”Speech is silver, but silence is golden”: improving speech-to-speech translation performance by slashing users input

Frederic Bechet, Benoit Favre, Mickael Rouvier

Aix Marseille Universite, CNRS-LIF

Abstract

Speech-to-speech translation is a challenging task mixing two of the most ambitious Natural Language Processing challenges: Machine Translation (MT) and Automatic Speech Recognition (ASR). Recent advances in both fields have led to operational systems achieving good performance when used in matching conditions with those of ASR and MT models training. Regardless of the quality of these models, errors are inevitable due to some technical limitations of the systems (e.g. closed vocabulary) and intrinsic ambiguities of spoken languages. However all ASR and MT errors don't have the same impact on the usability of a given speech-to-speech dialog system: some can be very benign, unconsciously corrected by users, some can damage the understanding between users and eventually lead the dialog to a failure. We present in this paper a strategy focusing on ASR error segments that have a high negative impact on MT performance. We propose a method that consists firstly in automatically detecting these erroneous segments then secondly estimating their impact on MT. We show that removing such segments prior to translation can lead to a significant decrease in translation error rate, even without any correction strategy.

Index Terms: speech-to-speech translation, confidence measures, OOV words, dependency parsing.

10.21437/Interspeech.2015-495

1. Introduction

In the domain of Speech-to-Speech Translation, not all Automatic Speech Recognition (ASR) and Machine Translation (MT) errors have the same impact on the usability of a given speech-to-speech dialog system: some can be very benign, unconsciously corrected by users, some can damage the understanding between users but can be corrected by adding some clarification sub-dialogs, finally some can lead the dialog into a dead-end, preventing the users to even know what to clarify for continuing the dialog. This issue of predicting what could be the impact of an error on the speech-to-speech process is the framework of this paper. We will focus here on detecting and characterizing *dubious* word segments that might have an impact on MT.

The applicative framework is a speech-to-speech system developed through the DARPA BOLT project [1]. The speech-to-speech system developed by the ThunderBOLT team, led by SRI, includes a dialog manager that can trigger a clarification dialog if an ASR erroneous segment is detected [1]. This system obtained good results on the 2012 and 2013 BOLT evaluations. However its drawback was to have a tendency to overgenerate clarification dialogs, slowing the interaction between users. We have shown in [2] that it was possible to retrieve the syntactic dependency structure of erroneous ASR transcriptions in order to ask more meaningful questions in the clarification dialogs.

This paper shows that the same kinds of features can be used in order to predict the impact of an ASR error segment on MT in order to prevent triggering a clarification dialog if no clear impact on MT is measured. Moreover we show in this paper that it is possible to reduce translation errors simply by removing these *dubious* segments. Following the principle that it is better to remain silent rather than translating erroneous words that can lead a dialog into a dead-end, we present a strategy that predicts the impact, positive or negative, of the deletion of a word segment on the Translation Error Rate (TER).

2. Related work

The relation between ASR and MT has been widely studied in the context of speech-to-speech translation. Most of the studies have focused on the coupling of ASR and MT processes [3, 4, 5, 6, 7]. They highlight the link between ASR and MT errors and show that an improvement in translation performance can be obtained by tuning ASR specifically for minimizing MT errors rather than ASR errors. This coupling is usually done by a rescoring of an ASR word lattice.

Another aspect of the relation between ASR and MT is the study of word errors produced by the two processes. In [8], the MT and ASR errors are categorized according to their Part-Of-Speech tags. It is interesting to notice that the largest part of both word error rates comes from nouns and verbs classes. In [9] it is noticed that the most important source of ASR errors are substitution errors (almost 60% of the total errors). However most of these substitution errors are due to morphological changes of the words such as plural/singular substitution, but the root form remains the same, limiting the impact on MT. Therefore not all ASR errors will have the same impact of MT, even for noun or verb substitution errors.

Detecting automatically errors in order to prevent a misunderstanding that could lead to a dialog failure in speech-to-speech translation application is a field of research directly related to this study. [10] shows that linguistic features can improve error detection in addition to ASR features. Within the context of the DARPA BOLT program, [11] and [12] present two studies on the detection of OOV errors using confusion network and neural network with a large set of ASR and lexical features. We have presented in [13] a similar study on the detection of error segment hypotheses that can be corrected with an interactive clarification dialog. The whole system is described in [1].

This current paper is a follow-up to this previous study on the localization of erroneous segments and their characterization in terms of syntactic nature and role [2]. The main originality of this work is to propose a strategy that can tune the error detection strategy directly on the minimization of translation errors rather than ASR errors. We show that this strategy can help reducing translation error rate by removing dubious word segments, even with no clarification dialogs.

3. Characterizing ASR errors

There are multiple sources of errors in ASR transcriptions, such as lexical and Language Model ambiguities (e.g. *I ran/ Iran*), Out-Of-Vocabulary words replaced by known words (e.g. *priest in / pristine*), non-canonical word pronunciation, voices differing from the training database (accent, voice quality, age, pathology), noise or sub-optimal search due to real-time constraints.

These different sources of errors lead to three kinds of errors in the automatic transcriptions: insertions, deletions and substitutions. It is interesting to characterize further these errors in order to predict their impact on the downstream processes, such as MT. When dealing with open domain applications, it is difficult to use semantic constraints in order to estimate this impact. One way to characterize them independently of any applicative context is to measure how an ASR errors can disrupt the syntactic structure of a sentence. With this criterion, an error on a verb will have a bigger impact than an error on an adjective for example.

To measure this impact we adopt the following process:

- the reference transcriptions of our corpus are tagged with a POS tagger and parsed with a syntactic dependency parser
- the same process is applied to the automatic transcriptions
- reference and automatic transcriptions are aligned at the word level
- this alignment is propagated at the POS and dependency level
- the strings of POS and dependency symbols in both reference and automatic transcriptions are then compared in order to compute an error rate.

We did this experiment on a corpus of spoken conversations collected during the DARPA BOLT project. We consider here the English-Iraqi Arabic speech-to-speech translation task presented in [1]. Only the English ASR side is considered in this paper. The ASR system used is the SRI *Dynaspeak* system [14] adapted to the task.

Table 1 presents the 10 most frequent errors at the word, POS and syntactic dependency level. Most of the errors at the word level are concentrated on small syntactic words. This is expected as they are short words, monosyllabic, very easily inserted or deleted by the ASR decoder, and also the most frequent words in the English language. The POS and dependency levels are more interesting. At the POS level, 17.1% of the ASR errors are nouns (NN+NNP) and 27.2% are verbs (VBP+VB+VBD), so about 44% of the errors are on *important* words (words not belonging to a potential stop-list). The same phenomenon can be seen at the syntactic dependency level: over 30% of the errors are on verbal dependencies (OBJ+SBJ). Since nouns and verbal dependencies are clearly crucial elements in the translation process of a sentence, it is important to detect such ASR errors before sending the transcription to the MT module. This detection process is presented in the next section.

4. Detecting and removing dubious word segments

This section presents the methodology proposed for detecting and estimating the impact of an ASR error on the MT task. This process consists first in detecting ASR errors, then characterizing each error segment hypothesis (*dubious segments*) according to its syntactic role within the sentence, and finally using prediction and characterization features to estimate its impact on the machine translation of the sentence. This last step is done using a classifier integrating directly MT performance (Translation Error Rate) as the objective function to optimize. The whole process is described in details in the next four subsections.

4.1. Word confidence measure

The first step in this process consists in estimating an ASR confidence measure for each word. We use in this study the method presented in [14] and developed at SRI for estimating ASR word confidence. The confidence model is a neural network predicting a binary feature error/non-error. Typical input features include maximum/mean/standard deviation of word posteriors from a confusion network produced by several decoders to exploit an observation that different ASR may hypothesize differently within an error region.

4.2. Sequence tagging for obtaining error segments

From these word-level confidence measures we use the method presented in [13] for obtaining error segments. This method uses a Conditional Random Field (CRF) tagger taking as input features, in addition to the discretized values of word confidence measures, lexical and syntactic features. The main intuition in using a sequence labelling tagger is that a misrecognized word often generates a sequence of ASR errors in the automatic transcription, especially when dealing with OOV errors. It was shown in [13] on a BOLT corpus containing a high density of OOV words that, on average, each misrecognized OOV generates a sequence of 4.8 erroneous words. The tagger and syntactic parser we use to provide the syntactic features to the CRF come from the MACAON tool suite [15].

The error segment CRF decoder we use is based on OpenFST and is also part of MACAON. It outputs multiple hypotheses represented as a transducer of possible tagged sequences: words are input labels, and error/non-error labels are output labels. By composing this transducer with a filter automaton that allows only one error segment per sentence, and enumerating the top n best paths, we obtain an n-best list of tagged sentence each containing exactly one error segment hypothesis. We need to have only one error segment for each hypothesis because we want now to estimate the potential impact of this segment on the MT process for the whole sentence, independently from the other error segments that can be found. But before estimating this impact we characterize each error segment according to its syntactic role in the sentence.

4.3. Characterizing error segment hypotheses

Following [13], we characterize each error segment hypothesis by replacing it with a dummy symbol, *XX*. We run the POS tagger and syntactic parser on the modified sentence in order to retrieve the syntactic role of symbol *XX*. The syntactic models have been retrain in order to predict a category and a dependency when facing this symbol. By extracting features about the syntactic structure of the sentence with and without the error segment as well as the nature and role of this segment as predicted by the syntactic model, the last step of this process is to predict the impact of deleting these dubious segments on the MT process.

word - **word error rate=5.6**										
word	THE	ARE	IN	'S	YOU	A	IS	DO	AND	CAN
%	3.8	3.8	2.8	2.6	2.5	2.3	2.2	2.2	1.8	1.6
part-of-speech - **POS error rate=4.6**										
POS	NN	VBP	VB	DT	VBD	IN	PRP	MD	NNP	WRB
%	12.1	11.9	8.4	7.5	6.9	6.5	6.2	5.0	4.1	3.7
syntactic dependencies - **dep error rate=10.0**										
DEP	OBJ	NMOD	SBJ	SUB	VC	PMOD	ADV	PRD	LOC	TMP
%	18.5	16.6	12.6	9.0	6.1	4.6	4.3	3.5	3.5	3.3

Table 1: 10 most frequent ASR errors at the word, POS and syntactic dependency levels in a corpus of English conversations with a WER of 5.6%

4.4. Predicting the impact of dubious segments

Each segment obtained with the previous processes can contain only ASR errors, a mix of ASR errors and valid words, or just valid words (false alarms). In ASR, a word is considered as an ASR error only if it differs from the reference transcription, regardless of its importance for the translation. For example the insertion of an '*a*' before a noun is harmless for the translation but it is considered as important as the substitution of a verb or a noun by the previous confidence models. To deal with this issue we train a classifier for estimating the impact of an error segment on a sentence rather than just considering its validity. This impact is modelled by computing the Translation Error Rate (TER) [16] between the automatic translation of the reference transcription, the translation of the ASR 1-best and the translation of the ASR 1-best without the error segment hypothesis.

We use the automatic translation of the reference transcription as our MT reference since we don't have a manual translation of our training corpus. Because we are dealing in this study with ASR errors, not MT errors, we consider that comparing automatic translations is a good approximation. Once both TER are computed, we can label each error segment with one of these three labels: *positive* if the TER is decreased by removing the error segment prior to translation; *neutral* if no change in TER is observed; *negative* if an increase of TER is observed.

A classifier (IcsiBoost [17]) is then trained to predict these labels, based on the following features: ASR confidence features, lexical and syntactic features on the ASR 1-best, nature and role predicted for the error segment, TER between the automatic translation of the original ASR 1-best and the same one without the error segment. A confidence score is attached to each label predicted.

For the sake of comparison, in the experiment section, we perform a comparative study by replacing the TER reduction criterion by the WER reduction criterion. We want to check the validity of using an objective function (TER) directly linked to the task (MT) rather than a general purpose one as WER. The strategy using the TER will be called **S-TER** and the one using WER will be **S-WER**.

4.5. Removing dubious segment

At decoding time the ASR 1-best hypotheses with word confidence measures are parsed in order to obtain the linguistic features needed by the CRF error segment tagger. An n-best list of sentences with one dubious segment per sentence is generated by the CRF. Each hypothesis is parsed in order to characterize the nature and role of the error segment, then the Icsiboost impact classifier is used to label the hypothesis as *positive*, *neutral* or *negative* w.r.t. the objective function chosen (WER or TER). For a given utterance, all dubious segments that have been labelled as *positive* or *neutral* with a confidence value above a given threshold are removed. By changing this threshold we can tune the risk of deleting a valid information or sending to the MT module erroneous words. A good operating point should aim at reducing the risk of missed detections by removing as many as possible dubious segments without impacting negatively the TER measure.

4.6. Example

Table 2 presents an illustration of each process on an example from the BOLT corpus. As we can see the ASR 1-best contains 3 erroneous words: *this candle is*, instead of the word *scandalous*. The CRF sequence tagger selects only *candle is* as the error segment. The predicted nature of this error segment is *noun* and its role is *modifier* of the word *issue*. By removing this segment and translating the slashed transcription, we obtain a better TER, therefore this error segment is labeled as *positive*.

type	**transcription**	**TER**
ref.	we have to avoid a scandalous issue	0%
1best	we have to avoid this *[candle is]* issue	10%
remove	we have to avoid this issue	6.6%

Table 2: Example of erroneous transcription with an error segment automatically detected (*candle is*) qualified as MT *positive*

5. Experiments

The experiments presented in this study have been done on an English-Iraqi Arabic speech corpus collected through the DARPA BOLT program and presented in section 3. We present results only for the English to Iraqi Arabic translation task. The ASR and MT modules used are those presented in [14] and [18]. We trained our CRF error detection model on a set of 11.2K English utterances. We followed a k-fold approach in order to use this same corpus to generate error segment hypotheses on which the impact classifier is trained. By generating a 20-best error segment hypothesis list for each utterance, we obtained a corpus of 200.4K sentences containing each one *dubious* segment. As presented in section 4, two classifiers have been trained to measure the impact of a dubious segment: one using the WER reduction as objective function (**S-WER**), the other one using the TER reduction (**S-TER**). The test corpus is made of 1879 English utterances corresponding to the evaluation data of phase 2 of the BOLT program. Each utterance is processed following the method described in 4.5.

The first evaluation is done at the word level. The task is to automatically detect erroneous words in the ASR 1-best. We compare 3 methods: *word conf* is simply a threshold on the

word confidence values provided by the ASR process; *segment conf S-WER* correspond to our selection method using WER as the objective function; *segment conf S-TER* is the same method with TER as the objective function. As we can see in figure 1, segment methods outperform word level confidence when recall is below 50%. This result confirms that methods taking into account the impact of an error on the whole sentence can improve accuracy. As expected the **S-WER** objective function is slightly better that **S-TER** as it is more directly linked to the decrease of errors in the ASR 1-best.

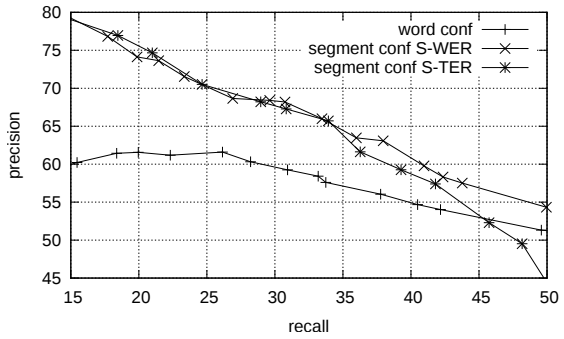

Figure 1: Precision/Recall curve on ASR error word detection obtained by tuning an acceptance threshold on confidence scores at the word or the segment level

transcription	wer	ter
***full* 1best**	6.6	13.4
1best w/o ASR errors (*oracle*)	4.9	9.8

Table 3: WER and TER evaluation of ASR 1-best with (***full* 1best**) and without (*oracle*) ASR errors (all true ASR errors are simply erased from the ASR 1-best)

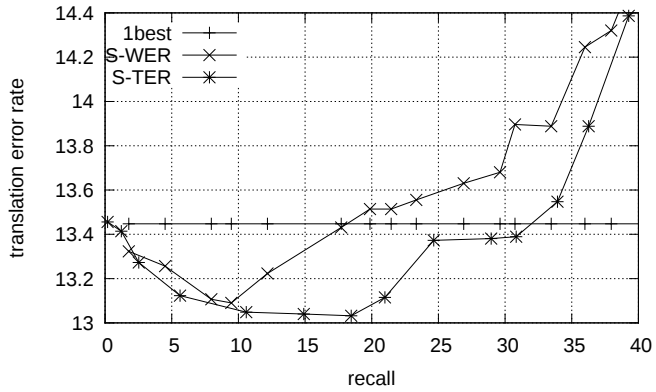

Figure 2: TER at several operating points of the segment deletion strategies

The second evaluation presents the variation of TER according to several operating points (rejection threshold) of our strategies. Table 3 shows the WER and TER of the 1-best ASR transcription (**1best**) and the **oracle** transcription corresponding to the same 1-best where all *true* erroneous words have been deleted. TER values are obtained by comparing the automatic translation from English to Iraqi Arabic of the reference transcription with the translation of the 1-best hypotheses. This approximation is acceptable here since we are focusing on ASR errors, not MT errors. However an evaluation with reference translations will be done when we will have the whole test corpus manually translated.

The **oracle** results validate the motivation of this paper: removing erroneous words can improve MT performance. An absolute gain of 3.6% (26.9% relative improvement) is achieved by removing true ASR errors prior to translation. The gain obtained with the fully automatic methods are much more modest, due to the difficulties of accurately detecting all ASR errors. Figure 2 shows how the TER evolves when the rejection threshold is tuned. As we can see, **S-TER** outperforms **S-WER** since an improvement in TER is observed from an operating point of 30% recall in erroneous word detection with **S-TER** and only 18% with **S-WER**. Since **S-TER** and **S-WER** have the same erroneous word detection performance at these operating points, according to figure 1, we can see that changing the objective function of the impact classifier toward a measure directly linked to MT performance is better than using a generic measure such as WER.

recall	S-WER	S-TER
30%	$27.8 \rightarrow 29.5 : +6.1\%$	$29.8 \rightarrow 29.5 : -1\%$
20%	$28.9 \rightarrow 29.3 : +1.3\%$	$32.8 \rightarrow 30.8 : -6.1\%$
10%	$37.3 \rightarrow 32.1 : -13.9\%$	$35.8 \rightarrow 31.4 : -12.3\%$

Table 4: Variation in TER at several operating points for strategies **S-WER** and **S-TER**

This result is confirmed by table 4 where **S-TER** reduces TER at all operating points, from 1% to 12.3% relative improvement. In this table TER is computed only on the utterances that have at least a segment deleted by **S-WER** or **S-TER**, unlike figure 2 where it was computed on the whole test corpus.

6. Conclusion

We have presented in this paper a strategy focusing on ASR error segments that have a high negative impact on MT performance. By directly modelling Translation Error Rate reduction in the objective function of our error segment classifier, we outperform a method based only on reducing Word Error Rate. Moreover, oracle results obtained on ASR transcriptions where all erroneous words have been deleted clearly validate the motivation of this paper, based on an intuitive observation: "speech is silver, but silence is golden", it is less risky to remove all dubious ASR word segments before translation rather than translating an error that could generate a misunderstanding between users of a speech to speech translation service.

7. Acknowledgements

This work was partially funded by DARPA HR0011-12-C-0016 as an AMU subcontract to SRI International.

8. References

[1] N. F. Ayan, A. Mandal, M. Frandsen, J. Zheng, P. Blasco, A. Kathol, F. Béchet, B. Favre, A. Marin, T. Kwiatkowski *et al.*, ""can you give me another word for hyperbaric?: Improving speech translation using targeted clarification questions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8391–8395.

[2] F. Bechet, B. Favre, A. Nasr, and M. Morey, "Retrieving the syntactic structure of erroneous asr transcriptions for open-domain spoken language understanding." in *ICASSP*, 2014, pp. 4097–4101.

[3] S. Stuker, M. Paulik, M. Kolss, C. Fugen, and A. Waibel, "Speech translation enhanced asr for european parliament speeches-on the influence of asr performance on speech translation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1293.

[4] R. Sarikaya, B. Zhou, D. Povey, M. Afify, and Y. Gao, "The impact of asr on speech-to-speech translation performance," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1289.

[5] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5632–5635.

[6] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation." in *INTERSPEECH*, 2005, pp. 3177–3180.

[7] P. R. Dixon, A. Finch, C. Hori, and H. Kashioka, "Investigation on the effects of asr tuning on speech translation performance," in *Proc. 8th International Workshop on Spoken Language Translation*, 2011, pp. 167–174.

[8] M. Popović and H. Ney, "Word error rates: decomposition over pos classes and applications for error analysis," in *Proceedings of the second workshop on statistical machine translation*. Association for Computational Linguistics, 2007, pp. 48–55.

[9] D. Vilar, J. Xu, L. F. dHaro, and H. Ney, "Error analysis of statistical machine translation output," in *Proceedings of LREC*, 2006, pp. 697–702.

[10] D. Xiong, M. Zhang, and H. Li, "Error detection for statistical machine translation using linguistic features," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 604–611.

[11] A. Marin, T. Kwiatkowski, M. Ostendorf, and L. S. Zettlemoyer, "Using syntactic and confusion network structure for out-of-vocabulary word detection." in *IEEE/ACL Spoken Language Technology workshop*, 2012, pp. 159–164.

[12] H.-K. Kuo, E. E. Kislal, L. Mangu, H. Soltau, and T. Beran, "Out-of-vocabulary word detection in a speech-to-speech translation system," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7108–7112.

[13] F. Béchet and B. Favre, "Asr error segment localization for spoken recovery strategy," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6837–6841.

[14] Y. C. Tam, Y. Lei, J. Zheng, and W. Wang, "Asr error detection using recurrent neural network language model and complementary asr," in *ICASSP 2014. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[15] A. Nasr, F. Béchet, J. Rey, B. Favre, and J. Le Roux, "Macaon: An nlp tool suite for processing word lattices," *Proceedings of the ACL 2011 System Demonstration*, pp. 86–91, 2011.

[16] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, 2006, pp. 223–231.

[17] B. Favre, D. Hakkani-Tür, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost.

[18] Y. Tam and Y. Lei, "Neural network joint modeling via context-dependent projection," in *ICASSP 2015. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015.