Rapid Prototyping of Domain-Specific Machine Translation Systems

Martha Palmer[1], Owen Rambow[2], and Alexis Nasr[3]

[1] IRCS, University of Pennsylvania, Philadelphia, PA 19104, USA
mpalmer@cis.upenn.edu
[2] CoGenTex, Inc., Ithaca, NY, USA
owen@cogentex.com
[3] LIA, Université d'Avignon, France
alexis.nasr@lia.univ-avignon.fr

Abstract. This paper reports on an experiment in assembling a domain-specific machine translation (MT) prototype system from off-the-shelf components. The design goals of this experiment were to reuse existing components, to enable the use of machine-learning tools for parser specialization and for transfer lexicon extraction, and to make the transfer as powerful as possible.

1 Introduction

This paper reports on an experiment in assembling a domain-specific machine translation (MT) prototype system from off-the-shelf components. The design goals of this experiment were as follows:

– By using state-of-the-art off-the-shelf components in parsing and generation, we proposed to profit from recent, independent research on training stochastic parsers on specialized corpora (exhibiting sublanguage effects) in a language for which the parser has been optimized with other corpora.
– By using "lexico-structural transfer" (a transfer based approach on a lexicalized predicate-argument structure), we proposed to avoid the disadvantages of a transfer that is too close to surface detail (for example, transfer at phrase-structure level) while also avoiding the need for devising an interlingua. More specifically, we proposed to profit from recent research aimed at automatically extracting transfer lexicons from bilingual corpora, while still allowing us to specify more complex transfer rules (involving "divergences") at a linguistically motivated level of generality.

To our knowledge, no existing MT system combines these design goals in this manner.

To show how these design goals can be met, we experimented with rapid prototyping of a machine translation system based on lexico-structural transfer (Rambow et al., 1997). We combined retrainable off-the-shelf components with semi-automated methods for transfer lexicon construction. In a six-month effort (with less than 12 man-months, about half of which were academic and half commercial), we were able to quickly develop a system that produces acceptable English to French translations in two limited domains, a battlefield message domain and a weather domain. We also demonstrated limited capability for Arabic (in the weather domain only). The staff included a French (native-speaker) computational linguist who worked on English-to-French transfer and French generation, as well as an Arabic (native speaker) linguist for the small Arabic system.

The structure of this overview paper is as follows. In Section 2, we detail the requirements that motivated our experiment. The system is presented in Section 3. We present the parsers, the transfer component, and the generation component in Sections 4, 5, and 6, respectively. We conclude with some observation for the next Phase of the project in Sectionsec-conc.

2 Special MT Requirements

The military has special machine translation (MT) needs which are not being met by currently available commercial MT systems. These needs center around the domain-specific nature of the data the military would like to be able to translate, e.g., battlefield messages traffic, medical diagnosis routines, military training manuals, intelligence reports and briefing slides, etc. In all of these applications, an accurate, efficient MT system would rely heavily on domain-specific vocabulary. In addition, the military often requires translation to or from "exotic" languages which are of little interest to commercial MT providers. For any specific language, for any specific military application, off-the-shelf products could potentially provide a portion of the necessary grammar and vocabulary, but they would have to be augmented extensively with additional domain-specific vocabulary and grammar rules.[1]

In addition to domain-specific requirements and language-specific requirements, the military has another special need which is not shared by the commercial world — the necessity of timely reaction to sudden crises, which can be in any spot in the world and can arise with no warning. A commercial enterprise can spend months gearing up for a new product launch in a new country, and this preliminary planning time can be spent developing support tools such as machine translation components. IN a world crisis this is not possible, so tools for quickly adapting an existing system to another language are just as essential to the military as domain-specific translation.

These special military requirements can be met, we believe, by an MT system which addresses the design goals outlined in Section 1. Specifically, the design goal of lexico-structural transfer will allow us to handle the domain-specific aspect of military translations, and to exploit machine learning tools for the acquisition of transfer lexicons. The design goal of using off-the-shelf trainable

[1] An example of domain-specific grammar is the use of certain types of telegraphic styles in military messages (omission of subjects and of function words). Furthermore, these telegraphic styles are also difficult to handle in target language generation.

components will allow us to meet the requirement of rapid configuration of new MT systems for new language pairs and domains.

3 Overview of the System

Skies were clear across the three maritime provinces early this morning.
⟶ Le temps était clair dans les trois provinces maritime ce matin tôt.
Behind this area a moderate flow will cause an inflow of milder air in southwestern Quebec producing mild temperatures on Sunday.
⟶ Une circulation modérée provoquera un afflux du air doux dans le sud-ouest du Québec à l'arrière de cette zone produisant des températures douces dimanche.
Loyalty of local civilian officials is questionable.
⟶ La loyauté des dirigeants locaux civils est douteuse.
The 175tr/9gtd is moving west on e4a48 Autobahn toward Berlin.
⟶ Le 175tr/9gtd se déplace vers l'ouest sur e4a48 autobahn vers Berlin.

Fig. 1. Some sample translations performed by TransLex

TransLex is an English-to-French translation system, with a small English-to-Arabic capability. Some sample French outputs can be seen in Figure 1; a sample Arabic output in Figure 2.[2] The main level of representation in TransLex is a syntactic dependency representation which we will call DSyntS, for *Deep Syntactic Structure* (roughly as defined in (Mel'čuk, 1988)). This level of representation contains all the meaning-bearing words of a sentence (nouns, verbs, adjectives, adverbs, and some prepositions), but no function words (determiners, auxiliary verbs, strongly governed prepositions, and so on). The grammatical contribution of function words (determination, tense, aspect, and so on) is represented through features. The meaning-bearing words are related syntactically using a small set of possible relation labels (essentially, different arguments and generic adjuncts). This level of representation is well suited for MT since it abstracts away from superficial grammatical differences between languages.

TransLex consists of the following components:

- Two parsers (the Collins parser and the SuperTagger from the University of Pennsylvania), each with a converter which converts the output from the parser to the DSyntS. (Two parsers are used only experimentally; in an operational context only one parser is needed, of course.)
- The core transfer component.

[2] No Arabic morphological component was implemented for the generator.

```
=========================================
RESULT OF REALIZATION:
=========================================

كان جو صافي عبر ال مقطع ال بحري ثلاث هذا صباح باكرا .
```

Fig. 2. Arab translation (without morphology) generated from *Skies were clear across the three maritime provinces early this morning.*

– The generator (RealPro) from CoGenTex.

To help us develop the transfer lexicons, we used Sable, a component developed at the University of Pennsylvania. The architecture is shown in Figure 3.

4 Two Parsers

We investigated the use of two parsers, both developed previously at Penn, namely the Collins parser (Collins, 1996) and the SuperTagger with Lightweight Dependency Analysis (Joshi and Srinivas, 1994; Srinivas, 1997). These parsers are rather different: the Collins parser is trained on a corpus annotated with phrase-structure parse trees, and uses the probability of specific word-word dependencies to determine the most likely parse. The SuperTagger is trained on a corpus where each lexical item has been annotated with the tree that is associated with it in a correct Tree-Adjoining Grammar parse - "supertags". It uses only these supertags to heuristically determine the most likely parse. We retrained both parsers on 450 messages from our original 500 message data set. The 50 test messages were selected by randomly removing a few messages from each topic of the training set, with the number of messages being proportional to the percentage of messages in that topic. This is very small training set by typical standards for empirical methods, and the performance of both parsers would improve dramatically given more training data. We also paid special attention to military terminology, and had a canonical expansion for military acronyms, many of which have multiple forms.

Neither parser produces an output in the format needed for our transfer module, which uses dependency structures, DSyntS, (see (Nasr et al., 1997) for details). Therefore, "converters" had to be implemented for both parsers. The Collins parser, which outputs a phrase-structure parse tree annotated with head information, uses the Generic Parse Analyzer (GPA) developed at Penn, which has been specialized for outputting a DSyntS during this project. The SuperTagger/LDA outputs a dependency tree which is based on the derivation structure of Tree Adjoining Grammar; while this representation is very close to the DSyntS, it is not identical (see (Rambow and Joshi, 1996)), so a small converter was needed to bridge the gap.

Fig. 3. TransLex system architecture

We hand corrected the parses for the 50 test messages to create a Gold Standard, and then evaluated our two parsers. The Collins parser achieved completely accurate parses for 72.4% of the sentences, with 85.7% having no more than two crossed brackets. We also evaluated the combination of the Collins parser and the GPA against the same messages annotated for deep-syntactic dependency relations (i.e., a DSyntS) and found that 69% of the correct head-argument and head-modifier relations had been found.[3]

there was very little decrease in performance. The best performance from the SuperTagger, 89% correct SuperTag assignments came from training it on a combined corpus of 200,000 WSJ words and the 5,000 word (450 messages) training set. At 65%, The performance of the SuperTagger-LDA-converter combination against on the weather corpus was slightly lower than that of the Collins parser with the GPA.

5 The Core Transfer Component

TransLex can draw on several separate transfer lexicons contained in separate files. These transfer lexicons are represented in an easily readable format, the Multi Lexical Base (MLB) format. Here is an example:

```
@ENGLISH: X [class:verb] (ATTR ALMOST)
@FRENCH: FAILLIR (II X [mood:inf])
```

First, the output of the automatic bilingual lexicon extractor (SABLE – see below) is converted into MLB format. At the current state-of-the-art, an automatically induced bilingual lexicon will not contain the detailed structural correspondences necessary for natural language generation in the target language. Thus, the resulting file is then hand-edited by a linguist or domain specialist. Additional MLB files containing translation lexicons can be entirely hand-crafted, or re-used from other related or even unrelated domains. The MLB files are ordered so that in case of multiple occurrence of a key, the different entries for that key are ranked. Finally, the MLB files are automatically processed to generate a fast loadable version of the transfer rules.

SABLE is a tool for analyzing bilingual corpora (or "bitexts") (Melamed, July 1997). SABLE can induce domain-specific bilingual transfer lexicons (Resnik and Melamed, 1997) using a fast algorithm for estimating a partial translation model. A translation model is a set of transfer pairs, consisting of one word from each language which are (in some context in the bitext) a translation of one another. The model's accuracy/coverage trade-off can be directly controlled via a threshold parameter. (By setting the threshold lower, more transfer pairs are proposed, but fewer of these are likely to be correct.) For example, on our battlefield message corpus of about 5,500 words (backed up by the Hansard corpus)

[3] Recall that our transfer representation does not include function words such as determiners, auxiliaries, and strongly governed prepositions, and is thus closer to a representation of propositional content than most syntactic representations.

we obtained a recall of 73% at a precision of 83%, or a recall of only 32% but at a precision of 91%. This feature makes the model suitable for applications that are not fully statistical such as TransLex.

6 Generation

For generation, we have used RealPro, CoGenTex's sentence realizer (Lavoie and Rambow, 1997). The input representation for RealPro is precisely the DSyntS formalism which we use for transfer. We have constructed a small French grammar. We based this grammar on the English grammar, which we adapted through successive modifications. We used an off-the shelf morphological component for French. (We did not integrate a morphological component for Arabic – which explains the lack of morphology in Figure 2.)

7 Outlook

Over the next two years, we will be building a more robust MT system based on the approach outlined in this paper, for the language pairs English/Korean and Korean/English. We will be choosing one or two domains, at least one of which will be military.

The work will be carried out in collaboration with Systran, Inc., which will enable us to reuse as much as possible existing resources. We will also be coordinating with other government funded English/Korean MT projects at New Mexico State University and Lincoln Labs to avoid duplication of effort. The crucial issues that we will be investigating include to what extent SABLE can be used to build or augment a lareg bilingual transfer dictionary, and to what extent we can rapidly develop a parser for Korean which can be retrained on different corpora.

Acknowledgments

The work reported in this paper was supported by contract DAAL01-97-C-0016 awarded by the Army Research Lab to CoGenTex, Inc., with the University of Pennsylvania as a subcontractor. The work was carried out while Nasr was with CoGenTex. We would like to thank Rich Kaste, project engineer, for his support and feedback. We would also like to thank Benoit Lavoie for his contributions to the implementation of the system, and Richard Kittredge for assistance during the project.

Bibliography

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA.

Joshi, A. K. and Srinivas, B. (1994). Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. In *Proceedings of the* 17^{th} *International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan.

Lavoie, B. and Rambow, O. (1997). RealPro – a fast, portable sentence realizer. In *Proceedings of the Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC.

Melamed, I. D. (July, 1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the ACL-97*, Madrid, Spain.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Nasr, A., Rambow, O., Palmer, M., and Rosenzweig, J. (1997). Enriching lexical transfer with cross-linguistic semantic features. In *Proceedings of the Interlingua Workshop at the MT Summit*, San Diego, California.

Rambow, O. and Joshi, A. (1996). A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena. In Wanner, L., editor, *Current Issues in Meaning-Text Theory*. Pinter, London.

Rambow, O., Nasr, A., Palmer, M., Bleam, T., Collins, M., Kipper, K., Melamed, D., Park, J., Rosenzweig, J., Schuler, W., and Srinivas, B. (1997). Machine translation of battlefield messages using lexico-structural transfer. Technical report, CoGenTex, Inc.

Resnik, P. and Melamed, I. D. (1997). Semi-Automatic Acquisition of Domain-Specific Translation Lexicons. In *Proceedings of the ANLP-97*, Washington, D.C.

Srinivas, B. (1997). *Complexity of Lexical Descriptions and its Relevance to Partial Parsing*. PhD thesis, Computer Science Department, University of Pennsylvania.