The IRST English-Spanish Translation System for European Parliament Speeches

Daniele Falavigna, Nicola Bertoldi, Fabio Brugnara, Roldano Cattoni, Mauro Cettolo
Boxing Chen, Marcello Federico, Diego Giuliani, Roberto Gretter, Deepa Gupta, Dino Seppi

Fondazione Bruno Kessler - IRST
I-38050 Povo (Trento), Italy

Abstract

This paper presents the spoken language translation system developed at FBK-irst during the TC-STAR project. The system integrates automatic speech recognition with machine translation through the use of confusion networks, which permit to represent a huge number of transcription hypotheses generated by the speech recognizer. Confusion networks are efficiently decoded by a statistical machine translation system which computes the most probable translation in the target language. This paper presents the whole architecture developed for the translation of political speeches held at the European Parliament, from English to Spanish and vice versa, and at the Spanish Parliament, from Spanish to English.

Index Terms: spoken language translation, automatic speech recognition, statistical machine translation.

1. Introduction

This paper describes our Spoken Language Translation (SLT) system [1] for the translation of political speeches recorded at the European and Spanish parliaments, from Spanish to English and vice versa. The system integrates state-of-the-art automatic speech recognition (ASR) and statistical machine translation (SMT) components through the use of confusion networks (CNs). CNs permit to represent a large number of transcription hypotheses, all provided with confidence scores. From the other side, CNs can be efficiently exploited by our SMT decoder, which searches the most probable translation along all possible transcription hypotheses contained in the CN.

Given an audio signal, the IRST SLT system computed the best translation through the following six steps: (i) speech segments are detected inside the audio signal; (ii) the ASR component computes for each speech segment a word-graph with multiple transcription hypotheses; (ii) the word-graph is transformed into a CN; (iii) punctuation information is inserted in the CN; (iv) the optimal translation is computed from the CN; (v) finally, case information is added to the translation.

The whole SLT system has been trained on both English and Spanish recordings of political speeches acquired during some European Parliament Plenary Sessions and in the Spanish Parliament. Translation has been performed in both directions: English to Spanish and Spanish to English.

The paper is organized as follows. Sections 2 and 3 present each processing step. Section 4 presents and comments experimental results obtained on the translation tasks of the 2007 TC-STAR Evaluation Campaign.

2. ASR Steps

2.1. Detection of Speech Segments

The audio signal is split into homogeneous non overlapping segments using an acoustic classifier, based on Gaussian Mixture Models (GMMs), followed by a segment clustering method based on the Bayesian Information Criterion (BIC) [1].

2.2. Speech Transcription

Detected speech segments are transcribed using the ASR system described below. The latter is formed by the following components: acoustic front-end, acoustic models, language models, pronunciation lexicon and decoding procedure.

2.2.1. Acoustic front-end

Acoustic observations for Hidden Markov Models (HMMs) consist of 13 Mel-frequency Cepstral Coefficients (MFCCs), including the zero order coefficient, computed every 10ms using a Hamming window of 20ms length. The filter-bank contains 24 triangular overlapping filters centered at frequencies between 125 and 6750 Hz.

Cluster-based Cepstral Mean and Variance Normalization (CMVN) is performed to ensure that for each segment cluster the 13 MFCCs have mean zero and variance one. First, second and third order time derivatives are computed after CMVN to form a 52-dimensional feature vector.

2.2.2. Acoustic models

Two sets of HMMs are trained and used in two different decoding steps.

In the first decoding step an environment normalization based on Constrained Maximum Likelihood Linear Regression (CMLLR) followed by Heteroscedastic Linear Discriminant Analysis (HLDA) projection were applied to acoustic observations as follows.

- A simple target model, that is a Gaussian mixture model (GMM) with 1024 components, was trained over the 52-dimensional acoustic observations.
- For each cluster of speech segments in the training data, a CMLLR transform [2] was estimated w.r.t. the target GMM.
- The CMLLR transforms were applied to the feature vectors. The resulting transformed/normalized feature vectors are supposed to contain less speaker, channel, and environment variabilities [3] than the corresponding non transformed vectors.

[1] This work was partially financed by the European Commission under the project TC-STAR - Technology and Corpora for Speech to Speech Translation Research (IST-2002-2.3.1.6).

- The HLDA transformation was estimated w.r.t. reference models. Reference models are triphone HMMs with a single Gaussian density per state, trained on normalized 52-dimensional acoustic observations [4].
- The HLDA transformation was applied to the normalized 52-dimensional vectors to obtain observation vectors with 39 components. These observation vectors are used to train HMMs employed in the first recognition step, as explained below.

A conventional Maximum Likelihood (ML) training procedure was used to initialize and train the HMMs used in the first recognition pass. These models are state-tied, cross-word, gender-independent triphone HMMs with diagonal covariance matrices. A phonetic decision tree was used for tying states and for defining the context-dependent allophones.

For the second decoding pass a different set of acoustic models was trained adopting the speaker adaptive training procedure described in [5]. More specifically, before performing the conventional ML training procedure, to reduce inter-speaker variability the following two passes were performed.

- For each cluster of speech segments in the training data, a CMLLR transform was estimated w.r.t. a set of target models. Target models are triphone HMMs with a single Gaussian density per state trained on normalized 39-dimensional observation vectors.
- The CMLLR transforms were applied to the feature vectors.

A set of state-tied, cross-word, gender-independent triphone HMMs with diagonal covariance matrices were estimated using the CMLLR transformed feature vectors. Similarly to HMMs used in the first decoding step a phonetic decision tree was used for tying states and for defining the context-dependent allophones.

It is worth noting that the same set of target models is used in both training and decoding stages to produce normalized acoustic features.

2.2.3. Decoder

The basic recognition process is based on two decoding stages, and is common to both English and Spanish systems.

A preliminary decoding pass is carried out with the first set of acoustic models on normalized, HLDA projected, 39-dimensional observation vectors. The preliminary transcriptions are exploited for adaptation/normalization purposes in the second decoding step.

Before the second decoding pass, cluster-based acoustic feature normalization is applied to normalized, HLDA projected, 39-dimensional observation vectors. For each cluster of speech segments, a CMLLR transform is estimated w.r.t. the set of target models used during training, then the CMLLR transform is applied to the feature vectors. The acoustic models used in the second decoding pass are also adapted to the cluster data before decoding. Means of Gaussian densities are adapted to the cluster data through the application of a number of simple ”offset” transformations estimated in the MLLR framework [6].

3. MT Steps

3.1. Extraction of Confusion Network

A word-graph contains several transcription alternatives considered during the ASR process, but its topology is very complex. A simpler and more compact way of representing these alternatves is achieved through a CN [7], also called as *sausage*. A CN is still a weighted directed graph with the peculiarity that each path from the start node to the end node goes through all the other nodes; words and posterior probabilities are associated to the graph edges.

The extraction of a CN from a word lattice is done by means of the `lattice-tool` by SRILM toolkit [8], after words are put in lowercase.

3.2. Punctuation Insertion

The ASR system does not provide punctuation information during recognition. In our system, punctuation is introduced by a procedure that enriches the input CN with possible punctuation marks computed by a statistical model (see companion paper).

3.3. Decoder

Since 2006 IRST has been contributing to the development of open source toolkit for SMT, called `moses` [9]. The `Moses` project started at a JHU Summer Workshop in 2006, and was jointly developed by several sites, including the University of Edinburgh, IRST, RWTH, University of Maryland, and MIT. The currently available release features a multi-stack, phrase-based, beam-search decoder able to process a CN as well as plain text.

`moses` implements a log-linear translation model including as feature functions: direct and inverted phrase-based and word-based lexicons, multiple word-based n-gram target language models, phrase and word penalties, and distance-based reordering model.

`moses` also includes facilities to train the bilingual lexicons given a word-aligned parallel corpus, and to optimize feature weights on a development set through a Minimum Error Rate training. `moses` is able to train, load and exploit very huge language models, through the exploitation of a software library developed at IRST [11].

Computational efficiency is obtained through pre-fetching and early recombining the translation alternatives of the source phrases. On-demand loading of lexicon and language models and quantization of language models [12] allows a big reduction of run-time memory usage.

A more detailed description of the decoder can be found in [13].

3.4. Capitalization

The final step of the translation process consists in the case restoration which is performed with the `disambig` tool by SRILM toolkit [8], fed with a n-gram case sensitive target language model.

4. Evaluation

We present performance achieved by our system on the benchmark provided for the TC-STAR 2007 Evaluation Campaign [14]. The task proposed in this evaluation consists in the translation from English to Spanish and from Spanish to English of speeches of EPPS and (only for the latter direction) of the Spanish Parliament (Cortes Generales). No distiction between EPPS and Cortes data were allowed.

The test sets consist of 3 and 6 hours of recordings in the English-to-Spanish and Spanish-to-English directions, respectively, covering the period June to September 2006. Two reference are available for both language directions.

Corpus	Description	English			Spanish		
		Words	Vocabulary	n-gram	Words	Vocabulary	n-gram
EPPS	Final Text Edition of EPPS	39M	116K	26M	40M	149K	26M
Parliaments	JRC-Acquis, EU-Bulletin, UN human transcription from EPPS and Cortes	150M	417M	39M	72M	317K	33M
GigaWord		1.8G	4.5M	289M	670M	1.7M	85M
Dev	dev data of 2005 and 2006 evaluations	320K	7.8K	201K	225K	9.3K	138K
LM1	news agencies	200M	49K	23M	80M	61K	5.4M
LM2	LM1 and Hansard corpus	674M	65K	27M			

Table 1: Statistics of the English and Spanish monolingual corpora exploited for the training. The number of running words, the size of the dictionary, and the number of the estimated n-gram probabilities are reported.

4.1. Training data

The specifications for the primary condition of the task imposes the use of a given English-Spanish parallel corpus, consisting of the Final Text Edition (FTE) of EPPS. The corpus contains a total of 37M Spanish and 36M English running words; Spanish and English dictionaries contain 143K and 110K words, respectively. No parallel data related to the Spanish parliaments were available.

As regards as monolingual resources any publicly available data were allowed for training both the ASR and SLT systems. Table 1 reports statistics on several English and Spanish corpora, respectively, used for training the ASR and SLT system. More details about these data can be found in the TC-STAR web page.

4.2. Training of the ASR module

4.2.1. Acoustic model training

The English audio training data set consists of about 301 hours of recordings: about 101h of them were transcribed, the remaining 200h are not transcribed. Similarly, the Spanish training audio corpus consists of about 285 hours of recordings: about 100h of them were transcribed, the remaining 185h are not transcribed. Untranscribed training data were transcribed automatically using early versions of the transcription systems.

English HMMs, for both decoding passes have about 9.4K tied states and about 300K Gaussian densities. Spanish HMMs have about 6.2K tied states and about 196K Gaussian densities.

4.2.2. English LM training

Two 4-gram LMs (LM1 and LM2) were trained for English, using the data reported in Table 2. In both cases, the resulting background LM was adapted to a text corpus consisting of the manual transcriptions of the EPPS audio data released for training of the acoustic models (about 0.8M words) plus texts, $\approx$4M words, corresponding to the EPPS FTE covering the same period of the acoustic training data.

Two pronunciation lexicon were adopted: USlex, generated by merging different source lexica for American English, and BEEPlex generated by exploiting the British English Example Pronunciations (BEEP).

The decoding network, used in the first decoding pass, is built exploiting the public 4-gram LM1 and the USlex: this results in a static decoding graph [15] with about 56M of states, 53M of labeled arcs and 88M empty arcs.

The decoding network, used in the second decoding pass, is built exploiting the public 4-gram LM2 and the BEEPlex: this results in a static decoding graph with about 81M of states, 79M of labeled arcs and 142M empty arcs.

4.2.3. Spanish LM training

For Spanish the same LM (denoted LM1 in Table 1) was exploited in both decoding passes. A 5-gram background LM was trained on the text data of the Spanish EPPS FTE, Spanish Parliament and parallel corpora. Similarly to English, the resulting background LM was then adapted with a 5-gram LM trained on the manual transcriptions of EPPS and Spanish Parliament audio data released for training the acoustic models (about 880K words) and 2005-2006 FTE corpora (about 3.8M words).

The pronunciations in the lexicon are based on a set of 31 phones. In addition, there is a model for silence and three models for filler words, breath and noises. The lexicon contains 61K words among those in EPPS domain. The phonetic transcriptions were automatically generated using a set of grapheme-to-phoneme rules for Spanish.

The 5-gram LM and the lexicon were used to build a static decoding graph with about 21M of states, 28M of labeled arcs and 34M of empty arcs.

4.3. Training of the SLT module

The parallel training corpus has been word-aligned symmetrically; 83M bilingual phrase pairs (48M Spanish and 44M English phrases) have been extracted and the four lexicon models introduced in Section 3 have been estimated. Phrases up to 8 words are exploited. The whole procedure has been performed by means of the GIZA++ software tool [16] and the training tools provided by `moses`.

Both English-to-Spanish and Spanish-to-English systems employ four 5-gram LMs estimated on the corresponding EPPS, Parliaments, GigaWord, and Dev corpora. Pruning of singletons was applied before for the estimation of the GigaWord LM. 5-gram probabilities have been smoothed according to the Kneser-Ney formula [17].

Feature weights of the log-linear model were optimized by applying a minimum-error-rate training procedure which tries to maximize the BLEU score over a development data set .

The modules for inserting punctuation and for case restoring rely on a 4-gram and a 3-gram LMs, respectively, which have been estimated on the EPPS corpus only.

4.4. Results

Table 2 reports the performance of our system on the English-to-Spanish and Spanish-to-English test sets in terms of four automatic case-sensitive evaluation measures, namely BLEU, NIST, Word Error Rate (WER), and Position Independent WER (PER). Moreover, the WER of input is reported; in the case of CN the Graph Word Error Rate has been computed, i.e. the

Input	English-to-Spanish					Spanish-to-English				
	ASR-WER	BLEU	NIST	WER	PER	ASR-WER	BLEU	NIST	WER	PER
CN	8.84	0.4049	8.96	48.02	36.86	7.98	0.3751	9.15	51.87	35.49
1-best	12.07	0.4047	8.95	48.11	36.97	10.67	0.3751	9.13	52.02	35.61
rover	9.02	0.4046	9.14	46.93	36.56	10.29	0.3844	9.30	50.49	34.78
human		0.5055	10.17	38.76	29.41		0.4686	10.17	42.46	30.28
best-system		0.5153	10.29	37.86	28.76		0.5000	10.83	38.82	27.54

Table 2: Performance of the FBK-irst system on the English-to-Spanish and Spanish-to-English task of the TC-STAR 2007 Evaluation Campaign. BLEU, NIST, WER and PER measures are reported together with the WER of the ASR input.

WER of the best path within the CN. The ASR-WER has been computed after a automatic re-segmentation of references.

We run four experiments for each translation direction. In the first experiment (CN) we apply the full system described above exploiting the CNs as interface between the ASR and SLT modules. In the second experiment (1best we fed the SLT module with the best transcription produced by the ASR module. A third experiment (rover) was performed by replacing the best transcriptions of our ASR system with the transcriptions obtained combining, using the ROVER algorithm [18], the best transcriptions of all of the participatnts at the TC-STAR 2007 evaluation campaign. It it worth noting that, in this case, the original punctuation has been maintained. Finally, for the sake of comparison we also translated the human transcriptions (human).

Figures show that the CN decoder performs very close than the text decoder. A possible explanation is that the CNs do not contain much better transcriptions than the best ones as shown by the closeness of the corresponding ASR-WER values. This result does not completely confirm the outcome reported in [13] where the former slightly outperforms the latter; but in this case the CN are much richer.

rover outperforms the 1-best, but the difference can be only partially explained with the better quality of the input. More probably, it is related to the different punctuation available in the input.

In terms of absolute performance, we can claim that FBK-irst system competes well with the best systems participating in the TC-STAR 2007 evaluation campaign. In Table 2 best-system reports the performance that the two (different) best systems achieve translating the human transcriptions.

Three weak points of the FBK-irst can be pointed out. First, the CN extracted from the word graph does not contain many different transcription hypotheses, and hence it is difficult to improve over the best transcriptions. Then, the second translation step is not employed because it gives no significant benefits at the moment. Finally, the case-restoring module has a low quality, because it causes a higher decrement of performance with respect to competitors. All these issues are under investigation.

5. References

[1] M. Cettolo, "Porting an audio partitioner across domains," in *Proc. of ICASSP*, Orlando, Florida, 2002, pp. I–301–304.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[3] G. Stemmer, et al., "Adaptive training using simple target models," in *Proc. of ICASSP*, Philadelphia, PA, 2005, pp. I–997–1000.

[4] G. Stemmer and F. Brugnara, "Integration of heteroscedastic linear discriminant analysis (hlda) into adaptive training," in *Proc. of ICASSP*, Toulouse, France, 2006, pp. I–1185–1188.

[5] D. Giuliani, et al., "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, pp. 107–123, 2006.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[7] L. Mangu, et al., "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[8] http://www.speech.sri.com/projects/srilm.

[9] http://www.statmt.org/moses.

[10] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *Proc. of ASRU*, San Juan, Puertorico, 2005.

[11] http://hermes.itc.it/opensource/irstlm.

[12] M. Federico and N. Bertoldi, "How many bits are needed to store probabilities for phrase-based translation?" in *Proc. of the Workshop on Statistical Machine Translation*. New York City, 2006, pp. 94–101.

[13] N. Bertoldi, et al., "Speech translation by confusion network decoding," in *Proc. of ICASSP*, Honolulu, Hawaii, USA, 2007.

[14] http://www.tc star.org.

[15] F. Brugnara, "Context-dependent search in a context-independent network," in *Proc. of ICASSP*, Hong Kong, China,2003, pp. 360–363.

[16] F. Och and H. Ney, "Improved statistical alignment models." in *Proc. of ACL*, Hong Kong, China, 2000.

[17] J. Goodman and S. Chen, "An empirical study of smoothing techniques for language modeling." Harvard University, Technical Report TR-10-98, 1998.

[18] J. G. Fiscus, "A post-Proc.essing system to yeld reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. of ASRU*, December 1997.