# Using the TED Talks to Evaluate Spoken Post-editing of Machine Translation

**Jeevanthi Liyanapathirana**
Copenhagen Business School
Dalgas Have 15
DK-2000 Frederiksberg, Denmark
jl.ibc@cbs.dk

**Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

**Abstract**

This paper presents a solution to evaluate spoken post-editing of imperfect machine translation output by a human translator. We compare two approaches to the combination of machine translation (MT) and automatic speech recognition (ASR): a heuristic algorithm and a machine learning method. To obtain a data set with spoken post-editing information, we use the French version of TED talks as the source texts submitted to MT, and the spoken English counterparts as their corrections, which are submitted to an ASR system. We experiment with various levels of artificial ASR noise and also with a state-of-the-art ASR system. The results show that the combination of MT with ASR improves over both individual outputs of MT and ASR in terms of BLEU scores, especially when ASR performance is low.

**Keywords:** machine translation, spoken post-editing, evaluation

## 1. Introduction

In this paper, we consider the task of voice-based post-editing of the output of a machine translation (MT) system. In other words, a human corrects the translation produced by a system simply by speaking out this correction. The paper presents a method for using an existing dataset to evaluate spoken post-editing methods without the need for costly experiments involving human post-editors. We use the French transcripts of TED talks (`www.ted.com`) as the source texts to be translated by MT, the original English audio recordings as the spoken corrections, and the English transcripts as a reference for the automatic evaluation of translation quality.

The proposed evaluation framework is applied to two original methods for combining the outputs of an automatic speech recognition (ASR) system and of an MT system, to generate a better translation than each of the outputs considered individually. The first method is heuristic-based and relies on the confidence of the ASR to select words, while the second method uses machine learning to learn selection rules based on a wider set of features, including word length, position and part-of-speech, and ASR confidence.

The paper is organized as follows. Section 2 presents the motivation of our system. Section 3 describes the setting of our experiments, the ASR and MT systems, and the use of TED data for spoken post-editing. Section 4 defines the two approaches to ASR/MT combination that we use to generate an improved translation. Section 5 describes the experiments and their results. Finally, Section 6 reviews previous work that has been done in this area, in comparison to our proposal.

## 2. Motivation

Voice-based post-editing is valuable whenever using a keyboard is not possible or practical, e.g. with mobile devices or impaired users. Therefore, the range of applications of our proposal is quite large, and covers the dissemination uses of MT rather than the assimilation ones, as defined by Hovy et al. (2002), because it helps producing a high-quality translation. The applications are intended for users who are proficient in both source and target languages, so that spoken post-editing is accurate, but who prefer using voice rather than a keyboard – e.g. on a mobile device such as a smartphone or a smartwatch. For instance, they could use spoken post-editing of MT to translate and disseminate to their colleagues an email originally written in a foreign language, or to re-tweet to their followers a message from a foreign contributor.

Mesa-Lao (2014) surveyed the post-editors' views and attitudes before and after the introduction of speech technology as a front-end to a computer-aided translation workbench. The survey shows that people tend to respond positively towards ASR used in post-editing, and they seem willing to adopt it as an input method for future post-editing tasks. In another user-oriented study, Dragsted et al. (2011) investigated the efficiency that can be achieved by using speech recognition software for translation tasks. With sufficient training and practice, the speech recognition's time consumption appears to approach that of sight translation, and speech recognition quality appears to approach that of written translation. These studies thus indicate the need for, and the potential acceptability of, voice-based post-editing of MT output, although more studies focusing on possible uses in multilingual social networks and/or access from mobile devices would be welcome.

## 3. Experimental Framework for Studying Spoken Post-Editing of MT

The proposed framework for voice-based post-editing is represented in Figure 1. The source text is translated by an MT system. The human post-editor views the source text and its automatic translation sentence-by-sentence, and utters for each sentence a corrected translation into a microphone coupled to an ASR system. An algorithm finally combines the output of the ASR with the automatic translation, to generate an improved translation by using as much as possible the correct fragments from each system.

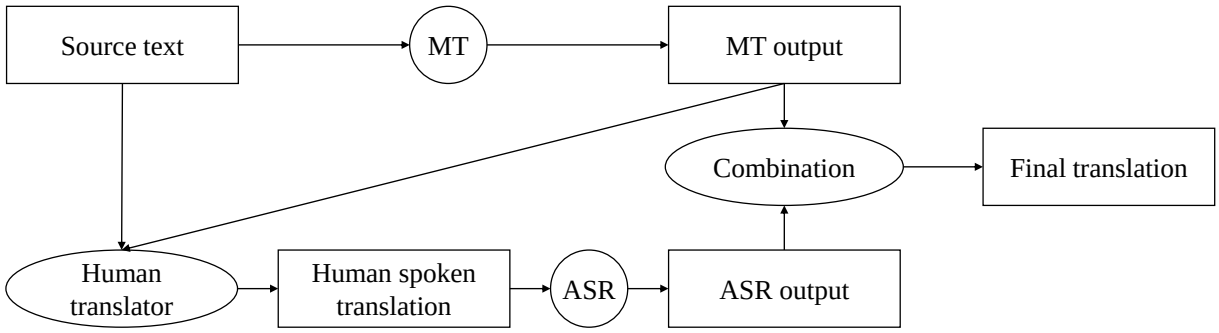


Figure 1: Workflow of the proposed system: the human translator views the source text and the MT output and speaks a corrected translation. The hypotheses of MT and ASR are combined into a final, improved translation.

The results of the system depend on the initial MT and ASR quality. Therefore, we will experiment with several ASR quality levels. Indeed, while state-of-the-art MT performance for a given language and domain can be considered quite stable, the ASR quality typically varies significantly with the speaker, microphone, surrounding noise, and computing resources. Moreover, we will assume that the human post-editor utters the entire corrected sentence, though in future work we will consider the possibility of uttering only a corrected fragment.

In this section, we explain how we make use of the TED data to train and test our proposal (3.1). Then, we present the third-party systems for MT and ASR used in our framework (3.2), along with a method to simulate ASR noise (3.3).

### 3.1. Data: TED Talks and Transcripts

One of the main experimental challenges is the availability of sufficient data, without the need to conduct costly experiments to collect it from human post-editors through recording and manual transcription. We propose thus the following approach.

We take advantage of the TED talks, available from `www.ted.com`, which are audio-visual recordings of prepared speeches, in English, on scientific, artistic and social topics, for a typical duration of 18 minutes. The speeches are transcribed, verified, and translated into a variety of languages by volunteers. TED talks data has been used in the IWSLT campaigns (Federico et al., 2014) and is distributed as a parallel corpus to train and test MT systems (Cettolo et al., 2012) or recommender systems (Pappas and Popescu-Belis, 2015). In our experiments, we use 20 talks (around 1200 sentences) obtained directly from the TED website.[1]

To emulate spoken post-editing, we consider each utterance spoken in English as a potential correction of the translation into English of a foreign-language version. Here, we use the French version as our source language. A French sentence is thus translated by our MT system, and then it is "corrected" by the English speaker, who utters exactly the reference translation.[2] In this way, the English transcript serves as a reference to MT, and allows us to score the output of our approach, in comparison to MT or ASR alone, using the BLEU score (Papineni et al., 2002).

As the speaker utters the exact reference sentence, if there were no ASR mistakes, the best strategy would be to consider the ASR output as the final translation (thus scoring 100% BLEU), simply discarding the MT output. However, ASR is far from perfect, and we therefore experimented with several levels of ASR accuracy to study the combination of ASR and MT. In such cases, our experimental framework and data set provide a testbed for ASR/MT combination algorithms.

### 3.2. MT and ASR Systems

We consider two MT systems. The Moses phrase-based statistical MT system (Koehn et al., 2007) was trained and tuned on the French/English parts of the Europarl corpus (Koehn, 2005), with an English language model generated using SRILM (Stolcke, 2002). The system's parameters were optimized using the Minimum Error Rate Training procedure (Och, 2003). However, we found that our Moses system was outperformed by the online Google Translate service (available at `http://translate.google.com`). Therefore, we will use the output of Google Translate in all the experiments presented below.

For ASR, we used the output over the TED English audio of two systems: (1) the baseline output provided by the IWSLT organizers (available at `http://workshop2013.iwslt.org/59.php`) from a system with around 15% word error rate, and (2) the ASR system presented by the University of Edinburgh at the IWSLT 2014 evaluation (Bell et al., 2012).

Figure 2 shows an example sentence from a TED talk, first in French and then in English, along with the corresponding ASR and MT outputs (respectively from the UEdin ASR and Google Translate).

[1] For training, we used 15 talks, with the following index numbers: 69, 93, 779, 789, 785, 790, 792, 799, 805, 837, 1090, 1100, 1131, 1133, and 1143. For testing, we used five talks: 227, 531, 535, 767 and 769. The actual URL where each talk can be viewed is obtained by appending the index number to the following prefix: `http://www.ted.com/talks/`.

[2] Following common practice in MT research, we do not attach importance to the fact that the sentence was originally generated in English, and that our French "source" is actually a translation. Instead, we take advantage of this situation to use the original English speech as a correction of the FR/EN machine translation.

**Source Sentence:** Vous savez, un des plaisirs intenses du voyage et un des délices de la recherche ethnographique est la possibilité de vivre parmi ceux qui n'ont pas oublié les anciennes coutumes, qui ressentent encore leur passé souffler dans le vent, qui le touchent dans les pierres polies par la pluie, le dégustent dans les feuilles amères des plantes.

**Reference Translation:** You know, one of the intense pleasures of travel and one of the delights of ethnographic research is the opportunity to live amongst those who have not forgotten the old ways, who still feel their past in the wind, touch it in stones polished by rain, taste it in the bitter leaves of plants.

**ASR:** intense pleasures of travel and one of the delights of ethnographic research is the opportunity to live amongst those have have not forgotten the old ways, to still feel their past the wind, touch and stones caused by rain, i tasted the bitter leaves of plants.

**MT:** You know, one of the intense pleasures of travel and one of the delights of ethnographic research is the ability to live among them who have not forgotten the old ways, still feel their past blowing in the wind, affecting the smooth stones in the rain, eaten in the leaves bitter plants.

Figure 2: A French source sentence, its reference translation, the speech recognition output from the English audio, and the MT output from the French sentence.

### 3.3. Simulating Variable ASR Error Rates

To study the robustness of our proposal with respect to the ASR error rate, which can vary greatly depending on the context of use, we experiment with several error rates, which were not, however, obtained from various systems (because this would be unpractical and hard to control), but rather using a simulation technique reproduced from Habibi and Popescu-Belis (2015). We introduce into the English reference transcripts three different types of simulated ASR noise: insertion, deletion, or substitution. We select random words from the text, and for each of them we perform one of the following edits: (1) *insert* a new word (randomly chosen from a dictionary) after the selected word; (2) *delete* the selected word; (3) *substitute* the selected word with a new word (again randomly chosen from a dictionary). These edits are performed on all occurrences of each selected word. The percentage of noise (i.e., word selection rate) is varied from 0.5% to 10%, which results in word error rates (WER) from 85% to 55%, as shown in Table 1 for a 3600-word sample. Overall, WER increases with the number of edits increases, though not linearly.

| **Number of edits** | 20 | 50 | 100 | 250 | 500 |
|---|---|---|---|---|---|
| **Errors in text (%)** | 0.4 | 1 | 2 | 5 | 10 |
| **Accuracy** | 85.1 | 81.2 | 79.2 | 69.6 | 54.5 |

Table 1: Percentage of errors in the text and accuracy scores (complement to 100% of word error rates) of degraded ASR output, over a 3600-word sample transcript, when varying the number of edits.

## 4. Methods for Merging ASR and MT

We define, implement, and evaluate two methods for merging the output of ASR and MT, with the goal of reducing the number of errors (i.e. differences from the reference translation) in the final merged output. The methods rely on the word-alignment of ASR and MT outputs, followed by a word selection procedure, for which we designed a heuristic-based algorithm and one based on machine learning. To perform ASR/MT alignment, a dynamic programming algorithm is used, from the HResults algorithm in the HTK toolkit (Young et al., 2006). Figure 3 shows an example of an alignment made between the ASR and MT output.

### 4.1. Heuristic-Based Approach

Starting from the word-level alignment between ASR and MT, we define several heuristics to decide which word (either from ASR or from MT) to select for the final merged output. The challenge is to select the word which is more likely to be correct, i.e. identical to the reference translation, without any particular knowledge of this reference. The heuristics loop over the pairs of aligned words, noted as $(W_{\text{ASR}}, W_{\text{MT}})$. The selection problem occurs only when the two words are different – when they are the same, this word is always selected. A specific case to consider is when one of the words is aligned with an empty word (noted $\emptyset$) in the other stream (but empty-to-empty alignments cannot occur).

The first heuristic (noted 'CONFIDENCE') selects the ASR word when the ASR confidence is higher than 0.8, while the second heuristic (noted 'CONFIDENCENODUPLICATES') adds rules to avoid selecting words that have neighboring duplicates, as follows:

- CONFIDENCE: if $W_{\text{ASR}} \neq \emptyset$ and $\text{Conf}_{\text{ASR}}(W_{\text{ASR}}) \geq 0.8$ then select $W_{\text{ASR}}$, else select $W_{\text{MT}}$.
- CONFIDENCENODUPLICATES: First, follow the same procedure as CONFIDENCE up to the end of the sentence. However, do not select in the resulting sentence a word which has neighboring duplicates, as follows. Let $W(-/+)n$ be the $n$-th word before/after the current word $W$. If $W \in \{W(-3), W(-2), W(-1), W(+1), W(+2), W(+3)\}$ then do not select $W$ in the resulting sentence.

We exemplify the action of each heuristic on the fragment shown in Figure 3. The CONFIDENCE heuristic does not select the ASR word "those" because it has a low ASR confidence score, but selects the MT word "them" instead. The output sentence is thus: "opportunity to live amongst them

```
MT:          ability#### to live amongst them# who have #### not
ASR:         opportunity to live amongst those ### have have not
Confidence:  .8         .8  .9   .7      .5       .8   .9   .8
```

Figure 3: English MT and ASR outputs aligned at the word level using HTK (excerpt from sentences in Figure 2). The third line indicates the confidence score of the ASR system for each word.

who have have not". The output of CONFIDENCENODUPLICATES is similar, except that the second occurrence of the word "have" is removed, which improves the result.

### 4.2. Machine Learning Approach

To leverage a larger set of features and find an optimal set of parameters, we apply supervised learning to perform the task of selecting one output word between the ASR and MT hypotheses, using the word-aligned representation shown in Figure 3, when these input words differ. In order to perform classification, a set of instances for training and testing were extracted, along with a set of features for each of them, inferred from the ASR and MT outputs. These features are noted using self-explanatory names as follows: ASR_has_characters, MT_has_characters, ASR_is_function_word, MT_is_function_word, Longer(ASR, MT), Levenshtein(ASR, MT), and Confidence(ASR). The latter two features are numeric, while the others are Boolean. Moreover, we use the lexical features ASR_word and MT_word but only include words with a frequency above 15; the other words are coded as 'OTHER'.

Since the new sentence is generated by choosing from each pair either the ASR word, or the MT word, or no word at all, the three possible classes (decisions) for the learning process are labeled as, respectively, 'ASR', 'MT' and 'NONE'. To construct the training data, we choose the class for each instance (word pair) as follows, after word-aligning the ASR and MT outputs with the reference translation.[3] If the aligned ASR word is equal to the word in the reference, then we label the pair with 'ASR'. Else, if the aligned MT word is equal to the reference, then we label the pair with 'MT'. If neither the ASR word nor the MT word match the reference word, we label the pair with 'NONE'. As for the heuristic-based method, we are only interested in instances where the ASR and MT words differ – if they are the same, that word is always selected.

For training a classifier, we experiment with several classification algorithms implemented in the WEKA toolkit (Hall et al., 2009): C4.5 Decision Trees, Support Vector Machines (SVMs), Random Forests, Decision Tables, and the Naive Bayes classifier.

## 5. Results of Experiments

In this section, we first present the results of the heuristic-based approach, on actual and simulated ASR output as defined in Sections 3.2 and 3.3, and then perform similar experiments with the machine learning based approach, with the training and testing data presented in Section 3.1.

[3]This three-way alignment is performed as follows: first we align separately the ASR output with the reference, and the MT output with the reference. Then these alignments are merged, using the reference words as a common ground.

### 5.1. Heuristic-Based Approach

We tested the heuristic-based approach with the baseline ASR output from the IWSLT workshops and MT from Google Translate. Table 2 shows the BLEU scores of the ASR and MT systems considered independently, and the scores of their combination using the CONFIDENCE heuristic. The results show that the CONFIDENCE heuristic clearly outperforms the use of MT only ($p < 0.001$). However, although CONFIDENCE has higher scores than those of the ASR correction only, the difference is not statistically significant over the five test talks.

| | **ASR** | **MT** | **CONFIDENCE** |
|---|---|---|---|
| Average BLEU | 64.4 | 45.26 | 66.7 |
| STD | 5.24 | 5.38 | 5.76 |
| $p$-value wrt. CONF. | 0.93 | 0.0001 | — |

Table 2: Average BLEU scores over five TED talks for the ASR output over the spoken correction (IWSLT baseline with a WER of about 15%), for the MT output (Google Translate), and for their combination using the CONFIDENCE heuristic.

When varying the quality of the ASR systems, we found that the CONFIDENCE and CONFIDENCENODUPLICATES heuristics have similar scores, and outperform the ASR only when its accuracy is considerably low; however, in that case, the use of MT alone offers a better option. In other words, *a priori* knowledge of the accuracies of ASR and of MT leads to simply selecting the output of one or the other, depending on their accuracy, with no need for a combination method.

### 5.2. Machine Learning Approach

Table 3 shows the obtained classification accuracy for various levels of ASR noise, in terms of number of correctly classified instances, for several types of classifiers, with 10-fold cross-validation over the training set. The features used were those listed above, except for the ASR confidence, which is not available for simulated ASR errors. The SVM classifier outperformed the others, and is therefore be used in the other experiments presented below.

Once the classifiers were trained on the training data and the optimal feature set was identified, the classifications were performed on the test set. We calculated the BLEU scores of the newly generated sentences using varying ASR error levels, and averaged the BLEU scores over the five TED talks of the test set, also computing confidence intervals over the 5-talk sample.

Figure 4 shows the mean BLEU scores of the ASR output, Google Translate output and the newly generated sentence

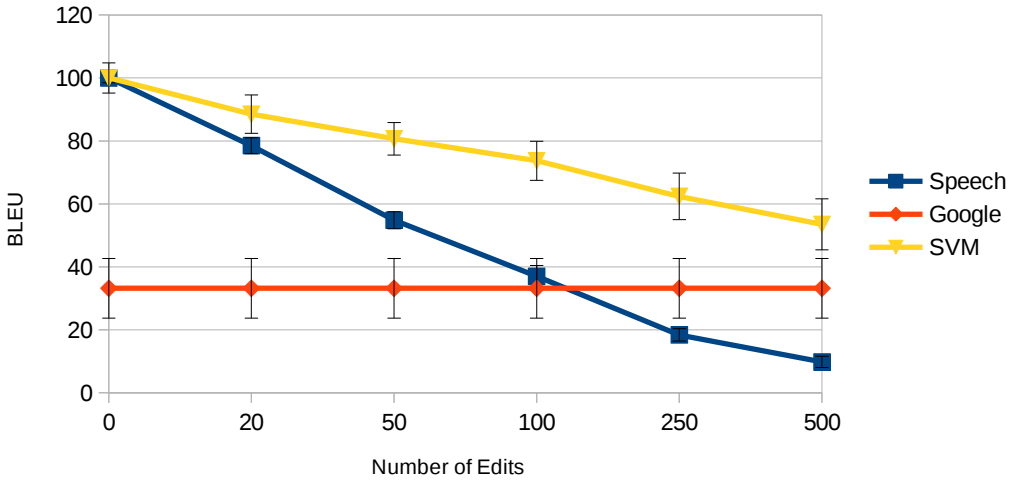


Figure 4: Mean BLEU score for various error levels of the ASR (in terms of number of edits, see Table 1) for three outputs: ASR only, MT only, and ASR combined with MT using the SVM classifier. The vertical error bar depicts the 95% confidence interval for each data point. Overall, the combination of ASR and MT outperforms both ASR and MT considered alone, except for the smallest ASR error levels.

| Percentage of errors in ASR | Classifier | | | |
|---|---|---|---|---|
| | J48 | N.Bayes | R.Forest | SVM |
| 0.4% | 99.4 | 99.5 | 99.6 | 99.7 |
| 1% | 97.5 | 97.6 | 97.9 | 97.9 |
| 2% | 96.1 | 96.3 | 96.3 | 96.8 |
| 5% | 93.2 | 93.3 | 93.4 | 93.5 |
| 10% | 91.0 | 91.4 | 91.5 | 91.8 |

Table 3: Accuracy (number of correctly classified instances) for several classifiers and varying levels of ASR noise.

| Method | Talk ID | | | | |
|---|---|---|---|---|---|
| | 227 | 531 | 535 | 767 | 769 |
| ASR | 72.7 | 62.6 | 75.0 | 58.4 | 65.01 |
| MT | 36.4 | 43.7 | 26.8 | 39.9 | 39.76 |
| C4.5 Dec. Trees | 71.5 | 62.2 | 75.0 | 59.5 | 64.47 |
| SVM | 73.5 | 63.2 | 77.0 | 59.0 | 64.18 |
| SVM + FW | 73.4 | 63.2 | 77.1 | 59.1 | 65.38 |

Table 4: BLEU scores for combination methods (along with ASR and MT alone) over five TED talks. The scores of the last line (SVM plus a heuristic favoring function words from MT) outperform both ASR and MT considered independently (first two lines).

(with 95% confidence intervals) against different quality levels of the ASR. When there is no ASR noise, the classifiers generates a sentence that is identical to the ASR sentence. As ASR noise levels increase, the BLEU score of the combination method remains higher than both the ASR output and the MT output, and this difference increases when the ASR quality decreases considerably. The combination method thus leads to an effective solution for spoken post-editing of MT, which outperforms both the non-edited MT and the voice-dictated translation.

Table 4 displays the scores of several machine learning methods along with ASR of spoken corrections (from the UEdin system) and MT (from Google Translate) for five TED talks. An error analysis on one of the TED talks (n. 769) shows that most function words of the MT translation are unaligned with an ASR word. Therefore, we experiment with inserting into the output translation of the combined system the MT words which (1) are function words and (2) are not aligned with an ASR word. This additional post-processing step, noted "SVM + FW" in Table 4, improves BLEU score in TED talks n. 535, 767, and 769.

To determine the optimal feature subsets, we performed feature selection on the training data with 10-fold cross-validation. The set with the highest information gain included the following features: Confidence(ASR), Longer(ASR, MT), ASR_has_characters and MT_has_characters. When compared to the heuristic-based method, the Confidence(ASR) feature plays an equally important role for machine learning approaches. The features ASR_has_characters and MT_has_characters fill up the requirement of checking for non-null words in the heuristics approach. Other features such as ASR_is_function_word and MT_is_function_word, Levenshtein(ASR, MT) have a lesser importance in the classification process.

## 6. Related Work

The combination in sequence of ASR and MT for spoken language translation has been extensively studied – see for instance the proceedings of the IWSLT workshop series started in 2004 (Federico et al., 2014). The parallel combination of ASR and MT, as it occurs in more particular contexts of use, has been studied comparatively less often. An example of parallel integration was presented by Khadivi and Ney (2008), who used different MT models for rescoring ASR n-best lists. As an alternative to n-best list rescoring, and to provide a tighter integration, they also used ASR word graphs, while Matsoukas et al. (2007) rescored word lattices when translating them. Khadivi et al. (2005) ex-

plored integrating speech recognition and translation models for automatic dictation. Similar experiments were also conducted by Rodríguez et al. (2012), using SMT model probabilities to update statistical language model probabilities in ASR, and achieving low computational complexity and error rate in the output.

The TransTalk project proposed to use the ASR and MT models together to generate a better translation (Dymetman et al., 1994; Brousseau et al., 1995). Since the speech recognizer has access to the source text as well as the spoken translation, the statistical translation model guides the recognition process in this work. The application of translation models was conducted before, during and after speech recognition, resulting in higher speed and accuracy. Reddy et al. (2007) experimented with a dictation system, using the combined SMT and ASR statistical models. The most significant performance improvement was obtained by rescoring ASR lattices from an initial recognition pass with a language model trained from the SMT output for the given document and speaker. A different approach was taken by Reddy and Rose (2010), where translation probabilities derived from SMT, along with named entity tags derived from named entity recognition, were used with acoustic phonetic information obtained from an ASR system. Reddy and Rose (2008) addressed issues related to task-independent ASR, by including domain information from the document in the form of named entity labels.

Instead of using n-best rescoring approaches, Khadivi et al. (2006) unified MT models and ASR models using finite state automata, which they claim is more suitable for a real-time prediction engine. Also, in a different manner, Matusov et al. (2005) exploited word lattices of ASR hypotheses as input to the translation system based on weighted finite-state transducers.

Interactive machine translation (IMT), where a human expert is integrated in the process of automatic translation, has been considered by a number of previous studies. In IMT, a human expert interacts with a system by partially correcting the errors of the system's initial output. Then, the system proposes a new solution, and the process can be repeated until the output meets the desired quality. Alabau et al. (2011) used keyboard, mouse and speech as input modalities and reported a significant performance boost in speed and quality. Vidal et al. (2006) followed a similar approach: the human translator utters part of a prefix of the final target sentence, and then either amends or validates it until the end of the sentence is reached. Using an interactive predictive process to correct the system generated errors was explored by Khadivi and Vakil (2012): an ASR n-best list is rescored by using translation models, thereby achieving better results. Ortiz-Martínez et al. (2012) presented a translator's workbench which takes into account the cognitive processes involved in human translation. The workbench determines what type of assistance is offered to the translator, and takes input from user by means of keyboard, mouse or e-pen.

Our work differs from the previously conducted work in two ways. On the one hand, we aim mainly at disseminating translated information when a keyboard is not available, mainly aiming at scenarios like mobile applications. On the other hand, in contrast to the integration methods conducted at the model level, we use surface level information from the ASR and MT outputs to generate the new translations. Our results show that a minimal amount of information is sufficient to improve the translation significantly.

## 7. Conclusion and Future Work

In this paper, we explored the possibility of applying spoken corrections to machine translation in order to generate a better translation. Additionally, we presented a framework to re-use the TED talks (audio, transcript and translation) to evaluate this integration, in particular using ASR error simulation to test our methods in various conditions. In contrast to the previous work conducted in this area involving integrating ASR and MT models, we reached an improvement by using the word level information from both outputs. Our results show that using machine learning approaches to generate a new translation out of the ASR and MT outputs is a successful approach. In the future, we expect to find more effective features and methods that could generate more efficiently an improved translation output.

The proposed methods can be applied to any setting where spoken correction of a written text is desired, for instance when correcting short messages dictated to smart personal assistants, especially smartphones or smartwatches, where keyboard-based correction is not practical or feasible. These methods can also be extended to improve the written translation of a spoken discourse (e.g. a lecture) using interpreted speech, when available, assuming that no transcripts are available.

In all the experiments presented above, the translator who spoke the translation was assumed to have seen the reference translation. We aim to expand our research by relaxing this assumption. In future experiments, the translator will translate or correct each sentence without looking at the reference translation, but only at the MT output. We will thus explore how useful the spoken translation is to improve the translation performance, when its content is not influenced by the reference translation. Furthermore, the translator will correct the MT by speaking out only the fragments of the sentence which need to be corrected. This way, we will investigate the performance improvement obtained by providing minimal spoken translation input.

## 8. Acknowledgments

The major part of this work was carried out while the first author was a visitor at the Idiap Research Institute. This work was partly supported by the Swiss National Science Foundation through the MODERN Sinergia project (`www.idiap.ch/project/modern`). We are grateful to ILCC/CSTR, University of Edinburgh, for the output of their IWSLT 2014 ASR system.

## 9. Bibliographical References

Alabau, V., Rodríguez-Ruiz, L., Sanchis, A., Martínez-Gómez, P., and Casacuberta, F. (2011). On multimodal interactive machine translation using speech recognition. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)*, pages 129–136, Alicante, Spain.

Bell, P., Swietojanski, P., Driesen, J., Sinclair, M., McInnes, F., and Renals, S. (2012). The UEDIN ASR systems for the IWSLT 2014 evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 26–33, Lake Tahoe, CA.

Brousseau, J., Drouin, C., Foster, G., Isabelle, P., Kuhn, R., Normandin, Y., and Plamondon, P. (1995). French speech recognition in an automatic dictation system for translators: the TransTalk project. In *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT$^3$: Web inventory of transcribed and translated talks. In *Proceedings of the 16$^{th}$ Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Dragsted, B., Mees, I., and Hansen, I. G. (2011). Speaking your translation: students' first encounter with speech recognition technology. *Translation & Interpreting*, 3(1):10–43.

Dymetman, M., Brousseau, J., Foster, G. F., Isabelle, P., Normandin, Y., and Plamondon, P. (1994). Towards an automatic dictation system for translators: the TransTalk project. In *Proceedings of the 3rd International Conference on Speech and Language Processing (ICSLP)*, pages 691–694, Yokohama, Japan.

Federico, M., Stücker, S., and Yvon, F. (2014). *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA.

Habibi, M. and Popescu-Belis, A. (2015). Keyword extraction and clustering for document recommendation in conversations. *IEEE Transactions on Audio Speech and Language Processing*, 23(4):746–759.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorer Newsletter*, 11(1):10–18.

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75.

Khadivi, S. and Ney, H. (2008). Integration of automatic speech recognition and machine translation in computer-assisted translation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(8):1551–1564.

Khadivi, S. and Vakil, Z. (2012). Interactive-predictive speech-enabled computer-assisted translation. In *Proceedings of the International Workshop on Spoken Language Technology (IWSLT)*, pages 237–243, Hong-Kong, China.

Khadivi, S., Zolnay, A., and Ney, H. (2005). Automatic text dictation in computer-assisted translation. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH)*, pages 2265–2268, Lisbon, Portugal.

Khadivi, S., Zens, R., and Ney, H. (2006). Integration of speech to computer-assisted translation using finite-state automata. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL), Poster Sessions*, pages 467–474, Sydney, Australia.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbs, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Matsoukas, S., Bulyko, I., Xiang, B., Nguyen, K., Schwartz, R. M., and Makhoul, J. (2007). Integrating speech recognition and machine translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1281–1284, Honolulu, HI.

Matusov, E., Kanthak, S., and Ney, H. (2005). On the integration of speech recognition and statistical machine translation. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH)*, pages 3177–3180, Lisbon, Portugal.

Mesa-Lao, B. (2014). Speech-enabled computer-aided translation: A satisfaction survey with post-editor trainees. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 99–103, Gothenburg, Sweden.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Ortiz-Martínez, D., Sanchis-Trilles, G., Casacuberta, F., Alabau, V., Vidal, E., Benedí, J.-M., González-Rubio, J., Sanchis, A., and González, J. (2012). The CASMACAT project: The next generation translator's workbench. In *Proceedings of the VII Jornadas en Tecnología del Habla and the III Iberian SLTech Workshop (IberSPEECH)*, pages 326–334, Madrid, Spain.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Pappas, N. and Popescu-Belis, A. (2015). Combining content with user preferences for non-fiction multimedia recommendation: A study on TED lectures. *Multimedia Tools and Applications (MTAP)*, 74(4):1175–1197.

Reddy, A. M. and Rose, R. C. (2008). Towards domain independence in machine aided human translation. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2358–2361, Brisbane, Australia.

Reddy, A. M. and Rose, R. C. (2010). Integration of statistical models for dictation of document translations in a machine-aided human translation task. *IEEE Transactions on Audio, Speech and Language Processing*,

18(8):2015–2027.

Reddy, A. M., Rose, R. C., and Désilets, A. (2007). Integration of ASR and machine translation models in a document translation task. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2457–2460, Antwerp, Belgium.

Rodríguez, L., Reddy, A. M., and Rose, R. C. (2012). Efficient integration of translation and speech models in dictation based machine aided human translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4949–4952, Kyoto, Japan.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, CO.

Vidal, E., Casacuberta, F., Rodriguez, L., Civera, J., and Hinarejos, C. (2006). Computer-assisted translation using speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):941–951.

Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book Version 3.4*. Cambridge University Press, Cambridge, UK.