Improving Spoken Language Translation by Automatic Disfluency Removal : Evidence from Conversational Speech Transcripts

Sharath Rao, Ian Lane and Tanja Schultz

interACT, Language Technologies Institute,
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{skrao, ian.lane,tanja} @cs.cmu.edu

Abstract

Machine translation of spoken language has made significant progress in recent years, however, translation quality is still limited due to specific idiosyncrasies of spoken language; including the lack of well-formed sentences and the presence of disfluencies. In this paper, we investigate the effect of disfluencies on Statistical Machine Translation (SMT) and introduce an Automatic Disfluency Removal scheme as a pre-processing step prior to translation. On Broadcast Conversation (BC) transcripts the proposed approach demonstrates that up to 8% relative improvement in BLEU can be obtained via Automatic Disfluency Removal. Furthermore, we show that the detrimental effect of disfluencies on SMT differs across disfluency types.

1. Introduction

Disfluencies are generally defined as “phenomena that interrupt the flow of speech and do not add propositional content to an utterance”. (Fox Tree, 1995). In other words, disfluencies are those parts of spontaneous speech which when removed yield sentences that were originally intended. These sentences are shorter and less ill-formed. Although several different types of disfluencies have been defined (Shriberg, 1994), in this paper we deal with 3 types of disfluencies.

- Fillers – These are words that the speaker uses to control the conversation indicating that she/he intends to continue with the utterance. These may also indicate hesitation on the part of the speaker.
- b) Repetitions - Speaker repeats some part of the utterance
- c) Corrections – Speaker modifies the utterance mid-way while generally maintaining original syntax.

In the past decade, speech disfluencies have become a subject of inter-disciplinary research. While linguists have focussed on aspects such as role of disfluencies in discourse and speech comprehension, researchers in human language technologies have worked on automatically removing disfluencies from conversational speech. Several approaches for automatic Disfluency Removal (DFR) have been proposed. A decision tree based classifier with a combination of prosodic features such as pitch, energy and fundamental frequency and lexical features has been shown to be successful for disfluency removal. (Liu et. al, 2006). Honal and Schultz use a noisy-channel approach where disfluency removal is modelled as the translation of disfluent speech to clean speech (Honal and Schultz, 2003). There has also been some work where deeper syntactic knowledge used within the noisy-channel model. Johnson and Charniak use a Tree Adjoining Grammar (TAG) to represent cross-serial dependencies between the reparandum and correction regions of disfluencies. (Johnson and Charniak, 2004). However, with the exception of results in (Harper et. al 2005) where it is shown that disfluency removal aids parsing of conversational speech, most of the work in disfluency removal has so far been motivated by rich transcription tasks with a view to generate well-formed sentences with improved readability. While it has been argued that disfluency removal might help downstream NLP tasks such as question answering and Spoken Language Translation (SLT), to the best of our knowledge there has been no quantitative study establishing the claim.

Presence of disfluencies can hurt translation quality in two different ways. Firstly, disfluencies such as fillers and repetitions make utterances longer without adding semantic information. On the other hand, corrections add spurious content and therefore produce inaccurate translations. Secondly, since SLT systems are trained on large amounts of text data that have well-formed sentences, disfluencies cause a mismatch between training and evaluation data and can result in poor translation. In this paper, we demonstrate the adverse effect of disfluencies on SLT and show how automatic DFR can improve SLT performance.

2 Disfluency Removal System

The end-to-end pipeline proposed for SLT tasks is shown in Fig 1. S1 and S2 denote 2 scenarios that we explore in this paper - namely translation of broadcast conversation transcripts and ASR first-best output. A detailed description of the ASR and SMT systems is provided in the next section.

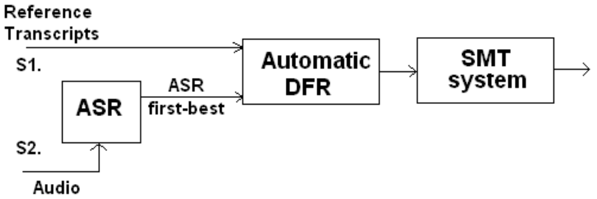

Fig. 1 : The end-to-end spoken language translation system

We use the CMU Disfluency Removal system (Honal and Schultz, 2003). This system is based on the noisy-channel model widely used in ASR and Statistical Machine Translation (SMT) systems. Here the problem of disfluency removal is viewed as a translation task where source language is the disfluent speech and the target language is non-disfluent speech. However, unlike in SMT systems, translation during DF removal involves only deciding which source words are to be deleted. No word insertions or reordering is permitted. Five translation models are used capturing features such as a) position of disfluency in a sentence b) position of the word within the current disfluent region c) words in the context of a disfluency d) distance from a fragment and e) information about previous deletions in the context of a disfluency. These models are combined log-linearly using weights learnt via gradient descent. (Honal and Schultz, 2005)

3 Data and System Description

The DFR system is trained on 46300 words (19 Al Jazeera shows) of Arabic BC) transcripts that were annotated for disfluencies by a native Arabic speaker. In addition, 2 shows designated as the development set were used to estimate weights for 5 disfluency models using gradient descent. Table 1 shows relevant corpus characteristics for training and evaluation data.

Dataset	Sentences	Words	Disfluencies (%)
Training	6370	46300	6.50
BCAD05	691	10570	3.48
GALE06	256	4480	5.86

Table 1. Training and Test Data Characteristics

The evaluation data consists of 2 Arabic BC datasets released by the LDC - a) BCAD05 consisting of 5 shows (Al Jazeera shows distinct from the training set) comprising 691 sentences and b) a subset of 5 shows from the GALE 2006 Evaluation data comprising 249 sentences. Both the training and evaluation data were manually annotated for disfluencies by the same annotator.

We use the CMU GALE 2006 Arabic speech recognition system trained on 190 hours of speech data consisting of 130 hours of broadcast news and 60 hours of BC (Naomany et al, 2007). A 4-gram language model trained on the Arabic Gigaword corpus was used. The output of the first pass speaker independent decoding was used in all our experiments. For our translation experiments, we use the ISL Arabic to English phrase based translation system (Eck et al, 2007). This alignment models are trained on a parallel corpus of 3.4 million sentences comprising UN data and data provided by the LDC and target language model is trained on 100 million words. The optimal model combination parameters were obtained by performing Minimum Error Rate Optimization on Arabic RT04 data (Och, 2003).

4 Experimental Results

4.1 Performance of Disfluency Removal

The performance of disfluency removal measured in terms of precision and recall is shown in Table 2. A combined single score of F-measure is also reported. We use a stricter criterion during evaluation where a disfluency is considered deleted only if every constituent disfluent word is deleted and no other words are deleted. While BCAD05 data consists of Al Jazeera TV chat shows, GALE06 data has 5 shows from channels not represented in the training set. We believe that this might explain the relatively lower recall of 57.57% for GALE06 data.

Dataset	Precision	Recall	F-measure
BCAD05	88.77	90.21	89.48
GALE06	95.59	57.57	71.86

Table 2. Performance of Automatic Disfluency Removal for Arabic broadcast conversation data

The translation performance is measured in terms of the BLEU score (Kishore et. al 2001). Higher the BLEU score, the better the translation quality. The absolute value of the BLEU itself depends on the number of reference translations used in computing n-gram similarity. The relatively low baseline scores (relative to the state-of-the-art systems) in this paper can be attributed to the use of a single reference translation per sentence. Reference translations used for evaluating translation performance were cleaned to ensure that there were no disfluencies on the English side.

In reporting the results of our translation experiments in the rest of this section, we use the following terminology. 3 different forms of each source sentence are translated to give 3 candidate English translations: a) "Unclean" input refers to a sentence that has not been subjected to disfluency removal and hence is expected to contain disfluencies. b) "Automatically Cleaned" input refers to translating the output of the DF removal system and might contain disfluencies and extra deletions due to imperfect DF removal c) "Manually cleaned" input is the target sentence which contains no disfluencies.

4. 2 Disfluency Removal and SLT Performance

Table 3 shows the impact of DF removal in terms of translation quality for the different datasets.

Dataset	Unclean input	Automatic DFR	Manual DFR
GALE06	11.13	11.55	11.68
BCAD05	12.16	12.36	12.30

Table 3. Effects of Disfluency Removal (DFR) on translation performance (BLEU) for Arabic broadcast conversation data.

The BLEU scores are first computed over each show and then averaged. In each case, we note that the automatic DF removal helps improve translation performance. An improvement of 0.154 BLEU points is obtained on the BCAD05 data. An even larger gain of 0.42 points is obtained on the GALE06 data. The translation corresponding to manually cleaned output can be considered as soft upper-bound and these results show that automatic DF removal approaches the upper-bound in terms of translation quality. Since the above analysis was conducted over the entire dataset, it also includes a large number of sentences that do not contain disfluencies.

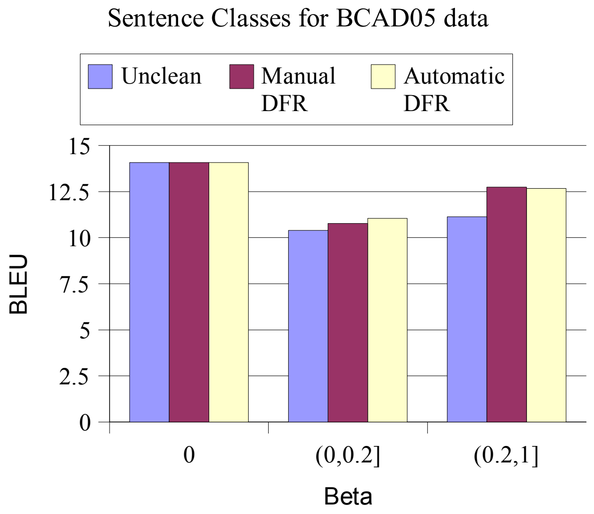

Fig 2. Effect of disfluency removal on sentence classes with different beta values for BCAD05 data; beta = ratio of number of disfluent words to sentence length

In particular, 48% of the sentences in GALE06 data contained no disfluencies. The corresponding number for BCAD05 data was much higher at 76%. However, we expect that maximum gain from disfluency removal is achieved for sentences that contain at least one disfluency, with larger gains to be expected when there are more disfluencies assuming they can all be removed. In order to study how the extent to which a sentence is disfluent effects translation, we divide sentences into groups based on β, the ratio of the number of disfluent words to the sentence length and compute BLEU scores for each group separately.

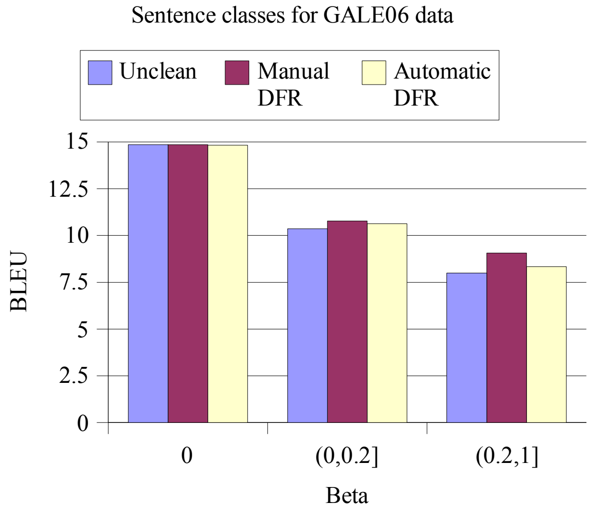

Fig 3. Effect of disfluency removal on sentence classes with different beta values for GALE06 data ; beta = ratio of number of disfluent words to sentence length

Figure 2 and 3 show how translation performance varies with β for BCAD05 and GALE06 data respectively. Firstly, as one might expect translation performance deteriorates with increasing β for both the datasets. More importantly, for a given β, translation after automatic DFR is better compared to unclean input translation. In general, the impact of DF removal on translation performance becomes more prominent as β increases. For e.g.: for BCAD05 data automatic DFR gives a BLEU score of 12.67 compared to 11.13 for the corresponding unclean input. Moreover, high precision of the DFR system ensures that there is no deterioration in translation performance for sentences without disfluencies (β=0).

4.3 Analysis of Disfluency Categories

The extent of degradation in translation quality is likely to vary with types of disfluencies such as fillers, repetitions and corrections. Differing lengths of disfluencies - fillers being one word long and corrections and repetitions that are typically longer - can partly explain this. In our work, we consider two classes of disfluencies - a) corrections and repetitions and b) filler words. Due to the presence of relatively fewer disfluencies in each of the categories, for the rest of the work in the paper, we combine the more complex disfluencies – repetitions and corrections - into a single category. We built 2 separate systems, one trained on data from which all corrections and repetitions were removed and the other system is trained on data with all filler words removed. Owing to fewer sentences in GALE06 data containing disfluencies, we only report results on BCAD05 data.

Sentence subset	Unclean input	Automatic DFR	Manual DFR
With fillers	10.39	11.11	11.18
Without fillers	13.93	13.93	13.92
Entire dataset	13.05	13.02	13.23

Table 4. Effect of fillers on translation performance (BLEU) for BCAD05 data

Table 4 shows the effect of fillers alone on translation performance. For each of the 3 translation scenarios, we report BLEU scores over the entire dataset. It is seen that there is a small improvement of 0.18 BLEU points in translation quality over the entire dataset when fillers are automatically removed. However, since less than 20 % of the sentences contain fillers, it is useful to report results separately for sentences that do contain fillers and those that do not. For sentences with one or more filler words, a significant improvement of nearly 0.80 BLEU points (8% relative increase) is achieved by removing fillers. Moreover, there is no decrease in BLEU scores for sentences without filler words.

Table 5 analyses the effect of corrections and repetitions on translation quality. While the overall trend is similar to that seen in the context of filler words, the relative improvement obtained is lower. There is no significant improvement seen over the entire dataset due to fewer than 5% of the sentences containing repetitions and corrections. However, there is a relative increase of about 5% in BLEU points for sentences with corrections and/or repetitions when automatic DF removal is used. Due to high precision, there is almost no degradation in translation performance seen on sentences without corrections and/or repetitions.

Sentence subset	Unclean DFR	Automatic DFR	Manual DFR
With C&R	11.77	12.78	11.88
Without C&R	13.33	13.33	13.35
Entire Set	13.22	13.27	13.23

Table 5. Effect of corrections and repetitions (C&R) on translation performance (BLEU) for BCAD05 data

It is interesting to note that BLEU scores with automatic DF removal can also exceed that with manually cleaned input. This implies that the DF removal system removes words that are not considered disfluent by the human annotator but whose absence nevertheless helps translation. One of the possible explanations for this is the ambiguous nature of what exactly constitutes a disfluency, i.e. it is likely that removal of certain words that are not obviously disfluent does cause the input to have more resemblance to text on which SLT systems perform better. To investigate this suspicion, we compared the perplexities of automatic DF cleaned and manually cleaned sentences obtained with a 4-gram Arabic language model. The perplexity of automatically DF removed sentences was found to be marginally lower which confirms our hypothesis that the cleaned text better matches training conditions.

Another observation which is also contrary to initial expectations is that removing fillers gives better improvement in comparison to removing corrections and repetitions. There are two possible reasons for this - a) there are relatively fewer corrections and repetitions as compared to fillers and b) when they do occur, they are harder to remove than fillers as is evidenced by nearly 100% recall for fillers versus 38% for corrections and repetitions. Thus one might expect that by improving hit rate in removing more complicated disfluencies, larger improvement in translation performance is achievable.

5. Experiments on ASR Output

5.1 Disfluency Removal on ASR output

Thus far the results presented involve DF removal and translation of manual transcriptions of Arabic BC shows. However, a more realistic scenario involves removing disfluencies from the output of an ASR system. Our ASR system uses acoustic models that detect non-verbal speech events such as hesitations and other non-human noise events. As a consequence, these acoustic models partially take care of disfluencies such as fillers. In order to show the effect of disfluency removal in general, we distinguish between 2 types of first-best output - a) unmodified first-best retains filler words hypothesized by the ASR decoder b) modified first-best removes these filler words. We use the automatic DF removal system to clean both the modified and unmodified first-best outputs.

The results of DFR on the modified and unmodified first-best output are shown in Table 6. It appears that DF removal helps in only 1 of the 4 cases. A higher gain of 0.89 BLEU is obtained by merely removing fillers hypothesized by ASR. No gains from DFR are seen for either of the 2 first-best output for the GALE06 data. In the next section we analyse the effect of ASR errors on DFR and argue why translation gains from DF removal on manual transcripts are not seen on ASR output.

Dataset	Unclean input	Automatic DFR
BCAD05(UFB)	10.45	10.58
BCAD05(MFB)	11.54	11.47
GALE06(UFB)	9.86	9.76
GALE06(MFB)	10.22	10.18

Table 6. Translation results (BLEU) for ASR first-best output ; UFB – Unmodified first best, MFB – Modified first best

5.2 Recognition Errors in Disfluent Regions

The ASR system introduces errors in transcribed output which can classified as deletion, substitution or insertion type errors. Each of these make disfluency removal challenging by either distorting the lexical context of a disfluency or altering the disfluency itself. For example, during ASR decoding the language model gives a low probability to a word repetitions and fillers and hence is likely to cause either a substitution or deletion error. In order to quantify the effects of recognition errors, we aligned the ASR first-best unmodified output to the DF annotated transcripts and observed what exactly happens to disfluent regions after ASR.

Table 7. shows recognition errors pertaining to different disfluencies for BCAD and GALE06 data. For both datasets, majority of filler words were either substituted or deleted although the exact proportion depends on specific data. In case of repetitions and corrections, a large majority are substituted or deleted. For GALE06 data, fewer than 10% (24 out of 320) were retained. Thus when presented with ASR output with word errors involving both disfluent words and their non-disfluent context, disfluency models trained on manually transcribed data perform poorly on the DF removal task.

DF classes and Dataset	Deletions (%)	Substitutions (%)	Retentions (%)
BCAD05(C&R)	33.33	24.50	42.10
GALE06(C&R)	40.60	50.00	9.40
BCAD05(fillers)	39.80	40.40	19.80
GALE06(fillers)	37.73	27.35	34.90

Table 7. Recognition errors for two disfluency categories; S – Substitutions, D – Deletions and R – retentions

6. Conclusion

From experiments on Arabic Broadcast conversation transcripts, it is seen that presence of disfluencies hurts the quality of Spoken Language Translation (SLT). We show that automatic disfluency removal prior to translation can give up 8% improvement in translation quality as measured by BLEU. However, similar gains were not achieved on ASR output due to recognition errors in disfluent regions and the resulting mismatch of training and test conditions. In future work, we plan to incorporate acoustic-prosodic cues while folding disfluency removal into the ASR decoding process.

7. Acknowledgements

This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-2-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. References

M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart and A. Krasnyanskaya (2005), “Final report : parsing speech and structural event detection,” http://www.clsp,jhu.edu/ws2005/groups/eventdetect/documents/ finalreport.pdf,

M. Johnson and E. Charniak, (2004) “A TAG-based noisy channel model of speech repairs”, Proc. of ACL.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D, Ostendorf, M. and Harper, M. (2006) Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. IEEE Trans. Audio, Speech and Language Processing. Vol 5. (pp. 1526-1540)

Liu, E. Shriberg, A. Stolcke, & M. Harper (2005), Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection, In *Proc. of Eurospeech*, Lisbon.

Matthias Honal and Tanja Schultz. (2003). Correction of disfluencies in spontaneous speech using a noisy-channel approach, Proceedings of the Eurospeech. Geneva

Matthias Honal and Tanja Schultz. (2005). Automatic disfluency removal on recognized spontaneous speech - Rapid adaptation to speaker-dependent disfluencies. Proceedings of International Conference on Acoustics, Speech and Signal Processing. Philadelphia

Mohamed Noamany, Thomas Schaaf and Tanja Schultz (2007) Advances in the CMU-InterACT Arabic Gale Transcription System. Proceedings of the HLT/NAACL. Rochester, NY

Franz Josef Och (2003). Minimum error rate training in statistical machine translation. Proceedings of the Association for Computational Linguistics, Japan

Shriberg, E.E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley.

Fox Tree, (1995), The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. Journal of Memory and Language, 34, pp 709-738

Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, Bing Zhao, Almut Silja Hildebrand, Stephan Vogel and Alex Waibel, “The UKA/CMU Statistical Machine Translation System for IWSLT 2006”, In Proc. of the IWSLT, Kyoto, 2006

K. Papineni and S. Roukos and T. Ward and W. Zhu, “Bleu: a method for automatic evaluation of machine translation”, Technical Report RC22176, IBM Research Division, Thomas J. Watson Research Center, 2001