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

Project Proposal

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

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 this past Spring. The performance of the system developed so far exceeds our expectations. We now 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. The system constructed 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 Project objectives

We propose a project whose objective is the construction of a preliminary Hebrew to English machine translation system. Machine translation is among the most ambitious natural language processing applications, and work on Hebrew is particularly challenging.

Computational processing of Hebrew is complicated by the high level of ambiguity posed by the language. The Hebrew script,[1] not unlike the Arabic one, attaches several short particles to the word which immediately follows them. These include, *inter alia*, the definite article *h* ("the"), prepositions such as *b* ("in"), *k* ("as"), *l* ("to") and *m* ("from"), subordinating conjunctions such as *$* ("that") and *k$* ("when"),

[1] To facilitate readability we use a transliteration of Hebrew using ASCII characters in this paper.

relativizers such as *$* ("that") and the coordinating conjunction *w* ("and"). The script is rather ambiguous as the prefix particles can often also be parts of the stem. Thus, a form such as *mhgr* can be read as a lexeme "immigrant", as *m-hgr* "from Hagar" or even as *m-h-gr* "from the foreigner".

An added complexity arises from the fact that there exist two main standards for the Hebrew script: one in which vocalization diacritics, known as *niqqud* "dots", decorate the words, and another in which the dots are omitted, but where other characters represent some, but not all of the vowels. Most of the modern printed and electronic texts in Hebrew use the "undotted" script. While a standard convention for this script officially exists, it is not strictly adhered to, even by the major newspapers and in government publications. Thus, the same word can be written in more than one way, sometimes even within the same document. For example, the word *niqion* "cleaning" can occur also as *nqion*. This fact adds significantly to the degree of ambiguity, and requires creative solutions for practical Hebrew language processing applications.

The challenge involved in constructing an MT system for Hebrew is amplified by the poverty of existing resources. The collection of corpora for Hebrew is still in early stages and all existing significant corpora are monolingual. Hence the use of aligned bilingual corpora for MT purposes is currently not a viable option. There is no available large Hebrew language model which could help in disambiguation. Good morphological analyzers are proprietary and publicly available ones are limited. No publicly available bilingual dictionaries currently exist, and no grammar is available from which transfer rules can be extracted.

The system we propose to develop is based on a transfer-based framework specifically designed for rapid MT prototyping for languages with limited linguistic resources. This framework has been under development for the past two years by the MT research group at Carnegie Mellon, under the leadership of Dr. Lavie and his colleagues. While the framework itself is still research work in progress, it is sufficiently well developed for serious experimentation. The framework includes a declarative formalism for symbolic transfer grammars. A grammar consists of a collection of transfer rules, which specify how phrase structures in a source-language correspond and transfer to phrase structures in a target language, and the constraints under which these rules should apply. The framework also includes a fully-implemented transfer engine that applies the transfer grammar to a source-language input sentence at runtime, and produces all possible word and phrase-level translations according to the grammar. This framework was specifically designed to support advanced research on methods for automatically acquiring transfer grammars from limited amounts of elicited word-aligned data. The framework also supports manual development of transfer grammars by experts familiar with the two languages.

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 this past Spring. The performance of the system developed so far is very promising and exceeds our expectations. The general architecture of the system is already in place and fully implemented. The extended system we propose will follow the same design principles, which we outline below.

The system consists of the following main components: a Hebrew input sentence is pre-processed, and then sent to a *morphological analyzer*, which produces all possible analyses for each input word, represented in the form of a lattice of possible input word lexemes and their morphological features. The input lattice is then passed on to the *transfer engine*, which applies a collection of lexical and structural *transfer rules* in order to parse, transfer and generate English translations for all possible word and phrase segments of the input. This comprehensive collection of output segments is stored in an output lattice data-structure. The lexical transfer rules used by the transfer engine are derived from a *bilingual dictionary*, while the higher-level structural transfer rules come from either a manually-developed or automatically-acquired *transfer grammar*. In the final stage, the English lattice is fed into a *decoder* which uses a *language model* of English to search and select a combination of sequential translation segments that together represent the most likely translation of the entire input sentence. A schematic diagram of the system architecture can be seen in Figure 1.

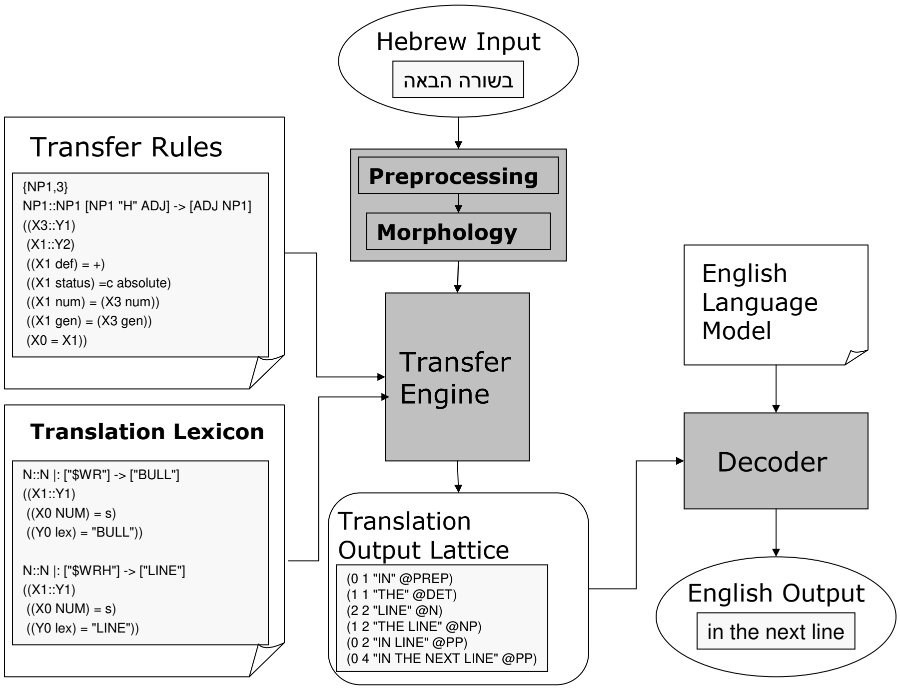

Figure 1: Architecture of the Hebrew-to-English Transfer-based MT System and its Major Components

maxwell anurpung comes from ghana for israel four years ago and since worked in cleaning in hotels in eilat
a few weeks ago announced if management club hotel that for him to leave israel according to the government instructions and immigration police
in a letter in broken english which spread among the foreign workers thanks to them hotel for their hard work and announced that will purchase for hm flight tickets for their countries from their money

Figure 2: Select Translated Sentences from the Development Data

The current system was developed over the course of a two month period, with a total labor-effort equivalent to about four person-months of development. Most of this time was devoted to the construction of the bilingual lexicon and stabilizing the front-end Hebrew processing in the system (Morphology and input representation issues). Once the system was reasonably stable, we devoted about two weeks of time to improving the system based on a small development set of data. For development we used a set of 113 sentences from the Hebrew daily *HaAretz*. Average sentence length was approximately 15 words. Development consisted primarily of fixing incorrect mappings before and after morphological processing and modifications to the bilingual lexicon. The small transfer grammar was also developed during this period. Given the limited resources and the limited development time, we find the results to be highly encouraging. For many of the development input sentences, translations are reasonably comprehensible. Figure 2 contains a few select translation examples from the development data.

3 Project start date and duration

Start date: October 2004. Duration: two years.

4 Participant list and roles

The project will involve two principal investigators, Dr. Shuly Wintner and Dr. Alon Lavie, and two research assistants, Yaniv Eytani in Haifa and Katharina Probst at CMU. Other members of the CMU MT research group will actively collaborate in the research.

Shuly Wintner is a Lecturer with the Department of Computer Science, University of Haifa. His areas of interest include computational linguistics and natural language processing, and in particular linguistic formalisms, formal grammars, mathematics of language and language engineering. In addition, he is involved in several projects whose aim is to develop tools and resources for processing Hebrew and Arabic. He is the author of several research articles in various areas of computational linguistics.

Alon Lavie is an Associate Research Professor in the Language Technologies Institute at Carnegie Mellon University. He holds BS and MS degrees in Computer Science from the Technion, and received a PhD in Computer Science from Carnegie Mellon. His main areas of research are Machine Translation of both text and speech, and Spoken Language Understanding. His current most active research is on the design and development of new approaches to Machine Translation for languages with limited amounts of data resources. He has also worked extensively on the design and development of Speech-to-Speech Machine Translation systems and on robust parsing algorithms for analysis of spoken language. Alon is a co-PI of the AVENUE project (funded by NSF/ITR), which is concerned with the design and rapid development of new Machine Translation methods for languages for which only scarce resources are available. In the past, he was a co-PI of the Nespole! project, funded jointly by the European Commission and the US NSF, whose main goal was to advance the state-of-the-art of speech-to-speech translation in a real-world setting of common users involved in e-commerce applications. The project was a collaboration between three European research labs (ITC-irst in Trento Italy, ISL at University of Karlsruhe in Germany, CLIPS at UJF in Grenoble France), Alon's research group at CMU, and two industrial partners (APT - the Trentino provincial tourism bureau, and AETHRA – an Italian tele-communications commercial company).

5 Innovation

5.1 Current state of the art

The current trend in machine translation is leaning towards *statistical MT*. Several research state-of-the-art MT systems for Chinese-to-English and for Arabic-to-English have been developed under this paradigm in the last few years. While statistical MT can produce impressive results in many cases, such approaches suffer from two limitations. First, they crucially rely on the existence of a very large scale parallel corpus of texts, consisting of aligned texts in the two languages. Such a corpus is unavailable for Hebrew. Second, statistical MT systems are notoriously dependent on the domain for which they were developed (and on which they are trained). Testing such systems on text from different domains yields extremely poor results. Transfer-based systems, on the contrary, are more robust to such changes and are hence more broad-coverage. Most existing state-of-the-art commercial MT systems follow a transfer approach, but are predominantly purely rule-based. Achieving high-levels of coverage and accuracy in these systems has typically required hundreds of person-years of manual labor. These labor-intensive transfer systems are therefore economically feasible for only a small set of language pairs, and the approach is not particularly appealing for Hebrew. The innovative ideas behind the underlying framework of our proposed system have the potential of achieving similar levels of translation performance, with orders-of-magnitude less human labor.

5.2 Innovative aspects of the proposal

As noted above, we are unaware of any machine translation system for Hebrew. This will be the first attempt to address the challenges involved in such a task. Additionally, the use of automatically acquired transfer rules, and an integration of manually crafted resources (including a morphological analyzer and transfer rules) with statistical based methods (including an English language model) sets this project apart from other machine translation systems currently under development.

5.3 Interdisciplinary aspects of the proposal

The construction of a broad coverage machine translation system requires both computational resources and linguistic knowledge. For example, transfer rules are best developed by a bilingual speaker with formal linguistic training and an understanding of the underlying formalism and its capabilities. Acquisition and adaptation of morphological analyzers and bilingual dictionaries require lexicographic work that is best done by linguists. However, the system is first and foremost a computer science endeavor, and state-of-the-art computational solutions will be utilized in its construction.

6 Project work plan, timetable, and milestones

6.1 List of tasks to be performed (with timetable)

- 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

6.2 Description of deliverables list

A publicly available Hebrew to English machine translation system, based on the resources and components described above, will be made available by October 2006.

6.3 Itemized expected expenditures

One research assistant (half time), $5000/year for two years	$10,000
One visit/year of Dr. Lavie to Israel at $2,500 per visit	$5,000
Lexicographic work, incl. manual translation of articles for evaluation, $1,000/year	$2,000
Part-time grammar developer for transfer rules, $1,500/year for two years	$3,000
Total	**$20,000**

7 Exploitation plan

In addition to the research objectives, a successful completion of this project will contribute both to the state of the art in Hebrew Computational Linguistics and to the CRI, and in particular to the CRI/IRST cooperation unit.

Contribution to the resources for processing Hebrew

Contribution to the CRI/IRST showcase

8 Financial agreement between C.R.I. and the project managers

(Milestones and transfers of funds, and distribution of income from the project, if applicable).