Finite-state acoustic and translation model composition in statistical speech translation: empirical assessment

Alicia Pérez(1), **M. Inés Torres**(2)
(1)Dep. Computer Languages and Systems
(2)Dep. Electricidad y Electrónica
University of the Basque Country UPV/EHU
Bilbao (Spain)
(1)alicia.perez@ehu.es
(2)manes.torres@ehu.es

Francisco Casacuberta
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Valencia (Spain)
fcn@iti.upv.es

Abstract

Speech translation can be tackled by means of the so-called decoupled approach: a speech recognition system followed by a text translation system. The major drawback of this two-pass decoding approach lies in the fact that the translation system has to cope with the errors derived from the speech recognition system. There is hardly any cooperation between the acoustic and the translation knowledge sources. There is a line of research focusing on alternatives to implement speech translation efficiently: ranging from semi-decoupled to tightly integrated approaches. The goal of integration is to make acoustic and translation models cooperate in the underlying decision problem. That is, the translation is built by virtue of the joint action of both models. As a side-advantage of the integrated approaches, the translation is obtained in a single-pass decoding strategy. The aim of this paper is to assess the quality of the hypotheses explored within different speech translation approaches. Evidence of the performance is given through experimental results on a limited-domain task.

1 Introduction

Statistical speech translation (SST) was typically implemented as a pair of consecutive steps in the so-called *decoupled approach*: with an automatic speech recognition (ASR) system placed before to a text-to-text translation system. This approach involves two independent decision processes: first, getting the most likely string in the source language and next, getting the expected translation into the target language. Since the ASR system is not an ideal device it might make mistakes. Hence, the text translation system would have to manage with the transcription errors. Being the translation models (TMs) trained with positive samples of well-formed source strings, they are very sensitive to ill-formed strings in the source language. Hence, it seems ambitious for TMs to aspire to cope with both well and ill formed sentences in the source language.

1.1 Related work

Regarding the coupling of acoustic and translation models, there are some contributions in the literature that propose the use of semi-decoupled approaches. On the one hand, in (Zhang et al., 2004), SST is carried out by

Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, pages 99–107,
Donostia–San Sebastián, July 23–25, 2012.

an ASR placed before a TM with an additional stage that would re-score the obtained hypotheses within a log-linear framework gathering features from both the ASR system (lexicon and language model) and the TM (eg. distortion, fertility) and also additional features (POS, length etc.).

On the other hand, in (Quan et al., 2005), the N-best hypotheses derived from an ASR system were next translated by a TM, finally, a last stage would re-score the hypotheses and make a choice. Within the list of the N-best hypotheses typically a number of them include some n-grams that are identical, hence, the list results to be an inefficient means of storing data. Alternatively, in (Zhou et al., 2007) the search space extracted from the ASR system, represented as a word-graph (WG), was next explored by a TM following a multilayer search algorithm.

Still, a further approach can be assumed in order to make the graph-decoding computationally cheaper, that is, confusion networks (Bertoldi et al., 2007). Confusion-networks implement a linear approach of the word-graphs, however, as a result, dummy hypotheses might be introduced and probabilities mis-computed. Confusion networks traded off between the accuracy and storage ability of word-graphs for decoding time. Indeed, in (Matusov and Ney, 2011) an efficient means of doing the decoding with confusion networks was presented. Note that these approaches follow a two-pass decoding strategy.

The aforementioned approaches implemented phrase-based TMs within a log-linear framework. In this context, in (Casacuberta et al., 2008) a fully integrated approach was examined. Under this approach, the translation was carried out in a single-pass decoding, involving a single decision process in which acoustic and translations models cooperated. This integration paradigm, was earlier proposed in (Vidal, 1997), showing that a single-pass decoding was enough to carry out SST.

Finally, in (Pérez et al., 2010) several SST decoding approaches including decoupled, N-best lists and integrated were compared. Nevertheless, the paper focused on the potential scope of the approaches, comparing the theoretical upper threshold of their performance.

1.2 Contribution

All the models assessed in this work relay upon exactly the same acoustic and translation models. It is the combination of them on which we are focusing. In brief, the aim of this paper is to compare different approaches to carry out speech translation decoding. The comparison is carried out using exactly the same underlying acoustic and translation models in order to allow to make a fair comparison of the abilities inherent to the decoding strategy. Apart from the decoupled and semi-decoupled strategies we also focus on the fully-integrated approach. While the fully integrated approach allows to provide the most-likely hypothesis, we explored a variant: an integrated architecture with a re-scoring LM that provided alternatives derived from the integrated approach and used re-scoring to make the final decision. Not only an oracle-evaluation is provided as an upper-threshold of the experiments but also an experimental set-up to give empirical evidence.

The paper is arranged as follows: Section 2 introduces the formulation of statistical speech translation (SST); Section 3 describes different approaches to put into practice SST, placing emphasis on the assumptions behind each of them. Section 4 is devoted to assess experimentally the performance of each approach. Finally, in Section 5 the concussions drawn from the experiments are summarized.

2 Statistical speech translation

The goal of speech translation, formulated under the probabilistic framework, is to find the most likely string in the target language ($\hat{\mathbf{t}}$) given the spoken utterance in the source language. Speech signal in the source language is characterized in terms of an array of acoustic features in the source language, $\mathbf{x}$. The decision problem involved is formulated as follows:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \tag{1}$$

In this context, the text transcription in the source language (denoted as $\mathbf{s}$) is introduced as a hidden variable and Bayes' rule applied:

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

Assuming $P(\mathbf{x}|\mathbf{s},\mathbf{t}) \approx P(\mathbf{x}|\mathbf{s})$, and using the maximum term involved in the sum as an approach to the sum itself for the sake of computational affordability, we yield to:

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

As a result, the expected translation is built relying upon both a translation model ($P(\mathbf{s},\mathbf{t})$) and an acoustic model in the source language ($P(\mathbf{x}|\mathbf{s})$). This approach requires the joint cooperation of both models to implement the decision problem since the maximum over $\mathbf{s}$ concerns both of them.

2.1 Involved models

Being the goal of this paper to compare different techniques to combine acoustic and translation models, it is important to keep constant the underlying models while varying the strategies to combine them. Before to delve into the composition strategies and due to the fact that some combination strategies are based on the finite-state topology of the models, a summary of the relevant features of the underlying models is given in this section.

2.1.1 Translation model

The translation model used in this work to tackle all the approaches consists of a stochastic finite-state transducer (SFST) encompassing phrases in the source and target languages together with a probability of joint occurrence. The SFST ($\mathcal{T}$) is a tuple $\mathcal{T} = \langle \Sigma, \Delta, Q, q_0, R, F, P \rangle$, where:

Σ is a finite set of input symbols;

Δ is a finite set of output symbols;

Q is a finite set of states;

$q_0 \in Q$ is the initial state;

$R \subseteq Q \times \Sigma^+ \times \Delta^* \times Q$ is a set of transitions. $(q, \tilde{s}, \tilde{t}, q') \in R$, represents a transition from the state $q \in Q$ to the state $q' \in Q$, with the source phrase $\tilde{s} \in \Sigma^+$ and producing the substring $\tilde{t} \in \Delta^*$, where $\tilde{t}$ might consist of zero or more target words ($|\tilde{t}| \geq 0$);

$F : Q \to [0,1]$ is a final state probability;

$P : R \to [0,1]$ is a transition probability;

Subject to the stochastic constraint:

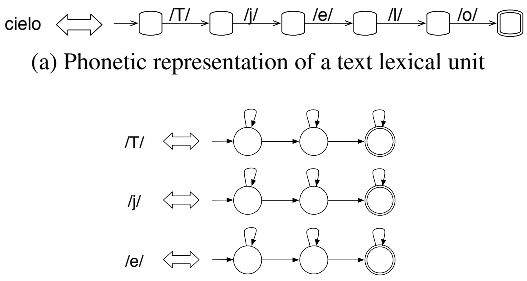
$$\forall q \in Q \quad F(q) + \sum_{\tilde{s},\tilde{t},q'} P(q,\tilde{s},\tilde{t},q') = 1 \tag{4}$$

For further reading on formulation and properties of these machines turn to (Vidal et al., 2005).

The SFST can be understood as a statistical bi-language implemented by means of finite-state regular grammar (Casacuberta and Vidal, 2004) (in the same way as a stochastic finite-state automaton can be used to model a single language): $\mathcal{A} = \langle \Gamma, Q, q_0, R, F, P \rangle$, being $\Gamma \subseteq \Sigma^+ \times \Delta^*$ a finite-set of bilingual-phrases. Likewise, bilingual n-gram models can be inferred in practice (Mariño et al., 2006).

2.1.2 Acoustic models

The acoustic model consists of a mapping of text-transcriptions of lexical units in the source language and their acoustic representation. That comprises the composition of: 1) a lexical model consisting of a mapping between the textual representation with their phone-like representation in terms of a left-to-right sequence; and 2) an inventory of phone-like units consists of a typical three-state hidden Markov model (Rabiner, 1989). Thus, acoustic model lays on the composition of two finite-state models (depicted in Figure 1).



(a) Phonetic representation of a text lexical unit

(b) HMM phone-like units

Figure 1: Acoustic model requires composing phone-like units within phonetic representation of lexical units.

3 Decoding strategies

In the previous section the formulation of SST was summarized. Let us now turn into practice and show the different strategies explored to combine acoustic and translation models to tackle SST. The approaches accounted are: decoupled, semi-decoupled and integrated architectures. While the former two are implementable by virtue of alternative TMs, the latter is achieved thanks to the integration allowed by finite-state framework. Thus, in order to compare the combination rather than the TMs themselves, all of the combinations shall be put in practice using the same SFST as TM.

3.1 Decoupled approach

Possibly the most widely used approach to tackle speech translation is the so-called serial, cascade or decoupled approach. It consists of a text-to-text translation system placed after an ASR system. This process is formally stated as:

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} \max_{\mathbf{s}} P(\mathbf{x}|\mathbf{s})P(\mathbf{s})P(\mathbf{t}|\mathbf{s}) \qquad (5)$$

In practice, previous expression is implemented in two independent stages as follows:

1st stage: an ASR system would find the most likely transcription ($\hat{\mathbf{s}}$):

$$\hat{\mathbf{s}} \approx \arg\max_{\mathbf{s}} P(\mathbf{x}|\mathbf{s})P(\mathbf{s}) \qquad (6)$$

2nd stage next, given the expected string in the source language ($\hat{\mathbf{s}}$), a TM would find the most likely translation:

$$\hat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P(\mathbf{t}|\hat{\mathbf{s}}) = \arg\max_{\mathbf{t}} P(\hat{\mathbf{s}}, \mathbf{t}) \qquad (7)$$

The TM involved in eq.(7) can be based on either posterior or joint-probability as the difference between both of them is a normalization term that does not intervene in the maximization process. The second stage has to cope with expected transcription of speech ($\hat{\mathbf{s}}$) which does not necessarily convey the exact reference source string ($\mathbf{s}$). That is, the ASR might introduce errors in the source string to be translated in the next stage. However, the TMs are typically trained with correct source-target pairs. Thus, transcription errors are seldom foreseen even in models including smoothing (Martin et al., 1999). In addition, TMs are extremely sensitive to the errors in the input, in particular to substitutions (Vilar et al., 2006).

This architecture represents a suboptimal means of contending with SST as referred in eq. (3). This approach barely takes advantage of the involved knowledge sources, namely, acoustic and translation models.

3.2 Semi-Decoupled approach

Occasionally, the most probable translation does not result to be the most accurate one with respect to a given reference. That is, it might happen that hypotheses with a slightly lower probability than that of the expected hypothesis turn to be more similar to the reference than the expected hypothesis. This happens due to several factors, amongst others, due to the sparsity of the data with which the model was trained.

In brief, some sort of disparity between the probability of the hypotheses and their quality might arise in practice. The semi-decoupled approach arose to address this issue. Hence, rather than translating a single transcription hypothesis, a number of them are provided by the ASR to the TM, and it is the latter that makes the decision giving as a result the most likely translation. The decoupled approach is implemented in two steps, and so is it the semi-decoupled approach. Details on the process are as follows:

1st stage: for a given utterance in the source language, an ASR system, laying on source acoustic model and source language model (LM), would provide a search sub-space. This sub-space is traced in the search process for the most likely transcription of speech but without getting rid of other highly probable hypotheses.

For what us concern, this sub-space is represented in terms of a graph of words in the source language ($\mathcal{S}$). The word-graph gathers the hypotheses with a probability within a threshold with respect to the optimal hypothesis at each time-frame as it was formulated in (Ney et al., 1997). The obtained graph is an acyclic directed graph where the nodes are associated with word-prefixes of a variable length, and the edges join the word sequences allowed in the recognition process with an associated recognition probability. The edges consist of the acoustic and language model probabilities as the ASR system handles throughout the trellis.

2nd stage: translating the hypotheses within $\mathcal{S}$ (the graph derived in the 1st stage) allows to take into account alternative translations for the given spoken utterance. The searching space being explored is limited by the source strings conveyed by $\mathcal{S}$. The combination of the recognition probability with the translation probability results in a score that accounts both recognition and translation likelihood:

$$\hat{\mathbf{t}} \approx \arg \max_{\mathbf{t}} \max_{\mathbf{s} \in \mathcal{S}} P(\mathbf{s}) P(\mathbf{s}, \mathbf{t}) \qquad (8)$$

Thus, acoustic and translation models would one re-score the other.

All in all, this semi-decoupled approach results in an extension of the decoupled one. It accounts alternative transcriptions of speech in an attempt to get good quality transcriptions (rather than the most probable transcription as in the case of the decoupled approach). Amongst all the transcriptions, those with high quality are expected to provide the best quality in the target language. That is, by avoiding errors derived from the transcription process, the TM should perform better, and thus get translations of higher quality. Note that finally, a single translation hypothesis is selected. To do so, the highest combined probability is accounted.

3.3 Fully-integrated approach

Finite-state framework (by contrast to other frameworks) makes a tight composition of models possible. In our case, of acoustic and translation finite-state models. The fully-integrated approach, proposed in (Vidal, 1997), encfompassed acoustic and translation models within a single model. To develop the fully-integrated approach a finite-state acoustic model on the source language ($\mathcal{A}$) providing the text transcription of a given acoustic utterance ($\mathcal{A}$:

$X \rightarrow S$) can be composed with a text translation model ($\mathcal{T}$) that provides the translation of a given text in the source language ($\mathcal{T} : S \rightarrow T$) and give as a result a transducer ($\mathcal{Z} = \mathcal{A} \circ \mathcal{T}$) that would render acoustic utterances in the source language to strings in the target language. For the sake of efficiency in terms of spatial cost, the models are integrated on-the-fly in the same manner as it is done in ASR (Caseiro and Trancoso, 2006).

The way in which integrated architecture approaches eq. (3) is looking for the most-likely source-target translation pair as follows:

$$\widehat{(\mathbf{s}, \mathbf{t})} = \arg\max_{(\mathbf{s},\mathbf{t})} P(\mathbf{s}, \mathbf{t}) P(\mathbf{x}|\mathbf{s}) \qquad (9)$$

That is, the search is driven by bilingual phrases made up of acoustic elements in the source language integrated within bilingual phrases of words together with target phrases.

Then, the expected translation would simply be approached as the target projection of $\widehat{(\mathbf{s}, \mathbf{t})}$, the expected source-target string (also known as the lower projection); and likewise, the expected transcription is obtained as a side-result by the source projection (aka upper projection).

It is well-worth mentioning that this approach implements fairly the eq. (3) without further assumptions rather than those made in the decoding stage such as Viterbi-like decoding with beam-search. All in all, acoustic and translation models cooperate to find the expected translation. Moreover, it is carried out in a single-pass decoding strategy by contrast to either decoupled or semi-decoupled approaches.

3.4 Integrated WG and re-scoring LM

The fully-integrated approach looks for the single-best hypothesis within the integrated acoustic-and-translation network. Following the reasoning of Section 3.2, the most likely path together with other locally close paths in the integrated searching space can be extracted and arranged in terms of a word graph. While the WG derived in Section 3.2 was in source language, this one would be bilingual.

Given a bilingual WG, the lower-side net ($WG.l$) can be extracted keeping the topology and the associated probability distributions while getting rid of the input string of each transition, this gives as a result the projection of the WG in the target language. Next, a target language model (LM) would help to make the choice for the most likely hypothesis amongst those in the $WG.l$.

$$\hat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P_{WG.l}(\mathbf{t}) P_{LM}(\mathbf{t}) \qquad (10)$$

In other words, while in Section 3.2 the translation model was used to re-score alternative transcriptions of speech whereas in this approach a target language models re-scores alternative translations provided by the bilingual WG. Note that this approach, as well as the semi-decoupled one, entail a two-pass decoding strategy. Both rely upon two models: the former focused on the source language WG, this one focuses on the target language WG.

4 Experiments

The aim of this section is to assess empirically the performance each of the four approaches previously introduced: decoupled, semi-decoupled, fully-integrated and integrated WG with re-scoring LM. The four approaches differ on the decoding strategy implemented to sort out the decision problem, but all of them rely on the very same knowledge sources (that is, the same acoustic and translation model).

The main features of the corpus used to carry out the experimental layout are summarized in Table 1. The training set was used to infer the

TM consisting of an SFST and the test set to assess the SST decoding approaches. The test set consisted of 500 training-independent pairs different each other, each of them was uttered by at least 3 speakers.

		Spanish	**Basque**
Train	Sentences	15,000	
	Running words	191,000	187,000
	Vocabulary	702	1,135
Test	Sentences	1,800	
	Hours of speech	3.0	3.5

Table 1: Main features of the Meteus corpus.

The performance of each experiment is assessed through well-known evaluation metrics, namely: bilingual evaluation under-study (BLEU) (Papineni et al., 2002), word error-rate (WER), translation edit rate (TER).

4.1 Results

The obtained results are given in Table 2. The performance of the most-likely or single-best translation derived by either decoupled or fully-integrated architectures is shown in the first row of Tables 2a and 2b respectively. The performance of the semi-decoupled and integrated WG with re-scoring LM is shown in the second row. The highest performance achievable by both the semi decoupled approach and the integrated WG with re-scoring LM is given in the third row. To do so, an oracle evaluation of the alternatives was carried out and the score associated to the best choice achievable was given as in (Pérez et al., 2010). Since the oracle evaluation provides an upper threshold of the quality achievable, the scope of each decoupled or integrated approaches can be assessed regardless of the underlying decoding algorithms and approaches. The highest performance achievable is reflected in the last row of Tables 2a and 2b.

4.2 Discussion

While the results with two-pass decoding strategies (either decoupled or semi-decoupled approach) require an ASR engine, integrated approaches have the ability to get both the source string together with its translation. This is why we have make a distinction between ASR-WER in the former and source-WER in the latter. Nevertheless, our aim focuses on translation rather than on recognition.

The results show that semi-decoupled approach outperforms the decoupled one. Similarly, the approach based on the integrated WG with the re-scoring target LM outperforms the integrated approach. As a result, exploring different hypotheses and making the selection with a second model allows to make refined decisions. On the other hand, comparing the first row of the Table 2a with the first row of the Table 2b (or equally the second row of the former with the second row of the latter), we conclude that slightly better performance can be obtained with the integrated approach.

Finally, comparing the third row of both Table 2a and Table 2b, the conclusion is that the eventual quality of the hypotheses within the integrated approach are significantly better than those in the semi-decoupled approaches. That is, what we can learn is that the integrated decoding strategy keeps much better hypotheses than the semi-decoupled one throughout the decoding process. Still, while good quality hypotheses exist within the integrated approach, the re-scoring with a target LM used to select a single hypothesis from the entire network has not resulted in getting the best possible hypothesis. Oracle evaluation shows that the integrated approach offers a leeway to achieve improvements in the quality, yet, alternative strategies have to be explored.

	ASR	target		
	WER	BLEU	WER	TER
D 1-best	7.9	40.8	50.3	47.7
SD	7.9	42.2	47.6	44.7
SD tgt-oracle	7.5	57.6	36.2	32.8

(a) Decoupled and semi-decoupled

	source	target		
	WER	BLEU	WER	TER
I 1-best	9.6	40.9	49.6	46.8
I WG + LM	9.3	42.6	46.7	43.9
I tgt-oracle	6.6	64.0	32.2	28.5

(b) Integrated and integrated WG with LM

Table 2: Assessment of SST approaches decoupled (2a) and integrated (2b) respectively.

5 Conclusions

Different approaches to cope with the SST decoding methodology were explored, namely, decoupled approach, semi-decoupled approach, fully-integrated approach and integrated approach with a re-scoring LM. The first two follow a two-pass decoding strategy and focus on exploring alternatives in the source language; while the integrated one follows a single-pass decoding and present tight cooperation between acoustic and translation models.

All the experimental layouts used exactly the same translation and acoustic models differing only on the methodology used to overcome the decision problem. In this way, we can assert that the differences lay on the decoding strategies rather than on the models themselves. Note that implementing all the models in terms of finite-state models allows to build both decoupled and integrated approaches.

Both decoupled and integrated decoding approaches aim at finding the most-likely translation under different assumptions. Occasionally, the most probable translation does not result to be the most accurate one with respect to a given reference. On account of this, we turned to analyzing alternatives and making use of re-scoring techniques on both approaches in an attempt to make the most accurate hypothesis emerge. This resulted in semi-decoupled and integrated-WG with re-scoring target LM approaches.

What we can learn from the experiments is that integrating the models allow to keep good quality hypotheses in the decoding process. Nevertheless, the re-scoring model has not resulted in being able to make the most of the integrated approach. In other words, there are better quality hypotheses within the word-graph rather than that selected by the re-scoring target LM. Hence, further work should be focused on other means of selecting hypotheses from the integrated word-graph.

However, undoubtedly significantly better performance can be reached from the integrated decoding strategy than from the semi-decoupled one. It seems as though knowledge sources modeling the syntactic differences between source and target languages should be tackled in order to improve the performance, particularly in our case, a strategy for further work could go on the line of the recently tackled approach (Durrani et al., 2011).

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation: through the *Tímpano* (TIN2011-28169-C05-04) and *iTrans2* (TIN2009-14511) projects; also through *MIPRCV* (CSD2007-00018) project within the Consolider-Ingenio 2010 program; by the Basque Government to PR&ST research group (GIC10/158, IT375-10), and by the Generalitat Valenciana under grants *ALMPR* (Prometeo/2009/01) and GV/2010/067.

References

[Bertoldi et al.2007] N. Bertoldi, R. Zens, and M. Federico. 2008. Efficient speech translation by confusion network decoding. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pg. 1696–1705

[Casacuberta and Vidal2004] F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2): pg. 205–225.

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

[Caseiro and Trancoso2006] D. Caseiro and I. Trancoso. 2006. A specialized on-the-fly algorithm for lexicon and language model composition. *IEEE Transactions on Audio, Speech & Language Processing*, 14(4): pg. 1281–1291.

[Durrani et al.2011] N. Durrani, H. Schmid, and A. Fraser. 2011. A joint sequence translation model with integrated reordering. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pg. 1045–1054

[Mariño et al.2006] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4): pg. 527–549

[Martin et al.1999] S. C. Martin, H. Ney, and J. Zaplo. 1999. Smoothing methods in maximum entropy language modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing* , vol. 1, pg. 545–548

[Matusov and Ney2011] E. Matusov and H. Ney. 2011. Lattice-based ASR-MT interface for speech translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4): pg. 721 –732

[Ney et al.1997] H. Ney, S. Ortmanns, and I. Lindam. 1997. Extensions to the word graph method for large vocabulary continuous speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pg. 1791 –1794

[Papineni et al.2002] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Annual Meeting on Association for Computational Linguistics*, pg. 311–318

[Pérez et al.2010] A. Pérez, M. I. Torres, and F. Casacuberta. 2010. Potential scope of a fully-integrated architecture for speech translation. *Annual Conference of the European Association for Machine Translation*, pg. 1–8

[Quan et al.2005] V. H. Quan, M. Federico, and M. Cettolo. 2005. Integrated n-best re-ranking for spoken language translation. *European Converence on Speech Communication and Technology, Interspeech*, pg. 3181–3184.

[Rabiner1989] L.R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): pg. 257–286

[Vidal et al.2005] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. 2005. Probabilistic finite-state machines - part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7): pg. 1026–1039

[Vidal1997] E. Vidal. 1997. Finite-state speech-to-speech translation. *International Conference on Acoustic, Speech and Signal Processing*, vol. 1, pg. 111–114

[Vilar et al.2006] David Vilar, Jia Xu, Luis Fernando D'Haro, and H. Ney. 2006. Error Analysis of Machine Translation Output. *International Conference on Language Resources and Evaluation*, pg. 697–702

[Zhang et al.2004] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, and W. K. Lo. 2004. A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. *International Conference on Computational Linguistics*, pg. 1168-1174

[Zhou et al.2007] B. Zhou, L. Besacier, and Y. Gao. 2007. On efficient coupling of ASR and SMT for speech translation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pg. 101–104