# The BM-I2R Haitian-Créole-to-English translation system description for the WMT 2011 evaluation campaign

**Marta R. Costa-jussà**
Barcelona Media Innovation Center
Av Diagonal, 177, 9th floor
08018 Barcelona
marta.ruiz@barcelonamedia.org

**Rafael E. Banchs**
Institute for Infocomm Research
1 Fusionopolis Way 21-01
Singapore 138632
rembanchs@i2r.a-star.edu.sg

## Abstract

This work describes the Haitian-Créole to English statistical machine translation system built by Barcelona Media Innovation Center (BM) and Institute for Infocomm Research (I2R) for the 6th Workshop on Statistical Machine Translation (WMT 2011). Our system carefully processes the available data and uses it in a standard phrase-based system enhanced with a source context semantic feature that helps conducting a better lexical selection and a feature orthogonalization procedure that helps making MERT optimization more reliable and stable. Our system was ranked first (among a total of 9 participant systems) by the conducted human evaluation.

## 1 Introduction

During years there has been a big effort to produce natural language processing tools that try to understand well written sentences, but the question is how well do these tools work to analyze the contents of SMS. For example, not even syntactic tools like stemming can bring to common stems words that have been shortened (like Xmas or Christmas).

This paper describes our participation on the 6th Workshop on Statistical Machine Translation (WMT 2011). The featured task from the workshop was to translate Haitian-Créole SMS messages into English. According to the WMT 2011 organizers, these text messages (SMS) were sent by people in Haiti in the aftermath of the January 2010 earthquake. Our objective in this featured task is to translate from Haitian-Créole into English either using raw or clean data.

We propose to build an SMT system which could be used for both raw and clean data. Our baseline system is an standard phrase-based SMT system built with Moses (Koehn et al., 2007). Starting from this system we propose to introduce a semantic feature function based on latent semantic indexing (Landauer et al., 1998). Additionally, as a total different approximation, we propose to orthogonalize the standard feature functions of the phrase-based table using the Gram-Schmidt methodology (Greub, 1975). Then, we experimentally combine both enhancements.

The only difference among the raw and clean SMT system were the training sentences. In order to translate the clean data, we propose to normalize the corpus of short messages given very scarce resources. We only count with a small set of parallel corpus at the level of sentence of chat and standard language. A nice normalization methodology can allow to make the task of communication easier. We propose a statistical normalization technique using the scarce resources we have based on a combination of statistical machine translation techniques.

The rest of this paper is organized as follows. Section 2 briefly describes the phrase-based SMT system which is used as a reference system. Next, section 3 describes our approximation to introduce semantics in the baseline system. Section 4 reports our idea of orthogonalizing the feature functions in the translation table. Section 5 details the data processing and the data conversion from raw to clean. As follows, section 6 shows the translation results. Finally, section 7 reports most relevant conclusions of this work.



*Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 452–456,
Edinburgh, Scotland, UK, July 30–31, 2011. 

## 2 Phrase-based SMT baseline system

The phrase-based approach to SMT performs the translation splitting the source sentence in segments and assigning to each segment a bilingual phrase from a phrase-table. Bilingual phrases are translation units that contain source words and target words, e.g. *unité de traduction — translation unit*, and have different scores associated to them. These bilingual phrases are then selected in order to maximize a linear combination of feature functions. Such strategy is known as the log-linear model (Och, 2003) and it is formally defined as:

$$\hat{e} = \arg\max_{e} \left[ \sum_{m=1}^{M} \lambda_m h_m (e, f) \right] \qquad (1)$$

where $h_m$ are different feature functions with weights $\lambda_m$. The two main feature functions are the translation model (TM) and the target language model (LM). Additional models include lexical weights, phrase and word penalty and reordering.

## 3 Semantic feature function

Source context information is generally disregarded in phrase-based systems given that all training sentences contribute equally to the final translation. The main objective in this section is to motivate the use of a semantic feature function we have recently proposed (Banchs and Costa-jussà, 2011) for incorporating source context information into the phrase-based statistical machine translation framework. Such a feature is based on the use of a similarity metric for assessing the degree of similarity between the sentences to be translated and the sentences in the original training dataset.

The measured similarity is used to favour those translation units that have been extracted from training sentences that are similar to the current sentence to be translated and to penalize those translation units than have been extracted from unrelated or dissimilar training sentences. In the proposed feature, sentence similarity is measured by means of the cosine distance in a reduced dimension vector-space model, which is constructed by using Latent Semantic Indexing (Landauer et al., 1998), a well know dimensionality reduction technique that is based on the singular value decomposition of a matrix (Golub and Kahan, 1965).

The main motivation of this semantic feature is the fact that source context information is actually helpful for disambiguating the sense of a given word during the translation process. Consider for instance the Spanish word *banco* which can be translated into English as either *bank* or *bench* depending on the specific context it occurs. By comparing a given input sentence containing the Spanish word *banco* with all training sentences from which phrases including this word where extracted, we can figure out which is the most appropriated sense for this word in the given sentence. This is because for the sense *bank* the Spanish word *banco* will be more like to co-occur with words such as *dinero* (money), *cuenta* (account), *intereses* (interest), etc., while for the sense *bench* it would be more likely to co-occur with words such as *plaza* (square), *parque (park)*, *mesa (table)*, etc; and the chances are high for such disambiguating words to appear in one or more of the training sentences from which bilingual phrases containing banco has been extracted.

In the particular case of translation tasks where multi-domain corpora is used for training machine translation systems, such as the Haitian-Creole-to-English task considered here, the proposed semantic feature has proven to contribute to a better lexical selection during the decoding process. However, in tasks considering mono-domain corpora the semantic feature does not improves translation quality as the most frequent translation pairs learned by the system are actually the correct ones.

Another important issue related to the semantic feature discussed here is that it is a dynamic feature in the sense that it is computed for each potential translation unit according to the current input sentence being translated. This makes the implementation of this semantic feature very expensive from a computational point of view. At this moment, we do not have an efficient implementation, which makes it unfeasible in the practice to apply this methodology to large training corpora.

As the training corpus available for the Haitian-Creole-to-English is both small in size and multi-domain in nature, it constitutes the perfect scenario for experimenting with the recently proposed source context semantic feature. For more details about im-

plementation and performance of this methodology in a different translation task, the reader should refer to (Banchs and Costa-jussà, 2011).

## 4 Heuristic enhacement

The phrase-based SMT baseline system contains, by default, 5 feature functions which are the conditional and posterior probabilities, the direct and indirect lexical scores and the phrase penalty. Usually, these feature functions are not statistical independent from each other. Based on the analogy between the statistical and geometrical concepts of independence and orthogonality, and given that, during MERT, the optimization of feature combination is conducted on log-probability space; we decided to explore the effect of using a set of orthogonal features during MERT optimization.

It is well know in both spectral analysis and vector space decomposition that orthogonal bases allow for optimal representations of signals and variables, as they allow for each individual natural component to be represented independently of the others. In linear lattice predictors, for instance, each filter coefficient can be optimized independently from the others while convergence to the optimal solution is guarantied (Haykin, 1996). In the case of statistical machine translation, the linear nature of feature combination in log-probability space suggested us that transforming the features into a set of orthogonal features could make MERT optimization more robust and efficient.

According to this, we used Gram-Schmidt (Greub, 1975) to transform all available feature functions into an orthogonal set of feature functions. This orthogonalization process was conducted directly over the log-probability space, i.e, given the five vectors representing the feature functions $h_1, h_2, h_3, h_4, h_5$, we used the Gram-Schmidt algorithm to construct an orthogonal basis $v_1, v_2, v_3, v_4, v_5$. The resulting set of features consisted of 5 vectors that form an orthogonal basis. This new orthogonal set of features was used for MERT optimization and decoding.

## 5 Experimental framework

In this section we report the details of the used data preprocessing and raw to clean data conversion.

### 5.1 Data preprocessing

The WMT evaluation provided a high variety of data. Our preprocessing consisted of the following:

- Lowercase and tokenize all files using the scripts from Moses.
- In the case of the haitian-Creole side of the data, replace all stressed vowels by their plain forms.
- Filter out those sentences which had no words or more than 120.

Table 1 shows the data statistics of the different sources before and after this preprocessing. The different sources of the table include: in-domain SMS data (SMS); medical domain (medical); newswire domain (newswire); united nations (un); state department (state depart.); guidelines for approapriate international disaster donations (guidelines); krengle senetences (krengle) and a glossary includes wikipedia name entities and haitisurf dictionary. The sources of this material are specified in the web page of the workshop.

All data from table 1 was concatenated and used as training corpus. The English part of this data was used to build the language model. As development and test corpus we used the data provided by the organization. Both development and test contained 900 sentences.

Finally, in the evaluation, we included development and tests as part of the training corpus, and then, we translated the evaluation set.

### 5.2 Raw to clean data conversion

This featured task contained two subtasks. One was to translate raw data and the other was to translate clean data. Therefore, we have to build two systems. Our raw data system was built using the training data from table 1. The clean data system was built using all training data from table 1 except in-domain SMS data. Particularly, a modified version of the in-domain SMS data was included in the clean data system. The modification consisted in cleaning the original in-domain SMS data using an standard Moses SMT system. We built an SMT system to translate from raw data to clean data. This SMT system was built with the development, test and evaluation data which in total were 2700 sentences. We

| | | Statistics | |
|---|---|---|---|
| | | before | after |
| SMS | sentences | 17,192 | 16,594 |
| | words | 386.0k | 383.0k |
| medical | sentences | 1,619 | 1,619 |
| | words | 10.4k | 10.4k |
| newswire | sentences | 13,517 | 13,508 |
| | words | 326.9k | 326.7k |
| wikipedia | sentences | 8,476 | 8,476 |
| | words | 113.9k | 113.9k |
| un | sentences | 91 | 91 |
| | words | 1,906 | 1,906 |
| state depart. | sentences | 56 | 14 |
| | words | 450 | 355 |
| guidelines | sentences | 60 | 9 |
| | words | 795 | 206 |
| krengle | sentences | 658 | 655 |
| | words | 4.2k | 4.2k |
| bible | sentences | 30,715 | 30,677 |
| | words | 946k | 944k |
| glossary | sentences | 49,990 | 49,980 |
| | words | 126.4k | 126.3k |

Table 1: Data Statistics before and after training preprocessing. Number of words are from the English side.

used 2500 sentences as training data and 200 sentences for development to adjust weights. The raw and clean systems were tuned with their respective developments and tested on their respective tests.

## 6 Experimental results

In this section we report the results of the approaches proposed in previous sections. Table 2 and 3 report the results on the development and test sets on the raw and clean subtask, respectively.

First row on both tables report the results of the baseline system briefly described in section 2. Second row and third row on both tables report the performance of the semantic feature function and on the heuristic approach of orthogonalization (orthofeatures) respectively. Finally, the last row on both tables report the performance of both semantic and heuristic features when combined.

Results shown in tables 2 and 3 do not show coherent improvements when introducing the new

| System | Dev | Test |
|---|---|---|
| baseline | 32.00 | **31.01** |
| +semanticfeature | **32.34** | 30.68 |
| +orthofeatures | 31.63 | 29.90 |
| +semanticfeature+orthofeatures | 32.21 | 30.34 |

Table 2: BLEU results for the raw data. Best results in bold.

| System | Dev | Test |
|---|---|---|
| baseline | 35.86 | 33.78 |
| +semanticfeature | 35.98 | 33.90 |
| +orthofeatures | 35.57 | **34.10** |
| +semanticfeature+orthofeatures | **36.28** | 33.53 |

Table 3: BLEU results for the clean data. Best results in bold.

methodologies proposed. The clean data seems to benefit from the semantic features and the orthofeatures separately. However, the raw data seems not to benefit from the orthofeatures and keep the similar performance to the baseline system when using the semantic feature. Although, this trend is clear, the results are not conclusive. Therefore, we decided to participate in the evaluation with the full system (including the semantic features and orthofeatures) in the clean track and with the system including the semantic feature in the raw track. Actually, we used those systems that performed best in the development set. Additionally, results with the semantic feature may not be significantly better than the baseline system, but we have seen it actually heps to improve lexical selection in practice in previous works (Banchs and Costa-jussà, 2011).

## 7 Conclusions

This paper reports the BM-I2R system description in the Haitian-Créole to English translation task. This system was ranked first in the WMT 2011 by the conducted human evaluation. The translation system uses a PBSMT system enhanced with two different methodologies. First, we experiment with the introduction of a semantic feature which is capable of introducing source context information. Second, we propose to transform the five standard feature functions used in the translation model of the PBSMT system into five orthogonal feature func-

tions using the Gram-Schmidt methodology. Results show that the first methodology can be used for both raw and clean data. Whereas the second seems to only benefit clean data.

## Acknowledgments

The research leading to these results has received funding from the Spanish Ministry of Science and Innovation through the Juan de la Cierva fellowship program. The authors would like to thank Barcelona Media Innovation Center and Institute for Infocomm Research for their support and permission to publish this research.

## References

R. Banchs and M.R. Costa-jussà. 2011. A semantic feature for statistical machine translation. In *5th Workshop on Syntax, Semantics and Structure in Statistical Translation (at ACL HLT 2011)*, Portland.

G. H. Golub and W. Kahan. 1965. Calculating the singular values and pseudo-inverse of a matrix. journal of the society for industrial and applied mathematics. In *Numerical Analysis 2(2)*, pages 205–224.

W. Greub. 1975. *Linear Algebra*. Springer.

S. Haykin. 1996. *Adaptive Filter Theory*. Prentice Hall.

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 *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic, June.

T. K. Landauer, D. Laham, and P. Foltz. 1998. Learning human-like knowledge by singular value decomposition: A progress report. In *Conference on Advances in Neural Information Processing Systems*, pages 45–51, Denver.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, July.