A client mobile application for Chinese-Spanish statistical machine translation

Jordi Centelles[1,2], *Marta R. Costa-jussà*[1,2], *Rafael E. Banchs*[2]

[1] Universitat Politècnica de Catalunya, Barcelona
[2] Institute for Infocomm Research, Singapore
{visjcs,vismrc,rembanchs}@i2r.a-star.edu.sg

Abstract

This show and tell paper describes a client mobile application for Chinese-Spanish machine translation. The system combines a standard server-based statistical machine translation (SMT) system, which requires online operation, with different input modalities including text, optical character recognition (OCR) and automatic speech recognition (ASR). It also includes an index-based search engine for supporting off-line translation.
Index Terms: machine translation, Chinese, Spanish

1. Introduction

Nowadays, machine translation technologies have become matured enough to support some basic commercial applications. In this sense, currently available systems are able to support basic applications such as cross-language information retrieval and webpage translation. This, combined with the increasing use of smart-phones, which provide portability and availability of internet almost everywhere, is allowing for most of conventional on-line applications to be deployed and used on mobile platforms.

In this show and tell paper we describe a client mobile application for Chinese-Spanish machine translation. This system, which is a server-based application, requires internet connection for its operation. However, as small index of most common translations is included in the client side so some basic translation capabilities are still available on off-line mode. The mobile application also incorporates a set of different functionalities oriented to ease the user-system interaction, as well as to improve the user experience. These functionalities include OCR and ASR input modalities, as well as language detection and image retrieval.

2. SMT system description

The server-side translation engine used by the system is based on the phrase-based SMT approach [1], in which the translation is performed by splitting the input sentence in fragments and each of these fragments is assigned a translation unit from a translation-table. These translation units are selected in order to maximize a linear combination of feature functions. The two main feature functions included in our system are the translation model and the language model. Additional models, such as lexical weights, phrase and word penalty and reordering are also included.

In general, Chinese-Spanish SMT systems follow a pivot approach, in which English is used as intermediate language for the full translation [2]. This is mainly done because of the lack of appropriate bilingual data to train a direct Chinese-Spanish translation system. The main difference of our system with respect to other commercial systems is that we use a direct approach. For this to be possible we have compiled a reasonable large Chinese-Spanish corpus by leveraging on and combining few existing resources. More specifically, we use the *Holy Bible* corpus [3], the United Nations corpus [4], a small subset of the European Parliament Plenary Speeches (which has been automatically translated into Chinese), a technical Chinese-Spanish parallel corpus from TAUS [5], and an small in-house developed corpus in the transportation and hospitality domains. The resulting combined Chinese-Spanish corpus has a total of 70 million words.

In addition, specific preprocessing pipelines have been used for each language. More specifically, for the case of Chinese, the Stanford segmenter [6] is used; while for Spanish, Freeling [7] is used. When Spanish is used as a input, it is additionally reduced to a lower-cased and unaccented form. Finally, for implementing the SMT engine, we have used MOSES [8] with a standard configuration: align-grow-final-and alignment symmetrization, 5-gram language model with interpolation and kneser-ney discount and phrase-smoothing and lexicalized reordering. For optimization purposes, we used our in-house developed corpus.

3. Mobile Application

This section describes the main features of the mobile applications.

The android app was created by using the Android Development Toolkit (ADT), which is a plug-in for the Eclipse IDE. The communication between the Android app and the server is handled by using the HTTPClient interface, which allows a client to send data to the server via the POST method. The Iphone app was developed by using xcode, in which the programming language used is Objective C.

The main features included in the mobile-based client application are the ASR and OCR input modalities, image retrieval, language detection and off-line translation mode.

3.1. Input modalities and language detection

In addition to text input, the application also incorporates ASR and OCR as input modalities.

For ASR, the application uses the native ASR engines of the used mobile platforms: Jelly-bean in the case of Android[1] and Siri in the case of iOS[2].

Regarding OCR, which allows for electronic conversion of scanned images into machine-encoded text, we adapted the open-source OCR Tesseract (released under the Apache license) [9].

[1] http://www.android.com/about/jelly-bean/
[2] http://www.apple.com/ios/siri/

The application also implements a very simple, but effective, language detection system, which allows for automatically distinguishing between Chinese and Spanish, without the need for the user to specify which input language is currently being used. The language detection algorithm is based on comparing code-average values with a numeric threshold for the UTF-8 encoding used by the application.

3.2. Image retrieval

The mobile application also implements a image retrieval function. For this, the popular website flickr [10] is used. This functionality allows for the user to retrieve images that are relvant to the concepts or entities being translated. Some specific Chinese and Spanish terminology can be better described with a picture than with a verbatim translation of the input. For instance, this is mostly the case for specific terminology referring to local geographic names, food names, etc.

3.3. Off-line mode

The mobile-based client application also provides some basic off-line translation capabilities. This is based on a simple index-based search strategy, in which previous translations to commonly used inputs are retrieved based on semantic similarity scores.

The basic operation of the feature is as follows. When a huge number of translations have been done, some statistics are computed to know which are the most common queries entered by the users. Then, these most commonly used inputs are stored, along with their corresponding translations, in a local index which can be directly accessed on the mobile when the device is off-line. A basic search engine is used to retrieve translations from the index based on the semantic similarity between the given input and the stored most-common inputs.

The index is updated on a periodic basis after a significant amount of new data has been collected and when the application is operating in online mode.

4. Future enhancement plans

A future enhancement plan for helping the translator improve its performance over time is to include an option in the mobile-based application for letting the users suggest a better translation than the one they have obtained. When the user thinks that the translation in the target language could be better, she or he can submit a suggestion for the better translation. A database that can store all these suggestions had to be included in the process on the server-application side.

Over time, this database of translation pairs can be used to re-train the main SMT engine and improve the coverage and the overall quality of the system performance.

5. Conclusions

In this show and tell paper, we described a mobile-based client application for Chinese-Spanish SMT. This service allows users to have an easy and convenient access to translations via a mobile app.

The main characteristics of the presented system are: the use of a direct translation engine between Chinese and Spanish (without the use of pivot language), the support of the mobile platform, the integration of alternative input modalities such as automatic speech recognition and optical character recognition, and the incorporation of additional supporting features such as image retrieval, language detection and off-line operation mode.

6. Acknowledgments

The authors would like to thank the Universitat Politècnica de Catalunya and the Institute for Infocomm Research for their support and permission to publish this research. This work has been partially funded by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951), the Spanish Ministerio de Econom´a y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER), and the HLT Department of the Institute for Infocomm Reseach.

7. References

[1] P. Koehn, F. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, 2003.

[2] M. R. Costa-jussà, C. A. Henríquez Q, and R. E. Banchs, "Evaluating indirect strategies for chinese-spanish statistical machine translation," *J. Artif. Int. Res.*, vol. 45, no. 1, pp. 761–780, Sep. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2444851.2444870

[3] R. E. Banchs and H. Li, "Exploring Spanish Morphology effects on Chinese-Spanish SMT," in *MATMT 2008: Mixing Approaches to Machine Translation*, Donostia-San Sebastian, Spain, February 2008, pp. 49–53.

[4] A. Rafalovitch and R. Dale, "United Nations General Assembly Resolutions: A Six-Language Parallel Corpus," in *Proceedings of the MT Summit XII*, Ottawa, 2009, pp. 292–299. [Online]. Available: http://www.uncorpora.org/

[5] TausData, "Taus data," accessed online May 2013 http://www.tausdata.org, 2013.

[6] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter," in *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[7] L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón., "FreeLing 2.1: Five Years of Open-Source Language Processing Tools," in *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valleta, Malta, May 2010.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, June 2007, pp. 177–180.

[9] Tesseract, "Ocr," accessed online May 2013 https://code.google.com/p/tesseract-ocr/, 2013.

[10] Ludicorp, "Flickr," accessed online May 2013 http://www.flickr.com/, 2004.