2004

Challenges in Using an Example-Based MT System for a Transnational Digital Government Project

Violetta Cavalli-Sforza
Carnegie Mellon University

Ralf D. Brown
Carnegie Mellon University

Jaime G. Carbonell
Carnegie Mellon University, jgc@cs.cmu.edu

Peter J. Jansen
Carnegie Mellon University

Jae Dong Kim
Carnegie Mellon University

Challenges in Using an Example-Based MT System for a Transnational Digital Government Project

Violetta Cavalli-Sforza, Ralf D. Brown, Jaime G. Carbonell, Peter J. Jansen, Jae Dong Kim

Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.
{violetta,ralf,jgc,pjj,jdkim}@cs.cmu.edu

Abstract. We describe ongoing efforts towards and challenges in using an Example-Based Machine Translation (EBMT) system in the context of a multi-national, multi-university and multi-agency transnational digital government project. The project is aimed at applying information technology to the problem of collecting and sharing information securely in a multilingual context. We report on a number of issues encountered in obtaining and using language data for the EBMT system, discuss our current solutions, and briefly describe ongoing enhancements to the system to meet some of the technical and practical challenges posed by using this machine translation approach in the project domain.

1. Background

We describe ongoing efforts towards and challenges in adapting and using an Example-Based Machine Translation (EBMT) system in the context of a transnational digital government project (Cavalli-Sforza, et al., 2003; Su et al., under review). The project represents an unusual collaboration between universities, government agencies, and an international organization aimed at applying information technology (IT) to a problem of international concern: detecting and monitoring activities related to the transnational movement of illicit drugs. The process is coordinated by the Organization of American States (OAS). The work is performed by a team of researchers from seven universities (U. of Belize, Pontificia Universidad Católica Madre y Maestra in the Dominican Republic, Carnegie Mellon U., North Carolina State U., U. of Colorado, U. of Florida, U. of Massachusetts) and experts from agencies in the three participating countries: the OAS's Inter-American Observatory on Drugs and Office of Science and Technology in the U.S., the National Drug Abuse Control Council (NDACC) of Belize's Ministry of Health, and the National Drug Council of the Dominican Republic. The motivation for this project and the choice of partner countries and institutions stems, in part, from an NSF-funded workshop for exploratory research on transnational digital government (May 9-11, 2001, Belize City, Belize). The specific choice of governmental domain at which to target our research activity – the collection, notification, and sharing of information regarding movement of people across borders – was the result of a collective decision in the early phases of the project. It speaks directly to one of the indicators ("Displacement") used by the Multilateral Evaluation Mechanism (MEM), a multinational effort that involves the collection and analysis of data by, and from, several government agencies and non-government organizations within each country. The MEM is managed by OAS's Inter-American Drug Abuse Control Commission (CICAD) with participation by 34 OAS member states.[1]

[1] This project, like the controversial U.S. Computer Assisted Passenger Prescreening System (CAPPS) and its international version APIS, raises issues concerning the security/privacy of information and touches the delicate areas of immigration and travel restriction policies. We emphasize that our project only uses IT to enable collection, notification and transnational sharing of information that is already used nationally within Belize and the DR in accordance to these countries' border control procedures. Our main concern is therefore with respect to security/privacy issues and with providing access to information in a way that a) complies with the different procedures and regulations of the two countries and b) will be extendable to other countries that may participate in the future. To address this we are: (1) using filters to keep information secure and private and (2) techniques for analysis of software requirements that align privacy and security. A possible further challenge for the future will be adherence to international agreements on related topics.

2. Project Challenges

Information systems that support international collaborations among governments face several research challenges in collecting and managing information across agencies and organizations without compromising the security and policies of the countries, interoperating transparently across countries with heterogeneous information networks, and sharing multilingual information. The methods and technology used by Belize and the Dominican Republic to collect, store, and share information currently differ. One aim of this project is to allow immigration agents to access information about travelers more uniformly and efficiently, so as to expedite processing of routine border-crossing situations and facilitate handling of potentially problematic cases.

In our prototype system, data is entered by each country in its own language, but can be queried by authorized individuals in a different language. Much of the information routinely collected through arrival/departure forms is stored in a database using set phrases whose translation requires only table lookup. Translation is needed for text stored in the comment field of an individual's record, where immigration agents can place additional information that results from observation and questioning of the traveler. The text may be a more detailed description of the traveler, the circumstances surrounding the (attempted) border crossing, or a transcribed dialogue. The text may be translated as a whole or searched for the presence of key phrases.

The principal scenario of use for which we are developing the system is one in which an immigration agent submits a query in order to access available information on an individual who is requesting entry into the country. The query is submitted from a computer terminal or, in the case of remote posts or crossings along a patrolled border, via radio or cell phone. The traveler may be on a watch list, display suspicious behavior at the time of crossing, have insufficient or questionable documentation, or provide inconsistent information. Information extracted from the natural language query is converted to a database query; results are returned and translated into the requested language, if necessary. More extensive details regarding system architecture and scenarios of use are given in Cavalli-Sforza et al. (2003) and Su et al. (under review). Future extensions of this project may require speech-to-speech translation of dialogues between immigration officials and travelers.

At present we are focusing on bidirectional Spanish-English translation of texts. English is the official language in Belize, with Spanish being an important second language due to borders with Guatemala and Mexico. The Dominican Republic is primarily and officially Spanish speaking.[2,3]

While our translation needs are presently limited to English and Spanish, one goal of the project is to provide a model that is extensible for use by different agencies in other countries and other domains. Lacking sufficient resources to develop a knowledge- or transfer-based machine translation (MT) system, and requiring an approach that can be quickly adapted to other domains and languages, we chose to use CMU's Pangloss-Lite (Panlite) Multi-Engine Machine Translation (MEMT) system (Frederking & Brown, 1996) and rely primarily on the Example-Based MT engine, which has been undergoing continuous enhancements (e.g., Brown, 1999; Brown, 2000; Brown et al., 2003). Panlite was the underlying translation system in the rapid-deployment speech-to-speech DIPLOMAT project (www.lti.cs.cmu.edu/Research/Diplomat/; Frederking et al., 1997); it therefore seemed to be an ideal choice. Nonetheless, the use of Panlite in the context of this project has been challenging for a variety of reasons.

In the first place, the lack of domain-specific parallel data has made it necessary to exercise a great deal of creativity in building the linguistic data resources. Secondly, after the data has been obtained, there are still challenges to be met in using and managing the data to satisfy the requirements of the application, while also providing reasonably predictable translation behavior. Finally, there are challenges in using the Panlite system as is for this project. Panlite's development in the last few years has been guided by translation tasks with requirements that were very different from this project's requirements. The use of the system in the context of this transnational digital government research has spurred the undertaking a number of enhancements that should increase the system's usability and its performance, especially in domains where data is limited and varied. In the remainder of this paper, we describe in some detail the current Panlite system and discuss ongoing efforts to meet the challenges posed by our application and to alleviate the problems we are encountering.

---

[2] Other written languages that may be added in the future include Haitian Creole, French and Portuguese.

[3] Spoken languages the region include: in Belize, Belizean Creole English, Spanish, Garífuna, three Mayan languages, and Plautdietsch (Memnonite German); in the Dominican Republic, Spanish, Haitian Creole French and Samaná English (a Creole language).

3. The Panlite MEMT System

The Panlite MEMT system (Frederking & Brown, 1996) provides a framework for using multiple translation engines in parallel. Given an input in the source language, each engine provides a translation into the target language for the full input or fragments of the input, along with a score for each translation. Translation candidates are placed in a chart as 'edges' covering the input or some portion of it. One component of Panlite, the Language Modeler, uses statistical knowledge of the target language, among other information, to select or piece together from the chart the best scoring translation(s) that cover the entire input.

The Panlite system supports the integration of widely different MT engines (for example, transfer-based, knowledge-based and statistical engines) but provides three built-in engines in addition to the language modeler: an EBMT engine, a Dictionary engine, and a Glossary engine. Each engine is described briefly below. A version of each engine's language-pair specific resources must be built for each direction of translation (e.g. English→Spanish and Spanish→English) but, for the most part, can be built automatically from the same unidirectional source file. Hand-construction and/or refinement of language resources may be used to improve performance in each direction of translation.

3.1 The EBMT Engine

At its simplest, the EBMT engine translates input phrases or sentences by matching new input against source text in previously seen source-target pairs. If it cannot find a match for the entire input, it looks for matches for all possible multi-word fragments of the input and posts to the chart what it believes to be the corresponding translations.

The essential 'training' data for the EBMT engine is a sentence-aligned corpus. The system does not 'learn' in the traditional machine learning sense; its training consists of preprocessing the parallel data and building an index so as to make retrieval of any part of the source text and corresponding part of the target text as fast as possible when the system is translating new input. The preprocessing also includes determining the correspondence between fragments of parallel source and target sentences. While this computation could be performed at translation time, it is more efficient to perform it at indexing time and to store a correspondence table in the index.

At runtime (translation time), the input sentence to be translated and its fragments are matched against the source-side of the indexed training corpus, with some flexibility in determining what is considered a to be a good – if not exact – match and some control over the extent of the search for candidate matches. Candidate translations are produced from the target side of the indexed corpus; they are scored and posted with their score to the chart, which stores the translations provided by all engines.

The EBMT engine includes two mechanisms for generalization of language data that allow a parallel corpus to go further in matching new input.[4] The first and older mechanism, "tokens," establishes classes of words or phrases in the source language, and corresponding translation in the target language, that are syntactically interchangeable (some restrictions apply). Examples of token class definitions are shown in Figure 1. The example on the right shows that classes can be defined based on other previously defined classes, and that multiple source and target expressions can be considered equivalent. In building a system for a given direction of translation, all expressions will be recognized, but only the first one on the target side will be generated. The class definitions are shown English→Spanish but are automatically reversible.

@@begin <weekday> Monday lunes ... Sunday domingo @@end <weekday>	@@begin <dates> <date>1 - <date>2 <date>1 to <date>2 <date>1 - <date>2 <date>1 a <date>2 ... @@end <dates>

Figure 1. Examples of token class definitions

The second generalization mechanism, "tagged entries," allows adding linguistic information directly into the corpus, creating corpus entries that define parallel source and target grammar fragments. Tagged entries also define classes whose members have syntactically equivalent behavior, as in the following simple example:

```
;;;(TOKEN <NP>)
<N-S> <adj-s>2 <adj-s>1
<adj-s>1 <adj-s>2 <N-S>
```

The first line specifies that a member of class **<NP>** is being defined. The first line (for Spanish) says that a noun phrase (**<NP>**) is a singular noun (**<N-S>**) followed by two singular adjectives. In English (the second line), the adjectives would

[4] These two mechanisms are actually being merged in the implementation but will continue to be specified separately in the language resources for a specific language pair using the formats shown in the examples.

appear before the noun and in the inverse order in which they appear in Spanish. For example, the phrase "curly blonde hair" would be expected to appear in Spanish as "pelo rubio rizado." Numeric suffixes are one way of specifying source-target alignment, but other ways are also available.

Tagged entries can refer to previously defined tagged entries and also to token classes, and vice versa. Tokens and tagged entries are used both at indexing time and at runtime. When indexing, the class name is recursively substituted for the literal text. Generalized examples are always stored in the indexed corpus, whereas the original examples are optionally stored. At runtime, a generalized training example can be used to match input that differs from it only by a word or expression in the same equivalence class. Tagged-entries have been shown to substantially increase the work performed by a parallel corpus of a limited size (Brown, 1999).

3.2 The Dictionary Engine

The Dictionary engine provides translations for single words on the source side; on the target side, the corresponding translation may be a word or a phrase. The engine is based on a dictionary, which may be constructed manually or automatically from machine-readable dictionaries or from parallel texts and may contain translation frequency information. If the dictionary is automatically constructed or inverted, hand-refinement is usually needed.

The EBMT engine uses the dictionary during corpus indexing to find sub-sentential alignments between source and target language pairs. Under certain conditions, the dictionary is also used at translation time to extend source-side matches: if a portion of the input matches a source-text fragment in the training corpus except for one word, and the word has one translation that is significantly more frequent than other translations, that translation will be substituted in the target fragment.

While not strictly necessary, a Dictionary engine is highly desirable, since it enables backing off to single-word translations in cases where other engines are not able to translate the entire input and leave holes in the coverage of the input.

3.3 The Glossary Engine

The Glossary engine is primarily intended to provide translations for source language phrases. The target language translation may be a single word or a phrase. If the source language is a word, there is an option to copy the word and its translation at runtime to the dictionary as well. Since glossaries are built manually, the translation is generally assumed to be a good one; placing the same translation in two engines will result in it receiving greater weight in the determination of the final translation for the input.

Both the Glossary and the EBMT engines work with translations of source language phrases. An important difference between them is that, while the source-target alignment is known to be correct for glossary phrase translations, it is only hypothesized in the EBMT engine. However, as the Glossary engine's modest generalization capability has fallen into disuse and the EBMT engine's has been enhanced, the overall trend has been to reduce the Glossary engine's importance, with an eye to eventually merging it with the EBMT engine.

3.4 Controlling Panlite's Behavior

In general, the behavior of each component of Panlite, and especially the EBMT engine and the Language Modeler, is governed by a complex set of parameters whose values are specified in a configuration file, although some can be overridden for an individual invocation of the system or changed at runtime.

In EBMT, parameters control input and output processing, automatic dictionary creation and refinement, use of generalization, corpus indexing, sub-sentential alignment and scoring of source-target language pairs, matching of input to known source text, and memory management at runtime. The Language Modeler component of the Panlite MEMT system performs the final selection of translation(s) for the input. The selection is based on a language model of the target language, the scores of individual fragments posted to the chart, the weights associated with each engine, whether fragments are allowed to overlap and by how much, and several other user-specifiable controls that interact in a complex manner.

Understanding and learning to control and adjust Panlite's overall behavior is not an easy endeavor. Making the system more accessible to developers of translation systems for specific language pairs was one of the challenges we encountered as we began to use the system for the transnational digital government project. See Section 6 below.

4. Building Suitable Language Resources

The EBMT system contained in Panlite is fundamentally a data-driven approach to translation. While it generally requires less parallel data than a statistical MT system does for comparable performance, the quality of the translation output is still heavily dependent on having sufficient data in

the domain and style of interest. Unfortunately, though there is an abundance of English-Spanish parallel corpora, the available resources are largely formal documents and some newswire. They diverge widely in content and style from text in our domain and we have found that they provide little translation help. A baseline version of the system, trained on a hand-refined version of the English-Spanish United Nations (U.N.) parallel corpus (Graff & Finch, 1994) with some additional parallel text obtained from the Pan American Health Organization (PAHO) and a statistically-derived dictionary based on that corpus, produces translations that are virtually useless in the context of our project, and is particularly ill-suited to the translation of dialogues.

If suitable monolingual data were available, it would be possible to create a parallel corpus through manual or semi-automatic translation, but very little of that data is to be found either. In the Dominican Republic, computerized access to a database of traveler information is available at major ports of entry (e.g., the international airport in Santo Domingo) though not always at remote border posts. The current process however, is not equipped to store information that goes beyond responses to questions on standard arrival/departure forms. In Belize, the immigration and emigration process currently does not make use of computer technology for storing traveler information, although some information is handwritten in a diary kept at each border station.

Overall, we have experienced significant difficulty in obtaining authentic documents to use in our translation system, whether monolingual or bilingual. Our Belizean and Dominican partners have provided some examples of dialogues and traveler descriptions after talking to immigration officials, but the data they have supplied so far has been nowhere near the needed amount. It is only very recently that we have seen authentic materials. These include an extract of comments found in a station diary and a few advisories that either instruct border immigration officials to be on the lookout for specific individuals (individuals on watchlists) or inform officials of new policies regarding allowing entry of travelers into the country. Examples of this kind of data, as provided to us (including errors, only the names have been changed), are shown in Figure 2 below.

In response to the lack of domain-relevant language resources, we have employed a variety of techniques to build our corpus and improve translation quality. A prototype version of the information system – including a database, network communication, capabilities for entering queries via a natural language interface (currently text-based but soon to become speech) or a forms interface, and translation service – has recently started running at universities in the U.S. and Belize and will soon run in the Dominican Republic too. The presence of the prototype system at universities will allow our colleagues in the client countries to collect data from actual future users prior to fielding the system at actual ports of entry. It will also give us an opportunity to collect more authentic data.

Ellen Garcia of 24 yrs, and Maria Vargas of 21 yrs, both Belizeans. Both in a state of drunkenness, made scandal inside Border facility. Detained and released 8 hours later.
Note: Juan Smith an American (old person) 30.6.29 is being deemed a prohibited Immigrant do not land under any circumstance.
In view of the SARS wpdemil affecting the far east countriesand Canada, the minister of home affairs has imposed a temporary Baan on the entry into Belize of persons in the following countries: Mainland China, Hong Kong, Singapore, Vietnam, India, Canada.

Figure 2. Examples of authentic text from Belize

In the meantime, so as to field a system that can translate in the domain to some extent, we have 'bootstrapped' it using the following techniques:

Translation. (Quasi)-native Spanish language speakers translate, from English into Spanish, sample dialogues and hypothetical descriptions of border crossings developed by project members. This technique allows us to obtain a broad range of Spanish translations for similar sentences, but also gives rise to some complexity in selecting the translations. This point is discussed further below.

Scenario Generation. Translators imagine circumstances surrounding border crossings and write, in both languages, descriptions of individuals and circumstances, and hypothetical dialogues in those situations. This technique helps us to extend the range of content of the texts.

System Use. Project members use the system to test its translation capabilities or to provide other examples of questions, answers, and situation descriptions. Their interactions with the system are logged and, when the translations are not correct, they are manually translated and used to augment the parallel corpus.

Interviews. During our last project meeting, in Belize, a Senior Immigration Official was interviewed and asked to recollect different experiences of problematic border crossings during

his career. He also answered several questions regarding the types of behaviors that might be considered suspicious and cause immigration officials to hold travelers for further questioning. The information collected during this interview and brief discussions with officials at border points was used as a basis for composing texts which, with their manually-produced translations, help augment the corpus. We have asked our colleagues in Belize and the Dominican Republic to apply this technique to gather additional data of this kind in their interaction with immigration officials. In addition to providing more text for translators, this technique aids in the generation of more authentic scenarios of border crossings.

News Briefs. A recent type of data that we have started acquiring from the Dominican Republic is a collection of news briefs concerning immigration incidents, ranging from 1-2 sentence notifications to 1-2 paragraph articles. An example of such a text is shown in Figure 3 below.

Autoridades de Dominicanas han confirmado que una red de contrabando humano ha tomado como base de operaciones la isla, y que desde ella, facilitan el tráfico humano hacia otras islas, como San Martín. Asimismo, confirmaron que la mayoría de las personas que son "traficadas" son ciudadanos dominicanos o haitianos, que viajan de manera legal mediante vuelos charter, y que reciben permiso para quedarse por una semana, pero que, aparentemente, "salen por la puerta trasera".

Figure 3. Example of domain-relevant news brief text from the Dominican Republic

In spite of all these initiatives, our parallel corpus still falls short of the minimally desired amount. The DIPLOMAT project found that approximately 50K words for each language was required in order to obtain decent translation performance and successfully create standard backoff trigram language models. At approximately 10 words per sentence, averaged over dialogue and descriptions, this implies having approximately 5,000 sentence pairs. We are still far short of that: up till now, our corpus includes only approximately 2,200 domain-specific pairs, of which a (decreasing) portion consists of alternative translations for the same English source sentence. The most recent material we have collected will add a few hundred sentence pairs, but the corpus will still fall well below the minimum target size. Thus increasing the size and coverage of the corpus, by any means available including the above techniques and sources of information, remains a high priority for improving the translation performance of the system.

The translation pairs in Figure 4 exemplify the quality of translation the system produces when it has seen similar but not identical source sentences. The source texts in Figure 4 are modifications of examples in the indexed (training) corpus. Specific words or phrases are underlined to draw attention to problems in the source and/or target. Usually the output is understandable even if grammar problems are present. Notice that here, as in earlier examples, sometimes it is the input that is not fully grammatical. Performance on sentences or phrases that deviate extensively from training examples is substantially lower.

El viajero se presentó en la frontera sin documentos. ***Passenger, presented himself at the border without documents.***
El viajero no entendía español. ***The traveler not understood Spanish.***
Spoke with central office. ***Habló con oficina central.***
Andrew Jones traveling from Flores, Peten, crossed at 11:25 a.m. ***Jones Andrew como Flores, Petén, cruzado A las 11 : 25 a.m.***

Figure 4. Translation Examples

5. Using Available Language Resources

In addition to the difficulty of developing suitable parallel text resources for use with the Panlite system, a number of issues have arisen when using the resources that were already available or that we have been able to collect.

5.1 Using Out-of-Domain Corpora

Because our domain-specific corpus is still small in size and coverage, in an effort to increase the robustness of the system, we have been layering it on top of out-of-domain corpora. These include a small corpus of general travel glossaries containing around 2800 source-target pairs and a wide variety of phrases and idioms, most of which are not particularly relevant to the domain. This corpus is itself layered on top of the U.N./PAHO parallel corpus, which also contains some general glossary material. The layering relies on a feature of the EBMT system that permits assigning a greater weight to examples in more recently added text (i.e. text added closer to the end of the indexed corpus) in the process of scoring and choosing translations. Alternatively we could also weight each corpus individually – and we may in the future – but it will

require re-indexing a very large corpus composed of several widely dispersed sub-corpora.

The layering of multiple corpora broadens the language coverage and increases somewhat the robustness of the system, particularly in view of the fact that the dictionary currently in use is automatically extracted from those same out-of-domain corpora and therefore does not provide good single-word translations (see next section). Still, most of the time the translations obtained from the out-of domain text are not acceptable. An interesting effect of using a large out-of-domain corpus as fallback, and allowing flexibility in matching and in the composition of fragments, is the introduction of irrelevant material into the translation: the system effectively "hallucinates" part of the translation. The following example shows the phenomenon; we underline incorrect translations and **bold** improperly inserted material.

Ellen Garcia of 24 yrs, and Maria Vargas of 21 yrs, both Belizeans.

Key Garcia **de Bahía** 24 **por** yrs, **consenso,** e Vargas Maria, **de Bahía** 21 **por** yrs, **tanto individual como** Belizeans.

The current system uses a dictionary that was automatically extracted from a corpus composed of the large U.N./PAHO corpus and a small general glossary. The language model is based on a large selection of similarly out-of-domain documents. In an attempt to improve the quality of sub-sentential alignment and translation of individual words, we could recompute the statistical dictionary including the domain-relevant data. We could similarly recompute the language model hoping to select a better combination of fragments in the final translation. However, the domain-relevant data we have available is so little in comparison to the data on which the dictionary and the language model are based that it is unlikely to make much difference in translation performance with the current algorithms.

Instead, we are currently in the process of hand-refining the dictionary, on one hand, and considering/developing better automatic dictionary extraction and alignment algorithms on the other. Past experience with the Panlite system shows that automatically extracted dictionaries work better for indexing than manually developed/refined ones; the latter work better at translation time. As a result, Panlite allows the use of different dictionaries at indexing and translation time. So, although hand-refinement of the dictionary will not produce better alignment, it will ensure that, at least for those fragments of the sentence where no (better-) matching phrase can be found, the system can supply a reasonable word-by-word translation. The resulting translation may be syntactically very poor and choppy, but should be semantically acceptable.

A small number of manual additions and refinements to the translation dictionary have yielded a qualitatively noticeable improvement in the translation of probable phrases and sentences for this domain. This approach, however, is not without its disadvantages. On one hand, separate and different manual updates must be prepared for the two directions of translation. Since the current dictionary is a full-form dictionary (no morphology information is kept in it) and Spanish is a highly inflected language, manually enriching the dictionary is both omission-prone and a significant amount of work. On the other hand, with the exception of a handful of words that we may imagine would be useful to place in the dictionary, the main impetus for augmenting the dictionary still comes from domain-relevant texts, bilingual or monolingual and in either language. Hence the labor we put into hand-refining the dictionary will give extra coverage for individual words seen in examples but will provide very little assistance with unseen input.

5.2 Using Newly Acquired In-Domain Texts

While the authentic and constructed in-domain texts that we have been using are the most important source of examples and vocabulary for the MEMT system, their use is not without problems. This section describes some of the linguistic and technical challenges that we have encountered in using these language resources.

Linguistic Variety. Our informants/translators use different Latin American Spanish dialects, which differ in common everyday words and idioms (e.g., a vague word like 'bag' might be found as *bulto*, *bolso*, *bolsa*, *cartera*, *maleta*, *maletín*, *valija*, *veliz*, *saco*; correspondingly, some of those words may be variously translated into English as 'bag', 'briefcase', 'handbag', 'sack', 'suitcase', and 'wallet'). American English and Belizean English also differ, and not only in spelling (Belizean English is influenced by British English). More importantly, in the authentic data we have seen, immigration agents tend to use an abbreviated form of English, frequently dropping pronouns and auxiliary verbs and using some acronyms and abbreviations (e.g., Figure 2). While the English is perfectly understandable, it cannot be translated into a similarly abbreviated Spanish. In general, the need to accommodate different dialects and linguistic variation is an

unavoidable aspect of our corpus and our project, one that we must be prepared to deal with if we are to field the system in a broader range of countries in the Americas.

Multiple translations for the same source. Any text can generally be translated in more than one way. In our case, multiple translations are largely an artifact of collecting parallel data by seeking Spanish translations for the same English source sentence from multiple translators. Figure 5 gives an example, a commonly occurring sentence in a dialogue between an immigration or customs officer and a traveler. More rarely, we also obtain multiple English translations for the same Spanish source. On one hand, multiple translations provide more examples that can be matched, especially for translating from Spanish into English. On the other hand, using multiple translations has the potential of making system output less predictable. We can order translations from worst to best and use the EBMT system's corpus weighting mechanism to favor later and therefore supposedly better translations. However, while the EBMT system allows us to express a preference for examples that are found later in the corpus, it does not guarantee that it will choose a particular example, since the final translation depends on translations posted by multiple engines and a complex scoring system.

Are these your bags?
¿Son suyas estas bolsas?
¿Estas bolsas son suyas?
¿Éstas son sus bolsas?
¿Es éste su equipaje?
¿Son éstas sus maletas?
¿Son éstos sus bultos?
¿Son éstas sus bolsas?

Figure 5. Multiple Translation Example

Unpredictable and open-ended data. The domain of border crossings involves many people and place names that are not restricted to Spanish and English, since both countries are strong tourism magnets. Recognizing people and place names so that they can be appropriately translated or not translated, as the case may be, is a general problem in machine translation and is particularly salient in this domain.

The initial approach we have adopted to supporting multiple translations and linguistic variety has been to accept different inputs on the source side but to generate a single output on the target side. In doing so, we are taking advantage of the need to develop different language resources for each direction of translation. For example, if we have multiple Spanish translations for an English sentence, the Spanish→English translator will use all the Spanish versions in the indexed corpus: they provide more match opportunities for the input but all map to the same translation. For English→Spanish, we order translations from worst to best based on how accurately the target reflects the meaning and structure of the source (translators frequently give approximate translations), and whether the structure of the source and target texts facilitates sub-sentential alignment, while taking into consideration word order preference of the target language. Only the best version is included in the training corpus. This asymmetric use of corpus materials for the two directions of translation has been one of the factors motivating the development of a corpus management system for this project (Cavalli-Sforza, Carbonell & Jansen, 2004). A similarly asymmetric treatment is used in hand-refining the dictionary resource.

The issue of dealing with unpredictable person and place names has not yet been seriously addressed. In the long run, we plan to use a named entity identifier for each language (e.g. BBN's IdentiFinder), which only needs training on monolingual data. The current system usually reproduces in the target, without translation, words that it has no knowledge of. In addition, person and place names often occur in specific contexts (e.g., for names, preceded by honorifics; for place names, preceded by specific verbs or prepositions), so names can also be captured via the EBMT system's generalization mechanisms. This is not a general solution, however, since there will always be new names and new places that have not yet been assigned to a generalization class.

6. Enhancing EBMT and Panlite

An important outcome of using the Panlite System in the context of our Transnational Digital Government project has been the stimulus to improve both the usability and the functionality of the EBMT and the Panlite systems. In this section we briefly describe the range of changes that are currently underway and planned.

Panlite was initially developed in the Pangloss project (www.lti.cs.cmu.edu/Research/Pangloss/) and used in the DIPLOMAT project to provide rapid-deployment of speech-to-speech translation systems for dialogues. In recent years, the system has been primarily enhanced in response to the

TIDES (www.darpa.mil/ipto/programs/tides/) program competitions with the goal of improving translation on newswire text, with training performed on parallel texts consisting largely of formal government documents and some newswire material. The Panlite system as is does not support particularly well the requirements of our application, which includes translation of both dialogues and brief third person descriptions tied to a database record for a specific individual, in a domain where training text is not abundant and is subject to a great deal of linguistic variation and input irregularity.

The experience of applying the Panlite system to translation in the border-crossing domain has made apparent the urgency of improving both the pre-processing and, to a lesser extent, post-processing capabilities of the system. Post-processing – in the sense of cleaning up ugliness in capitalization, punctuation and spacing – is lower priority, since it affects the aesthetics more than the clarity of the output. Pre-processing is a more necessary and, in our case, a more complex task. By pre-processing we refer to the following types of processing:

Regularization of the Input. This processing includes expansion of contractions (e.g., "don't" → do not) and normalization of alternative spellings (e.g., 'color' vs. 'colour').

Treatment of Abbreviations. We have been warned that immigration agents at border locations in Belize are likely to use abbreviations to expedite the process of information storage and retrieval. There is not much evidence of this in the little authentic data we have seen but, if true, we will need to treat abbreviations either as input to be regularized and/or as dictionary entries.

Spelling Correction. As seen in Figure 2, real future users of the system cannot be realistically expected to be careful about spelling. At least simple spell-checking and correction of the input will be needed before attempting to match it against the corpus. Spell-checking and correction depends on dictionary contents, regularization and treatment of inflected word forms.

Morphological analysis and generation. The early versions of the Panlite system used in the DIPLOMAT project included a fairly simple table-driven capability for morphological analysis, written in Lisp (the current Panlite system uses C/C++ for speed). Since that time, the system has evolved without that functionality, using optional external files of stems or roots of inflected words but largely relying on vast amounts of parallel data to provide the required contexts for correct inflection. In the borders domain, we are translating both dialogues and third person descriptions and therefore need to use the full range of person-number-gender combinations in Spanish. The ability to use morphological knowledge to both match examples and generate translations is important in a domain where data is limited and with highly inflected languages like Spanish and Arabic. (There has been some use of the EBMT system with Arabic and more is planned for the near future.) Recent in-house experiments have shown that simple stemming can significantly improve performance for translating from Arabic. We are therefore currently investigating how to re-integrate morphological analysis and generation into the Panlite system to improve the quality of all types of translation and also as a prelude to exercising better control of inflectional features (person, gender and number) in the translation of dialogues and descriptions that, in the borders domain, are tied to database records for individuals of known sex.

The Panlite system has been used for translation with a variety of language pairs, including English-Spanish, and continuously modified to keep pace with new technological developments and resources. Thus we have been able to take advantage to some extent of both existing language resources and code enhancements, including the tagged entries described in Brown (1999) and the use of overlapping fragments in composing the final translation (Brown et al., 2003). However, as the system has grown and changed, many control parameters have been added to run specific experiments. The interface has grown somewhat "organically" and has become virtually incomprehensible to all but its primary developer. One relatively small but important improvement is therefore the redesign of the EBMT engine's interface first, and soon thereafter Panlite's. The new design, which is now partially implemented, is intended to support a GUI control panel style interface. For backwards compatibility and in deference to some users' tastes, it will also continue supporting the system's current interface. The new interface design partitions control inputs by engine, by component within the engine, and by its use at training and/or translation time; it also explicitly states dependencies and co-dependencies among control inputs. The design will aid the user to achieve a better understanding and control of Panlite's power and flexibility; it will also aid the developer in maintaining interface documentation as enhancements to the system's functionality give rise to changes in the control interface.

Further planned enhancements to the system include: a new corpus indexing scheme to improve lookup speed, improved sub-sentential alignment and dictionary extraction and refinement algorithms, a new approach to generalization in EBMT, and an improved decoder for Panlite.

7. Summary

In this paper we have described our ongoing experience in applying a primarily example-based MT system to a transnational digital government project. The project's requirements differ significantly from the majority of previous uses of the system, and the parallel data required by a data-driven MT approach has been difficult to obtain. We have developed, and are still developing, a number of techniques for gathering language resources and for handling the heterogeneity of the data that both results from our collection efforts and is inherent in the project domain. In response to the challenges of using the MT system in this and other projects, we are also in the process of augmenting and improving many of the system's capabilities.

The transnational digital government project is approximately halfway through its initial funding cycle. Prototype versions of the system will soon be installed at university sites in all of the participating countries (currently they are installed at two sites). There, they will undergo testing with authentic users in a controlled environment before being placed in the field. The next few months should prove quite revealing from the perspective of development of language resources. We also expect to have many of the system improvements in place before the final version is fielded and to be able to perform larger scale and more formal evaluations in the near future. As the system is put to the test in real settings and with more authentic data, we also expect to acquire a greater understanding of the real weight of the issues discussed above, to encounter new challenges, and to devise solutions that are better informed by the needs and constraints of actual use.

Acknowledgments

Research reported in this paper is funded in part by National Science Foundation (NSF) award EIA-0131886. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

We gratefully acknowledge the collaboration of our many Belizean, Dominican and U.S. colleagues who are participating in this project.

References

Brown, R.D. (1999). Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99) (pp. 22–32). Chester, UK.

Brown, R.D., Hutchinson, R., Bennett, P.N., Carbonell, J.G., & Jansen, P. (2003). Reducing Boundary Friction Using Translation-Fragment Overlap. In Proceedings of MT Summit IX (pp. 24–31). New Orleans, LA.

Brown, R.D. (2000). Automated Generalization of Translation Examples. Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)(pp. 125–131). Saarbrücken, Germany.

Cavalli-Sforza, V., Antón, A.I., Brooks, O., Carbonell, J., Cole, R., Connolly, R., Fortes, J., Herrera, M., Krsul, I., McSweeney, C., Ortega, C., Su, S., Towsley, D., Ventura, J., & Ward, W. (2003). Enabling Transnational Collection, Notification, and Sharing of Information. In Proceedings of the 2003 National Conference on Digital Government Research (dg.o2003). Boston, Mass. (http://www.diggov.org/archive/library/dgo2003/#CD)

Cavalli-Sforza, V., Carbonell, J.G., and Jansen, P.J. (2004). Developing Language Resources for a Transnational Digital Government System. To be presented at the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal.

Frederking, R.E. & Brown, R.D. (1996). The Pangloss-Lite Machine Translation System. In Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96) (pp. 268–272), Montreal, Canada.

Frederking, R., Rudnicky, A., & Hogan, C. (1997). Interactive Speech Translation in the DIPLOMAT Project. Presented at the Spoken Language Translation Workshop at the 35th Meeting of the Association for Computational Linguistics (ACL-97). Madrid, Spain.

Graff, D. & Finch R. (1994). Multilingual Text Resources at the Linguistic Data Consortium. In Proceedings of the 1994 ARPA Human Language Technology Workshop. Morgan Kaufmann.

Su, S., Fortes, J., Kasad, T., Patil, M., Matsunaga, A., Tsugawa, M., Cavalli-Sforza, V., Carbonell, J., Jansen, P., Ward, W., Cole, R., Towsley, D., Chen, W., Antón, A.I., He, Q., McSweeney, C., deBrens, L., Ventura, J., Taveras, P., Connolly, R., Ortega, C., Piñeres, B., Brooks, O., & Herrera, M. (under review). A Prototype System for Transnational Information Sharing and Process Coordination. Submitted to the 2004 National Conference on Digital Government Research. Seattle, WA.