On-line and Off-line Chinese-Portuguese Translation Service for Mobile Applications

Jordi Centelles[1,2], Marta R. Costa-jussà[1,2], Rafael E. Banchs[2], and Alexander Gelbukh[3]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain

[2] Institute for Infocomm Research, Singapore

[3] Centro de Investigación en Computación, Instituto Politécnico Nacional, México D.F., Mexico

jordi.centelles.sabater@alu-etsetb.upc.edu, marta.ruiz@upc.edu, rembanchs@i2r.a-star.edu.sg, gelbukh@gelbukh.com

Abstract. We describe a Chinese-Portuguese translation service, which is integrated in an Android application. The application is also enhanced with technologies such as Automatic Speech Recognition, Optical Character Recognition, Image Retrieval, and Language Detection. This mobile translation application, which is deployed on a portable device, relies by default on a server-based machine translation service, which is not accessible when no Internet connection is available. For providing translation support under this condition, we have developed a contextualized off-line search engine that allows the users to continue using the application. The system includes a search engine that is used to support our Chinese-Portuguese machine translation services when no Internet connection is available.

Keywords. Online communications, structure, user-generated content, emotions.

1 Introduction

Machine Translation services have become quite popular over the Internet in recent years. Additionally, the first mobile translation applications that offer automatic translation services on portable devices are also starting to appear. Currently, statistical approaches to machine translation are dominating the market, as they allow for automatically learning translation tables from parallel corpora [1, 5]. The main problem for these approaches is the high amount of resources they consume regarding to memory and computational power. Due to this, most translation applications operate under a client-server architecture in which the client only provides a dummy interface while all the computations are carried out on a remote server. The main limitation of this scheme is that the client requires Internet connection for the service to be available.

In this paper, we describe a Chinese-Portuguese Machine Translation Android Application, which integrates technologies such as Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), Image Retrieval, and Language Detection. Additionally, we present a search-based strategy for supporting machine translation services when Internet connection is not available. More specifically, our proposed off-line strategy is designed to support our Chinese-Portuguese translation service that has been deployed at the client side as a mobile application. The off-line mode also includes contextualization strategies that allow improving the system performance based on user preferences, location, and time.

This paper compiles and puts together our previous research and development works [2, 3, 4]. The main contribution of the present paper in comparison with the previous work is the step-by-

Computación y Sistemas Vol. 18, No. 3, 2014 pp. 603–610
ISSN 1405-5546
DOI: 10.13053/CyS-18-3-2047

step description of the mobile application methods in Section 3.

The rest of the paper is structured as follows. In Section 2, we describe the Chinese-Portuguese on-line translation service supported by a standard phrase-based statistical machine translation system. In Section 3, we describe the full on-line translation application integrated with ASR, OCR, image retrieval, and language detection. In Section 4, we present the proposed off-line mode strategy and its contextualization capabilities. Finally, in Section 5, we present our conclusion and propose future directions of research.

2 The Chinese-Portuguese On-line Translation Service

In this section, we describe the Chinese-Portuguese on-line translation service, which follows a statistical approach. First, we describe the standard theory behind the statistical machine translation approach based on phrases. Second, we describe details of how we build our system.

2.1 Statistical Machine Translation based on Phrases

In the development of our machine translation system, we have followed the standard phrase-based statistical machine translation approach based on Moses [6]. Next, we briefly describe the original mathematical equations that support the phrase-based standard system.

The basic theory behind this paradigm is the popular noisy-channel model. For a given sentence *s* in a source language to be translated into an target sentence *t*, we can model the possible translations with a probability distribution $p(t|s)$. This probability distribution can be rewritten (by using Bayes) into the translation probability for translating a source sentence *s* into target *t* as:

$$p(t|s) = \frac{p(s|t)p(t)}{p(s)}.$$

Obviously, there is not only one correct translation. An acceptable way to find the best translation from *s* to *t* is to find the *t,* which maximizes this conditional probability. Moreover, as *s* is fixed, then:

$$arg\max_t p(t|s) = arg\max_t p(s|t)p(t),$$

where $p(s|t)$ is referred to as the translation model and $p(t)$ is referred to as the language model. The former is trained on a parallel corpus at the level of sentences. The latter is trained on a monolingual corpus. This original approach combining two models has been extended to the log-linear model [7], which combines additional models to the translation and language model, e.g. the reordering model.

2.2 Data and System Description

To build our Chinese-Portuguese phrase-based statistical machine translation system, we require a Chinese-Portuguese parallel corpus. Different domain corpora were concatenated into a single training corpus. In particular, we have used the following:

- TAUS. Data provided by this organization include translation memories of technical content. This corpus has 5 million sentences and around 60 million words.
- In-house. This corresponds to a small corpus in the transportation and hospitality domains. This corpus has 729 sentences and around 4.5 thousand words.

As in many natural language processing tasks, data preprocessing is an essential step to do in order to obtain better models. Regarding data preprocessing we have done the following:

- For Chinese, we have segmented the data using the Stanford Segmenter tool [9].
- For Portuguese, we have true cased the data and tokenized it with Moses tools.

After preparing the data, we proceeded to train the phrase-based system with Moses, which we used with the standard configuration.

For fine-tuning the translation engine, we have used the TAUS development dataset (808 sentences) and, then, we have tested with the TAUS (721 sentences) and the in-house test.

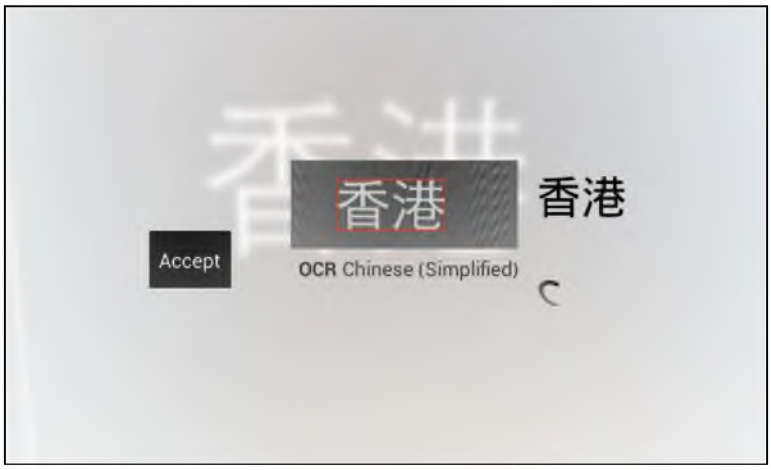

Fig. 1. OCR screenshot

With the reported data and configuration, we show the automatic and human evaluation results for Chinese-Portuguese in terms of the standard metric BLEU in Table 1. The system obtains the better quality in terms of BLEU for the Portuguese-to-Chinese direction.

Table 1. Translation results

Translation direction	Domain / Dataset	BLEU
Chinese-to-Portuguese	TAUS	37.97
	In-house	4.49
Portuguese-to-Chinese	TAUS	39.58
	In-house	6.48

3 Chinese-Portuguese On-line Translation Application

In this section, we present a detailed description of the mobile application that connects to this service (the client side). The Android application for the Chinese-Portuguese translation client was programmed with the Android Development Toolkit (ADT). It is a plug-in for the Eclipse IDE that provides the necessary environment for building Android applications.

For the communication between the Android application and the server, we use the HTTP client interface. Among other things, it allows a client to send data to the server via, for instance, the *post* method.

In addition to the base translation system, the application also incorporates Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) technologies as input methods, as well as Image Retrieval and Language Detection functionalities [2]. Additional functions of the application are the copy button and full-screen.

3.1 Input Methods

For ASR, the application relies on the native ASR engine of the used mobile platform, which is Jelly-bean[1] for Android.

Regarding the OCR (see screenshot in Figure 1), the application adapts the open-source OCR Tesseract[2]. The OCR integration was conducted by using the Android *intent* method. *Intent* allows opening a secondary application inside the main application and when the secondary one is done, its result is sent to the main application with the *putExtra* method.

In order to send the sentence entered by the user from the application to the translation system, there is stored one PHP file in the web-server waiting for the sentences. Thereby, the

[1] http://www.android.com/about/jelly-bean/
[2] https://code.google.com/p/tesseract-ocr/

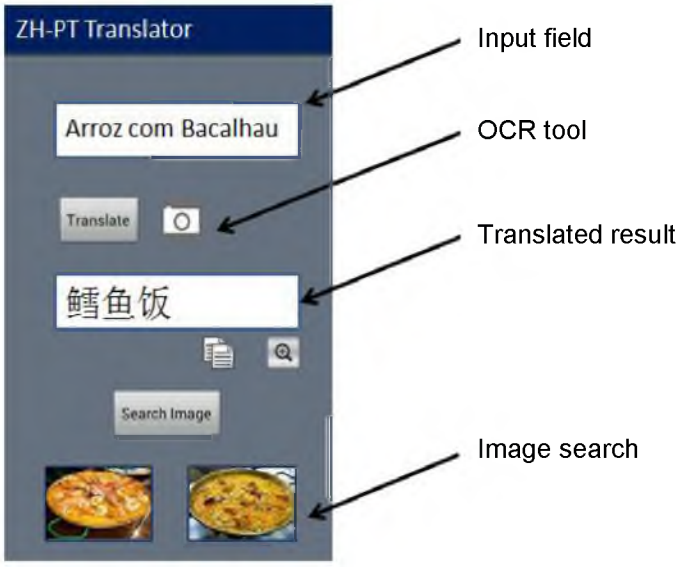

Fig. 2. Android-based Chinese-Portuguese translation client application

communication with the PHP file is performed with the HTTP client Interface as mentioned earlier. After creating a new HTTP client, it is necessary to specify the URL of the server, the data that has to be sent, and its codification. Finally, with the HTTP entity you settle down that the application has to be kept waiting for a server's reply. When the translation is received, the handler method is in charge of displaying it in the appropriate text field. Furthermore, it is important to include a conversion function from a stream-to-bits to a string. We needed to define two HTTP client functions: one for the translations from Portuguese to Chinese and another for the other way around.

3.2 Image Retrieval and Language Detection

For image retrieval, we used the popular website flickr[3] and the following procedure:

- When the "search Image" button is pressed, we send an URL (using the Java *HTTPclient* method) to a flickr server with all the needed information to receive the needed metadata.
- In the URL we specify the tag (i.e. the topic of the images we want), the number of images, the secret key (needed to interact with flickr) and also the type of object we expect (in our case, a JSON object).
- When the response of the server is received, the JSON object is parsed. There are two main fields in the JSON object: *photos* and *stat*. *Stat* only reports whether the *url* sent was correct ("ok") or not ("fail"). *Photos* contains the information for each one of the 5 images we requested, if they are found.
- To avoid the application from crashing when there are no images for a tag, we had included an *if* sentence that closes the thread and displays this sentence on the screen: "No available images".
- With the *HTTPConnection* method and the information parsed, we send the URL again to the server and we retrieve the images requested in the application.

It is worth mentioning that the Java class that implements all these methods extends an *AsyncTask* in order to not block the user interface while the application is exchanging information with the flickr server. The Android-based application is shown in Figure 2.

For language detection, the application

[3] http://www.flickr.com/

Fig. 3. Full-screen mode

implements an effective technique. The application encoding is UTF-8. For this type of coding codes for most characters used in Portuguese are in the range from 40 to 255, and for Chinese are in the range from 11,000 to 30,000. The procedure consists in computing the average code for the sequence of characters to be translated and use a threshold to determine the language.

3.3 Additional Functionality: Copy Button and Full-screen Mode

There are two additional features: the copy button and the full-screen. When the copy button is pressed, it copies the output of the translation on the clipboard. This means that you can paste the translation in every other text field (for instance, you may open your smartphone's browser and paste in a Google search field the Chinese sentence you already translated and thus, obtain the search results).

Another feature of our system is the full-screen (see screenshot in Figure 3). It allows the user to show in full-screen mode the translation of his sentence for easier display to others. For example, you can write "Eu quero ir para a rua X" (*I want to go to the street X*). The sentence is translated and if you press the full-screen button, the Chinese sentence will be displayed in the whole screen, which is easier to show to the taxi driver.

In addition, the system uses a database to store the translation performed by the system and keep track of the most used translations. To create the databases we used the popular open source database management system: MySQL.

4 Off-line Search-based Translation

The previous described system requires having Internet connection available. In this section, we describe a search-based off-line strategy to support the Chinese-Portuguese translation service as initially proposed in [3]. We describe our search engine implementation for translation, and then, we present the developed contextualization strategy for improving the performance of the system.

4.1 Search Engine for Translation

In the majority of information retrieval applications, the user provides a query aiming at recovering documents that are relevant to the query. The translation task can be seen from a similar perspective: the user provides a source sentence to be translated (a query) aiming at obtaining the most meaningful translation for it.

In our proposed approach to translating by means of retrieval search engine, we construct two composed indexes, one for each language (source and target), in which pointers to each other are also included. This index construction is performed in three steps:

1. Common translation collection: we collect the most commonly Chinese and Portuguese sentences and their respective translations from the translation service. This bilingual data collection is updated on a monthly basis according to the activity of the on-line registered users.
2. Bilingual dictionary match: form the collected bilingual sentence pairs, a bilingual dictionary is used to identify Chinese and Portuguese

term translations simultaneously occurring in the sentence pairs, which are replaced by entry codes in the dictionary. The entries of the used bilingual dictionary correspond with nouns and adjectives that are commonly observed in the translated pairs.

3. Index construction: a Chinese index is constructed by using the processed Chinese sentences and, in the same way; a Portuguese index is constructed by using the processed Portuguese sentences. The two indexes include pointers to each other so each Portuguese sentence points to its corresponding Chinese translation and each Chinese sentence points to its corresponding Portuguese translation.

These indexes are implemented by using the scheme of bag-of-words, for which the TF-IDF weighting scheme is used [8]. For searching across the indexes, cosine similarity metric is used for ranking the retrieved outputs. Given a user input in the source language, the retrieval process is implemented in two steps:

1. Dictionary match: the input sentence is evaluated for occurrences of terms from the bilingual dictionary. In case a term is detected, it is replaced by its corresponding entry code.
2. Source search: two searches are performed over the source language index, the first one involves the original sentence provided by the user, and the second one involves the processed sentence (if terms have been found on it). The retrieved sentence with highest cosine similarity score is then selected.

Finally, the translation is constructed by using the corresponding sentence pair from the target language index:

- Sentence extraction: the target sentence corresponding to the selected source sentence is extracted from the target index if the obtained cosine similarity is high enough (current threshold value is 0.85).
- Sentence post edition: if the selected target sentence includes one or more dictionary entry codes on it, they are replaced by their corresponding dictionary forms before providing the final translation to the user.

4.2 Contextualized Translation Services

For providing the system with contextualization capabilities, each requested translation and its corresponding result from the online service are logged in the system along with the following four types of metadata:

- User information, which offers a unique identification number for the user requesting the translation.
- Location information, which provides spatial coordinates as offered by the GPS service of the mobile device at the moment the translation was requested.
- Time information, which offers time stamp for the specific hour and day at which the translation was requested.
- Semantic information, which gives a semantic categorization of the specific topic the requested translation belongs to.

These types of metadata are used to train a personalized predictive model able to estimate which are the most probable translations the current user might be requesting in the next 24 hours, based on the current context (user-location-time) and previous translation history.

This model is updated every time the system is using the online mode, and the corresponding translation indexes and dictionaries are refreshed based on the model predictions. In this way, when going off-line, a personalized and contextualized translation service is locally available for the user.

5 Conclusions and Future Work

In this work, we have described the on-line and off-line Chinese-Portuguese translation service designed for an Android application. Our on-line system is a corpus-based approach where the translation quality depends on the quality and quantity of the corpus used for training. Generally speaking, systems for translating between distant language pairs, such as Chinese and Portuguese, typically follow pivot approaches through English

(or other major-resourced language) because of the lack of parallel data to train the direct approach. The main advantage of our system is that we have implemented the direct approach by relying on a quite large corpus, which has been properly preprocessed.

Our mobile translation application, which is deployed on a portable device, integrates ASR, OCR, image retrieval, and language detection technologies. However, the application relies by default on the server-based machine translation engine, which is not accessible when no Internet connection is available. For providing translation support under this condition, we have developed a contextualized off-line search engine that allows the users to continue using the application. The off-line translation system is implemented by means of a search engine.

As future work we plan to improve our off-line solution by incorporating predictive suggestions, so the system can suggest source sentences to the user by using partial inputs as queries for searching across the source index. We also want to improve the contextualization capabilities by including user dependent models for spatial and time localization.

Acknowledgements

This work is supported by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER).

References

1. **Brown P. F., Della Pietra S. A., Della Pietra V. J., & Mercer R. L. (1993).** The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, 263–311.
2. **Centelles, J., Costa-jussà, M. R., & Banchs, R. E. (2014).** CHISPA on the GO. A mobile Chinese-Spanish translation service for travelers in trouble. In *Proc of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2014), *Demo Track*.
3. **Centelles, J., Costa-jussà, M. R., Banchs, R. E., & Gelbukh, A. (2014).** An IR-based strategy for supporting Chinese-Portuguese translation services mode. In *15th International Conference on Intelligent Text Processing and Computational Linguistics*, *CICLing 2014*, *Part II*, *Lecture Notes in Artificial Intelligence*, Vol. 8404, 324–330.
4. **Centelles, J., Costa-jussà, M. R., & Banchs, R. E. (2014).** A Client Mobile Application for Chinese-Spanish statistical machine translation. In *Proc. of the InterSpeech 2014*, *Demo Track*.
5. **Koehn P., Och F. J., & Marcu D. (2003).** Statistical phrase-based translation. In *Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL'13)*, 127–133.
6. **Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007).** Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, 177–180.
7. **Och, F. J. & Ney, H. (2002).** Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, 295–302.
8. **Salton G. & Buckley C. (1988).** Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, Vol. 24, No. 5, 513–523.
9. **Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005).** A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

Jordi Centelles is a Telecommunications Engineer from Polytechnic University of Barcelona (UPC), currently working as a Consultant at PriceWaterhouseCoopers (PwC). Besides, He is pursuing a BSc in Business Management from University of Barcelona. He developed his final degree project about Statistical Machine Translation in the Institute for Infocomm Research (Singapore). In addition, he

participated in the Google Summer of Code 2013 programming the basis of a Chinese-Spanish Rule-Based translation system.

Marta R. Costa-jussà received her PhD from the Universitat Politécnica de Catalunya (UPC, Barcelona) in 2008. Her research experience is mainly in Machine Translation (MT). She has worked at LIMSI-CNRS (Paris), Universitat Politècnica de Catalunya (Barcelona), Barcelona Media Innovation Center (Barcelona), Universidade de São Paulo (São Paulo), Instituto Politécnico Nacional (Mexico), and Institute for Infocomm Research (Singapore). She has received prestigious fellowships, such as Juan de la Cierva (from the Spanish Government), FAPESP Visiting Professor (from São Paulo Research Foundation), and an IOF Marie Curie (from the European Commission). She has participated in 12 European and National (Spanish, French, and Brazilian) projects. She has organized 5 conferences or workshops in the areas of MT and IR, has given more than 20 invited talks, and published over 90 papers in international scientific journals and conferences; her papers received several awards. She has been cooperating with companies (TaUYou and UniversalDoctor) as a consultant.

Rafael E. Banchs is currently a Research Scientist at the Institute for Infocomm Research in Singapore. He received his Ph.D. in Electrical Engineering from the University of Texas at Austin in 1998. He was awarded a Ramon y Cajal fellowship from the Spanish Ministry of Education and Science from 2004 to 2009. His recent areas of research include Machine Translation, Information Retrieval, Cross-language Information Retrieval and Dialogue Systems. More specifically, he has been working on the application of vector space models along with linear and nonlinear projection techniques to improve the quality of statistical machine translation, cross-language information retrieval systems, natural language understanding and automated dialogue systems. He has been author and co-author of more than 80 technical papers, some of which have been published in indexed journals and international conferences.

Alexander Gelbukh received his Ph.D. degree in computer science from VINITI, Russia. He is Research Professor and Head of the Natural Language Processing Laboratory of the Centro de Investigación in Computación (CIC) of the Instituto Politécnico Nacional (IPN), Mexico; President of the Mexican Society of Artificial Intelligence (SMIA); member of the Mexican Academy of Sciences; and National Researcher of Mexico (SNI) at excellence level 2. He has received various prestigious awards and fellowships, including various best paper awards. He has organized over 30 international conferences and has led over 20 research projects. He is Editor in Chief of two international research journals. He is author or co-author of more than 500 research publications in natural language processing and artificial intelligence.

Article received on 14/01/2014, accepted on 01/02/2014.