Machine Translation Meets Frequent Case Generation in Query Translation Based CLIR

Kimmo Kettunen
Department of Information Studies, University of Tampere
Kanslerinrinne 1, FIN-33014
Tampereen yliopisto
kimmo.kettunen@uta.fi

Abstract

In this paper we introduce evaluation results of Cross-language information retrieval for two small languages, Finnish and Swedish. Our approach is based on machine translation of topics and usage of the Frequent Case Generation method for management of query term variation in translated topics. Retrieval results of more standard query term variation management approaches, such as stemming and lemmatization of translated topics, are also shown.

1 Introduction

Cross-language information retrieval (CLIR) has become one of the research areas in information retrieval during the last 10 years. The development and success of WWW has been one of the key factors that has increased interest in retrieval tasks where the language of queries is other than that of the retrieved documents. There are vast amounts of textual data in various languages available electronically and the textual and linguistic abundance increases constantly. Thus there is and will be a social need for retrieval systems, where the user can state his/her search request in native language and get the documents in another language that he/she is capable of understanding to the extent that some information need is satisfied. Although real finished applications of CLIR in the Web still mostly don't exist (despite Google's Translated Search), it could be approximated that some sort of CLIR applications may reach maturity during 5-10 years.

CLIR has many approaches. One of the most popular approaches to CLIR has been query translation. When queries are translated, different methods can be used: either the queries are translated with electronic dictionaries or word lists, with machine translation programs or using large parallel corpora as translation's knowledge source. All these query translation methods have been successful and they can also be mixed. Recently much research has been done using parallel corpora as translation resource, but also all the older methods flourish. (Abusalah et al., 2005; Kishida, 2005; Oard and Riekema, 1998).

2 Frequent Case Generation and MT based CLIR

In this paper we shall combine available machine translation programs of two small languages into FCG, Frequent Case Generation, a recent method for management of query term variation. Machine translation has been used in CLIR as a query translation tool e.g. for English, German, French and Spanish, but not much for small languages like Swedish or Finnish. FCG, on the other hand, has been quite recently introduced to monolingual management of query term variation (Kettunen, 2008; Kettunen et al., 2007). It has proven quite successful in management of query term variation for morphologically complex or moderately complex languages. Thus it is of interest to verify, if the method can be used in CLIR of these same languages. Airio and Kettunen (2008) have tried FCG successfully in CLIR, but in this context it was used with a dictionary-based query translation tool, Utaclir (Hedlund, 2003; Hedlund et al., 2004).

We shall report evaluation results of machine translated queries from English to Finnish and Swedish. Materials of CLEF 2003 are used in the tests and the process of query translation and retrieval is arranged as follows:

1) English CLEF 2003 topics are first translated to target languages with available machine translation programs for each language. We translated separately title and title and description fields from the topics. Some of the used MT programs are free web versions, some commercial programs that have been used under test license. Used programs for En à Sv translation are Systran's web translator (http://www.systran.co.uk/), Google Translate Beta (http://translate.google.com/translate_t) and Tolken99 (http://www.tolken99.net/, version 4.2), a MT program for PCs. En à Fi translations are done with Sunda's MT program (www.sunda.fi), Google Translate and Teemapoint's MT program (www.teemapoint.fi, version 1.3).
2) After translation the translated topics are normalized morphologically with FINTWOL and SWETWOL lemmatizers respectively. Lemmatized translated topics are sent to FCG procedures that generate variant keyword forms for nouns and adjectives of each language's queries. The final translated FCG queries are run in the textual database of the target language in Lemur query engine and results are evaluated with *trec.eval*. For comparison also IR results of lemmatized, stemmed and plain query translations are shown.

3 Conclusion

Results of our tests show, that at best the proposed MT+FCG CLIR technique works at least as well as usage of a more standard dictionary-based query translation approach combined with FCG (Airio and Kettunen, 2008). Achieved IR results depend mostly on the quality of the MT program: some of the translation programs used in the tests seem to translate topics much better, while some produce quite low level translations. Worst results are achieved with Systran's Swedish web translator. PC based MT program for Swedish, Tolken99, is able to translate the queries quite well, and Google Translate succeeds really good in Swedish. Sunda's Finnish MT program, Google Translate and Teemapoint's translator are more even in their translation capabilities, at least from the query point of view, although Google Translate seems to get the best results most of the times. Translated Finnish queries yield at best very good performance that many times outperforms performance of a dictionary-based query translation method.

Acknowledgments

This work was supported by the Academy of Finland grant number 1124131. We wish to thank Ms. Eija Airio, Dept. of Information Studies, University of Tampere, for implementing all the Unix scripts for the query processes.

References

Mustafa Abusalah, John Tait and Michael Oakes. 2005. Literature Review of Cross Language Information Retrieval. *Transactions on Engineering, Computing and Technology* V4: 175–177.

Eija Airio and Kimmo Kettunen. 2008. Does Dictionary Based Bilingual Retrieval Work in Non-Normalized Index? Submitted.

Turid Hedlund. 2003. Dictionary-based Cross-language Information Retrieval. Acta Universitatis Tamperensis 962.

Turid Hedlund, Eija Airio, Heikki Keskustalo, Raija Lehtokangas, Ari Pirkola and Kalervo Järvelin. 2004. Dictionary-based Cross-language Information Retrieval: Learning Experiences from CLEF 2000-2002. *Information Retrieval* 7: 99–119.

Kimmo Kettunen. 2008. Automatic Generation of Frequent Case Forms of Query Keywords in Text Retrieval. In Bengt Nordström and Aarne Ranta (eds.), *Advances in Natural Language Processing*, GoTAL 2008, LNAI 5221: 222–236.

Kimmo Kettunen, Eija Airio and Kalervo Järvelin. 2007. Restricted Inflectional Form Generation in Management of Morphological Keyword Variation. *Information Retrieval* 10(4-5): 415–444.

Kazuaki Kishida. (2005). Technical Issues of Cross-Language Information Retrieval: a Review. *Information Processing & Management* 41: 433–455.

Douglas Oard and Anne Riekema. 1998. Cross-language information retrieval. In Martha E. Williams (ed.), *Annual Review of Information Science and Technology* (ARIST), volume 33, 223–256.