Spoken Language Translation Using Automatically Transcribed Text in Training

Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, Hermann Ney

Human Language Technology and Pattern Recognition
Computer Science Department
RWTH Aachen University
Aachen, Germany
surname@cs.rwth-aachen.de

Abstract

In spoken language translation a machine translation system takes speech as input and translates it into another language. A standard machine translation system is trained on written language data and expects written language as input. In this paper we propose an approach to close the gap between the output of automatic speech recognition and the input of machine translation by training the translation system on automatically transcribed speech. In our experiments we show improvements of up to 0.9 BLEU points on the IWSLT 2012 English-to-French speech translation task.

1. Introduction

Spoken language translation (SLT) connects automatic speech recognition (ASR) and machine translation (MT) by translating recognized spoken language into a target language. In general, the speech translation process is divided into two separate parts. First, an ASR system provides an automatic transcription of spoken words. Then, the recognized words are translated by a machine translation system.

However, a difficult part of SLT is the interface between the ASR system and the MT system, due to the mismatch between the output of the ASR system and the expected input of the MT system. A standard MT system expects grammatically correct written language as input, because it is usually trained on written bilingual text with punctuation marks and case information. In contrast, the output of an ASR system is automatically transcribed natural speech containing recognition errors. Thus, the expected input of the MT system does not match the actual ASR output. Furthermore, ASR systems recognize sequences of words and do not provide punctuation marks or case information.

In this paper, we describe how the inconsistency between the ASR output and the SMT input is solved by replacing the source language data of a bilingual training corpus with automatically transcribed text. In a first approach, we keep the target language including case information and punctuation, because our goal is to improve the translation quality directly in an SLT task. On this new corpus, we train a statistical machine translation (SMT) system and use the system to translate the recognized speech into another language. Furthermore, case information and punctuation are restored during the translation process.

As a second approach, we built a bilingual training corpus with ASR output as source language data and the corresponding manual transcription with case information and punctuation marks as target language data. In the next step, an SMT system is trained on this corpus. Before translating the recognized speech into the target language, the ASR output is translated into manual transcription. Thus, the postprocessing of the ASR output is modelled as machine translation and we are able to translate the postprocessed ASR output with a standard translation system which is trained on written bilingual text.

On the English-French SLT task from IWSLT 2012, we show that our presented approaches improve the translation quality by up to 0.9 BLEU and 0.9 TER.

The paper is organized as follows. In the next section, we give a short overview of related work. In Section 3, we describe the usage of automatically transcribed text in the training process of an SMT system. Finally, we discuss the experimental results in Section 5, followed by a conclusion.

2. Related Work

In [1], an approach is presented to improve automatic call classification by training an SMT system on a bilingual corpus with ASR output as source language data and the corresponding manual transcribed text as target language data. The SMT system cleans the automatically transcribed text before the call classification. For further improvement of their framework, n-Best lists of the recognition were used. They performed experiments using IBM model 2 on live data collected from an enterprise call center and showed improvements in class classification accuracy.

A similar approach is presented in [2]. The authors describe a statistical transformation model which transforms spoken language into written language. Further, they compare the approach with a rule-based transformations model

in terms of precision and recall.

Another approach to transform spoken language into written language is described in [3]. A transduction model based on weighted finite-state transducers is trained on a parallel corpus of automatic transcription and manual transcription. In the experiments, Cantonese speech was transformed to standard written Chinese. The authors report improvements in Word Error Rate.

In [4], the use of automatically transcribed text as training data was described. The authors recognized audio recordings of parallel speech with an ASR system to create additional monolingual as well as bilingual corpora. They showed improvements by training a language, an acoustic and a translation model including the additional data.

In [5] different methods for punctuation prediction were analyzed. By using a translation system to translate from unpuncated to punctuated text the translation quality was improved on the IWSLT 2011 English-to-French Speech Translation of Talks task.

In our work, we revisit the idea of building a new corpus using automatically transcribed text as source language data. However, instead of cleaning the ASR output, we translate from ASR output into a target language directly, i.e. we replace the source language data of the bilingual corpus only. Furthermore, we do not want to collect additional monolingual or bilingual data, but the goal is to improve the quality of spoken language translation by using automatically transcripted text in the training process of a translation system. By training a phrase-based machine translation system on the new corpus, we want to close the gap between the output of an ASR system and the expected input of an SMT system. Moreover, we combine the original and the new corpus in various ways and extract *n*-Best lists from lattices to create a larger corpus. In addition, based on the idea of modeling punctuation prediction as machine translation, we train a translation system on a bilingual corpus with ASR output as source language data and corresponding manual transcription as target language data. This system translates from ASR output to manual transcription, i.e. the postprocessing of the ASR output is performed with a machine translation system. The main advantage of this method is that a standard text translation system can be used to translate the postprocessed ASR output.

3. Automatically Transcribed Text in Training

The starting point of this work is a data source which provides audio recordings, the corresponding manual transcriptions and the translation of these transcriptions. The online-available TED talks are such a kind of source [1]. This website provides manually transcribed and translated lecture-type talks presented at TED conferences. Furthermore, WIT^3 (Web Inventory of Transcribed and Translated Talks) redistributes the original content published by the TED website for the machine translation community [6]. The transcriptions and the translations are processed as parallel bilingual corpus to be able to train an SMT system. Further, development and test sets are provided.

In an SLT application, the development and test sets are automatically transcribed speech, which have to be translated into a target language. We assume in this work that the recognitions of the development and test sets do not contain punctuation and casing and the segmentation is given and corresponds to sentence-like units. With an SMT system, the automatically recognized speech is translated. Furthermore, the punctuation and the case information are restored during the translation process as described in [7]. In order to train such an SMT system, the punctuation and the case information of source language data in the bilingual training corpus are deleted to create a *pseudo ASR output*. In our work, we train an SMT system on a bilingual corpus with real ASR output instead of pseudo ASR output as source language data.

Due to the fact that WIT^3 also specifies the talks which were used to create the provided bilingual corpora, we are able to recognize the relevant audio recordings with our ASR system. About 1028 relevant talks are available on the web. In sum, roughly 250 hours of speech have to be recognized. Using the automatically transcribed recordings as source language data, we build a new bilingual corpus to train an SMT system for an SLT task.

3.1. Sentence Alignment

In general, an ASR system does not provide sentence-wise segmentation. However, a bilingual corpus, which is used to train an SMT system, consists of parallel sentences. In order to align automatic transcriptions sentence-wise to a given segmented manual transcription, we employ an automatic re-segmentation algorithm as described in [8].

The re-segmentation algorithm calculates the Levenshtein alignment between the recognition and its manual transcription. By backtracing the decisions of the edit distance algorithm, an alignment between a given sequence of words and an already sentence-wise segmented manual transcription as reference can be found. Thus, the sentence segmentation of the reference is transferred to the recognition. The re-segmentation algorithm is solved by dynamic programming.

As mentioned, WIT^3 provides manually transcribed text as well as the corresponding translation. First, we align our recognized training data to the manual transcription, which is already segmented on sentence level. In a second step, we replace the manual transcription with its translation. This results in a parallel bilingual corpus with ASR output as source language data and its translation with punctuation and case information as target language data.

Table 1 shows an example of an aligned bilingual sentence pair with various source language sentences. Starting with the given *manual transcription*, the *pseudo ASR output* is created by removing the full stop at the end of the sen-

[1] http://www.ted.com/

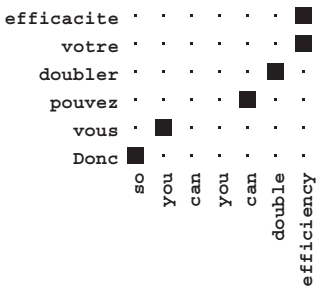

Figure 1: Partial alignment between *automatic transcription* and *manual translation* (Table 1).

tence and lowercasing the very first word. This transformed sentence is grammatically correct. In contrast, the *automatic transcription* of the sentence contains the repetition of the phrase "*you can*". Furthermore, "*60*" is transcribed as written number "*sixty*".

In Figure 1 a part of the corresponding alignment between the automatic transcription and its translation is shown. During the training procedure of the SMT system, phrase pairs such as

- ⟨you can you can, vous pouvez⟩
- ⟨sixty percent, 60 %⟩

are learned. With these phrase pairs, the SMT system is able to correct ASR output and to rewrite written numbers as digits during the translation process. Instead of translating the phrase "*you can*" twice, the SMT system has got the option to translate the phrase into "*vous pouvez*" directly, if such an error occurs in a given ASR output.

3.2. ASR Output Postprocessing

Another approach to make use of automatically transcribed text is to set up an SMT system which translates from ASR output into manually transcribed text. Therefore, we do not replace the manual translation with its translation as described before, but an SMT system is trained on a corpus with automatically transcribed text as source language data and manual transcriptions as target language data. Before the actual translation of the recognized speech, the SMT system performs a postprocessing of the ASR output. The ASR output is translated and during the translation process punctuation marks and case information are restored. Considering the bilingual sentence pair in Table 1 and the corresponding alignment in Figure 2, during the training of the SMT system phrase pairs such as

- ⟨you can you can, you can⟩
- ⟨sixty percent, 60 %⟩

are extracted. The main advantage is that the postprocessed ASR output can be used as input for an existing standard text translation system. Thus, we do not have to modify the training data of the translation system to translate ASR output.

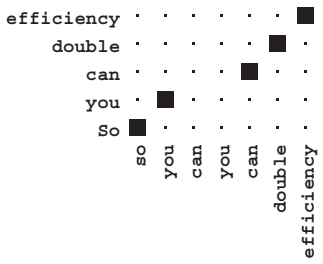

Figure 2: Partial alignment between *automatic transcription* and *manual transcription* (Table 1).

4. System Description

In this section, we describe our ASR and MT system, which are employed in this work. With the ASR system, we recognize the source language data of the new bilingual corpus as well as the development and test sets in a given SLT task. We train a MT system on the different corpora and combination to verify the impact of automatically transcribed text in the training. All setups are tuned on a development set and are compared on a test set.

4.1. ASR System

The ASR system is based on our English speech recognition system that we successfully applied in Quaero evaluations [9].

The recognizer is a generative statistical classifier that maps a sequence of acoustic observations x_1^T to a word sequence w_1^N via Bayes decision rule:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}}\, p(w_1^N)^{\gamma} p(x_1^T | w_1^N)\,. \tag{1}$$

The prior probability $p(w_1^N)$ is the *language model*, $p(x_1^T | w_1^N)$ is the *acoustic model*, and γ is the *language model scale*.

In the acoustic feature extraction, the system computes Mel-frequency cepstral coefficients (MFCC) from the audio signal, which are transformed with a vocal tract length normalization (VTLN). In addition, a voicedness feature is computed. Acoustic context is incorporated by concatenating nine feature vectors in a sliding window. The resulting feature vector is reduced to 45 dimensions by means of a linear discriminant analysis (LDA). Furthermore, bottleneck features derived from a multilayer perceptron (MLP) are concatenated with the feature vector.

The acoustic model is based on hidden Markov models (HMMs) with Gaussian mixture models (GMMs) as emission probabilities. The GMM has a pooled, diagonal covariance matrix. It models 4500 generalized triphones which are derived by a hierarchical clustering procedure (CART). The parameters of the GMM are estimated with the expectation-maximization (EM) algorithm with a splitting procedure according to the maximum likelihood criterion.

Table 1: Example of a bilingual sentence pair. *pseudo ASR output* is created by removing punctuation and case information of the *manual transcription*. The *automatic transcription* was recognized with our ASR system and *manual translation* is the corresponding given translation.

Corpus	
manual transcription	So you can double efficiency with a 60 percent internal rate of return .
pseudo ASR output	so you can double efficiency with a 60 percent internal rate of return
automatic transcription	so you can you can double efficiency with a sixty percent internal rate of return
manual translation	Donc vous pouvez doubler votre efficacite nergtique avec un Taux de Rendement Interne de 60 % .

The language model is a Kneser-Ney smoothed 4-gram. Several language models are trained on different datasets. The final language model is obtained by linear interpolation. The vocabulary of the recognition lexicon is obtained by applying a count-cut-off on the language model data. Each word in the lexicon can have multiple pronunciations. Missing pronunciations are derived with a grapheme-to-phoneme tool.

The recognition is structured in three passes, In the first pass, a speaker independent model is used. The recognition result of the first pass is used for estimating feature transformations for speaker adaptation (CMLLR). The second pass uses the CMLLR transformed features. Finally, a confusion network decoding is performed on the word lattices obtained from the second pass.

Table 2: Acoustic training data of ASR system

Corpus	Amount of data [hours]
quaero-2011	268h
hub4+tdt4	393h
epps	102h

Table 3: Language model training data of ASR system

Corpus	Amount of data [running words]
Gigaword 4	2.6B
Ted	2.7M
Acoustic transcriptions	5M

The acoustic model of the ASR system is trained on 793 hours of transcribed acoustic data in total, see Table 2. The acoustic training data consists of American broadcast news data (hub4+tdt4), European parliament speeches (epps), and British broadcast conversations (quaero). The MLP is trained on the 268 hours of the quaero corpus only. We use 4500 triphone states and perform eight EM splits, resulting in a GMM with roughly 1.1 million mixture components.

The language model is trained on a large amount of news data (Gigaword), the transcriptions of the audio training data, and a small amount of in-domain data (ted), see Table 3. The recognition lexicon consists of 150k words.

4.2. MT System

The decoder of the phrase-based translation system which is used in this work is described in [10] and is part of RWTH's open-source SMT toolkit Jane 2.1 [2]. We use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based distortion model, a 4-gram target language model and three binary count features. The features $h_m(f_1^J, e_1^J)$ are combined in a weighted log-linear model to find the best translation $\hat{e}_1^{\hat{I}}$

$$\hat{e}_1^{\hat{I}} = \arg\max_{e_1^I} \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^J). \quad (2)$$

The weights are optimized using standard MERT [11] on 200-best lists with BLEU as optimization criterion.

5. Experimental Evaluation

The proposed approach was evaluated on the IWSLT 2012 English-to-French spoken language translation task based on the already mentioned TED talks. For the evaluation, WIT[3] provides in-domain bilingual training data based on manually transcribed text and its translation. The 1028 talks (around 250 hours of speech), which corresponds to the bilingual training data, were recognized with the described ASR system.

For the baseline model, we removed punctuation and case information of the source language to create pseudo ASR output (Table 7) as we assume that the source language as produced by the speech recognition system does not contain any punctuation marks or case information. Punctuation and case information are restored during the translation process. To indicate that an SMT system was trained on this corpus, we mark the setup with MANUAL-TRANSCRIPTION.

In Table 8, the data statistics for the bilingual corpus with ASR output as source language data are shown. The number of sentences and running words differs from the

[2] http://www-i6.informatik.rwth-aachen.de/jane/

original bilingual corpus in Table 7, because a small number of recordings were not accessible. In the following, setups based on this data are tagged with AUTOMATIC-TRANSCRIPTION.

As a first approach, we only consider the output of the ASR system based on a confusion network decoding on the word lattices obtained from the second pass. Setups trained on the corpus are marked with AUTOMATIC-TRANSCRIPTION (cn-decoding).

To extend the training corpus, we further extracted n-Best lists from the resulting lattices of the second pass. We hope that the MT system could gain by using more ASR output in training. For the extraction of the n-Best lists, we used the LATTICE-TOOL from the SRI toolkit [12]. The n-Best lists were sentence-aligned to the corresponding manual translation as described before. In our experiments, we chose $n = \{1, 10, 20\}$. Thus, the size of the corpus was multiplied by n. Setups using corpora based on n-Best lists are labelled with AUTOMATIC-TRANSCRIPTION (n-Best). Note that AUTOMATIC-TRANSCRIPTION (1-Best) differs from AUTOMATIC-TRANSCRIPTION (cn-decoding). In contrast to 1-Best decoding which extracts the maximum probability sentence from the search space, cn-decoding approximates the minimization of the expected WER and is closer to the theoretical WER optimal decision rule for ASR. Therefore cn-decoding in practice always performs better than 1-Best output.

For the spoken language translation task in the IWSLT 2012 evaluation campaign, ASR output is provided as development set and test set (Table 5). However, to be consistent with the recognized training data, we used our own recognitions of the development and test sets in all experiments (except for one of the baseline experiments). In Table 4, we compare the word error rate (WER) of the provided sets (IWSLT 2012) with our recognitions (RWTH). A lower WER indicates a better recognition quality. The data statistics for RWTH (cn-decoding) are shown in Table 6.

Table 4: Comparison of the development and test sets in terms of WER

	dev	**test**
IWSLT 2012	18.0	16.7
RWTH (pass 1)	20.0	18.4
RWTH (pass 2)	17.5	15.9
RWTH (cn-decoding)	17.3	15.7

For all experiments, we used a 4-gram language model with modified Kneser-Ney smoothing which was trained with the SRILM toolkit on the monolingual version of the in-domain bilingual training data and on the Europarl and News Commentary data. Further, GIZA++ [13] was employed to train word alignments for each setup.

Table 5: Data Statistics for the provided development and test set (IWSLT 2012)

	dev	**test**
Sentences	934	1 664
Running Words	17 755	27 754
Vocabulary	3 133	3 698

Table 6: Data Statistics for development and test set recognized by our ASR system (RWTH (cn-decoding))

	dev	**test**
Sentences	934	1 664
Running Words	17 804	27 514
Vocabulary	3 149	3 689

5.1. Phrase Table and Data Combination

In this work, we analyze three different approaches to combine both corpora AUTOMATIC-TRANSCRIPTION and MANUAL-TRANSCRIPTION. We hope to further improve the translation quality by augmenting our baseline system with the original data. Due to the fact, that a small amount of the recordings were not accessible or were recognized with a low quality, the system could gain from adding the manually transcribed data.

5.1.1. Union

As first approach, we built the union of the phrase tables of AUTOMATIC-TRANSCRIPTION and MANUAL-TRANSCRIPTION. If a phrase pair occurs in both phrase tables, the phrase probabilities and lexical probabilities of both phrase pairs are interpolated linearly. In all other cases, we just keep the phrase pair. This method is denoted by AUTOMATIC-TRANSCRIPTION $\cup$ MANUAL-TRANSCRIPTION.

5.1.2. Two Phrase Tables

We augmented the phrase table of our baseline system, which was trained on AUTOMATIC-TRANSCRIPTION, with an additional phrase table based on MANUAL-TRANSCRIPTION. The phrase tables were connected by a bi-

Table 7: Data Statistics for pseudo ASR output as source language data (MANUAL-TRANSCRIPTION)

	English	French
Sentences	140 537	
Running Words	2 361 366	2 894 364
Vocabulary	47 159	64 627
Singletons	18 722	27 696

Table 8: Data Statistics for ASR output as source language data (AUTOMATIC-TRANSCRIPTION (cn-decoding))

	English	French
Sentences	135 603	
Running Words	2 311 602	2 803 745
Vocabulary	37 886	63 558
Singletons	12 715	27 211

nary feature, i.e phrases from AUTOMATIC-TRANSCRIPTION got the feature value 1 and phrases from MANUAL-TRANSCRIPTION the value 0. Setups using two phrase tables are marked as AUTOMATIC-TRANSCRIPTION ∘ MANUAL-TRANSCRIPTION.

5.1.3. Training Data Concatenation

In contrast to the other two methods, the training corpora MANUAL-TRANSCRIPTION and AUTOMATIC-TRANSCRIPTION were combined before the phrase extraction. In particular, MANUAL-TRANSCRIPTION and AUTOMATIC-TRANSCRIPTION were concatenated and the translation model was re-trained. This setup is named AUTOMATIC-TRANSCRIPTION + MANUAL-TRANSCRIPTION.

5.2. Results

Table 9 shows the comparison between different setups. We measured the translation quality of all systems in BLEU [14] and TER [15] on the development set as well as on the test set. First, we ran two baseline experiments. Both systems were trained on MANUAL-TRANSCRIPTION. The first setup was tuned and tested on the provided development and test sets (IWSLT 2012) and the second one on our own recognitions. It seems that a better WER results in a higher translation quality.

Using AUTOMATIC-TRANSCRIPTION (cn-decoding) performs only slightly better then the baseline. The biggest improvement was achieved by AUTOMATIC-TRANSCRIPTION (cn-decoding) ∘ MANUAL-TRANSCRIPTION in comparison to MANUAL-TRANSCRIPTION (baseline, RWTH (cn-decoding)). The translation quality was improved by 0.5 points in BLEU and 0.4 points in TER on the test set. With AUTOMATIC-TRANSCRIPTION + MANUAL-TRANSCRIPTION, we get an improvement of 0.4 points in BLEU and 0.7 points in TER. AUTOMATIC-TRANSCRIPTION (cn-decoding) ∪ MANUAL-TRANSCRIPTION performs worst of all combination methods.

The idea to improve the SLT system by using a larger corpus based on *n*-Best lists does not help. At least the system trained on 20-Best lists performs similar to the baseline. It seems that there is a mismatch between the development and test sets, which are based on confusion network decoding, and the *n*-Best lists extracted with the LATTICE-TOOL.

Finally, we employed the idea of ASR output postprocessing with an MT system. For a robust baseline, we used an existing text translation system trained on TED data, Europarl and News Commentary data, Multi-UN data and Gigaword data. This system was chosen to show the impact of this method even in a large setup. In Table 10, we compare the IMPLICIT method as described in [7] with our approach (POSTPROCESSING).

The training data for IMPLICIT setup was preprocessed by removing all punctuation marks and case information from the source language data, while the target language is kept untouched. The removal was done after the word alignment. The punctuation marks in the target sentence which were aligned with punctuation marks in the source sentences become non-aligned.

For POSTPROCESSING, we set up a standard phrase-based system trained on a bilingual corpus with ASR output as source language data and manual transcription as target language data. As development and sets we used again our recognitions. The system was tuned on the development using standard MERT on 200-best lists with BLEU as optimization criterion. The output of this system was the input of the existing text translation system.

With our proposed method, we achieve an improvement of 0.9 points in BLEU and 0.9 points in TER.

Table 11 shows an example of different input (English) and their translations (French). During the postprocessing of the ASR output repetition such as "*i i*" and "*i 'm i 'm*" are transformed to "*I*" and "*I 'm*". With the IMPLICIT approach, "*i 'm i 'm*" is translated twice. In the translation of post-processed ASR output, the phrase "*je suis*" is obtained only once. It seems that the postprocessing of the ASR output helps the text translation system to translate automatically transcripted input.

6. Conclusion

In this paper, we have introduced an approach to close the gap between automatic speech recognition and machine translation in the application of spoken language translation. In a speech translation setting, we showed that using automatically transcripted text in the training process of a machine translation system can improve the translation quality.

Further, we modelled the ASR output postprocessing as machine translation. The main advantage is that the translation system used in speech translation does not require any preprocessing. On the IWSLT 2012, we got an improvement of up to 0.9 points in BLEU and TER.

In future work, we would like to improve the WER of an ASR system directly by applying a machine translation system as postprocessing step.

7. Acknowledgements

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innova-

Table 9: Comparison of results for the SLT task English-French (IWSLT 2012), including data used to train the translation model.

setup	dev		test	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
MANUAL-TRANSCRIPTION (baseline, IWSLT 2012)	18.0	69.1	20.8	62.7
MANUAL-TRANSCRIPTION (baseline, RWTH (cn-decoding))	18.5	68.4	21.1	62.5
AUTOMATIC-TRANSCRIPTION (cn-decoding)	18.4	68.8	21.3	62.3
AUTOMATIC-TRANSCRIPTION (cn-decoding) $\cup$ MANUAL-TRANSCRIPTION	18.6	68.1	21.2	62.2
AUTOMATIC-TRANSCRIPTION (cn-decoding) $\circ$ MANUAL-TRANSCRIPTION	18.7	68.0	21.6	62.1
AUTOMATIC-TRANSCRIPTION (cn-decoding) + MANUAL-TRANSCRIPTION	18.6	67.9	21.5	61.8
AUTOMATIC-TRANSCRIPTION (1-Best)	18.4	68.7	21.1	62.4
AUTOMATIC-TRANSCRIPTION (10-Best)	18.4	68.8	21.0	62.3
AUTOMATIC-TRANSCRIPTION (20-Best)	18.5	68.6	21.2	62.4

Table 10: Comparison between the methods IMPLICIT and POSTPROCESSING on the SLT task English-French (IWSLT 2012).

method	dev		test	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
IMPLICIT	19.2	67.8	22.5	61.6
POSTPROCESSING	20.1	67.2	23.4	60.7

Table 11: Comparison of different input sentences and the corresponding reference and translation. POSTPROCESSING is the output of the SMT which postprocesses the automatic transcription.

Input/Translations	
automatic transcription	and you know i i thought well i 'm i 'm like living in a science fiction movie
manual transcription	and I thought like , " Wow . I am like living in a science fiction movie .
POSTPROCESSING	and , you know , I thought , " Well , I 'm like living in a science fiction movie .
IMPLICIT translation	et , vous savez , je me suis dit : " Eh bien , je suis comme je suis vivant dans un film de science-fiction .
POSTPROCESSING translation	et , vous savez , j' ai pens : " Eh bien , je suis vivant dans un film de science-fiction .
reference translation	et l j' ai pens : " Wow . c' est comme si je vivais dans un film de science-fiction .

tion. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

8. References

[1] T. Faruquie, N. Rajput, and V. Raj, "Improving automatic call classification using machine translation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, april 2007, pp. IV–129 –IV–132.

[2] T. Kawahara, K. Shitaoka, and H. Nanjo, "Automatic transformation of lecture transcription into document style using statistical framework," in *INTERSPEECH*. ISCA, 2004.

[3] P. Xu, P. Fung, and R. Chan, "Phrase-level transduction model with reordering for spoken to written language transformation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 4965 –4968.

[4] M. Paulik and A. Waibel, "Spoken language translation from parallel speech audio: Simultaneous interpretation as slt training data," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, march 2010, pp. 5210 –5213.

[5] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in *International Workshop on Spoken Language Translation*, San Francisco, California, USA, Dec. 2011, pp. 238–245.

[6] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[7] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *International Workshop on Spoken Language Translation*, Kyoto, Japan, Nov. 2006, pp. 158–165.

[8] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *International Workshop on Spoken Language Translation*, Pittsburgh, PA, USA, Oct. 2005, pp. 148–154.

[9] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR evaluation system for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, 2011, pp. 2212–2215.

[10] J. Wuebker, H. Ney, and R. Zens, "Fast and scalable decoding with language model look-ahead for phrase-based statistical machine translation," in *Annual Meeting of the Assoc. for Computational Linguistics*, Jeju, Republic of Korea, July 2012, pp. 28–32.

[11] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[12] A. Stolcke, "Srilman extensible language modeling toolkit," in *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 901–904.

[13] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, IBM Research Report RC22176 (W0109-022), Sept. 2001.

[15] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.