Broadcast news speech-to-text translation experiments

Sylvain Raybaud **David Langlois** **Kamel Smaïli**
LORIA - Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex
givenname.lastname@loria.fr

Abstract

We present S2TT, an integrated speech-to-text translation system based on POCKETSPHINX and MOSES. It is compared to different baselines based on ANTS — the broadcast news transcription system developed at LORIA's Speech group, MOSES and Google's translation tools. A small corpus of reference transcriptions of broadcast news from the evaluation campaign ESTER2 was translated by human experts for evaluation. The Word Error Rate (WER) of the recognition stage of both systems are evaluated, and BLEU is used to score the translations. Furthermore, the reference transcriptions are automatically translated using MOSES and GOOGLE in order to evaluate the impact of recognition errors on translation quality.

Index Terms: speech-to-text translation, speech recognition

1 Introduction

Our goal is to build a large vocabulary speech-input machine translation system. While it is designed to be versatile, we first develop it for and test it on the broadcast news corpus of ESTER2 evaluation campaign (Galliano et al., 2006). The audio data (Section 4) consists in recorded French broadcast news. A fraction of it was held out and translated by a professional translator who is a native English speaker, and by a bilingual lecturer, for evaluating the performances of the system.

A straightforward approach to speech translation is to first transcribe speech using an Automatic Speech Recognition system (ASR), then translate it using a Machine Translation system (MT) (Ney et al., 2000). But speech translation is more than the combination of these two tasks because one have to deal with disfluencies, hesitations, repetitions, filler words and ill-formed sentences, which make spontaneous speech difficult to translate. Alternatively, one can develop an integrated architecture which eliminates the need for a full intermediate transcription (for example by directly translating the word lattice which also makes the system more robust to recognition errors) and allows the use of information specific to spoken language for the translation process. A longer discussion about the difference between the approaches can be found in (Seligman, 2000). We will not deal with all these issues at once. This version of our system is designed to be a stepping stone to a more elaborate, more integrated system. It is still very similar to a serial system: CMU POCKETSPHINX and SPHINXBASE libraries are used to decode chunks of audio data, which are then merged and resegmented before they are translated using MOSES libraries (Koehn and al., 2007). This integrated system is described in section 3. In the rest of the paper it will be called S2TT (**S**peech **to** **T**ext **T**ranslation System). This system will be compared to a pure serial baseline: ANTS (Illina et al., 2004), the broadcast news transcription system developed at LORIA, passes its one-best hypothesis to MOSES. We will call this system ANTS-MOSES. We compare the results of these two systems to translations generated by MOSES and GOOGLE of the reference transcription of the audio corpus, in order to estimate how much translation can be improved through improvement of the recognition sys-

tem. For the sake of comparison, we also used GOOGLE to translate the 1-best hypothesis generated by the recognition step of S2TT. Details about the evaluation (protocol, test data and results) are provided in section 5, before the discussion and conclusion.

2 The ANTS-MOSES system

2.1 Description

ANTS is based on JULIUS decoder (Lee et al., 2001). ANTS (Illina et al., 2004) is the combination of two main components: first, a speech segmentation tool (broad-band/narrow-band speech segmentation, speech/music classification, detection of silences and breaths) splits audio files into short overlapping segments. ANTS uses four acoustic models, one for each combination of broad- or narrow-band and male or female speakers. After the recognition step, overlapping transcriptions segments are merged using common words at the end of a segment and the beginning of the following one. JULIUS uses two passes: a bigram model is used during the first pass, and a 4-gram model is used for the second pass. The lexicon contains 127K pronunciations for 63K words. The training corpus for the language model (LM) is composed of French newspapers (*Le Monde* and *L'Humanité*, 580M words) and news broadcast transcriptions (*ESTER* and *TNS*, 130M words) which were used for the ESTER2 evaluation campaign in which ANTS participated and ranked #4 (for more details see (Illina et al., 2004; Galliano et al., 2006)). Finally, the best hypotheses produced by ANTS are simply piped into MOSES (see Section 4 for more details about the translation models).

3 S2TT: An integrated system using SPHINX and MOSES

3.1 Description

While we mostly test it on recorded audio files, the system is designed to eventually allow real time translation of microphone input. It is designed with the idea that audio must be transformed into translated text by passing through several intermediate states:

1. *Raw audio*: raw audio file or audio stream from the microphone.
2. *Audio chunks*: overlapping segments of this stream short enough to be efficiently decoded by the recognition engine.
3. *Rich transcription of audio chunks*: result of the automatic transcription of audio chunks. By *rich transcription* we mean words, non word events (noise, hesitation), timing, confidence measure, and whatever useful information the transcription engine can provide.
4. *Rich stream of transcribed items*: result of the merging of the transcription of the overlapping chunks.
5. *"Machine-translation friendly" segments*: previous stream resegmented into shorter segments appropriate for machine translation (ideally, sentences or phrases).
6. *Translated segments*: translation of the aforementioned segments.

In order to achieve modularity and high reactivity, the transformation from one state to the next is performed by five concurrent threads.

3.2 Speaker segmentation and adaptation

We used the system developed for ANTS (Section 2) for segmenting speech files according to sampling rate and speaker gender. The segmentation was performed off-line, which is not suitable for microphone input. We plan to include it in the integrated architecture in future versions.

3.3 Segmenting speech for translation

Resegmentation of ASR hypothesis for translation is a complex task of uttermost importance (Matsoukas et al., 2007). We used GIZA++ (Och, 2000) for generating the translation models. Translation tables are learnt on sentence aligned bilingual corpora and reordering takes place within whole sentences, but no further. It is therefore important that recognised utterances are as close as possible to well formed, single, complete sentences. ASR systems typically rely on speaker changes and silences for segmentation. While speaker change guarantees that one sentence ends and another begins (as long as they do not overlap), silences may take place in the middle of a sentence. Several sentences may also be spoken without a pause. The very concept of sentence is not sound when it comes to spontaneous speech, let alone the concept of well formed sentence.

The algorithm for segmenting ASR output in an MT friendly fashion goes as follows: first we set the maximum length of the sentence L_{max}[1]. Then we look for a position i in the L_{max} first items of the stream where an n-gram LM (we use the same LM as for the recognition process, generally with $n = 3$) generates an End-of-Sentence event. If such an event is found, then the segment $w_0, .., w_i$ is extracted. If no such position is found we look for a silence. If a silence is found at position i, then the segment $w_0, .., w_i$ is extracted. If no silence is found, then we look for the position i where an End-of-Sentence event is most likely to occur according to the LM, and extract the segment $w_0, .., w_i$. Then the process starts again at position $i + 1$ until depletion of the audio source.

4 Data and models

The acoustic models for the recognition subsystem were trained on data used in campaign ESTER2 (Section 2.1, Table 1), using SPHINXTRAIN. The source LM (French) is a 3-gram model trained on the data described in Section 2.1. The translation system is based on the classical tools of literature: MOSES for decoding, GIZA++ for producing the translation and reordering table and SRILM (Stolcke, 2002) for creating an English 3-gram LM with Kneser-Ney smoothing (Kneser and Ney, 1995). The bilingual training corpus is composed of roughly 1.6 million aligned sentences (50 million words) extracted from EUROPARL (Koehn, 2005) [2]. In order to adapt our system to broadcast news, we tuned MOSES' parameters on a 3,000 sentences (roughly 80,000 words) bilingual news corpus distributed for the WMT 2009 evaluation campaign. Table 1 summarises the details about the corpora.

Recognition	audio duration (hours)	words	
Train	100	800k	
Dev	5	40k	
French LM	-	700M	
Translation	Sentences pairs	words	
		French	English
Train	1.7M	53M	48M
English LM	1.6M	-	45M
Dev	3K	86K	77K

Table 1: Corpora sizes

System	WER	B-1	B-2	B-1&2
S2TT	25.1	18.4	25.7	30.6
S2TT-GOOGLE	25.1	19.4	34.4	38.0
ANTS-MOSES	22.3	19.3	27.3	32.3
Ref-MOSES	0	23.6	34.1	40.5
Ref-GOOGLE	0	24.3	48.2	51.7
Expert-1	0	-	34.8	-
Expert-2	0	30.6	-	-

Table 2: Translation quality with different systems

5 Evaluation

5.1 Test corpus

We extracted 252 French sentences from the different broadcast news recordings of ESTER2 test corpus (35'39" of audio). These sentences were translated by two experts who will be called *Expert-1* and *Expert-2*. These two sets of 252 reference translations will be called R1 and R2. There are important differences of style between these two sets which must be discussed because they have influence on the results of the evaluation. *Expert-1* is a native English speaker and a professional translator. Translations in R1 are therefore in literary style. On the other hand *Expert-2* is a French native speaker and an English teacher. She used the *The Corpus of Contemporary American English*[3] as an online help to elaborate R2.

5.2 Evaluation

Table 2 summarizes the results obtained by different systems on this data. The BLEU scores are computed with the script `multi-bleu.perl` provided with MOSES. B-1 is BLEU computed against reference set R1, B-2 is the score computed against R2, and B-1&2 is the score computed using the two sets. These results call for a number of comments. *S2TT* is the integrated speech translation system we developed (Section 3). It uses the same models as ANTS-MOSES but is not as polished. For all systems, we observe a large difference between B-1 and B-2. This is not surprising because of how R2 was produced: the use of a translation database makes this set more standard and homogeneous than R1. For the sake of comparison, we also translated using GOOGLE the ASR hypothesis generated by S2TT

[1] In our experiments we set $L_{max} = 40$ to match the limit set in the baseline of WMT evaluation campaigns

[2] Description of training and tuning stages can be found at http://statmt.org/wmt09/baseline.html

[3] http://corpus.byu.edu/coca/

(this is the line *S2TT-GOOGLE*).

The system called ANTS-MOSES uses ANTS to produce one-best recognition hypothesis which are then translated using MOSES. This systems improves WER by three points absolute (11% relative) and improves BLEU by one to two points depending on the test (5% to 6% relative). However it is not integrated and it is difficult to make the recognition and translation systems interact in this system.

The systems called *Ref-MOSES* and *Ref-GOOGLE* are actually just MOSES (with the same models as the one used for S2TT) and GOOGLE used for translating the reference French transcriptions. This helps to distinguish how much improvement can be brought by improving separately the recognition system and the translation system. It shows that improving the recognition can boost the overall translation score by 5 to 10 points absolute (30% relative), the rest of the improvement must be achieved by working on the translation system or by better coupling the recognition and translation engines.

Finally, the last two lines are meant to measure the "difference" between the two sets of reference translations: we compute BLEU score of R1 using R2 as reference, and the other way around. Note that in some cases, translations proposed by experts get lower BLEU scores than automatic translations. This clearly highlights the limits of this metric.

6 Perspectives and conclusion

In this paper we have presented the prototype of an integrated speech translation system. It achieves results comparable to those obtained by combining state of the art speech recognition systems and state of the art translation systems and implement an original method for re segmenting the hypothesis generated by the recognition system in a way that should make them easier to translate for the translation system. We will use it as a stepping stone to implement methods specific to speech translation, which ought to be more than a recognition stage followed by a translation stage. Our next step will be to implement translation of word lattices, with a focus on efficiency. We also plan to use information from the recognition system (detection of pauses, confidence estimations, ...) to help the segmentation step and the translation system. We also plan to use a Hidden Event LM (Stolcke and Shriberg, 1996) to improve recognition and segmentation. The speaker segmentation step, which is currently performed offline, must also be implemented directly into S2TT in order to make it truly integrated. Finally, we plan to carry out a thourough analysis of how recognition errors propagate and affect the translation process.

References

S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, and K. Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC*, pages 315–320, Genoa, Italy.

I. Illina, D. Fohr, O. Mella, and C. Cerisara. 2004. The Automatic News Transcription System: ANTS some Real Time experiments. In *8th International Conference on Spoken Language Processing - ICSLP' 2004*, page 4, Jeju, Corée du Sud.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *icassp*, pages 181–184. IEEE.

P. Koehn and al. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

A. Lee, T. Kawahara, and K. Shikano. 2001. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of the EUROSPEECH conference*, pages 1691–1694.

S. Matsoukas, I. Bulyko, B. Xiang, K. Nguyen, R. Schwartz, and J. Makhoul. 2007. Integrating speech recognition and machine translation. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pages 1281–1284.

H. Ney, S. Nießen, F. J. Och, C. Tillmann, H. Sawaf, and S. Vogel. 2000. Algorithms for Statistical Translation of Spoken Language. *IEEE Trans. on Speech and Audio Processing, Special Issue on Language Modeling and Dialogue Systems*, 8:24–36, January.

F. Och. 2000. Giza++ tools for training statistical translation models.

M. Seligman. 2000. Nine issues in speech translation. *Machine Translation (Springer)*, 15(1-2):149–186.

A. Stolcke and E. Shriberg. 1996. Statistical language modeling for speech disfluencies. In *ICASSP*, pages 405–408. IEEE.

A. Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *ICSLP*, pages 901–904, Denver, USA.