Improving Embedded Machine Translation with User Interaction

Ralf D. Brown
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
ralf+@cs.cmu.edu

Abstract

Automated machine translation (MT) of unrestricted texts is known to be error-prone, and thus typically requires human involvement in the translation process to achieve acceptable quality. In a conversational system performing interactive speech translation between two speakers, neither pre-editing nor conventional post-editing is possible, and the users will generally have no direct access to the embedded MT component. Since most translation errors are the result of incorrect selections, one approach to solving the dilemma is to provide the user with access to various alternative translations and allow him to select the one which appears correct in context. If the system remembers the manual selections, it can make the correct selection automatically whenever the same context is encountered thereafter.

1 Introduction

Machine translation (MT) of texts is known to be error-prone, but how can one recover from errors when the translation program is embedded in a larger system in which the user may have no direct access to the translator? Because translation errors are typically the result of making an inappropriate selection when there is ambiguity, one approach is to present the user with various alternative translations from which to select the correct one when the MT program decides on a different, incorrect alternative.

2 The Translation System

The bidirectional DIPLOMAT Rapid-Development Speech-to-Speech translation system[6] consists of four primary components: a speech recognizer for each language, a speech synthesizer for each language, a machine translation program for each translation direction, and a graphical user interface (see Figure 1). The machine translation is performed by Pangloss-Lite [7], a multi-engine translation system[5; 4]. By the simple expedient of providing all candidate translations produced by the various translation engines (in addition to Pangloss-Lite's selection of the best translation) to the user interface, the user is able to perform simple editing of the translated text when the backtranslation contains some errors. If the user wishes, the modified translation result can be stored for use in future translation requests.

Pangloss-Lite uses multiple translation engines (taking advantage of various translation techniques such as dictionary lookup and Example-Based MT) and selects what it considers the best overall translation from among the candidates proposed by the engines. The selection is performed by placing all proposed translations into a chart or lattice structure – one edge per candidate – and using a statistical trigram model of the target language, together with additional weighting factors, to select the best path through the lattice[2]. (This language model is also useful for other translation-related tasks such as OCR error correction of the input text[8].) In ordinary stand-alone operation, the selected edges are concatenated into a string which is output as the final translation of the input text.

When embedded within the DIPLOMAT system, Pangloss-Lite is instead configured to output the entire lattice, with the edges selected by the language model added to the lattice as special arcs. The user interface reads in the entire lattice and extracts the language-model arcs (the lines with the label **:ELM** in Figure 5) to form the initial translation displayed to the user. For verification, the initial translation is back-translated and again the entire lattice is read, with the language-model arcs displayed initially.

Pangloss-Lite also supports on-the-fly updates of the knowledge bases for its Dictionary, Glossary, and Example-Based Machine Translation (EBMT) [3] engines. When sent the appropriate command instead of a string to be translated, it will add a new dictionary translation, phrase-pair, or sentence pair to the exist-

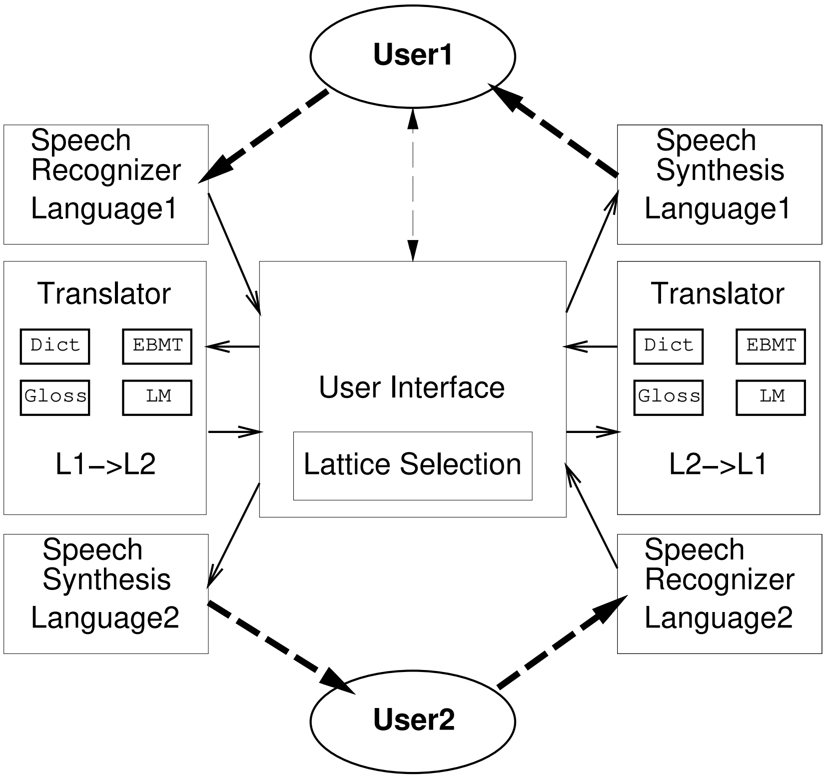

Figure 1: DIPLOMAT system block diagram

ing data for the appropriate engine. These updates are then immediately available for any further translations requested of the system. In the case of the Glossary and EBMT engines, the newly-added translations are given priority over existing data in the knowledge base, since they were presumably added to correct an erroneous translation or supply a translation for an unknown word or phrase.

3 User Interaction

The DIPLOMAT user interface is asymmetric, with all the complexity on the side of a primary user ("interviewer") who is assumed to have had a few hours of training in the system's use (Figure 2). The other user ("interviewee") is presented with a very simple interface containing a text display of what was recognized and verification dialogs with large "Yes" and "No" buttons designed for use with a touch-screen (Figure 3). Should a portion of the back-translation of the interviewer's utterance (or of the forward translation of the interviewee's utterance) which is displayed on screen look "odd" or incorrect to the primary user, (s)he can invoke an editing function by clicking the right-hand mouse button on the appropriate portion of the text. This pops up a list of all the candidates present in the lattice beginning at the same position as the arc on whose word(s) the user clicked (see Figure 4). On selecting one of the items in the pop-up menu, which includes the untranslated word or phrase in case of erroneous translation of a proper name, the selected item is substituted for the arc on which the user clicked. If the new arc covers a different number of words than the one which was previously used, the remainder of the text is adjusted as necessary. If the new arc is longer, it will also replace one or more of the arcs following the one being replaced. If the new arc is shorter, one or more arcs are added, using language-model arcs if possible.

In many cases, such as in Figure 5 (an English-to-Croatian translation), the correct term is among the alternatives available for the back-translation, which gives the user some confidence that the error occurred during back-translation and that the original forward translation is probably correct. If the correct term is not present, and the user has some rudimentary knowledge of the other language, he may be able to guess at the appropriate alternative translation for the text in the foreign-language field. After editing that text, a new back-translation can be generated to verify the guess.

On-the-fly updates to the Pangloss-Lite engines can be used to memorize the results of the user's modifications to the translation output. A very early version of the DIPLOMAT user interface provided a dialog with

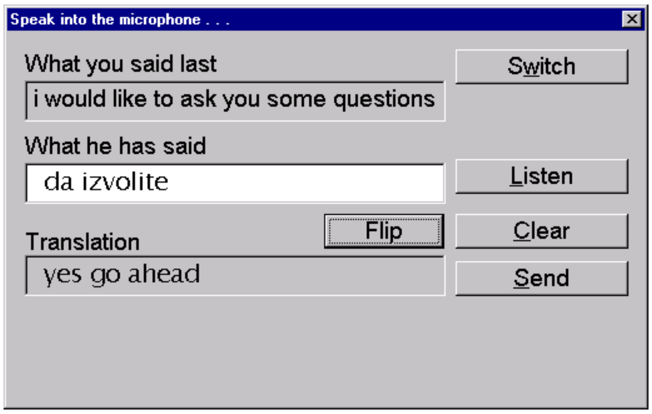

Figure 2: User Interface: Primary User (Interviewer)

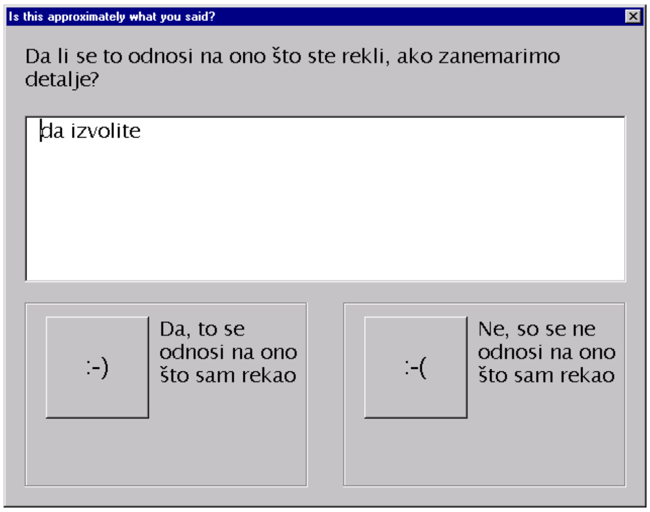

Figure 3: User Interface: Interviewee

which the user could specify a word or phrase and its translation to be added to the dictionary or glossary (as appropriate) knowledge base. The current user interface allows the user to click a button to store the results of any changes made by selecting alternatives from the lattice. Whenever the same utterance is processed thereafter, the correction will already have been made, saving the user time and effort.

These two capabilities – selecting alternatives from the translation lattice and updating the knowledge base with a corrected translation – allow the system's output to be improved without requiring either the user or the user-interface code to know details of the machine translation system's inner workings. This text-driven updating is made possible by the translation technologies – both glossaries and the Example-Based MT engine's knowledge base require nothing more for an update than a phrase or sentence together with its translation. Additional knowledge sources such as word roots and language models can remain hidden from the end user.

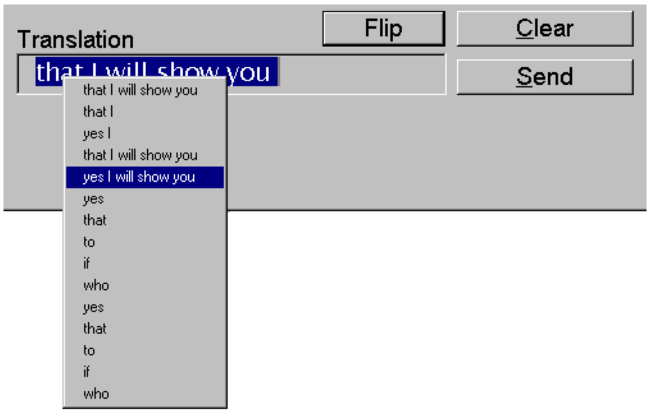

Figure 4: The Lattice Editor

4 User Interaction aiding Rapid Development

Because of the simplicity of the data required by the dictionary and EBMT engines, a new system can be generated extremely rapidly – one of the aims of the DIPLOMAT project – given sufficient parallel text (at least one million words for general coverage, but more is always better). Table 1 shows the time required to produce a bidirectional English-French system using Pangloss-Lite, beginning with the Hansard corpus of Canadian parliamentary proceedings available from the Linguistic Data Consortium. Note that French had never previously been used with Pangloss-Lite – the given time-line is the actual elapsed time for generating a new language for the very first time, and includes a total of nearly twelve hours of interruptions in processing (i.e. subsequent rebuilds "from scratch" will go even more quickly). Figure 6 shows a sample translation from that initial rapid-development system. Where parallel text is unavailable, such as for low-density languages like Haitian Creole, it can be produced by hiring native speakers to translate selected texts; these translators need not have any particular knowledge of linguistics, just as the person who generates an MT system from the parallel text need not have any particular knowledge of the language.

The multi-engine MT approach relies on a statistical language model as a primary knowledge source for deciding between competing translations, but language models typically require enormous amounts of text by the standards of low-density languages – 50 megabytes of text would be considered a small training set, yet is likely to be more than the sum total of all electronically-available text in the language. As a result, when only a limited amount of text is available even monolingually, the language model is of lower quality and more prone to making wildly erroneous selections. Here, too, the user

```
Input:       were any people hurt
Translation: su kojekakav narodski povrijedjen
Backtrans1:  are any people hurt
Lattice:     0 0 :ELM 0.019813 "are"
             0 0 :GLOSS 1 "were"
             0 0 :GLOSS 1 "are"
             1 1 :ELM 0.0105649 "any"
             1 1 :GLOSS 1 "any"
             1 1 :GLOSS 1 "some"
             1 1 :GLOSS 1 "various"
             2 2 :ELM 0.00663629 "people"
             2 2 :GLOSS 1 "people"
             2 2 :GLOSS 1 "nation"
             2 2 :GLOSS 1 "ethnic group"
             2 2 :GLOSS 1 "the common people"
             2 2 :GLOSS 1 "the common folk"
             2 2 :GLOSS 1 "the masses"
             2 2 :GLOSS 1 "crowd"
             2 2 :GLOSS 1 "population"
             3 3 :ELM 7.11531e-05 "hurt"
             3 3 :DICT 0.9 "wounded"
             3 3 :DICT 1 "hurt"
             -1 -1 :INPUT -1 "su kojekakav
                       narodski povrijedjen"
Backtrans2:  were any people hurt
```

Figure 5: Using the lattice to modify a backtranslation

interaction comes into play, with the user's selections providing what amounts to a supplementary language model. Thus the system becomes usable earlier in the development phase when the corpus must be manually generated than it would have been without user interaction. All new sentence pairs resulting from the user's choices are logged (stored in a file), which provides valuable feedback to the development team of precisely what kind of text is being translated in practice, and to focus on improving the system's performance where it matters most.

A number of issues relating to orthography emerged during work on the Haitian Creole DIPLOMAT system (see Jeffrey Allen's paper[1] elsewhere in this volume for details), since Haitian Creole is a vernacular language which is still undergoing standardization – a single Creole word typically has two or more distinct spellings considered correct by various native speakers. Of interest here is that the manual selection among alternatives can be used to normalize the system's orthography to the user if the MT component outputs the variant spellings rather than forcing a particular orthography selected *a priori* by the developers. As the user makes selections, the system adapts to his particular orthography. Logging these selections provides feedback on actual usage patterns among the user community. Similary, if automated techniques are used in an effort to associate variant spellings with each other (and conflate them to improve the system's coverage), distinct words with similar spellings may erroneously be grouped[1], and user interaction can correct such errors during use just as orthography can be adapted.

00h 00m	Insert LDC Hansard Corpus CD-ROM into drive.
01h 01m	All files from the IBM Hansard corpus have been copied, decompressed, and prepared for use, and several missing lines have been corrected.
08h 56m	Initial statistical bilingual dictionary created.
19h 33m	Extracted 400,000-pair maximally-diverse subsample of IBM corpus for use as EBMT corpus.
21h 42m	French trigram language model complete.
23h 31m	English-to-French EBMT complete.
23h 39m	First English-to-French translation done.
25h 07m	French-to-English EBMT complete.
25h 25m	First French-to-English translation done.
…	Two hours of manual translation to create a file of word equivalence classes for EBMT while dictionary-refinement process executes.
31h 33m	Refined statistical dictionary complete.
33h 32m	Improved EBMT complete in both directions.

Table 1: Generating an English↔French MT System

5 Evaluating User Interaction

While having the user select from among the candidate translations shows great promise and has multiple uses, it has not yet been formally evaluated. Clearly, there are a number of areas to evaluate:

- What percentage of the translations need to be edited? What percentage of input words? How does this vary as more data is added to the translation engines?
- How much does augmenting the knowledge base improve future translations? How much does it reduce the need for user interaction in selecting alternate translations?
- Does memorizing modified translations cause any interference with the translations of other input when EBMT performs a partial match on the newly-added sentence pair?
- Will methodological enhancements (e.g. an improved language model or modeler) reduce the need

French Input: Comme on m'a donné avis de ce rappel au Règlement hier, j'ai pu examiner à fond tous les précédents cités dans ce commentaire et, pour ceux qui n'auraient pas eu le temps de le faire, je crois utile de les résumer.
System's Initial Translation: As we gave me notice to the point of order yesterday, I know seem to melts all the precedents quoted this citation is and, for those who have to had the time in which to do, I think useful to summarize.
Corrected using Alternate Selections: As one gave me notice of this on points of order yesterday, I have been able to examine to bottom all the precedents quoted in this comment and, for those who have to not had time to do, I believe useful to summarize.
Manual Translation of French: Since I was given notice of this point of order yesterday, I have been able to completely examine all of the precedents that were cited in the commentary and, for those who haven't had the time to do so, I believe it would be useful to summarise their contents.
Hansard English Version: The committee will appreciate that since notice of this point of order was given yesterday, I have, in anticipation, reviewed in detail all the precedents menti oned in that citation, and for the benefit of those who may not yet have time to do so, I believe it would be useful to take the time to summarize them.

Figure 6: Sample translation from rapid-development French-English system.

for user interaction?

Determining the answers to the above questions will require the creation of a "gold standard" corpus in which the appropriate selections from among the translation engines' outputs have already been made by a human. The primary evaluation tool – which must still be implemented – is a "wrapper" for the embedded MT system which

1. provides each sentence of a test corpus to Pangloss-Lite and retrieves the resulting translation.
2. compares the translations for each sentence to the human selections logged during an initial manual run using the DIPLOMAT user interface, and computes statistics on the number of mismatches.
3. automatically provides those pre-recorded selections as updates on all, none, or some subset (fixed or random) of sentences that do not match the expected output.

The above evaluation wrapper will most likely underestimate the number of "correct" translations, since (just as for any evaluation of translation correctness) there will often be multiple acceptable translations. In this case, however, the problem is more constrained since only a limited set of possibilities is available, and adding alternate selections to the translation system as corrections will not expand the range of possible translations[1].

Clearly, if the input is identical to the input for a previously-corrected translation, that corrected translation will be retrieved in its entirety, and thus the MT system will produce the most correct translation possible any time that input is encountered after the system is augmented with that sentence by the user. What is less certain is how well augmentation will help sentences which are similar but not identical; as the portions of the new input which match a previous augmentation become smaller, pre-existing candidate translations become more likely to be picked in preference to the user-selected translation.

References

[1] Jeffrey Allen. Lexical variation in Haitian Creole and orthographic issues for MT and OCR applications. In *Proceedings of the 1998 AMTA Workshop on Embedded MT Systems*, 1998.

[2] Ralf Brown and Robert Frederking. Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 221–239, 1995. Available at `http://www.cs.cmu.edu/~ralf/papers.html`.

[3] Ralf D. Brown. Example-Based Machine Translation in the PANGLOSS System. In *Proceedings of the Sixteenth International Conference on Computation Linguistics*, pages 169–174, 1996. Available at `http://www.cs.cmu.edu/~ralf/papers.html`.

[4] Robert Frederking et al. Integrating Translations from Multiple Sources with The Pangloss Mark III Machine Translation System. In *Proceedings of the First Conference for Machine Translation in the Americas (AMTA)*, Columbia, Maryland, October 1994.

[5] Robert Frederking and Sergei Nirenburg. Three Heads are Better than One. In *Proceedings of the fourth Conference on Applied Natural Language Processing (ANLP-94)*, Stuttgart, Germany, 1994.

[6] Robert Frederking, Alexander Rudnicky, and Christopher Hogan. Interactive Speech Translation

[1]There is one partial exception to this: because EBMT can match just a portion of a sentence, it may be able to generate additional candidate translations for a particular input after augmentation, but it will not generate any fundamentally new translations.

in the DIPLOMAT Project. In Steven Krauwer et al., editors, *Spoken Language Translation: Proceedings of a Workshop*, pages 61–66, Madrid, Spain, July 1997. Association of Computational Linguistics and Eurpoean Network in Language and Speech.

[7] Robert E. Frederking and Ralf D. Brown. The Pangloss-Lite Machine Translation System. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 268–272, Montreal, Quebec, Canada, October 1996. Available at `http://www.cs.cmu.edu/~ralf/papers.html`.

[8] Christopher Hogan. Correcting OCRed Text for Machine Translation. In *Proceedings of the 1998 AMTA Workshop on Embedded MT Systems*, 1998.