LANGUAGE MODELING FOR VERBATIM TRANSLATION TASK

Maxim Khalilov and José A.R. Fonollosa

TALP Research Center, Universitat Pollitècnica de Catalunya, Barcelona

ABSTRACT

In this paper we present the first results towards finding the better TC-STAR[1] 2006 verbatim transcription system configuration by means of improving the quality of language model performance.
There is a present lack of research devoted to special techniques of verbatim translation, therefore we have made an attempt to improve translation accuracy by combining the Final Text Edition (FTE) system with supplementary verbatim corpus. Our work was focused on finding the best combination of the baseline (FTE) and verbatim language models for Spanish-English and English-Spanish language pairs. In order to improve the overall system performance standard n-gram based statistical machine translation (SMT) system was supplemented with a log linear combination of some additional feature functions and linguistically motivated word reordering technique.
In the final part of the study we report the results of the baseline system translation accuracy in comparison with the FTE-verbatim interpolated language model systems for various proportions of the language models linear combination.

1. INTRODUCTION

Statistical approach to the machine translation is comparatively recent area of scientific study, however, due to its promising potential it is growing relatively fast during last years.

A great advance in terms of translation system accuracy was done moving from the first systems based on the noisy channel approach model realizing word-to-word translation [1] to the phrase-based systems dealing with aligned bilingual corpora and implementing translation of the bilingual units [2,3]. However, this change entails correlation of the bilingual translation model, as well as additional language model, to the morphological nature of the corpus. Verbatim language peculiarities, such as non-grammatical or bad structured input, are not duly reflected in the regular translation systems. Hence, applying bilingual and language models properly calculated in accordance with semantically and grammatically corrected phrase alignment to the spontaneous speech input data, may cause translation errors. Because of the common point of view that SMT systems are robust to the non-grammatical input data, the majority of the up to date developed text-to-text translation systems address the problem of the pre-edited text translation, whereas systems of recognized speech translation like translation of the automatic speech recognizer output were set aside.

In this work we try to adapt the classical FTE system to the verbatim task by optimizing and adapting target language model which is an integrated part of the statistical translation system.

The paper is organized as follows. In Section 2 the n-gram translation model is briefly outlined, as well as additional models. Section 3 describes the verbatim task specific character and framework of the TC-STAR 2006 evaluation (www.elda.org/en/proj/tcstar-wp4). The performed experiments, results and discussions are described in the Section 4, while conclusions and future work are detailed in the Section 5.

2. STATISTICAL MACHINE TRANSLATION

As already mentioned, first SMT models were based on the noisy-channel paradigm [1], modeling translation of the target language sequence as argmax operation as described by the following equation:

$$\hat{e} = \arg\max_{e} P(e)\, P(f \mid e) \qquad (1)$$

where e refers to the sequence in target language and f to the sequence in the source language, and, consequently, translation model probability $P(f|e)$ and target language model probability $P(e)$.

2.1. N-gram translation model

The noisy channel approach was applied to the word-to-word alignment, where both models are estimated independently. The translation accuracy has been improved by switching to the phrase-based approach, in parallel, the n-gram based approach has appeared, operating with bilingual units extracted from the aligned bilingual corpus, referred to as tuples [4]. Tuples are extracted from a word-to-word aligned corpus according to certain constraints [5].

[1] Technology and Corpora for Speech to Speech Translation (http://www.tc-star.org)

The tuple n-gram translation model determines joint probability of the source and target language units as shown in the equation (2):

$$p(e,f) = \prod_{k=1}^{K} p(t_k \mid t_{k-N+1}, \ldots, t_{k-1}) \qquad (2)$$

where t_k refers to the k-*th* tuple of the given bilingual sentence pair segmented to *K* tuples, *N* refers to n-gram order.

Given a bilingual corpora, GIZA++ toolkit was used to generate word-to-word alignments in source-to-target and target-to-source directions [6], then tuples are extracted from these alignments.

Recently, the source channel model has been supplemented by the maximum entropy approach [7], implementing the posterior probability *p(e|f)* definition as a log-linear combination of the set of feature functions [8]. Speech-to speech translation task and finite state transducer model underlie this approach [4,9]. This technique allows to simplify feature models combinations under the translation hypothesis determination, as described below (3):

$$\hat{e}^I = \arg\max_{e^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I) \right\} \qquad (3)$$

where the feature functions h_m refer to the system models as bilingual translation model, target language model, etc. and λ_m to weights corresponding to these models.

The weight coefficients normally are to be optimized to provide largest value of scoring function (BLEU score in the system under consideration [10]).

2.2. Additional feature models

Following the maximum entropy approach, the baseline system implements the log-linear combination of the translation model and the following feature models:

2.2.1. Target language model

The first additional feature is not obligatory for the systems based on n-gram approach in contrast to word-based systems is used to improve system translation quality. The model is computed according to the following equation:

$$P_{LM}(t_k) \approx \prod_{n=1}^{K} p(w_n \mid w_{n-N+1}, \ldots, w_{n-1}) \qquad (4)$$

where t_k refers to the partial translation hypothesis, w_n to the n-*th* word in this partially translated sentence.

2.2.2. Word penalty model

This feature was implemented to compensate system's striving for short output sentences. This phenomenon is due to the target language model. Technically, the penalization depends on the total number of words contained in the partial translation hypothesis, and can be found as follows:

$$P_{WP}(t_k) = \exp(number\ of\ words\ in\ t_k) \qquad (5)$$

2.2.3. Source-to-target lexicon model

This model uses word-to-word IBM model 1 probabilities [11] to estimate lexical weights of each tuple, as follows:

$$P_{IBM1}((e,f)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p\left(e_n^i \mid f_n^i\right) \qquad (6)$$

where f_n^j and e_n^i are the *j-th* and *i-th* words in the source and target parts of the tuple $(e,f)_n$, being *J* and *I* the corresponding total number words in each side of it. Giza++ word-to-word source-to-target alignment was used.

2.2.4. Target-to-source lexicon model

The same as the previous model but for opposite direction, backward lexicon model. Giza++ word-to-word target-to-source alignment was used.

2.2.5. POS target language model

The feature represents an n-gram model of the POS tags training corpus. English POS tagger TNT [12] and for Spanish Freeling [13] were used for monolingual corpora tagging [14].

2.3. Reordering

For Spanish to English a linguistically motivated word reordering technique was applied in order to decrease the number of errors caused by the difference in word order between two languages. The idea of the algorithm is to detect blocks of alignments which, if swapped, produce a monotone translation, then to classify the alignments blocks to groups and identify the consecutive blocks in the source corpora to swap them. Detailed description of the reordering procedure can be found in [15, 16]. The technique was applied to all the source corpora and the reordered training set was realigned to build the final translation system.

2.4. Decoding

MARIE decoder was used as a search engine for the translation system, the details can be found in [17]. The decoder implements a beam-search algorithm with

pruning capabilities. Versions starting with 1.2 support POS tags target language models, thereby all five additional feature models described above were taken into account.

2.5. Optimization

Optimization of the weight coefficients of the scoring function is based on the simplex optimization method [18]. Given development set and references, the log-linear combination of the weights is adjusted to maximize score function (see eq. 3) according to the highest BLEU score. Experiments on log-linear combination of BLEU and NIST scores are planned for the future work.

3. TC-STAR EVALUATION FRAMEWORK

Translations provided by the baseline system has been evaluated in the framework of the TC-STAR 2006 evaluation campaign. Our institution participated in the Spanish-English and Spanish-English language directions of FTE, verbatim and ASR (Automatic Speech Recognition output) shared tasks.

3.1. Task description

Verbatim transcription includes spontaneous speech phenomena (hesitations, false-starts, half-words, etc). FTE is a manually corrected text slightly different from the verbatim text. TC-STAR 2006 verbatim evaluation was performed in true case and with punctuation marks.

The difference between FTE and verbatim texts can be seen from the example below:

FTE**:** *I am starting to know what Frank Sinatra must have felt like*
Verbatim: *I'm I'm I'm starting to know what Frank Sinatra must have felt like*

3.2. Corpora

The data provided for shared tasks are European Parliament Plenary Sessions (EPPS) database for English-Spanish and Spanish-English language pairs.

Additionally, monolingual EPPS Verbatim Transcription corpora of the EPPS was used for English and Spanish specific verbatim language modeling. Development and test sets have 2 references for both languages. First 500 phrases of the provided official development corpora for both direction have been used for to maximize score function.

Basic corpora statistic can be found in the Tables 1 and 2. Training corpora are unique for all the tasks, development and test corpora vary.

EPPS	Spanish	English
Training set		
Sentences	1.3 M	1.3 M
Words	36.57 M	34.9 M
Vocabulary	153 K	107 K
Development set		
Sentences	500	500
Words	15 K	12 K
Vocabulary	2.5 K	2.3 K
Test set		
Sentences	699	1.155
Words	31 K	30 K
Vocabulary	3.9 K	4 K

Table 1. *EPPS corpora (M stands for millions, K stands for thousands).*

EPPS-Vbt	Spanish	English
Training set		
Sentences	70 K	73 K
Words	512 K	781 K
Vocabulary	20 K	17 K

Table 2. *EPPS-Verbatim corpora.*

4. EXPERIMENTS AND RESULTS

The mixing procedure was implemented with *n-gram* tool from the SRI Language Modeling toolkit [19], allowing to read the second model for interpolation purposes and adjust weight coefficients when interpolating with the main model. We considered six alternative configurations for each target language. They include various weights of the minor verbatim model from 0 (FTE model only) to 1 (verbatim model only) with step 0.2.

The BLEU scores obtained on the development corpus as a result of the simplex optimization procedure (refer to Dev) were investigated. Tables 3 and 4 represent BLEU score of Spanish-English and English-Spanish translation experiments under the baseline system and systems using interpolated language models. Graphical representation of the obtained results can be found in the Figures 1 and 2.

Maximizing this value, the best performance point is provided by the system corresponding to 0.1 FTE – 0.9 VBT configuration for Spanish to English translation task and to 0.9 FTE – 0.1 VBT for English to Spanish task. It allows gaining 0.47 BLEU points on the Test for Spanish to English translation and 0.27 BLEU points for English to Spanish task comparing to the baseline system.

Moreover, we investigated the behavior of the BLEU score obtained on the Test corpus considering the same system configurations (Tables 5 and 6) and optimized models weights (refer to Test). Results also can be found in the Figures 3 and 4. It can be seen that despite absolute maxima achieved on the Test and Dev vary, the shapes of the curves are similar.

In the framework reported in the paper, we did not conduct any experiments focused on the translation model training based on the interpolated models, which can be an important point of investigation in future.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented a brief study of the possible ways of verbatim texts translation improvement by means of applying modified language model to the translation system. The baseline translation system, represents translation system used for TC-STAR 2006 evaluation and implements some new features, namely additional tagged target language model and lexically motivated reordering technique.

It is seen from the reported results that the described technique works for both of the considered languages and corresponding translation tasks, for English to Spanish and Spanish to English translations. Despite the slight difference on the translation accuracy was observed, the smoothed shapes of the translation BLEU score curves for Test and Dev corpora are similar.

Additional verbatim language model can be also included as a feature to the log-linear combination of the set of functions with a corresponding weight, as shown in (3).

Further work includes applying of new additional feature optimization, namely the algorithm adjusting weight coefficients so that to maximize a log-linear combination of NIST and BLEU over the development set and to apply the described technique in combination with phrase-based approach [20] which could imply further BLEU improvement due to different algorithm of taking advantage of each feature including the target language model.

6. ACKNOWLEDGMENTS

This work has been partially funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation – (IST-2002-FP6-506738, http://www.tc-star.org) and by the Spanish Government under grant AP20042835 (FPU grant).

7. REFERENCES

[1] P. Brown, S. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. “The mathematics of statistical machine translation: parameters estimation”. Computational Linguistics, 19(2):263-311.

[2] P. Koehn, F. J. Och, and D. Marcu. 2003. “Statistical phrase-based translation”. Proc. of the 2003 Meeting of the North American chapter of the ACL, Edmonton, Alberta.

[3] R. Zens, F. J. Och, and H. Ney, “Phrase-based statistical machine translation” in KI – 2002: Advanced in artificial intelligence, M. Jarke, J. Koehler, and G. Lakemeyer, Eds. Springer Verlag, September 2002, vol. LNAI 2479, pp. 18-32.

[4] A. de Gispert, and J. B. Mariño. 2002. “Using X-grams for speech-to-speech translation”. Proc. Of the 7th Int. Conf. on Spoken Language Processing.

[5] J. M. Crego, J. B. Mariño, A. de Gispert. 2004. “Finite-state-based and phrase-based statistical machine translation”. Proc. Of the 8th Int. Conf. on Spoken Language Processing, :37-40.

[6] F. J. Och and H. Ney. 2000, “Improved statistical alignment models”. Proc. of the 38th Ann. Meeting of the ACL, Hong Kong, China, October.

[7] A. Berger, S. Della Pietra, and V. Della Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics, vol. 22, no.1, pp. 39-72, March 1996

[8] F. J. Och and H. Ney. 2000, “Improved statistical alignment models”. Proc. of the 38th Ann. Meeting of the ACL, Hong Kong, China, October.

[9] E. Vidal, “Finite-state speech-to-speech translation”, Proc. of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 111-114, 1997.

[10] K. Papineni, S. Roukos, T. Ward, and Wei Jing Zhu. 2002. “Bleu: a method for automatic evaluation of machine translation”. Proc. of the 40th Ann. Conf. of the ACL, Philadelphia, PA, July.

[11] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Zhen Jin, and D. Radev, “A smorgasbord of features for statistical machine translation”, Proc. of the Human Language Technology Conference, HLT-NAACL’ 2004, pp. 161-168, May, 2004.

[12] T. Brants. 2000. “Tnt – a statistical part-of-speech tagger” Proceedings of the Sixth Applied Natural Language Processing.

[13] X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. “Freeling: An open-source suite of language analyzers”. 4th Int. Conf. on Language Resources and Evaluation, LREC’04.

[14] M. Popović and H. Ney. “POS-based Word Reorderings for Statistical Machine Translation”. 5th International Conference on Language Resources and Evaluation (LREC), pages - to appear, Genova, Italy, May 2006.

[15] M. R. Costa-jussà, J. M. Crego, A. de Gispert, P. Lambert., M. Khalilov, R. E. Branchs, J. B. Mariño, J. A. R. Fonollosa. 2006. “TALP Phrase-based statistical translation system for European language pairs”. NAACL 2006 Workshop on Statistical Machine Translation. Pages – to appear. 2006

[16] M. R. Costa-jussà and J. A. R. Fonollosa. 2006. “Statistical Machine Reordering”. Proceedings of EMNLP 2006. pp. 70-76.

[17] J. M. Crego, J. B. Mariño, and A. de Gispert. 2005. "An Ngram-based statistical machine translation decoder". Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'05.

[18] J.A. Nelder and R. Mead. 1965. "A simplex method for function minimization". The Computer Journal, 7:308-313

[19] A. Sholcke. 2002. "SRILM: an extensible language modelling toolkit". Proc. of the Int. Conf. on Spoken Language Processing: 901-904, Denver, CO, September.

[20] R. Zens, F. J. Och, and H. Ney, "Phrase-based statistical machine translation" in KI – 2002: Advanced in artificial intelligence, M. Jarke, J. Koehler, and G. Lakemeyer, Eds. Springer Verlag, September 2002, vol. LNAI 2479, pp. 18-32.

FTE	Vbt	BLEU Dev
Baseline (1.0 FTE-0.0 Vbt)		50.53
0.9	0.1	50.66
0.8	0.2	50.84
0.5	0.5	50.76
0.2	0.8	50.92
0.1	0.9	50.99
0.0	1.0	50.97

Table 3. *Results of Spa-to-Eng translation (Dev).*

FTE	Vbt	BLEU Dev
Baseline (1.0 FTE-0.0 Vbt)		44.05
0.9	0.1	44.39
0.8	0.2	44.01
0.5	0.5	43.90
0.2	0.8	43.52
0.1	0.9	43.17
0.0	1.0	42.03

Table 4. *Results of Eng-to-Spa translation (Dev).*

FTE	Vbt	BLEU Test
Baseline (1.0 FTE-0.0 Vbt)		52.24
0.9	0.1	52.49
0.8	0.2	52.83
0.5	0.5	52.81
0.2	0.8	52.87
0.1	0.9	52.72
0.0	1.0	52.63

Table 5. *Results of Spa-to-Eng translation (Test).*

FTE	Vbt	BLEU Test
Baseline (1.0 FTE-0.0 Vbt)		44.19
0.9	0.1	44.46
0.8	0.2	44.48
0.5	0.5	44.75
0.2	0.8	44.06
0.1	0.9	43.62
0.0	1.0	42.71

Table 6. *Results of Eng-to-Spa translation (Test).*

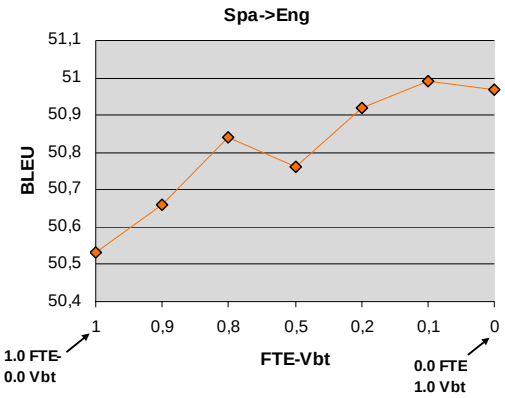

Figure 1. *Results of Spa to Eng translation (Dev).*

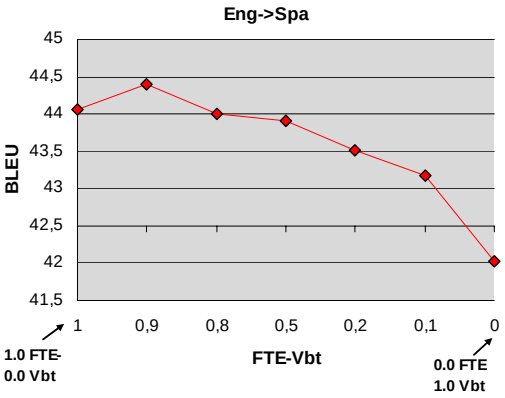

Figure 2. *Results of Eng to Spa translation (Dev).*

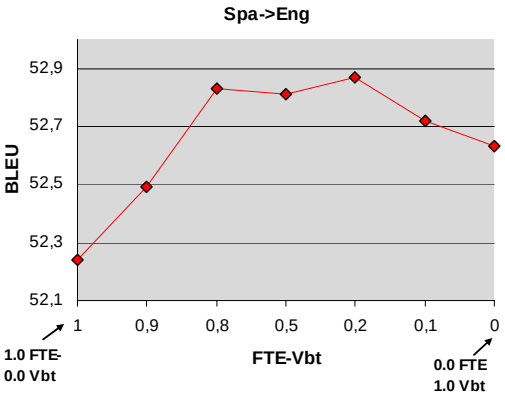

Figure 3. *Results of Spa to Eng translation (Test).*

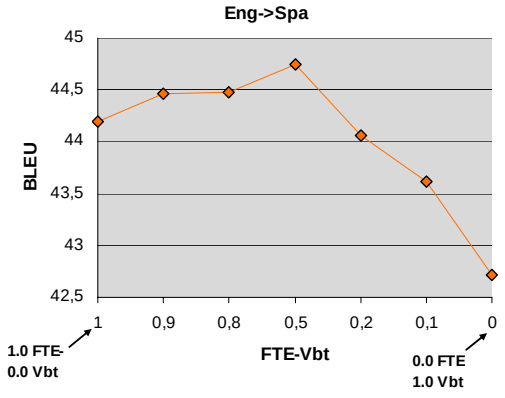

Figure 4. *Results of Eng to Spa translation (Test).*