2004

Developing Language Resources for Transnational Digital Government Systems: A Case Study

Violetta Cavalli-Sforza
Carnegie Mellon University

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

Peter J. Jansen
Carnegie Mellon University

Published In
.

Developing Language Resources for a Transnational Digital Government System

Violetta Cavalli-Sforza, Jaime G. Carbonell, Peter J. Jansen

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

Abstract

We describe ongoing efforts towards developing language resources for a transnational digital government project 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 project seeks to support information sharing, coordination and collaboration among government agencies within a country and across national boundaries by combining a variety of technologies including a distributed query processor with form-based and conversational user interfaces, a language translation system, an event server for event filtering and notification, and an event-trigger-rule server. The prototype system is being developed by U.S. universities in collaboration with an international agency and with universities and government agencies in Belize and the Dominican Republic. This paper focuses on the linguistic resources and their use in Example-Based Machine Translation (EBMT). We are in the process of developing an English-Spanish parallel corpus, focused on the domain of information elicited and used at border crossings, to fuel the EBMT system. While significant parallel corpora are available for these two languages in the newswire domain, they were found to be of very limited use for the border crossings application, spurring the need to develop our own resources.

Introduction

We describe ongoing efforts towards developing language resources for a transnational digital government project (Cavalli-Sforza *et al*., 2003; Su et al., under review), 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 in two Caribbean countries and the U.S. (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, the National Drug Abuse Control Council (NDACC) of Belize's Ministry of Health, and the National Drug Council of the Dominican Republic.

Information systems that support international collaborations among governments face several research challenges in managing information across agencies and organizations without compromising the security and policies of the countries, interoperating transparently across heterogeneous information networks, and sharing multilingual information. Our task within this project is to provide access to the data managed by the system in different languages using machine translation. Within the framework of this project, it was jointly decided to target the system at information regarding movements of individuals across borders and, in particular, on travelers requesting entry into Belize and the Dominican Republic.[1]

The methods and technology currently used by countries such as Belize and the Dominican Republic to collect, store, and share this type of information differ. One aim of this project is therefore to allow immigration agents to access information about travelers more uniformly and efficiently, so as to expedite processing of routine border-crossings and facilitate handling of difficult cases. In our system, each country enters data in its own language, but authorized individuals may query the data in a different language. Much of the information routinely collected through arrival and departure forms can be stored in a database using set phrases and table lookup for translation. Machine translation is needed for text stored in the comment field of an individual's record, where immigration agents can place information that results from observation and questioning of the traveler. The text may be a description of the traveler, the circumstances surrounding the border crossing, or a transcribed dialogue.

Immigration officials agree that storing and viewing such data can be of great value in identifying and handling suspicious travelers. Unfortunately, due in part to the recent introduction and limited use of computers by government agencies in countries such as Belize and the Dominican Republic, and in part to a different distribution of responsibilities across government agencies in the two countries, authentic text data exemplifying the type of translation required by the project has been extremely difficult to locate and obtain. The current processing of travelers at point-of-entry stations, especially in the Dominican Republic, is not set up to store information beyond responses to preset questions on arrival/departure forms. Colleagues from Belize have been able to provide a small number of examples of text that might be recorded in the diary kept at a border station, but we have found it necessary to extend that small sample in a number of ways. Below, we present our approach to gathering, developing and managing language resources in support of English-Spanish bi-directional translation, after briefly describing the overall information system and the machine translation system we are using for the project.

System Architecture and Functionality

The architecture of the prototype system includes a Host site and the sites of participating countries and their agencies (e.g., one in Belize and the other in the Dominican Republic) to represent the countries' agencies.

[1] Our work only seeks to support existing national immigration policies in Belize and the Dominican Republic.

Both countries have local databases, managed by their own local heterogeneous Database Management Systems (DBMSs), that store immigration, border control and government process-related data. The agents at ports-of-entry in each country use the local DBMS to enter, access and manipulate their data. Data that a country is willing to share with the agencies of another country are specified in an export schema, and the integration of all the export schemas forms a global schema. The global schema is used to generate query forms in different languages, through which personnel at port-of-entry stations, government agencies and other authorized users in participating countries can query against the distributed databases stored in any of these countries.

A Distributed Query Processor (DQP) allows users to query the distributed data using a form-based interface or a conversation-based interface. A machine translation system translates between English (used in Belize) and Spanish (used in the Dominican Republic) so that a user can issue a query and receive the query's results in the same language, regardless of the language of the data.

Another main function provided by the prototype system is event-trigger-rule processing. Authorized users in the participating countries can define and register events of common interest (e.g., a person wants to enter a country, or a person is on a watch list) at the Host site by using its Event Registration and Subscription Facility. Other users can browse and subscribe to these events and specify event filtering condition(s) (e.g., the person entering the country or the person on the watch list is from a certain country) for receiving event notifications when the subscribed events occur and the filtering conditions are satisfied. The subscribers also specify the desired means of notification (e.g., by emails, short messages to cell phones, and/or activation of application programs or processes defined as Web-services). The subscription and filtering information of an event is sent to the Event Server of participating sites where the event may occur. When an event occurs, the Event Server processes the registration and filtering information to decide which subscribers to notify. It will also notify the local Event-Trigger-Rule (ETR) Server to trigger rules associated with the event. Additionally, it will also notify the Event Servers of other collaborating countries, which in turn notify their ETR Servers to process the rules that are associated with the event that occurred. Since events are parameterized, the values of the parameters are data relevant to each occurrence of an event (i.e., event data). The event data can be passed to rules for examining the data conditions associated with the event.

Distributed rules are used to specify different countries' policies, regulations, constraints and security and privacy rules and are enforced by replicas of the ETR Server. The Event Registration and Subscription Facility at the host site and the Event Server and the ETR Server at participants' sites together enable the close communication, coordination and collaboration of participating countries and their agencies. Su et al., (under review) provides further details regarding system architecture, implementation, and use scenarios.

The Machine Translation System

In recent years, the field of machine translation has witnessed a marked shift away from knowledge-based approaches and towards the use of fully or partially empirically-based (corpus-based) approaches, particularly in situations where there is no time or budget for manual development of extensive lexical, syntactic and semantic knowledge resources. Examples of such situations include the rapid deployment of translators for new languages, new domains, or in cases of urgent need.

In order to build an MT system quickly, we chose to start with Carnegie Mellon University's Panlite system (Frederking & Brown, 1996), which was used as the translation engine in DIPLOMAT (Frederking, Rudnicky & Hogan, 1997), a rapid-deployment speech-to-speech MT project. Panlite is a multi-engine machine translation framework. Given a sentence to translate, each engine provides a translation (along with a score for each translation) for either the full sentence or fragments of the sentence. Translation candidates are placed in a chart as 'edges' covering the input or some portion of it. One component of the system, the language modeler, uses statistical knowledge of the *target* language (the language the system is translating into) to select or piece together from the chart the best scoring translation(s) that cover the entire input. Recent work has enabled a more effective use of overlapping fragments in composing the final translation (Brown et al. 2003).

The Panlite system supports the integration of widely different MT engines, but provides three built-in engines in addition to the language modeler: an Example-Based MT (EBMT) engine, a Glossary engine, and a Dictionary engine. At its simplest, the EBMT translates by matching new input in the *source* language (the language the system is translating from) against source sentences in previously seen examples of source-target sentence pairs.[2] If it cannot find a match for the entire input sentence, it tries to match all possible multi-word input fragments and posts to the chart what it believes to be the corresponding translations. At times, pieces of the input to be translated cannot be matched against any previously seen source sentences, so there will be holes in the translations produced by the EBMT system and it is useful to back off to the Dictionary engine to obtain single-word translations. Finally, a Glossary engine can supplement the translations provided by EBMT with human-supplied translations for phrases. Translation improvements can also be achieved by using the system's generalization capabilities, which allow the examples to work in a broader range of situations (Brown, 1999; Brown, 2000).

Corpus Collection and Management

Just as the quality of knowledge-based machine translation depends heavily on domain-specific lexical, syntactic, and semantic knowledge, the quality of translation for corpus-based approaches such as EBMT is strongly tied to the availability and coverage of domain-

[2] A corpus of such translation pairs, called a parallel corpus, is the essential 'training' data for CMU's EBMT system. The system does not 'learn' in the traditional machine learning sense; its training consists of processing the parallel data in such a way as to make retrieval of any part of the source sentences and corresponding part of the target sentences as fast as possible when the system is translating new input. The processing also includes determining the correspondence between fragments of parallel source and target sentences.

specific text resources. Unfortunately, though some English-Spanish parallel corpora are available, they are usually formal government documents and some newswire; and while monolingual text resources are abundant, most are out-of-domain, so there is virtually no domain-specific monolingual data from which to create parallel corpora by manual or semi-automatic translation. Because available resources diverge widely in content and style from the text we expect in our domain, we have found that they provide rather little translation help. Our Belizean and Dominican partners have provided some examples of dialogues and descriptions in the border crossing domain, and more recently a few examples of authentic text (e.g., comments recorded in a station diary, advisories to immigration officials regarding individuals on watchlists), but nowhere near the amount needed.

In response to this combined lack of domain-relevant language resources and scarcity of authentic data, we have been employing a variety of techniques to build our domain-specific corpus. The corpus is being used to 'bootstrap' the translation capability of the system prior to demonstrating it to government agencies in collaborating countries and prior to placing it in the field for actual use and authentic data collection. One demonstration, at the ministerial level, took place in Belize in December 2003. The demonstration of a more advanced prototype is planned to take place in the Dominican Republic in 2004. The prototype will also be demonstrated at dg.o2004 in May of 2004. At present and for the near future, the prototype is available to project members in a distributed but benign (university) environment. Actual field tests are planned for late 2004 and 2005.

At present the corpus contains approximately 2,200 domain-specific pairs, of which a decreasing fraction consists of alternate translations for the same English source sentence. We are in the process of adding a few hundred more pairs of recently acquired data.

Corpus Collection: Techniques and Issues

The following techniques are being used to 'bootstrap' the system's translation capability:

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 the same sentences, but also gives rise to some issues in selecting the translation(s) to use as training data for the EBMT system. The solution we have adopted is to select and use only the translation that best matches the style and content of the English source in the English→Spanish translation in order to facilitate intra-sentential alignment. For translation in the Spanish→English direction, we retain all translations in order to provide more match possibilities.

Scenario Generation. Translators imagine circumstances surrounding border crossings and write, in both languages, descriptions of individuals and situations, and hypothetical dialogues that might occur. This technique allows us to obtain a range of content for 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, where the translation is incorrect, it is manually corrected and used to augment the parallel corpus. Using the system for creating demonstration has been a particularly effective way of identifying system weaknesses.

Interviews. During our last project meeting in Belize, a Senior Immigration Official was interviewed and asked to recollect different experiences of problematic border crossings that he encountered 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, is used as a basis for composing texts which, with their manually-produced translations, are used to augment the corpus. In addition to directly providing more parallel text, 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.

While none of the above techniques creates texts that use exactly the kind of language we find in the few station records we have seen, they do largely address the need for domain-specific vocabulary[3] and constructs that is not satisfied by available linguistic resources. They also aid in representing the different dialects that are present in the region and that would be used if the system were fielded in a broader range of countries in the Americas. In fact, the need to accommodate linguistic variety is an unavoidable aspect of our corpus and our project. Our informants use different Latin American dialects, which differ in common everyday words, constructions, and idioms. American English and Belizean English also differ, and not only in spelling (which is influenced by British English). More importantly, in the more authentic data we have seen, immigration agents tend to use an abbreviated form of English, frequently dropping pronouns and auxiliary verbs and using a fair number of abbreviations and acronyms. While the English is perfectly understandable, it cannot be translated into a similarly abbreviated Spanish form.

Another issue for translation is the unpredictability of the data. The domain of border crossings involves many people and place names, not restricted to Spanish and English, since both countries are strong tourism magnets. We are only now starting to address this issue in the context of language resources.

Corpus Management

As the variety of types and sources of data began growing, it became apparent that simply adding new text as source-target pairs to older and less on-point text resources

[3] The vocabulary extracted from the texts is used to enrich and correct the dictionary engine of the Panlite system, the baseline dictionary resource having been automatically extracted from parallel texts that are not in the border crossings domain.

(including the U.N. parallel corpus, other formal documents, and general glossary files to which the MT system backs off) was not sufficient, so we began to develop a representation for the corpus and a set of tools to manage it.

The basic unit of data storage is the phrase or sentence pair, the level at which MT systems, and in particular EBMT and statistical MT systems operate. Different pairs may share a common source or target. For example:[4]

```
((:ENG ((:TEXT "A hand gun was also recovered.")
        (:QUALITY 1)))
 (:SPA ((:TEXT "Una pistola también fue recuperada.")
        (:QUALITY 2)))
 (:ATTRS (:ORIGIN "Demo Belize"
          :LAST-REVISION 0311252323
          :ORIGINAL-DIRECTION e2s
          :TOPIC border
          :FORM description
         )))
```

Most of the information stored in the above representation is self-explanatory. The **:QUALITY** field indicates the order of preference (higher is better) in generating the text as target for the EBMT training corpus. E.g., the above pair could have a companion pair with the same English text and a different Spanish text (**"Una pistola también fue encontrada."**) with **(:QUALITY 1)**. If we were generating an English-to-Spanish corpus, we would prefer **"Una pistola también fue recuperada."**. Keeping **:QUALITY** on both source and target allows us to reverse the corpus and select the preferred target text independently for each direction of translation (the English text could occur as one of a few translations for the Spanish text). The **:FORM** field captures information about the form of the text (e.g., is it a question, an exclamation, a command, etc), which is not always obvious from the characters in the text string; it is used to select subsets of the corpus for review and testing. The **:TOPIC** and **:FORM** fields are also used to extract parts of the corpus when creating training and testing data for the system.

We plan to augment the corpus metadata by adding information regarding the scenario(s) with which each text pair is associated. This information allows the recovery of the "story context" of the text, be it a dialogue or a description, when there is one. It is useful for creating system demos, because translation performance con be shown for a full document. It is also needed for more accurate translation of texts and dialogues containing anaphoric references, which is a future goal of the project.

Summary and Conclusions

We have described several techniques and some issues pertaining to collection and management of a corpus for use with a transnational digital government project. While the availability of linguistic resources, and especially parallel texts, in support of corpus-based approaches to translation has been growing, they span a limited topics and genres that are not helpful for our domain. Therefore we hope that our efforts will eventually result in a corpus that will be of benefit to others as well.

This work is still very much in progress. With the fielding of a pilot system, the next few months should prove quite revealing from the perspective of development of language resources. We expect to acquire greater understanding of the real weight of the above issues, 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 NSF award EIA-0131886. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We gratefully acknowledge the collaboration of all our colleagues in the U.S. and abroad who are participating in this project and in particular we thank our Belizean colleagues – Mr. Rodolfo Bol and Mr. Charles McSweeney – and our Dominican colleagues – Mr. Pedro Taveras and Mr. Juan Luis Ventura – who have assisted in collecting and processing the data.

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. In 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

Frederking, R., Rudnicky, A., and 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.

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.

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.

[4] Because the corpus representation is still not completely stable, we chose to use CommonLisp as the programming language and data representation, since it supports prototyping and facilitates rapid changes to both code and corpus representation.