Chinese-Spanish Statistical Machine Translation Experiments

Rafael E. Banchs, Josep M. Crego, Patrik Lambert, José B. Mariño

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

{rbanchs, jmcrego, lambert, canton}@gps.tsc.upc.edu

Abstract

This article presents some experimental results on Chinese-Spanish machine translation. The implemented translation system is based on the statistical framework and, more specifically, it implements the bilingual n-gram approach. Since, as far as we know, no Chinese-Spanish parallel corpus is currently available for training purposes, an alternative method for artificially constructing a training corpus is proposed and evaluated. This method is compared, in terms of translation quality, to the simpler approach of using English as a bridge language for performing Chinese-to-Spanish and Spanish-to-Chinese translations. Additionaly, the translation system is also compared with a rule-based translation engine which is available on-line.

Motivation

Language	*Native Speakers*
Chinese	**874.000.000**
Spanish	**332.000.000**
English	322.000.000
Hindi	291.500.000
Arabic	193.000.000

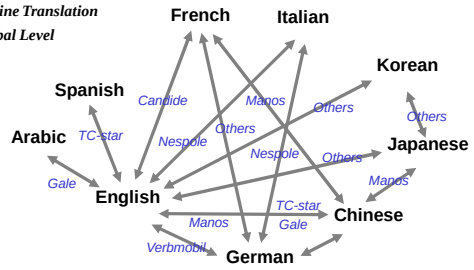
Language	*Including 2nd Language*
Chinese	**1.051.000.000**
Hindi	594.000.000
English	510.000.000
Spanish	**420.000.000**
Russian	255.000.000

The most spoken languages in the world

+

Growing interest in cultural commercial and technological relations between China and Spanish speaking countries

Main Research Efforts in Machine Translation at a Global Level

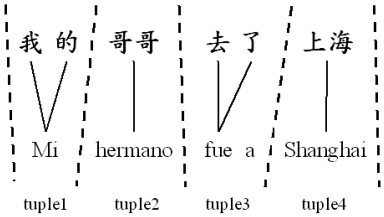


Nobody is working on Spanish-Chinese Machine Translation !!!

The Bilingual n-gram SMT Approach

- This approach to SMT implements a translation model which is based on 3-grams of bilingual units which are referred to as tuples. For a more detailed description refer to Mariño *et al*. (2005).

Translation Probability $\Rightarrow$ $p(T,S) \approx \prod_{k=1}^{K} p((t,s)_k \mid (t,s)_{k-1}, (t,s)_{k-2})$



- Tuples are extracted from the union of source-to-target and target-to-source Viterbi alignments in such a way that a maximal monotonic segmentation of each bilingual sentence pair is achieved.

Log-Linear Feature Function Combination

The bilingual n-gram SMT system implements four additional feature functions which are log-linearly combined with the translation model for decoding purposes:

- a target language model implemented by means of word 4-grams,
- a word bonus model that compensates the system preference for short translations over large ones, and
- two complementary translation models (source-to-target and target-to-source) which are implemented by using the IBM-1 lexical parameters.

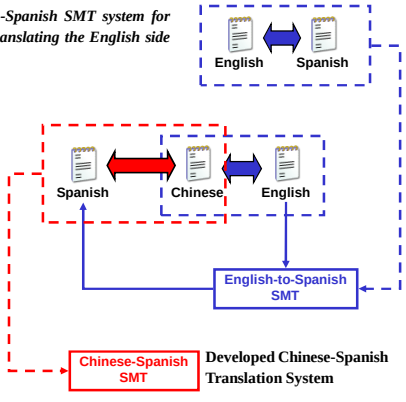
Decoding and Optimization

- A specific n-gram based SMT decoder was developed by Crego *et al*. (2005). This decoder implements a beam-search strategy based on dynamic programming and allows for non-monotonic search (however, in order to maintain computational time manageable, monotonic decoding is used here).
- An optimization tool based on a downhill simplex method was also developed. This algorithm allows for estima-ting log-linear weights for each feature so that the translation *BLEU* is maximized over a development set

Chinese-Spanish Parallel Corpus Construction

The proposed method relies on using an English-to-Spanish SMT system for constructing a Chinese-Spanish parallel corpus by translating the English side of a Chinese-English parallel corpus into Spanish.

- An English-to-Spanish translation system was trained and optimized.
- This system was used to translate into Spanish the English side of the Chinese-English training corpus.
- A development corpus was created by manually translating into Spanish the English side of a 330-sentence set extracted from additional Chinese-English parallel data.
- Similarly, a test corpus was created by manually translating into Spanish the English side of a 100-sentence set extracted from additional Chinese-English parallel data.



Developed Chinese-Spanish Translation System

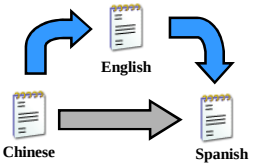
Original Chinese-English Spanish-English Corpora Statistics

Corpus	Language	Sentences	Words	Vocab.
ZH-EN	Chinese	105 K	1.9 M	29.5 K
	English	105 K	2.1 M	34.8 K
ES-EN	Spanish	105 K	2.0 M	40.0 K
	English	105 K	2.0 M	27.0 K

Constructed Chinese-Spanish Parallel Corpus Statistics

Corpus	Language	Sentences	Words	Vocab.
TRAIN	Chinese	105 K	1.9 M	29.5 K
	Spanish	105 K	2.0 M	34.8 K
DEV	Chinese	330	6.0 K	1.6 K
	Spanish	330	6.8 K	2.0 K
TEST	Chinese	100	1.9 K	813
	Spanish	100	2.1 K	908

Direct Translation vs. Indirect Translation

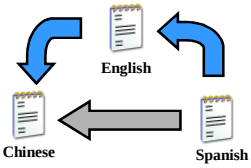


Chinese-to-Spanish Translation Results

Strategy	*BLEU*	*NIST*	*WER*	*PER*
Direct	0.1087	4.157	83.81	62.14
Indirect	0.1145	4.413	78.04	58.21

Spanish-to-Chinese Translation Results

Strategy	*BLEU*	*NIST*	*WER*	*PER*
Direct	0.0391	3.946	76.16	58.54
Indirect	0.0397	3.378	75.62	59.22



Banchs *et al*. (2006) have already confirmed that artificially constructing a Chinese-Spanish bilingual corpus does not necessarily conduce to significant improvements in translation accuracy with respect to the simpler approach of performing Chinese to Spanish translations by using English as a bridge.

Chinese-Spanish Statistical Machine Translation Experiments

Rafael E. Banchs, Josep M. Crego, Patrik Lambert, José B. Mariño

{rbanchs, jmcrego, lambert, canton}@gps.tsc.upc.edu

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Corpus-Filtering

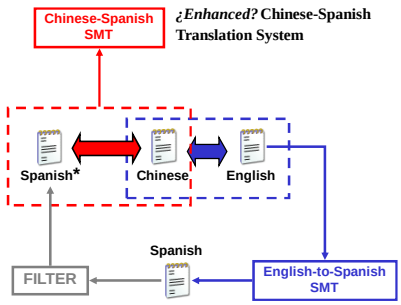
A corpus filtering strategy is proposed aiming at improving the direct Chinese-Spanish translation results:

- The filtering strategy consists on using a Spanish language model for selecting those best Spanish sentences in the Chinese-Spanish parallel corpus (the noise expected to occur in the Chinese-Spanish corpus should be related to the translation errors produced by the English to Spanish translation system).
- The filter is implemented by means of a 3-gram language model, trained from the Spanish side of the original English-Spanish corpus.
- Since language model probabilities are affected by sentence length, filtering is performed independently for each subset of Spanish sentences of equal length.

However, to opposite effects are expected to occur:

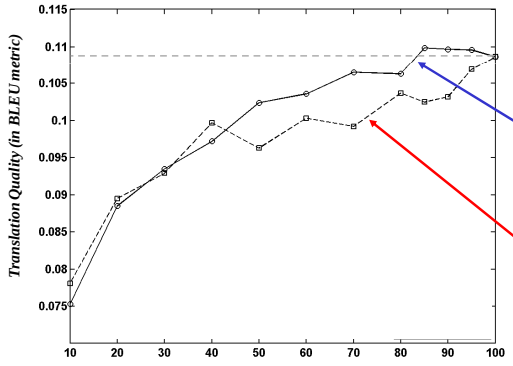
Corpus quality improvement → **Translation quality increases**

Bilingual corpus size reduction → **Translation quality decreases**



Proposed Corpus-Filtering Technique

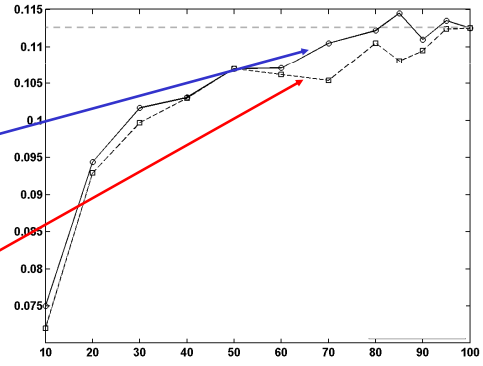
Experiment 1: Target Language Model from Training Data



Systems trained with filtered corpus

Control systems (with training corpus selected at random)

Experiment 2: Common Independent Target Language Model



Comparison with a Rule-Based MT System

System Comparison by using Automatic Evaluation Metrics

As an additional evaluation of the developed translation system, a comparison with a rule-based translation system was performed. In order to do this, a translation system which is publicly available on-line* was used.

In order to avoid any possible bias resulting from the optimization process, this evaluation was carried out over the manually constructed test set.

System	*BLEU*	*NIST*	*WER*	*PER*
Statistical MT	0.1336	4.3101	57.99	79.73
Rule-Based*	0.0697	2.8355	74.15	93.71

Note: this comparison is not completely fair because, first, the statistical system was adapted to the task under consideration while the rule-based one was not, and second, it is known that accuracy scores such as BLEU and NIST tend to favor statistical systems over rule-based systems.

* http://www.worldlingo.com/en/products_services/worldlingo_translator.html

Some Translation Examples from a Turistic Domain Task

Original Sentence	*Meaning*	*Statistical MT*	*Rule-Based MT*
请 给 我 看看 菜单 。	*Please show me a menu*	Le ruego me demuestran el menú, por favor.	Tenga por favor una mirada el menú a mí.
请 给 我 一 杯 白 葡萄酒 。	*Please give me a glass of white wine*	Le ruego me conceda un vaso blanco vino.	Por favor déme el cristal de los vinos blancos.
请 叫 服务生 搬 行李 。	*Call someone to carry my bags please*	Por favor, pedimos bellboy llevar el equipaje.	Por favor llámese el servidor para mover el bagaje.
我 想要 导游 。	*I want to have a travel guide*	Me gustaría recibir una guía, por favor.	Deseo el tourguide.
请 稍 等 。	*Just a moment please*	Hace un momento, por favor.	Espere por favor un pedacito.

Conclusions and Further Work

This work presented some experimental results on Chinese-Spanish SMT:

- It was confirmed that artificially constructing a Chinese-Spanish bilingual corpus does not produces significant improvements in translation accuracy with respect to the simpler approach of performing Chinese to Spanish translations by using English as a bridge.
- A method for artificially constructing and filtering a Spanish-Chinese parallel corpus was presented and evaluated. Results demonstrated that the negative effect resulting from data reduction is at least as relevant as the positive effect resulting from filtering for the overall system performance.
- The constructed translation system was compared with a rule-based translation system which is publicly available on-line. The developed system outperformed the on-line one.

For further research we are planning to work in two main directions:

- We will attempt improving the Chinese-Spanish parallel corpus construction technique. In this sense, different alternatives for filtering the artificially constructed data set should be designed and evaluated, such as the SMT confidence estimation techniques proposed by Ueffing and Ney (2004).
- Additional efforts will also be devoted to improve the translation system, by including additional features and allowing for non-monotonic search in the translation task under consideration.

Acknowledgements

This work has been partly funded by TALP (Centre de Tecnologies i Aplicacions del Llenguatge i la Parla) and by the Spanish Department of Education and Science.

References

- R.E. Banchs, J.M. Crego, P. Lambert, J.B. Mariño, 2006, "A feasibility study for Chinese-Spanish statistical machine translation", in *Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing.*
- F. Casacuberta, E. Vidal, 2004, "Machine translation with inferred stochastic finite-state transducers", *Computational Linguistics*, vol 30, pp. 205-225.
- J.M. Crego, J.B. Mariño, A. de Gispert, 2005, "A ngram-based statistical machine translation decoder", *in Proc. of the 9th European Conference on Speech Communication and Tech-nology*, Interspeech.
- J. M. Crego, A. de Gispert, and J. Mariño, "The TALP ngram-based SMT system for IWSLT'05" Pittsburgh, USA, October 2005, pp. 191–198.
- J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, M. Ruiz, 2005, "Bilingual n-gram statistical machine translation", *in Proc. of the X MT-summit*, pp. 275-282.
- N. Ueffing, H. Ney, 2004, "Bayes Decision Rules and Confidence Measures for Statistical Machine Translation", in *EsTAL - España for Natural Language Processing*, Springer Verlag, LNCS, Alicante, Spain, pp. 70-81.