific models on the test set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85	47.10	25.90	99.90
Baseline	**44.09**	**24.72**	86.67	40.76	22.31	70.61	36.82	19.92	70.76	36.67	**20.60**	76.21
SMT	43.18	23.96	86.97	**42.58**	**22.98**	90.45	36.06	19.24	85.30	**37.12**	19.69	89.24
SMT+Rank	42.88	23.90	87.12	40.76	22.31	89.70	**38.33**	**20.57**	91.52	36.52	20.16	90.91
ALL	43.33	23.71	88.48	40.00	21.80	88.64	37.73	20.16	90.00	36.21	20.49	88.03

3.2 Translation system

To translate the queries into the collection language (English), we use the SMT system developed within the Khresmoi project[2] [1] and KConnect project[3]. The SMT system is based on Moses, a state-of-the-art phrase-based SMT system [7]. It is trained on a mixture of data from medical and general domain and adapted to translate search queries rather than full sentences.

Queries are typically short sequences of terms with free order and no additional context. Such properties make translating queries difficult for the normal SMT systems that are optimized to translate normal sentences; Therefore feature weights in the SMT system are optimized towards PER (Position-independent word Error Rate [14]) rather then the traditional BLEU [**?**]. The Khresmoi SMT system includes models to translate from Czech, French, German, Hungarian, Polish, Spanish, and Swedish into English.

3.3 Baseline

We use the Terrier open source search engine [**?**] to index the collection and to perform the retrieval. Queries are generated by translating the multilingual queries into English using the Khresmoi SMT system and taking the *1-best-list* translation. Then Terrier's implementation of the language model with the Bayesian smoothing and Dirichlet prior is used to retrieve the 1000 highest ranked documents.

Results are evaluated using *trec_eval*.[4] Mainly we consider P@10 as the main metric, also we report MAP [9] and coverage (CVG) which is the percent of assessed documents among the top 10 retrieved documents. For significance tests, we use the paired Wilcoxon signed-rank test [6] with $\alpha = 0.05$. In the oracle experiment, first we translate each query in the training set and take the *15-best-list* translation hypotheses, then we select the hypothesis that gives the highest P@10, finally we take the average of P@10 for all training queries. Oracle results, as shown in Table 3, proves that selecting the best hypothesis for the CLIR has a wide potential space of improvements and can outperform the baseline system for all languages. We also report monolingual results of the system which uses the reference English queries.

3.4 Hypotheses reranking

The SMT system returns a list of translation alternatives, we will refer to it as *n-best-list*. The translations in this list are sorted ascending from the best translation (*1-best-list*). The presented approach is based on our previous work [11], in which we presented a machine learning model that is trained on SMT features and features extracted from *n-best-list* translations, document collection, retrieval system, Wikipedia articles and UMLS Metathesaurus. These features are used to train generalized linear regression with logit link function model, and optimized to predict the translation that gives the highest P@10. The model is trained using leave-one-out-cross-validation (LOOCV) approach by excluding one query (with its translations) each iteration, to tune the experiment parameters.

4. EXPERIMENTS AND RESULTS

To train the reranker, we first create training set using *15-best-list* translations for each query from the 100 training queries. These translations are used to create one training file contains up to 1500 instances. Then we use these instances separately to build a language-specific models for each language. We experiment with three systems: 1) A system which only uses features derived from the SMT system (denoted as SMT). 2) A system which combines the SMT features and features that are based on the original ranking of the translations (denoted as SMT+Rank) 3) A system exploiting all features that are described in detail in our previous work [11].

Results of the LOOCV evaluation of language-specific models on the training set are shown in Table 1. The italics font refer to results statistically significantly different from the baseline system. All systems are able to significantly outperform the baseline system for the Spanish language only, and the system which uses all features (ALL), gives the best result. When testing the model against the test set, see Table 2, we do not observe any improvement in any language.

[2] http://www.khresmoi.eu/

[3] http://www.kconnect.eu/

[4] http://trec.nist.gov/trec_eval/

Table 3: Cross-validation of language-independent models on the training set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90
Oracle	52.10	25.86	86.50	51.80	24.55	79.40	47.40	22.16	77.50	50.10	23.23	78.70
Baseline	41.90	0.2202	78.80	40.10	20.67	74.40	37.30	**19.55**	72.80	39.60	20.07	73.80
SMT	43.40	22.36	89.10	38.40	19.13	84.00	**38.70**	18.41	81.80	36.10	17.91	85.10
SMT+Rank	43.10	22.49	88.00	40.70	20.78	88.50	36.50	18.87	82.10	39.00	19.75	86.60
ALL	***46.90***	**24.05**	90.80	***42.20***	**21.91**	89.00	36.90	18.61	82.60	**40.00**	**20.12**	86.70

Table 4: Final evaluation results of language-independent models on the test set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85	47.10	25.90	99.90
Baseline	44.09	24.72	86.67	40.76	22.31	70.61	36.82	19.92	70.76	36.67	20.60	76.21
SMT	43.79	23.83	87.42	40.00	22.54	89.09	35.61	19.76	85.61	38.33	19.85	88.64
SMT+Rank	43.64	24.28	86.36	38.94	21.91	89.09	**38.18**	20.21	89.24	36.21	20.02	91.97
ALL	***46.36***	**25.30**	90.15	***43.18***	**23.88**	91.67	36.67	**20.38**	89.39	**38.79**	**21.06**	91.97

The features we presented are source-language-independent, which makes merging data from different languages possible in order to train the machine learning model and expand the training data. For each of the new languages (Spanish, Hungarian, Polish, and Swedish), we merge the translations from that specific language with the available training data for the the original language (Czech, German, French) to create a richer training data set.

Results for all systems that use the merged data in the LOOCV experiments are shown in Table 3. For the Spanish and Hungarian languages, the system combining all features (ALL) significantly outperforms the baseline system. A small and not statistically significant improvement is observed for Swedish by the system based on all the features and for Polish by the system based on the SMT features only (SMT).

Table 4 shows the results for the models that are trained on expanded data set against the test set. Systems exploiting all the features (ALL) in Spanish and Hungarian outperforms the baseline system significantly. We observe in the test results by our best system (ALL) 11 queries improved in Spanish, 8 in Hungarian, 5 in Polish and 7 in Swedish. Also there are degradations of 5 queries in Spanish, 3 in Hungarian, 5 in Polish and 3 in Swedish. The impact of untranslated terms appears mostly in the Polish language.

For example, query *2015.37: łuszcząca skin* has P@10 = 00.00, (reference translation: *scaly skin*). It contains the untranslated term *łuszcząca*, which means *scaly* in English. The monolingual query has P@10 = 99.00, the difference in performance is caused by the untranslated (out-of-vocabulary, OOV) words only.

A similar situation appears in query *2015.35* (Monolingual English query: *lot of irritation with contact lenses*), its P@10 = 00.00. The translated query is *significant irritation szkłami kontaktowymi.* It contains two untranslated terms: *szkłami* (lenses) and *kontaktowymi* (contact). These two untranslated words destroy the query.

Query *2015.29* in Spanish has P@10 = 30.00 in the baseline, its translation is *red patch on the skin and dry pus.* The (ALL) system improves it to P@10 = 90.00 and selects the translation *red patch on the skin and dry pus blister.*

Another example of improvement is observed in query *2013.32*, the baseline translation is *dyspnoea* with P@10 = 60.00, the selected translation is *shortness of breath* with P@10 = 90.00. The reference translation is *SOB* with P@10 = 50.00, and this is one case in which the best system outperforms not only the baseline system but also the monolingual one.

We find in the translated Spanish queries that we have a total of 10 terms which the SMT system was unable to translate (OOV) which harm the performance. There are also 20 OOVs in the Hungarian queries, while in the Swedish and Polish the case is worse, where there are 40 OOVs in Swedish and 54 OOVs in Polish queries. Usually the SMT system can not translate some terms because they did not appear in the parallel data. So for our case, having this OOVs translated into English correctly definitely improves the retrieval results. The problem of OOVs in the CLIR can be a subject of further investigation in the future.

5. CONCLUSION

In this paper, we presented our approach to adapt an SMT query translation reranker in the cross-lingual information retrieval task to new languages. The new languages suffer from low assessment coverage in their baseline systems, leading to low quality data to train the reranker model with. Our approach tackled this problem by exploiting data which were taken from languages whose retrieval systems were fully assessed (Czech, French and German) by medical experts we asked them to do so in our previous work. Data were merged from the fully assessed languages together with the data from one new language in order to train the model. Firstly, we created one training set for each of the new languages to build a source-language-specific model. This approach could not bring significant improvement to the baseline system. Then, we used the expanded data to train reranker models for the new languages. This approach significantly outperformed the baseline systems for Spanish and Hungarian. However, it could not significantly improve the baseline in Polish and Swedish, because the SMT system produces high number of OOVs in these languages comparing to Spanish and Hungarian.

Acknowledgments

This research was supported by the Czech Science Foundation (grant n. P103/12/G084) and the EU H2020 project KConnect (contract n. 644753).

6. REFERENCES

[1] O. Dušek, J. Hajič, J. Hlaváčová, M. Novák, P. Pecina, R. Rosa, and et al. Machine translation of medical texts in the Khresmoi project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228, Baltimore, USA, 2014.

[2] A. Fujii and T. Ishikawa. Applying machine translation to two-stage cross-language information retrieval. In *Envisioning Machine Translation in the Information Future*, volume 1934, pages 13–24. Springer, Berlin, Germany, 2000.

[3] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones, and H. Mueller. ShARe/CLEF eHealth evaluation lab 2014, Task 3: User-centred health information retrieval. In *Proceedings of CLEF 2014*, 2014.

[4] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth evaluation lab 2015. In *The 6th Conference and Labs of the Evaluation Forum*, Berlin, Germany, 2015. Springer.

[5] F. Hieber and S. Riezler. Bag-of-words forced decoding for cross-lingual information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, Denver, Colorado, 2015.

[6] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, Pittsburgh, USA, 1993.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, and et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007.

[8] V. Nikoulina, B. Kovachev, N. Lagos, and C. Monz. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 109–119, Avignon, France, 2012.

[9] P. Pecina, O. Dušek, L. Goeuriot, J. Hajič, J. Hlavářová, G. J. Jones, and et al. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165–185, 2014.

[10] S. Saleh and P. Pecina. CUNI at the ShARe/CLEF eHealth Evaluation Lab 2014. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, volume 1180, pages 226–235, Sheffield, UK, 2014.

[11] S. Saleh and P. Pecina. Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016*, Évora, Portugal, 2016. Springer.

[12] A. Sokolov, F. Hieber, and S. Riezler. Learning to translate queries for CLIR. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1179–1182, Gold Coast, Australia, 2014.

[13] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, and et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer, Berlin, Germany, 2013.

[14] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based search for statistical translation. In *In European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece, 1997.

[15] F. Ture and E. Boschee. Learning to translate: A query-specific combination approach for cross-lingual information retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 589–599, Qatar, 2014.

[16] D. Wu and D. He. A study of query translation using Google machine translation system. In *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*, pages 1–4, Wuhan,China, 2010.