An IR-based strategy for supporting Chinese-Portuguese translation services in off-line mode

Jordi Centelles, Marta R. Costa-jussà, Rafael E. Banchs and Alexander Gelbuck⊥

Institute for Infocomm Research, Singapore
⊥National Polytechnic Institute, Mexico City
{visjcs,vismrc,rembanchs}@i2r.a-star.edu.sg; gelbuck@geclbuck.com

Abstract. This paper describes an Information Retrieval engine that is used to support our Chinese-Portuguese machine translation services when no internet connection is available. Our mobile translation app, 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 app.

Keywords: online communications, structure, user generated content, emotions

1 Introduction

Machine translation applications have gained a lot of popularity in recent years. Currently, statistical approaches to machine translation are dominating the market, as they allow for automatically learning translation tables from parallel corpora (Brown et al 1993, Koehn et al 2003). The main problem for this 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 required internet connection to be available.

In this work, we present a search-based strategy for supporting machine translation services when internet connection is not available. More specifically, our proposed strategy, which is based on Information Retrieval technologies, is designed to support our Chinese-Portuguese translation service that has been deployed at the client side as a mobile app. The proposed strategy, allows for the mobile app to continued operating, with limited capabilities, on off-line mode when no internet connection is available. The off-line mode also includes contextualization strategies that allow improving the system performance based on user preferences, location and time.

The rest of the paper is structure as follows. In section 2, we describe the original Chinese-Portuguese on-line translation service. In section 3, we present the proposed off-line mode strategy and its contextualization capabilities. Finally, in section 4, we present our conclusion and proposed future directions of research.

adfa, p. 1, 2011.

2 Chinese-Portuguese On-line Translation Services

In this section we describe the original Chinese-Portuguese on-line translation service. First, we present a brief overview on the Chinese-Portuguese machine translation engine (the server side), and then we present a detailed description of the mobile app that connects to this service (the client side).

2.1 Chinese-Portuguese Translation System

In order to build our machine translation system, we have used a standard phrase-based statistical machine translation based on Moses (Koehn et al., 2007). This well-known approach splits the source sentence to translate in segments and it assigns to each segment a bilingual phrase from a phrase-table. Bilingual phrases are translation units that contain source words and target words. These bilingual phrases have different scores associated to them (including conditional, posterior and lexical probabilities). Among the list of bilingual phrases, the decoder is in charge of selecting the ones that maximize the linear combination of feature functions. Such strategy is known as the log-linear model (Och and Ney, 2002). The two main feature functions are the translation model and the target language model. Additional models include phrase and word penalty and reordering.

Our system is a corpus-based approach where the key for translation quality is regarding the quality and quantity of the corpus used for training. Generally speaking, translation between distant language pairs follows 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 are using the direct approach and at the same time we rely on a pretty large corpus which has been properly preprocessed.

Regarding data preprocessing we have done the following:

- For Chinese, we have segmented the data using the Stanford Segmenter tool (Tseng et al., 2005).
- For Portuguese, we have true cased the data and tokenized it with Moses tools.

Moses was used with the standard configuration. Different training domain corpus where concatenated to a single training corpus. We have corpora from different domains available. In particular we have used the following ones:

- TAUS. Data provided by this organization include translation memories of technical content.
- In-house. This corresponds to a small corpus in the transportation and hospitality domains

Corpus statistics for the training corpus are presented in Table 1.

Just to give an idea of the quality of our translation system we report the automatic and human evaluation results for Chinese-Portuguese. For fine-tuning the translation engines, we have used the TAUS development dataset and, then, we have tested with

the TAUS and In-house test. Results are shown in terms of the standard metric BLEU in Table 2.

Dataset	Parameter	Chinese	Portuguese
TAUS Train	Number of sentences	5 M	
	Running words	57 M	62 M
	Vocabulary	648 K	200 K
TAUS Dev	Number of sentences	808	
	Running words	11 K	12 K
	Vocabulary	3.0 K	3.4 K
TAUS Test	Number of sentences	721	
	Running words	9.9 K	10.9 K
	Vocabulary	2.8 K	3.3 K
In-house	Number of sentences	729	
	Running words	4.1 K	4.7 K
	Vocabulary	737	890

Table 1. Corpus details

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

Table 2. Translation results

2.2 Chinese-Portuguese Translation App

The android app for the Chinese-Portuguese translation client was programmed with the Android development tools (ADT). It is a plug-in for the Eclipse IDE that provides the necessary environment for building an app.

The Android-based app is depicted in Figure 1. For the communication between the Android app and the server we use the HTTPClient interface. Among other things, it allows a client to send data to the server via, for instance, the POST method, as used on the website case.

In addition to the base translation system, the app also incorporates Automatic Speech Recognition (ASR), Optical Character Recognition technologies as input methods (OCR), Image retrieval and Language detection (Centelles et al., 2013).

Also, 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.

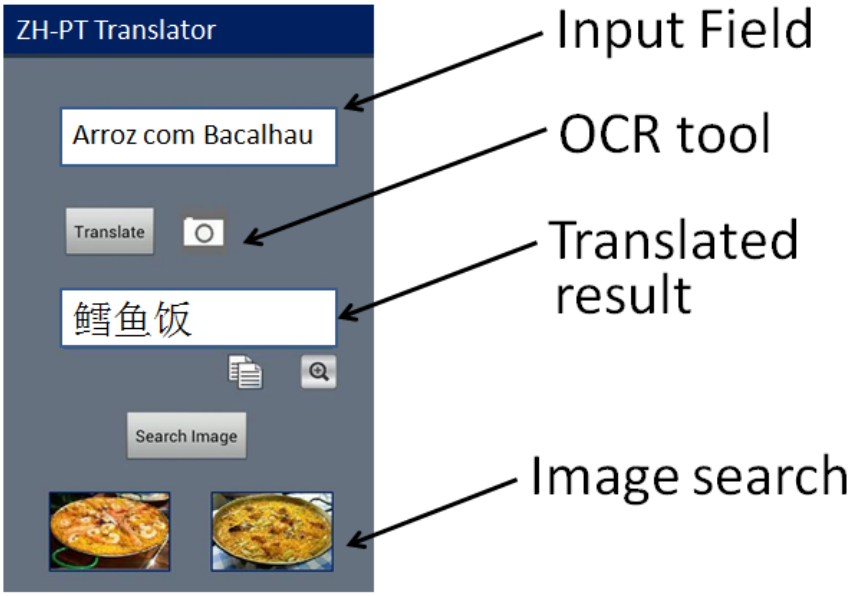

Fig. 1. Android-based Chinese-Portuguese translation client application

3 Off-line Search-based Translation System

In this section we describe our proposed search-based off-line strategy to support the Chinese-Portuguese translation service. First, we describe our search engine implementation for translation, and then, we present the developed contextualization strategy for improving the performance of the system.

3.1 Search Engine for Translation

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

In our proposed approach to translating by means of information retrieval we construct two composed indexes, one in each language, in which pointers to each other are also included. This index construction is performed in three steps:

- 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.

- 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.
- A Chinese index in 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 bag-of-words approach, for which the TF-IDF weighting scheme is used (Salton and Buckley 1988). 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:

- 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.
- 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.

Figure 2, illustrates the index construction, search, and translation generation processes used for the off-line translation system implementation.

3.2 Contextualized Translation Services

Finally, in this section we describe our contextualization strategy for improving the quality of the off-line translation service.

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 types of metadata:

- User information: unique identification number for the user requesting the translation.

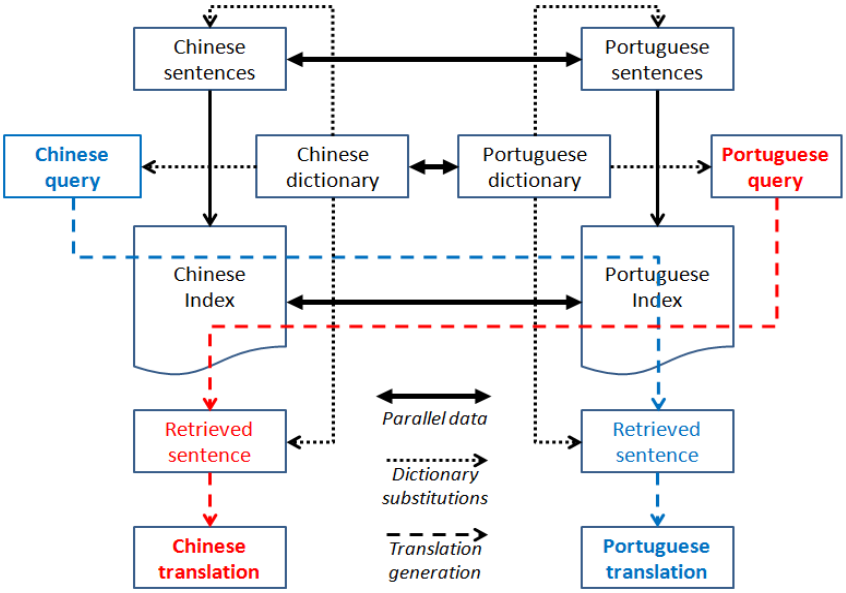

Fig. 2. Proposed approach for off-line translation by means of an information retrieval strategy over the collection of the most commonly requested translation pairs.

- Location information: spatial coordinates as provided by the GPS service of the mobile device at the moment the translation was requested.
- Time information: time stamp for the specific hour and day at which the translation was requested.
- Semantic information: a semantic categorization of the specific topic the requested translation belongs to.

These four 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.

4 Conclusions and Future Work

In this work we have described an Information Retrieval engine that is used to support our Chinese-Portuguese machine translation services when no internet connection is available. Our mobile translation app, 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 app.

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). The authors also want to thank the Institute for Infocomm Research for its support and permission to publish this research.

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*, 19(2): 263–311
2. Centelles, J., Costa-jussà, M.R and Banchs, R. E. (2013) CHISPA on the GO A mobile Chinese-Spanish translation service for travelers in trouble. Submitted to Demo track of IJCNLP 2013
3. Koehn P., Och F. J., Marcu D. (2003) Statistical phrase-based translation. In *Proc of HLT/NAACL'03*, pp 127–133
4. 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. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Associa-tion for Computational Linguistics (ACL'07), pages 177–180, Prague, Czech Republic, June
5. H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter. In Fourth SIGHAN Workshop on Chinese Language Processing
6. Salton G., Buckley C. (1988) Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5): 513–523.