Potential scope of a fully-integrated architecture for speech translation

Alicia Pérez and M. Inés Torres
Dep. de Electricidad y Electrónica
Universidad del País Vasco
manes.torres@ehu.es

Francisco Casascuberta
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia.
fcn@iti.upv.es

Abstract

The classical approach to tackle speech translation assembles a text-to-text translation system placed after a speech recogniser, yielding the so-called decoupled architecture. In this regard, there are two issues to bear in mind: first, what is translated in the decoupled architecture is the most likely transcription of the spoken utterance; second, translation systems are sensitive to errors in the source string, and speech recognition systems are still far from being flawless.

In this paper we promote the use of an architecture to carry out speech translation that allows to build up the most likely translation relying upon both acoustic and translation models in a cooperative manner, that is the so-called integrated architecture. The integrated architecture is implemented in the finite-state framework by virtue of the composition of finite-state acoustic models of the source language within a stochastic finite-state transducer that would encompass source and target languages.

The potential performance of the integrated architecture is assessed quantitatively in relation to the decoupled one. We conclude that while the single-best approach for both decoupled and integrated architectures show similar performance, an oracle evaluation reveals that the potential scope of the integrated architecture would offer statistically significant differences.

1 Statistical speech translation

The goal of statistical speech translation is to seek the most likely string in the target language, $\widehat{\mathbf{t}}$, given the acoustic representation of a speech signal in the source language, $\mathbf{x}$.

$$\widehat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}}\, P(\mathbf{t}|\mathbf{x}) \tag{1}$$

The source string, $\mathbf{s}$, that is the transcription of the speech utterance $\mathbf{x}$, can be introduced as a hidden variable the Bayes' decision rule applied (Ney, 1999).

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} \Pr(\mathbf{t},\mathbf{s}) \Pr(\mathbf{x}|\mathbf{t},\mathbf{s}) \tag{2}$$

Assuming that the target string $\mathbf{t}$ does not affect the probability of the source acoustic representation $\mathbf{x}$ and applying the maximum approach, previous expression turns into:

$$\widehat{\mathbf{t}} \approx \underset{\mathbf{t}}{\operatorname{argmax}} \max_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{3}$$

Note the parallelism of this expression with respect to that used in automatic speech recognition:

$$\widehat{\mathbf{s}} \approx \underset{\mathbf{t}}{\operatorname{argmax}} \max_{\mathbf{s}} P(\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{4}$$

Where $P(\mathbf{s})$ and $P(\mathbf{x}|\mathbf{s})$ are approached by the language model and the acoustic model respectively.

1.1 Related work

A simple approach to tackle speech translation consists of an automatic speech recognition (ASR) system followed by a text to text translation system leading the so-called decoupled architecture. What is translated in this approach is the most likely source string derived from the ASR system. On

this account there are two issues to bear in mind: on the one hand, the ASR systems are still far from being ideal, and thus might produce errors in the output; on the other hand, the translation systems are usually very sensitive to input errors. On the contrary, it might also happen that for certain spoken utterances (such as spontaneous or grammatically ill-formed ones) the exact transcription would yield a worse translation than it would do a less accurate transcription with the same meaning and following the grammar expected for the source-language. The underlying reason is that the text translation system is typically trained with grammatically correct pair of sentences, while the conventional syntactic structure of speech might differ from that in text. The arising question is how to take advantage of both acoustic and translation knowledge sources in order to produce translations that convey as well as possible the meaning of the spoken utterance. It seems as though the series of individual decisions made within the decoupled architecture should be avoided in favour of joint decision schemes. Indeed, recent efforts in speech translation aim at making both acoustic and translation knowledge sources cooperate (Casacuberta et al., 2008).

Taking a step forward to the decoupled (or cascade) architecture, the n-best hypotheses can be explored, instead of simply exploring the single-best hypothesis, and next, by re-ranking recognition and translation scores one of the hypotheses selected (Quan et al., 2005). In (Hasan et al., 2007) an efficient method to extract large amount of n-best lists is proposed, nevertheless, it is also claimed that there is no significant gain in translation quality on the use of very large n-best lists. The drawback of making use of large n-best lists is the fact that they usually entail highly redundant hypotheses resulting in high computational cost.

Confusion networks offer an efficient means of storing information that help to overcome the redundancy issue. It was in (Bertoldi and Federico, 2005) that the hypotheses from the ASR system were represented by means of confusion networks for further translation. While it is true that due to their inner structure they allow for efficient decoding strategies (Bertoldi et al., 2007), it is also true that confusion networks generalise on the hypotheses originally provided by the ASR system and it is not clear whether this is a shortcoming or not as far as translation accuracy is concerned.

In (Saon and Picheny, 2007; Mathias and Byrne, 2006) it was proposed to translate the original lattice derived from the ASR system. In lattice-based decoding reordering represents a problem that was not present in confusion networks and methods to tackle it are proposed in (Saon and Picheny, 2007). In (Matusov et al., 2008b) a methodology to deal with lattices that allow for efficient decoding strategies was presented yielding reduced runtime cost.

In short, the aforementioned strategies are built in a two-pass strategy (as the decoupled architecture was) however, they extract translation hypotheses by combining information from both ASR and translation systems more efficiently than decoupled architecture does. For the sake of notation we shall refer to the two-pass decoding strategies as semi-decoupled architectures.

By contrast to either decoupled or semi-decoupled architectures, in (Vidal, 1997) it was presented an integrated architecture involving both the acoustic and the translation models within a single finite-state network. At decoding stage the searching network was a stochastic finite-state transducer (SFST). While an SFST used for text-to-text translation encompasses source strings along with target strings, for speech translation source strings were replaced by their acoustic representation. By virtue of this integrated architecture both the translation is produced together with the transcription of the speech (in the source language).

1.2 Contributions

In this paper the potential of the integrated architecture built with phrase-based SFSTs (instead of those presented in (Vidal, 1997)) is assessed by an oracle-like evaluation metric. The performance is compare to either decoupled or semi-decoupled architectures. Note that it is far from the range of this article to tackle re-scoring strategies. Our aim is simply to focus on the architectures themselves regardless of additional functions that might blur the assessment.

Admittedly, the ability of the transducers has evolved since (Vidal, 1997) towards the phrase-based framework, as it is our case, however, similar algorithms can be applied (Pérez et al., 2007). Thus, we are apparently using a similar technique to that in (Vidal, 1997) but within recent framework for SFSTs and we would like to asses its

potential with respect to semi-decoupled architectures.

As a second contribution, we would like to add the materialisation of speech translation from Spanish-into-Basque. It must be noted that speech translation between these two languages entails a range of challenges. To begin with, the training material is limited due to the fact that Basque is a minority language, and on the other hand, it must be noted that Basque is a highly inflected language that shows little resemblance with Spanish or English in either syntax or morphology (Pérez et al., 2008).

2 Decoupled architecture

If we express the joint probability $P(\mathbf{t}, \mathbf{s})$ in terms of posterior and prior probabilities, equation (3) can be rewritten as:

$$\widehat{\mathbf{t}} \approx \underset{\mathbf{t}}{\operatorname{argmax}} \max_{\mathbf{s}} P(\mathbf{t}|\mathbf{s})\ P(\mathbf{s})\ P(\mathbf{x}|\mathbf{s}) \quad (5)$$

Decoupled architecture implements previous expression in a sub-optimal way involving two independent stages as follows:

1. Given the acoustic representation of the utterance in the source language, $\mathbf{x}$, its expected text transcription is first obtained by an ASR system:

$$\widehat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}}\ P(\mathbf{s})\ P(\mathbf{x}|\mathbf{s}) \quad (6)$$

2. Next, the translation of $\widehat{\mathbf{s}}$ is obtained using SFSTs. Thus:

$$\widehat{\mathbf{t}} \approx \underset{\mathbf{t}}{\operatorname{argmax}}\ P(\mathbf{t}|\widehat{\mathbf{s}}) = \underset{\mathbf{t}}{\operatorname{argmax}}\ P(\mathbf{t}, \widehat{\mathbf{s}}) \quad (7)$$

The decoupled architecture, depicted in Figure 1, is the most widely used approach due to the fact that it is independent of the sort of translation paradigm used, as both the speech recognition and the translation systems are decoupled. Figure 1 shows a joint probability translation model instead of a posterior probability model since the object of our work is the former.

2.1 Semi-decoupled architecture

Instead of simply translating the best output from the ASR system, the MT system could get access to the N-best strings and produce, as a result, the M-best translations for each hypothesis, as depicted in Figure 2, in an attempt to benefit from

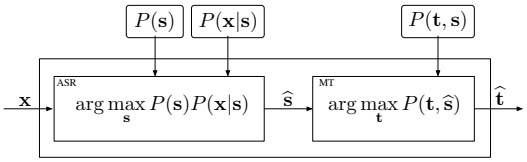


Figure 1: Decoupled architecture for speech translation. The system consists of two stages one after the other: the speech decoder and the text to text translator. The input is the speech signal, $\mathbf{x}$, in the source language, while $\widehat{\mathbf{s}}$ stands for the expected transcription of speech into text. This is translated into a text string in the target language, $\mathbf{t}$. The overall system relies on three knowledge sources, namely, the language model of the source language $P(\mathbf{s})$, the acoustic model $P(\mathbf{x}|\mathbf{s})$ and the text translation model $P(\mathbf{t}, \mathbf{s})$.

both knowledge sources (Quan et al., 2005). The re-scoring with other models, usually specialised language models, is a common practice as well. The drawback of N-best lists is that they tend to be redundant.

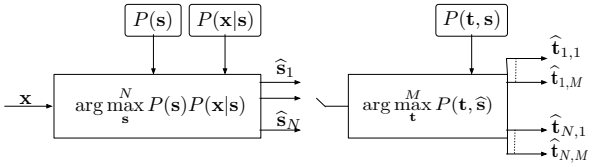


Figure 2: Semi-decoupled architecture based on a list of hypotheses. The output of the ASR is a set of N strings, for each of them M different translations are proposed.

In formal terms, instead of the single-best argument in equation (6) the N-bests are obtained. Next, for each string in the source language, the text-to-text translation could derive a list of the most likely M target strings as an alternative to the single-best proposed in equation (7). This way, for each input utterance a set of $N \times M$ translations could be derived, yet, some of the obtained translations might result to be equal.

Probabilistic word graphs offer a compact means of representing the set of hypothesis traced from the ASR. The word graph contains the hypotheses in source language and could be composed with the stochastic finite-state transducer. This procedure would parse the source graph with the translation graph, giving as a result a minimised transducer involving only paths that are compatible with those in the source graph. Never-

theless, this approach has in practice a high computational cost (Kolss et al., 2008). Instead, we have tackled it in terms of translating virtually all the strings whose probability value exceeds a given threshold rather than a fixed amount of them, in contrast with the N-best approach leading to the semi-decoupled approach.

3 Integrated architecture

By virtue of classical composition techniques associated to finite-state models, acoustic and transition models in equation (3) can be efficiently composed leading to an integrated network. This problem has a resemblance with statistical ASR, where a tight integration is achieved between acoustic and language models (Caseiro and Trancoso, 2006). In practice, for speech translation, the same sort of integration is obtained by replacing the language model with the SFST as suggested in Figure 3, since $P(\mathbf{t}, \mathbf{s})$ of equation (3) could be understood as a bilingual language.

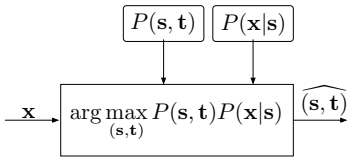


Figure 3: Integrated architecture for speech translation. The output of the system is both the text translation of an input speech signal and its transcription. The system is supported on two knowledge sources: the acoustic model and the translation model.

By means of a Viterbi-like search through the integrated network made of the SFST with the acoustic model (as depicted in Figure 4), the most likely path (or translation form) is obtained. Note that associated to each path there are both the expected source and target strings. As a result, a Viterbi-like searching engine with SFSTs would approach speech translation as follows:

$$\widehat{(\mathbf{s},\mathbf{t})} \approx \underset{(\mathbf{s},\mathbf{t})}{\operatorname{argmax}}\, P(\mathbf{s},\mathbf{t})\; P(\mathbf{x}|\mathbf{s}) \qquad (8)$$

This process turns out the expected string in the target language, $\widehat{\mathbf{t}}$, and as a by-product, the expected transcription of the spoken utterance in the source language, $\widehat{\mathbf{s}}$, all in a single decoding step.

It has already been mentioned that the integrated architecture, in practice, can be implemented by replacing the language model in the ASR system with an SFST and arranging the lexical model, which happens to be bilingual. As far as the lexical model is concerned, all the items having the same source phrase would display the same acoustic model regardless of the associated target phrase. As illustrated in Figure 4, the search is driven by the text-to-text phrase-based SFST. Nevertheless, when an edge is being explored the decoder turns to the lexical model that entails the phonetic representation of the source phrase along with the text representation of the target phrase. Next, each phoneme is replaced by its corresponding HMM. On-the-fly integration has shown to be an efficient technique in speech-recognition (Caseiro and Trancoso, 2006) and it can also be implemented similarly for speech translation. As an outcome, in the integrated architecture finite-state acoustic and translation models are neatly composed on the fly, i.e., the integration is not static, but it is carried out on demand at decoding time.

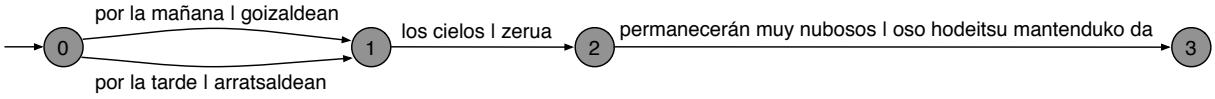
The main feature of the integrated architecture is its ability to construct the hypotheses on the basis of the cooperation of acoustic and translation model. Along with it, it is remarkable the ability of the integrated architecture to carry out both the transcription of source signal and its translation simultaneously in a single decoding stage.

3.1 Word-graph from the integrated speech translation decoder

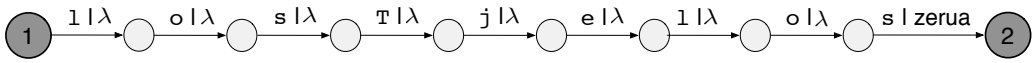
N-best lists can also be extracted under the integrated architecture as if it were the ASR system. Nevertheless, this procedure would provide both the source and the target strings jointly, in a single decoding, in contrast to the decoupled architecture, which requires two systems.

As a natural alternative to the integrated architecture providing N-best lists, we propose the use of probabilistic word graphs derived from the trellis of the integrated speech translation search engine. In the same way as an ASR can give a word graph as an output, within the integrated architecture we can do exactly the same as depicted in Figure 5.

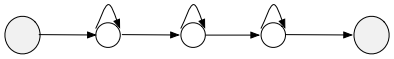
By contrast to the ASR process, speech translation within the integrated architecture does not rely on a language model but on an SFST. As a result, the edges of the resulting graph are labelled with input and output substrings. In addition, they also have an associated score, which accounts not only for the probability in the associated SFST, but



(a) Finite-state translation model.



(b) Integration of a lexical unit consisting of phone-like units for source phrase along with text representation of target phrase.



(c) Integration of HMMs within each phone-like unit.

Figure 4: Speech translation with integrated architecture involves the on-the-fly integration of several knowledge sources within a single finite-state network. (a) Finite-state transducer. (b) The lexical model consists on the phonetic transcription (SAMPA is used here) of the input substring by means of a left-to-right topology. (c) Phone-like units are modelled by typical three-state continuous hidden Markov models (HMMs).

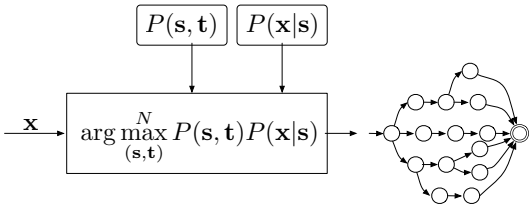


Figure 5: A word-graph from the integrated architecture.

contain, as well, the contribution of the acoustic model. The word graph relies on information associated to acoustics, such as time segmentation and acoustic scores, and also on scores related to translation model. Thus, the most likely hypothesis from the word graph coincides with the hypothesis given by the integrated approach without word graph layout in its single-best approach.

The big concern here is the fact that the most likely hypothesis is not always in agreement with the best hypothesis. That is, the most likely hypothesis might report a worse accuracy with respect to the reference than other hypothesis with lower probability. Thus, it might happen that other less likely hypotheses match better with the reference than the expected one. On this account, for each input utterance, the word-graph entails the entire set of successful translation forms associated to the SFST that have been derived from the speech translation process. Measuring the quality of the resulting word-graphs would allow us to get to know the upper threshold of the underlying translation model. That is, by measuring the quality of the derived word-graph for each utterance, it is being measured the best quality that can be obtained with the integrated architecture.

Finally, it is well worth mentioning that the obtained word graph can be composed with other models, such as a specialised target language model, that could help to filter the hypotheses in an efficient manner. Moreover, due to the thorough information gathered in the graph (such as acoustic probabilities, time synchronisation between input signal and translation model, translation probabilities etc.) a wide variety of re-scoring methods can be explored for each task. Nevertheless, re-scoring with more accurate models is out of the scope of this work.

4 Experimental results

In this Section several sets of experiments aiming to compare the speech translation architectures proposed in previous sections are presented. The corpus used for these experiments is the only multilingual corpus for MT in Basque that includes speech, namely *M*eteus. It is a weather forecast corpus (Pérez et al., 2008) consisting of daily weather forecast reports issued over 28 consecutive months. The main features of the corpus are summarised in Table 1. A training-independent subset made of 500 different pairs was extracted and then recorded by at least 3 bilingual speakers resulting in a speech evaluation set of $1,800$ utterances by 36 speakers.

For the purposes of these experiments, linguistic phrases that considered both syntax and se-

		Spanish	Basque
Train	Sentences	14,615	14,615
	Running words	191,156	187,195
	Vocabulary	702	1,135
Test	Sentences	1,800	1,800
	Hours of speech	3.0	3.5

Table 1: Main features of METEUS corpus.

mantics jointly were used (Pérez et al., 2008). They were identified by the Ametzagaña group, a non-profit organization working on R+D regarding with Basque language processing. A preliminary series of experiments on text translation over a text test yielded a BLEU of 66.1 for the phrase-based SFST while Moses (Koehn et al., 2006) provided a BLEU of 67.0 (being 64.0 in its monotonic decoding). Moreover, on average, text translation with Moses took 0.65 seconds per sentence, while with SFST it only took 0.25 seconds per sentence.

4.1 Semi-decoupled architecture

The system consisted of two consecutive stages: the speech decoder and the text-to-text translator. A word graph representing the set of hypothesis traced from the ASR was extracted, and from it, the N-best lists for different values of N were derived. Table 2 shows the speech translation results obtained in these experiments as well as the transcription WER derived from the ASR system. That is, both BLEU and WER are scores associated to the target language, while ASR-WER is associated to the errors in the source language which the translation system has to deal with.

The results obtained with the N-best lists being $N = 10^0$ corresponds to the single-best hypothesis, in other words, these results correspond to the fully decoupled architecture. Note that as the number of explored hypotheses grow, better accuracy is obtained. Nevertheless, from $N = 10^4$ onwards the improvements are not statistically significant (Bisani and Ney, 2004; Zhang and Vogel, 2004; Koehn, 2004). That is, the benefits obtained are less and less important. In fact, the performance has its upper value in the results derived with the word-graphs.

The graph-score defined in (Zens and Ney, 2005) is the score of the optimum sequence over all possibilities. Note that the most likely hypothesis, the single-best (that with $N = 10^0$), does not have the highest quality, meaning that other less-likely hypotheses could yield a better quality. In this regard, the results reported in Table 2 show the upper threshold of translation quality that the models can provide and permit us to evaluate the potential quality of the translation system. We would like to mention that in the literature there are efficient algorithms to find the oracle BLEU (the hypothesis with the highest attainable BLEU score) under different constraints (Li and Khudanpur, 2009; Leusch et al., 2008).

A further re-scoring criterion (Matusov et al., 2008a) would allow to re-rank the hypotheses and, hopefully, obtain improvements with respect to the most likely hypothesis. Yet, it is not the aim of this article to focus on re-scoring for either semi-decoupled or integrated architectures.

4.2 Integrated architecture

The integrated SFSTs allowed us to obtain both the translation of the speech utterance as well as its transcription, in a single decoding step. Hence, both speech transcription and translation results are jointly derived. The speech translation results in the target language, WER and BLEU, along with the source WER (or transcription WER) are summarised in Table 3. Both the quality of the N-best lists with different values of N are explored and compared to the potential quality that can be obtained with the integrated word-graph derived from the integrated architecture (denoted as WG). Note that the results with the single-best hypothesis (with $N = 10^0$) correspond to the integrated architecture proposed by (Vidal, 1997). Here, a step forward is taken by deriving either N-best lists or word-graphs and also thanks to the phrase-based framework recently defined for SFSTs.

Once again, as the number of explored hypotheses grow, better accuracy is obtained. Nevertheless, the results are saturated later than in the case of semi-decoupled architecture. By comparing Table 2 with Table 3, it can be derived that while for the single-best hypothesis the integrated architecture offered a slightly better performance for speech translation than the decoupled one, as the number of hypotheses under consideration grows, the difference between both, in terms of performance, increases significantly. As far as the quality of the transcribed source strings is reasonably better with the decoupled architecture than with the integrated one.

Thus, the integrated architecture can provide a-

		Semi-decoupled architecture					
		N-best					WG
		10^0	10^1	10^2	10^3	10^4	
Speech Translation	BLEU	40.8	49.7	54.8	56.7	57.2	57.6
	WER	50.3	42.2	38.2	36.8	36.4	36.2
	ASR-WER	7.9	7.0	6.3	5.0	4.6	4.5

Table 2: Speech translation results with semi-decoupled architecture using N-best lists and word graphs (WG). Particularly, the result with $N = 10^0$ corresponds to the fully decoupled architecture. The WER associated to the ASR is also included.

		Integrated architecture					
		N-best					WG
		10^0	10^1	10^2	10^3	10^4	
Speech Translation	BLEU	40.9	50.3	59.4	62.9	63.5	64.0
	WER	49.6	41.5	35.3	33.2	32.4	32.2
	source-WER	9.6	8.6	7.5	6.8	6.6	6.6

Table 3: Speech translation results with integrated architecture using N-best lists and word graphs (WG). The WER associated to the source language (source-WER) is also included.

high quality speech translation word graph. In this sense, higher speech translation scores can then be expected when generating a speech translation word graph under the integrated architecture, than when translating the word graph generated by an ASR system under the semidecoupled architecture.

5 Conclusions

In this article both decoupled, semi-decoupled and integrated architectures have been assessed using the same framework and task, and with the same underlying phrase-based SFSTs. That is, we are focusing on the potential quality of several architectures for speech translation applications rather than on translation models.

The use of a fully-integrated architecture made up of word-graphs comprising acoustic models within phrase-based SFSTs has been proposed. The word-graph derived from an integrated architecture that was presented in (Vidal, 1997), allows to measure the potential of the integrated architecture with respect to others.

Above all, it is remarkable the fact that the integrated architecture with word-graphs makes it possible a full cooperation between acoustic and translation models leading to a combination of both knowledge sources when it comes to searching for speech translation hypotheses. It has been shown that the quality of the obtained word-graph is significantly better than with either semi-decoupled or decoupled architectures. Moreover, as opposed to semi-decoupled architecture, a single decoding step is involved. Given the potential performance of the integrated architecture, it could be an adequate candidate for further re-ranking or post-processing operations with other models, such as particular language models in the target language. Anyway, those techniques could be of benefit regardless of the architecture used.

Finally, a manual inspection suggests that the integrated architecture with word-graphs allows to explore less likely paths that showed lower probability due to long distance alignments but have finally result to be more successful for automatic metrics while not always for human judgements.

Once it has been shown the ability of the integrated architecture to obtain much more accurate translations, for future work we will focus on exploring both reordering and re-ranking techniques that would allow to select the most accurate hypotheses from the integrated word-graph. This could be probably carried out by turning to another composition, that is, to the composition of the derived word-graph with a more accurate target language model.

References

Bertoldi, N. and M. Federico. 2005. A New Decoder for Spoken Language Translation Based on

Confusion Networks. *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 141–146.

Bertoldi, N., R. Zens, and M. Federico. 2007. Speech translation by confusion network decoding. In *Proceedings of ICASSP*, Honolulu, HA.

Bisani, Maximilian and Hermann Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412.

Casacuberta, F., M. Federico, H. Ney, and E. Vidal. 2008. Recent efforts in spoken language translation. *IEEE Signal Processing Magazine*, 25(3):80–88.

Caseiro, Diamantino and Isabel Trancoso. 2006. A specialized on-the-fly algorithm for lexicon and language model composition. *IEEE Transactions on Audio, Speech & Language Processing*, 14(4):1281–1291.

Hasan, Saša, Richard Zens, and Hermann Ney. 2007. Are very large N-best lists useful for SMT? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 57–60, Rochester, New York, April. Association for Computational Linguistics.

Koehn, Philipp, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondřej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Moran, and Evan Herbst. 2006. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. Technical report, Johns Hopkins University, Center for Speech and Language Processing.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Kolss, Muntsin, Stephan Vogel, and Alex Waibel. 2008. Stream decoding for simultaneous spoken language translation. In *Proc. of the interspeech*, pages 1052–1054, Brisbane, Australia, Sep.

Leusch, G, E Matusov, and H Ney. 2008. Complexity of finding the bleu-optimal hypothesis in a confusion network. In *Conference on Empirical Methods in Natural Language Processing*, pages 839–847, Waikiki, Honolulu, Hawaii, October.

Li, Z and S Khudanpur. 2009. Efficient extraction of oracle-best translations from hypergraphs. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 9–12, Boulder, Colorado, June. Association for Computational Linguistics.

Mathias, L. and W. Byrne. 2006. Statistical phrase-based speech translation. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1:561–564.

Matusov, E., G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. S. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008a. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

Matusov, Evgeny, Björn Hoffmeister, and Hermann Ney. 2008b. Asr word lattice translation with exhaustive reordering is possible. In *Interspeech*, pages 2342–2345, Brisbane, Australia, September.

Ney, Hermann. 1999. Speech translation: Coupling of recognition and translation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, Phoenix, AR, March.

Pérez, Alicia, M. Inés Torres, and Francisco Casacuberta. 2007. Speech translation with phrase based stochastic finite-state transducers. In *Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, Hawaii USA, April 15-20. IEEE.

Pérez, A., M.I. Torres, and F. Casacuberta. 2008. Joining linguistic and statistical methods for Spanish-to-Basque speech translation. *Speech Communication*. http://dx.doi.org/10.1016/j.specom.2008.05.016.

Quan, VH, M. Federico, and M. Cettolo. 2005. Integrated n-best re-ranking for spoken language translation. *Proceedings of Interspeech 2005*, pages 3181–3184.

Saon, George and Michael Picheny. 2007. Lattice-based viterbi decoding techniques for speech translation. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, pages 386–389, Kyoto, Japan, Dec.

Vidal, Enrique. 1997. Finite-state speech-to-speech translation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 111–114, Munich, Germany, April.

Zens, R. and H. Ney. 2005. Word graphs for statistical machine translation. In *ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 191–198, Ann Arbor, MI, June.

Zhang, Ying and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, October.