THE AVIVAVOZ PHRASE-BASED STATISTICAL MACHINE TRANSLATION SYSTEM FOR ALBAYZIN 2008

Carlos A. Henríquez Q.[1], *Maxim Khalilov*[1],
José B. Mariño[1], *Nerea Ezeiza*[2]

[1]Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
{carloshq|khalilov|canton}@gps.tsc.upc.edu

[2]Department of Language and Computer Systems
University of the Basque Country
n.ezeiza@ehu.es

ABSTRACT

This paper describe the SMT system developed by the TALP Research Center of the UPC (Universitat Politècnica de Catalunya) for the Albayzin 2008 evaluation campaign (Spanish to Basque translation task). Apart from a standard set of feature models, the system introduces two target language models: one based on lemmas and the second based on linguistic classes (Part-of-Speech). The word-to-word alignment was obtained using a segmented version of the Basque corpus. The results obtained over the development and test sets are analyzed and discussed.

1. INTRODUCTION

Nowadays, the available bilingual material for automatic translation between Spanish and Basque is limited. Aiming to overcome this issue, we present our first approach to phrase-based Statistical Machine Translation (SMT) for this pair of languages trying to reach acceptable translation quality under conditions of smaller training material. The developed SMT system uses a POS language model and a lemmatize language model for Basque in order to improve the translations obtained by a baseline system. System configuration is discussed and the results obtained with the provided parallel corpus are presented.

This paper is organized as follow. Section 2 gives a brief description of the phrase-based model that the system is based on. Section 3 describes the corpus statistics, alignment procedure, features models and the case restoration method. Section 4 reports the main results of our system performance during development and testing. Finally section 5 sums up the main conclusions from our institution's participation in the evaluation.

This work has been funded by the Spanish Government under grant TEC2006-13694-C03 (AVIVAVOZ project).

2. PHRASE-BASED SMT SYSTEM

The phrase-based translation system[4] implements a log-linear model in which a foreign language sentence $f^J = f_1, f_2, \ldots, f_J$ is translated into another language sentence $e^I = e_1, e_2, \ldots, e_I$ by searching for the translation hypothesis $\hat{e}^I$ maximizing a log-linear combination of several feature models[2]:

$$\hat{e}_I = \arg\max_{e^I}\{\sum_{m=1}^{M} \lambda_m h_m(e^I, f^J)\}$$

Where the feature function h_m refers to the system models and λ_m refers to the corresponding optimized model weights.

The main system models are the translation model and the language model. The first one deals with the issue of which target language phrase f_j translates a source language phrase e_i and the latter model estimates the probability of translation hypothesis. Apart from these two models, a set of additional models was used in the AVIVAVOZ system. They are presented in section 3.3.

The development of the AVIVAVOZ SMT system is based on the MOSES toolkit[3].

3. ALBAYZIN 2008 EVALUATION FRAMEWORK

3.1. Corpus

For the system design, the corpus used was the one provided for the evaluation campaign. It is a set of 61104 sentences, divided in three subsets: a training corpus containing 58202 sentences, a development corpus of 1456 sentences and a test corpus with the remaining 1446 sentences. Only one reference for each set was supplied. The basic corpus statistics can be found in table 1.

The tokenization, Part-of-Speech (POS) tags and lemmatization for both languages was also provided altough only the Basque information was used during development. According to the campaign documentation, the POS tags for Basque were obtained with the Eustagger[1] tool and only the category and subcategory of each tag were provided.

An additional preprocessing consisted in changing the text encoding from ISO88591 to UTF8, removing all the sentences bigger than 100 words and lowercasing the entire corpus. The removing step was performed only over the train corpus due to a restriction implied by the alignment tool. The removed set was smaller than 1% of the training corpus.

3.2. Alignment

Altough the baseline system used the lowercased corpus to obtain the word-to-word alignment, the final system used a different approach.

A segmentation tool was developed which splitted the Basque words using the POS information and a suffix dictionary; wherever a verb, an adjective or a name was found, the word was checked with the dictionary, and if the word ended with any of the listed suffix, it was splitted in two e.g. "publikoen", which is a Basque adjective, ended with the suffix "en" (listed in the dictionary), therefore it was splitted into "publiko+ +en".

With the segmentation tool, the Basque lowercase corpus was segmented and the alignment was computed on this corpus. Once the alignment was completed, and using a developed desegmentation tool which joins the words with their splitted suffixes, the corpus was desegmented to its original version and the links were properly relocated to the original words.

The Spanish lowercased corpus remained the same during the process. The alignment was automatically computed by the GIZA++[5] toolkit.

3.3. Features

The SMT system developed uses the following feature functions during translation:

- Phrase translation probability on both directions, based on a joint probability model[4].
- Lexical weighting on both directions, based on word-to-word IBM Model 1 probabilities[6].
- Phrase penalty features which compensate the system's preference for short output sentences.
- Target language model of order 5.
- POS target language model of order 7.
- Lemma target language model of order 7.

	Baseline	POS LM	Seg. Align	Lemma LM
BLEU	12.66	12.76	13.01	13.26
NIST	4.77	4.72	4.80	4.90
WER	77.90	78.38	78.61	77.61
PER	60.15	60.61	60.36	59.37

Table 2. Results obtained on the development set

	Baseline	POS LM	Seg. Align	Lemma LM
BLEU	11.43	11.68	11.89	11.95
NIST	4.68	4.66	4.65	4.74
WER	78.74	78.71	79.11	78.50
PER	60.38	60.64	60.65	59.93

Table 3. Results obtained on the test set

Beeing the first four the features used in our baseline system. The reordering model is based on lexicalized reordering[8] in all the performed experiments. This distortion model takes into account the relative movement between a given phrase and its adjacent phrases.

3.4. Case Restoration

Because all the design used a lowercased corpus, a final case-restoration tool is needed to stablish a truecase translation. For this matter, the *disambig* and *ngram-count* tools from the SRI Language Model toolkit[7] were used.

4. EXPERIMENTS AND RESULTS

For the Spanish-Basque task, four different systems were developed in a progressive fashion. The first one, called the *baseline*, has the default feature functions and parameters of MOSES. From that starting point, a *POS target language model* was add to the SMT system. In the third system we performed a modified *alignment*, which consisted in performing the alignment with a Basque segmented corpus as commented in section 3.2.

The final system also included a *target language model based on lemmas*. As mentioned in section 3.1, the lemmas for Basque were computed automatically and were provided by the organizers of the evaluation campaign. Both language models (of the POSs and the lemmas) were 7-gram models and the order of the surface words language model was 5. The maximum phrase size was set to 5 for all the systems.

Table 2 and 3 show the different results obtained with the systems developed. Each column corresponds to a different system, starting with the baseline and ending with the system that was submitted to the evaluation. It can be seen that the addition of the different features resulted in final improvement of 0.5% BLEU points over the test set and 0.6% over de development set.

	Train Corpus		Devel. Corpus		Test Corpus	
	Spanish	Basque	Spanish	Basque	Spanish	Basque
Number of sent	58202		1456		1446	
Max. sent size	242	236	93	76	317	232
Avg. sent size	19.77	15.20	23.33	18.70	22.32	17.85
Vocab size	97558	140931	7170	9031	6926	8691

Table 1. Basic corpus statistics

5. CONCLUSIONS

In this paper we introduced the AVIVAVOZ phrase-based SMT system which partipated in the Albayzin 2008 evaluation campaign. Starting with a brief introduction to the phrase-based statistical translation modelling, the corpus and preprocessing description were presented. Further, we described the set of feature models that were taken into account for the design of the system and the results obtained with different systems configurations.

The main conclusion which can be drawn from the results is that despite the additional features were useful, a different approach better dealing with global word reordering and the agglomerative characteristic of Basque is needed to obtain an improved system performance.

6. REFERENCES

[1] Itziar Aduriz, Nerea Ezeiza, and Ruben Urizar. Euslem: A lemmatiser/tagger for basque. pages 17–26, Göteborg. Göteborg University, Department of English, 1996.

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics.*, 16(2):79–85, 1990.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, 2007.

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL*, pages 48–54, 2003.

[5] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[6] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[7] A. Stolcke. Srilm - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*.

[8] Christoph Tillman. A block orientation model for statistical machine translation. In *HLT-NAACL*, 2004.