Proceedings of HLT-NAACL 2003
Demonstrations , pp. 1-2
Edmonton, May-June 2003

TIPS: A Translingual Information Processing System

Y. Al-Onaizan, R. Florian, M. Franz, H. Hassan, Y. S. Lee, S. McCarley, K. Papineni, S. Roukos, J. Sorensen, C. Tillmann, T. Ward, F. Xia

IBM T. J. Watson Research Center
Yorktown Heights

Abstract

Searching online information is increasingly a daily activity for many people. The multilinguality of online content is also increasing (e.g. the proportion of English web users, which has been decreasing as a fraction the increasing population of web users, dipped below 50% in the summer of 2001). To improve the ability of an English speaker to search mutlilingual content, we built a system that supports cross-lingual search of an Arabic newswire collection and provides on demand translation of Arabic web pages into English. The cross-lingual search engine supports a fast search capability (sub-second response for typical queries) and achieves state-of-the-art performance in the high precision region of the result list. The on demand statistical machine translation uses the Direct Translation model along with a novel statistical Arabic Morphological Analyzer to yield state-of-the-art translation quality. The on demand SMT uses an efficient dynamic programming decoder that achieves reasonable speed for translating web documents.

Overview

Morphologically rich languages like Arabic (Beesley, K. 1996) present significant challenges to many natural language processing applications as the one described above because a word often conveys complex meanings decomposable into several morphemes (i.e. prefix, stem, suffix). By segmenting words into morphemes, we can improve the performance of natural language systems including machine translation (Brown et al. 1993) and information retrieval (Franz, M. and McCarley, S. 2002). In this paper, we present a cross-lingual English-Arabic search engine combined with an on demand Arabic-English statistical machine translation system that relies on source language analysis for both improved search and translation. We developed novel statistical learning algorithms for performing Arabic word segmentation (Lee, Y. et al 2003) into morphemes and morphological source language (Arabic) analysis (Lee, Y. et al 2003b). These components improve both mono-lingual (Arabic) search and cross-lingual (English-Arabic) search and machine translation. In addition, the system supports either document translation or convolutional models for cross-lingual search (Franz, M. and McCarley, S. 2002).

The overall demonstration has the following major components:

1. Mono-lingual search: uses Arabic word segmentation and an okapi-like search engine for document ranking.
2. Cross-lingual search: uses Arabic word segmentation and morphological analysis along with a statistical morpheme translation matrix in a convolutional model for document ranking. The search can also use document translation into English to rank the Arabic documents. Both approaches achieve similar precision in the high precision region of retrieval. The English query is also morphologically analyzed to improve performance.
3. OnDemand statistical machine translation: this component uses both analysis components along with a direct channel translation model with a fast dynamic programming decoder (Tillmann, C. 2003). This system

achieves state-of-the-art Arabic-English translation quality.

4. Arabic named entity detection and translation: we have 31 categories of Named Entities (Person, Organization, etc.) that we detect and highlight in Arabic text and provide the translation of these entities into English. The highlighted named entities help the user to quickly assess the relevance of a document.

All of the above functionality is available through a web browser. We indexed the Arabic AFP corpus about 330k documents for the demonstration. The resulting search engine supports sub-second query response. We also provide an html detagging capability that allows the translation of Arabic web pages while trying to preserve the original layout as much as possible in the on demand SMT component. The Arabic Name Entity Tagger is currently run as an offline process but we expect to have it online by the demonstration time. We aslo include two screen shots of the demonstration system.

Acknowledgments

This work was partially supported by the Defense Advanced Research Projects Agency and monitored by SPAWAR under contract No. N66001-99-2-8916. The views and findings contained in this material are those of the authors and do not necessarily reflect the position of policy of the Government and no official endorsement should be inferred.

References

Beesley, K. 1996. Arabic Finite-State Morphological Analysis and Generation. *Proceedings of COLING-96*, pages 89– 94.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. 1993. The mathematics of statistical machine translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263–311.

Franz, M. and McCarley, S. 2002. Arabic Information Retrieval at IBM. *Proceedings of TREC 2002*, pages 402–405.

Lee, Y., Papineni, K., Roukos, S., Emam, O., and Hassan, H. 2003. Language Model Based Arabic Word Segmentation. Submitted for publication.

Lee, Y., Papineni, K., Roukos, S., Emam, O., and Hassan, H. 2003b. Automatic Induction of Morphological Analysis for Statistical Machine Translation. Manuscript in preparation.

Tillmann, C., 2003. Word Reordering and a DP Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1): 97-133.