A Transfer-based Hebrew-to-English Machine Translation System

Progress report and request for renewal

Alon Lavie
LTI, School of Computer Science
Carnegie Mellon University
alavie@cs.cmu.edu

Shuly Wintner
Department of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

July 17, 2005

1 Project summary

We propose to develop a preliminary Hebrew-to-English Machine Translation (MT) system under a transfer-based framework specifically designed for rapid MT prototyping for languages with limited linguistic resources. The task is particularly challenging due to two main reasons: the high lexical and morphological ambiguity of Hebrew and the dearth of available resources for the language. We will use existing, publicly available resources and adapt them in novel ways to support the MT task. The methodology behind the approach is based on two separate modules: a transfer engine which produces a lattice of possible translation segments, and a decoder which searches and selects the most likely translation according to an English language model.

We have already constructed an initial end-to-end Hebrew-to-English system under this framework, in the course of Dr. Lavie's CRI-funded two-month visit to Haifa University in 2004. We wish to extend this initial work into a two-year small-scale research project, for which we request funding in this proposal. The main work will involve significantly scaling up the coverage and accuracy of the language resources (translation lexicons and morphological analyzer). We will also develop a manually crafted broad-coverage transfer grammar and augment it with automatically acquired transfer rules. Performance will be evaluated on translation of Hebrew newspaper articles, using state-of-the-art measures for translation quality. This system will be the first large-scale Hebrew-to-English MT system ever to be developed. We believe it will have broad, although not immediate, commercial application, and will bootstrap serious future MT and NLP research involving Hebrew.

2 Original work plan

- Morphological analysis: acquisition and adaptation of an existing morphological analyzer, including a disambiguation module. M1–M4

- Acquisition of a bilingual dictionary: we will make use of whatever resources will be available to us from the Hebrew WordNet project, currently under development in Haifa. M1–M6

- Development of transfer rules, including both manually crafted rules and automatically acquired ones. M4–M12
- Generation: adaptation of an English language model. M12–M16
- Integration, including testing and evaluation. M16–M20
- Exploitation: prototype integration of the system in a larger project, probably the CRI/IRST showcase. M16–M20
- Dissemination of results: M20–M24

3 Current status

We adapted, improved and incorporated into the system the morphological analyzer of Yona and Wintner (2005). This is a much improved system, especially in terms of its coverage. It is based on a lexicon of over 20,000 entries, which is constantly being updated and expanded, so that the coverage improves over time. Our machine translation system now works flawlessly with this analyzer.

Work on morphological disambiguation is still underway. We will employ machine learning techniques for this module, and the bottleneck is human annotation of data. We are currently in the midst of annotating a 30,000 word corpus, and we expect the disambiguation tool to be ready by the end of 2005. Note that the disambiguator is fully compatible with the morphological analyzer, and no special adjustments will be needed once it is ready for use.

The bilingual dictionary we use was also greatly improved. The most important feature is that it is now fully integrated with the morphological analyzer: each analysis produced by the analyzer includes a pointer to the lexicon, which stores, among other data, also word translations. Currently, approximately 9,000 Hebrew lexical items have English translations. We intend to extend this number significantly using both automatic processes and manual lexicographic work.

Work on transfer rules did not progress significantly during this year. We currently have a set of approximately 30 manually crafted rules. We experimented with automatically induced rules (Probst and Lavie, 2004) but the results were not very encouraging.

The transfer engine we use was significantly rewritten during the past year. It is now more robust and more flexible. We are also in the process of compiling and incorporating a new language model for English generation.

Preliminary results of this project were published as Lavie et al. (2004).

4 Future plans

Our work in the second year of this project will proceed in two main tracks. One is the further development of resources, and in particular of manually crafted transfer rules. We intend to improve the lexicon, morphological analyzer and bilingual dictionary and integrate a disambiguation module; but we expect that the greatest improvement in the quality of the translation will stem from the incorporation of a partial transfer grammar in the form of transfer rules.

The second track will include integration, exploitation and dissemination of results. Specifically, we will use the machine translation system as a component in a cross-lingual information retrieval system that we currently develop. We also intend to evaluate the quality of the translations using several measures, including BLEU and METEOR (Lavie, Sagae, and Jayaraman, 2004).

Incidentally, we were contacted by the owner of an Israeli translation company who is interested in this project. We are currently negotiating with him the possibility of obtaining external funding for future development of the system, in return for some share of the intellectual property.

4.1 Budget

Our 2005 budget was divided between salaries for research assistants ($5000) and travel ($2500 for a visit of Alon Lavie to Haifa, $2000 for a visit of Shuly Wintner to CMU). For the second year we expect similar expenses, as follows:

One research assistant (half time), including lexicographic work, translations and grammar development	$5,000
One visit of Dr. Lavie to Haifa	$2,500
One visit of Dr. Wintner to CMU	$2,500
Total	**$10,000**

References

Frederking, Robert E. and Kathryn Taylor, editors. 2004. *Machine Translation: From Real Users to Research, 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004, Proceedings*, volume 3265 of *Lecture Notes in Computer Science*. Springer.

Lavie, Alon, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The significance of recall in automatic metrics for mt evaluation. In Frederking and Taylor (Frederking and Taylor, 2004), pages 134–143.

Lavie, Alon, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Probst, Katharina and Alon Lavie. 2004. A structurally diverse minimal corpus for eliciting structural mappings between languages. In Frederking and Taylor (Frederking and Taylor, 2004), pages 217–226.

Yona, Shlomo and Shuly Wintner. 2005. A finite-state morphological grammar of Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 9–16, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Limited machine translation and automatic speech recognition training data will be provided from multiple low resource languages to enable performers to learn how to quickly adapt their methods to a wide variety of materials in various genres and domains. As the program progresses, performers will apply and adapt these methods in increasingly shortened time frames to new languages.Â Performers will be evaluated, relative to a baseline system, on their ability to accurately retrieve materials relevant to an English domain-specific query from a database of multi-domain, multi-genre documents in a low resource language, and their ability to convey the relevance of those documents through summaries presented to English speaking domain experts. Machine translation, the process of automatically translating text from a source lan-guage (e.g. English) to a target language (e.g. French), has achieved impressive results in recent years. However, modern machine translation methods rely heavily on paral-lel data â€“ millions of sentences translated from the source to the target language. Such parallel data is not readily available for most pairs of source and target languages. The goal of this thesis is to explore ways of using other types of data to improve the trans-lations generated by machine translation systems. We consider two main type... Facebook says it has developed the first machine learning model to translate between 100 languages without going into English first.Â Facebook says the new multilingual machine translation model was created to help its more than two billion users worldwide. The company is still testing the translation system â€“ which it calls M2M-100 - and hopes to add it to different products in the future. The social media service says it has made the system open source -- meaning its computer code will be freely available for others to copy or change. Angela Fan, a research assistant at Facebook, explained the new machine translation model this week on one of the companyâ€™s websites. To translate a corpus of English text to French, we need to build a recurrent neural network (RNN). Before diving into the implementation, letâ€™s first build some intuition of RNNs and why theyâ€™re useful for NLP tasks. RNN Overview. RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both. Theyâ€™re called recurrent because the networkâ€™s hidden layers have a loop in which the output and cell state from each time step become inputs at the next time step. This recurrence serves as a form of memory. It allows contextual information to flow through the network so that relevant outputs from previous time steps can be applied to network operations at the current time step.