A Corpus of Spontaneous Speech in Lectures : The KIT Lecture Corpus for Spoken Language Processing and Translation

Eunah Cho[1], **Sarah Fünfer**[1], **Sebastian Stüker**[1,2], **Alex Waibel**[1]
[1]Interactive Systems Laboratories
[2]Research Group 3-01 'Multilingual Speech Recognition'
Karlsruhe Institute of Technology
Karlsruhe, Germany
{eunah.cho|sarah.fuenfer|sebastian.stueker|alex.waibel}@kit.edu

Abstract

With the increasing number of applications handling spontaneous speech, the needs to process spoken languages become stronger. Speech disfluency is one of the most challenging tasks to deal with in automatic speech processing. As most applications are trained with well-formed, written texts, many issues arise when processing spontaneous speech due to its distinctive characteristics. Therefore, more data with annotated speech disfluencies will help the adaptation of natural language processing applications, such as machine translation systems. In order to support this, we have annotated speech disfluencies in German lecture data collected at KIT. In this paper we describe how we annotated the disfluencies in the data and provide detailed statistics on the size of the corpus and the speakers. Moreover, machine translation performance on a source text including disfluencies is compared to the results of the translation of a source text without different sorts of disfluencies or no disfluencies at all.

Keywords: speech translation, speech disfluency annotation, corpus

1. Introduction

In our days, spontaneous spoken language processing is one of the most challenging tasks of natural language processing. Especially recently, the study of spontaneous speech has attracted a great deal of attention, as more and more natural language processing applications are introduced. However, due to the enormous differences between spontaneous speech and written texts, e.g. when it comes to style, using written texts for building such applications has a major drawback: it simply does not match the actual data.

There has been extensive work accomplished for English spontaneous speech annotation (Fitzgerald and Jelinek, 2009). Also some attempts of spontaneous speech annotation were made in other languages (Maekawa et al., 2000) as well. But the number of works done specifically in the domain of university lectures is considerably smaller.

The analysis performed on the KIT lecture translation project showed that disfluencies can have a severe effect on the translation performance. We therefore decided to look more deeply into this domain. By manually annotating disfluencies in our data, we aim to provide a more in-depth analysis of these phenomena and hope to initiate more efforts on annotating and analyzing speech in other languages and in other domains. Using the disfluency-annotated German lecture data, automatic disfluency-modeling techniques shown in other works in English (Johnson and Charniak, 2004; Fitzgerald et al., 2009) can be applied to German and the performances can be compared. This data also motivates the development of new techniques of disfluency detection.

As one of our research focuses is the machine translation of university lectures, we are not only establishing a disfluency annotated German lecture corpus but also provide an English reference sentence for each source sentence. This allows us to evaluate our machine translation performance and thereby gain insight into the impact of disfluencies on subsequent applications.

This paper is organized as follows. In Section 2, a brief description of spontaneous speech in general and its impact on language processing are given. The data annotation and reference translation process is described in Section 3, followed by Section 4 which contains the corpus details and statistics. Section 5 describes our experimental setups and their results along with an analysis. Finally, Section 6 concludes our paper.

2. Spontaneous Speech

In this section, we describe the characteristics of spontaneous speech which usually do not appear in written texts, such as filler words, repetitions and corrections, false starts, abortions of words or sentences, hesitations, incorrectly pronounced or used words, as well as an imperfect grammar.

Filler words/sounds are words or sounds that a speaker utters while thinking about what he/she is going to say next or how he/she is going to finish a sentence. Some people insert them constantly in their speech. Obvious filler words or sounds like *uh* or *uhm* in English, or *äh*, *ähm* or *hmm* in German are relatively easy to detect.

Filler words and discourse markers, however, are occasionally more difficult to distinguish as it depends on the context whether they are considered filler words or not. Examples are *like*, *well* or *and* in English, or *ja*, *und* or *nun* in German.

Repetitions of words or phrases as well as the correction of the latter are another characteristic of spoken language. The speaker copies exactly what he/she said before or utters a rough copy, only changing a part of a word or a phrase. There are various reasons for this: stuttering, bridging a gap that occurs while thinking, or simply the correction of a word or a phrase.

Another recurrent part of spontaneous speech are false

Table 1: *An example sentence of the **rough copy** class*

Disfluency annotation	... solche Dinge, die **+/werden da/+** werden da vorgestellt, was ein ganz neues **+/Ka=/+** Kapitel ist ...
English gloss	... such things, which **+/will be here/+** will be here introduced, which a totally new **+/cha=/+** chapter is ...
Reference	... things like that will be introduced there, which is a totally new chapter ...

Table 2: *An example sentence of the **non-copy** class*

Disfluency annotation	**-/Mit dem recht=/-** er würde wieder zurückgehen.
English gloss	**-/With the right=/-** it would again go back.
Reference	It would go back again.

starts, where speakers begin a sentence but change their plan of what they want to say and continue differently, or aborted words or sentences. In extreme cases of false starts, a new context is introduced, putting an abrupt end to the previously discussed idea. An additional problem of false starts is that pronunciation is often unclear and grammatically incorrect sentences occur.

2.1. Disfluencies and Language Processing

Due to context, intonation, the situation, and experience, humans are nevertheless able to understand such non-fluent spoken language. For machines, however, it is much more difficult to handle spontaneous speech. The above mentioned characteristics hinder language processing and cause a major performance drop. One reason is the mismatch between well-structured training data and the actual test data, showing all signs of spontaneous speech - training data for machine translation usually does not contain any disfluencies. Automatic segmentation, another component of speech translation, is also based on language-model features learned from relatively clean data consisting of well-formed sentences.

One of the current research focuses of our laboratory is the translation of speeches and academic lectures (Fügen et al., 2007; Fügen, 2008; Cho et al., 2013a). Especially lectures often show characteristics of spontaneous speech, as most of the people who give lectures tend to speak freely and do not read out a script. In the process of analyzing the output of automatic speech recognition and machine translation we realized that our performance occasionally suffers not only from less predictable spoken tokens which are hard to process for the automatic speech recognition systems but also from disfluencies and pauses that hinder correct n-gram matches. Moreover, disfluencies obstruct correct reordering and phrase-pair matches in machine translation. Incorrect grammar and repetitions and corrections also make translation difficult. So these characteristics of spoken language inhibit all the different automatic language processing processes from automatic speech recognition to machine translation and therefore have a negative effect on the understandability of the output.

3. The Disfluency Annotated KIT German Lecture Corpus

In this section, we describe how the data is annotated, and the different disfluency classes used. In some cases, we had to modify the English reference; the process is also explained.

3.1. Annotation

The disfluency annotation has been done manually and on lectures that were previously recorded and transcribed, as described in (Stüker et al., 2012). The manual transcripts of the lecture data contain all words, partial words, sounds and utterances of the speaker, including disfluencies.

Prior to the annotation of the lecture corpus, we carefully examined the manual transcripts and explicitly chose lecture sets with a relatively high amount of disfluencies. In some rare cases, lectures showed characteristics of prepared speech and thus had to be filtered out: the utterances of the lecturers were relatively clean and lacked repetitions, corrections, filler words and so on, or showed very little of those.

Then, human annotators were asked to work on the data. Their first task was to read the transcripts in order to understand and follow the train of thought of the speaker. Afterwards, they marked disfluencies and characteristics of spontaneous speech by using different tags presented in the following chapter.

3.1.1. Disfluency Classes

The annotators distinguished several categories of disfluencies, namely repetitions and corrections, filler words and sounds, false starts, aborted sentences, and unfinished words.

Filler words and sounds often occur when a speaker hesitates, for example *uh*, or *uhm* in English or *ähm* and *äh* and *hmm* in German. In order to enhance the performance of automatic processing of these various versions of fillers were unified into *uh*, or *uhm* repectively in our work. Words that only in some contexts are considered filler words remained unchanged. The class of **filler words** also includs discourse markers such as *you know* or *I mean* in English. These expressions do not generally carry a meaning. German examples are *nun* (*now*, *well*, in English) or *ja* (*yes*, *right* in English).

Table 3: *An example sentence from the disfluency-annotated German corpus*

Manual transcript	Wenn Sie natürlich in der Vorlesung sitzen und *der Vorlesung* folgen, dann ist **Sprache**, die gesprochene Sprache, ein Problem.
English gloss	When you of course in the lecture sit and *the lecture* follow, then is **speech** the spoken speech a problem.
Disfluency annotation	Wenn Sie natürlich in der Vorlesung sitzen und der Vorlesung folgen, dann ist **+/Sprache/+** die gesprochene Sprache ein Problem.
English gloss	When you of course in the lecture sit and the lecture follow, then is **+/speech/+** the spoken speech a problem.
Disfluency annotation with reconstruction	Wenn Sie natürlich in der Vorlesung sitzen und *ihr* folgen, dann ist die gesprochene Sprache ein Problem.
English gloss	When you of course in the lecture sit and *it* follow, then is the spoken language a problem.
Reference	Obviously, when you are sitting in the lecture and are following it, then spoken speech is a problem.

In spontaneous speech, repetitions and corrections occur when a speaker repeats his/her words. Repetition can either be identical to the first utterance, or slightly different, because a certain part of a sentence is corrected. Such disfluencies are grouped together as **rough copy** in our work. Partial words can also occur in this class. An example of a repetition and a partial word is shown in Table 1, along with the literal translation and reference translation of the sentence. In this example, the verb and an additional word next to it *werden da* (engl. *will be here*) are forming an identical repetition. In the same sentence, a partial word *Ka* is also annotated as a rough copy, as it is a partial, but repetitive fragment of its next word *Kapitel* (engl. *Chapter*).

Another class that we use in our work is **non-copy**, which is reserved for false starts or aborted sentences. This tag covers the case when a speech fragment is dropped and a new fragment is introduced, which is often observed at the start of a sentence. An example of this class is shown in Table 2. Here, we can observe that a different topic is introduced after the previous topic is dropped. The last token of the non-copy disfluency is tagged as a partial word as it is one from the full word *rechteste* (engl. *furthest to the right*).

3.1.2. Sentence Reconstruction

Even after disfluent words and filler sounds had been removed, many of the spoken fragments are grammatically imperfect. Although this lies in the nature of human speech, it causes problems for subsequent applications such as machine translation systems that are normally trained on well-formed, grammatically correct sentences. Therefore, we wanted another version of annotation offering a grammatically correct utterance.

So after the first version of disfluency annotation was done, the annotators corrected the given sentence. They deleted repetitions, corrections and filler words and formed a correct sentence. If necessary, they were allowed to reorder words, and if there were no other possibilities to form a correct sentence, they could even leave out parts and change or add words. By doing so, we hope to get a grammatically correct version that is easier to understand and more fluent, while still containing the meaning of the original, thus suitable for use in subsequent automatic processes such as machine translation.

Despite its extreme difficulty, we consider it our ultimate goal to perform this level of disfluency detection and correction automatically. This second version of annotation is therefore inevitable. As a result, we hence get two German versions: an unchanged one only completed by tags, and another one considered to be a grammatically and linguistically correct German reproduction of the original sentence.

Table 3 is an excerpt of our annotation corpus, which shows sentence reconstruction process described in this section and reference translation. Words considered to be or causing disfluency are in bold letters. The first two rows show the original manual transcript of a German sentence along with its literal English translation. It contains a repetition of the word *Sprache* (engl. *speech*). Therefore, in the next two rows representing the first annotated version, the word is marked with a repetition tag. Moreover, the fluency of this sentence can be clearly improved by replacing the word *der Vorlesung* (engl. *the lecture*) with a pronoun, as the noun is already used in the first part of the sentence. Finally, the last line offers a correct English reference translation.

3.2. Reference Translation

The transcripts had been translated prior to the disfluency annotation as described in (Stüker et al., 2012). Annotators, however, were also asked to check the English translation against the German source text, thereby completing their task. Although repetitions and other characteristics of spontaneous spoken language in the source sentence were not supposed to have been taken into account for the translation, and moreover are not needed for a readable reference translation, we found that sometimes the English translations still contained filler words or sounds, repetitions and corrections or unfinished or aborted sentences and words. In this case, we asked our annotators to also tag them, in order to make the reference more fluent.

No additional reference is created for the reconstructed sentences. The reference translation is based on the first version of the annotation. Therefore, it is possible that reference sentence does not exactly match the reconstructed

Table 4: *Data statistics on classes of disfluency for each speaker*

Speaker ID	Filler words		Rough copy		Non-copy		Non-disfluency		All tokens	(hh:mm:ss)
Speaker 1	2,991	10.00%	1,072	3.58%	368	1.23%	25,486	85.19%	29,917	03:01:50
Speaker 2	633	2.88%	504	2.29%	413	1.88%	20,465	92.96%	22,015	02:21:26
Speaker 3	550	3.97%	320	2.31%	97	0.70%	12,870	93.01%	13,837	01:27:04
Speaker 4	607	6.14%	490	4.96%	76	0.77%	8,715	88.14%	9,888	00:59:29
Speaker 5	601	6.66%	308	3.41%	79	0.88%	8,040	89.06%	9,028	01:25:47
Speaker 6	126	2.28%	192	3.47%	64	1.16%	5,145	93.09%	5,527	00:46:53
Speaker 7	229	5.43%	33	0.78%	17	0.40%	3,938	93.38%	4,217	00:35:09
Speaker 8	418	12.67%	287	8.70%	83	2.52%	2,510	76.11%	3,298	01:13:52
Speaker 9	74	4.66%	34	2.14%	26	1.64%	1,455	91.57%	1,589	00:12:47
Speaker 10	41	4.27%	43	4.48%	7	0.73%	869	90.52%	960	00:05:19
Speaker 11	56	6.50%	71	8.25%	24	2.79%	710	82.46%	861	00:06:18
Speaker 12	15	1.82%	11	1.34%	14	1.70%	782	95.13%	822	00:05:47
Speaker 13	43	6.22%	8	1.16%	1	0.14%	639	92.47%	691	00:04:33
Speaker 14	26	4.01%	17	2.62%	2	0.31%	603	93.06%	648	00:04:56
Speaker 15	41	6.65%	48	7.78%	37	6.00%	491	79.58%	617	00:05:21
SUM	6,451	6.21%	3,438	3.31%	1,308	1.26%	92,718	89.22%	103,915	12:36:31

sentences.

4. Corpus Details and Statistics

In this section, we will provide a detailed analysis of the disfluencies occurring in the lecture data. Relevant statistics on disfluencies will be given, including the amount of each sort of disfluency present in the corpus. Moreover, the proportions of different categories of disfluencies used by different speakers will be compared and discussed.

Table 4 shows the data statistics on disfluency classes for each speaker. Talk duration for each speaker is also shown, as well as the number of tokens of different disfluency classes and their proportion in each talk. Tokens include all words as well as punctuation marks. Therefore, one word or a punctuation mark is considered as one token.

Most of the talks are from computer science lectures given in our university. Currently we have altogether 20 lectures from 15 speakers, and plan to extend this to 29 lectures given by 19 speakers. Therefore, some of the talks are merged lectures from one speaker while some talks are only short excerpts from a lecture. Each talk has a different length, therefore we have largely varying number of tokens gathered.

From this statistics, it is clear that several speakers do use filler words very frequently, while others show fewer disfluencies.

Looking at the summed number, up to now we have annotated around 104K tokens including punctuation marks, which correspond to 4,611 parallel sentences in German and English. The English reference consists of 92K tokens. The biggest part of disfluencies is filler words and discourse markers, which represent around 6% of all tokens. Rough copy tokens correspond to 3% of all tokens. Among them, 375 tokens are partial words, which is 0.4% of all tokens. Non-copy disfluency equals 1.3% of the whole corpus. Among them, only 12 tokens are partial words. Almost 89% of the whole corpus are tokens without disfluencies.

We are currently working on the annotation of more German lecture data, planning to gather altogether around 130K tokens.

5. Experiments

In this section, we present the performance of our current machine translation system, using as input different data, with different classes of disfluencies, no disfluencies at all or the grammatically correct version. Different sorts of disfluencies and their impact on machine translation thus can be analyzed.

5.1. System Description

We use a phrase-based machine translation system to measure the impact of different disfluencies. We used 1.76 million sentences of German-English parallel data for training. In order to make our models fit the lecture data better, we extracted parallel data from TED talks [1] and adapted the models to the domain. Preprocessing including text normalization, tokenization, and smartcasing was applied before training. Additionally, we applied compound splitting for the German side.

We used the Moses package (Koehn et al., 2007) in order to build the phrase table. A language model was trained on 462 million English words using the SRILM Toolkit (Stolcke, 2002). In addition to this language model, we used a bilingual language model (Niehues et al., 2011), which improves the translation performance by extending source word context. Scaling factors were optimized with minimum error rate training (Venugopal et al., 2005), using an in-house decoder (Vogel, 2003).

5.2. Results

In order to compare the system performance with or without disfluencies, we conducted nine different experiments on the data shown in Table 4.

In the baseline experiment, we translated all transcribed words, including words that were annotated as disfluency.

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

The experiment "No *uh*" shows the score without obvious filler words, previously unified, so either *uh* or *uhm*. For the next experiment we removed all filler words including the obvious filler words and translated the text. The same experiment is applied for other classes of disfluency; we measure the machine translation performance when we do not have rough copy tokens or non-copy tokens. These disfluencies were later removed altogether for comparison. Finally, we removed all annotated tokens and showed the translation score. Additionally, we translated the second version of the annotation, where reconstruction of sentences was allowed in order to generate grammatically correct sentences.

Table 5: *Translation performance comparison with disfluency*

System	BLEU
Baseline	18.85
No *uh*	19.78
Filler words removed	21.18
Rough copy removed	19.45
Non-copy removed	19.08
Filler words & rough copy removed	21.97
Rough copy & non-copy removed	19.69
All disfluencies removed	22.26
Annotation with reconstruction	22.52

Table 5 shows the results of the experiments. The scores are reported in case-sensitive BLEU (Papineni et al., 2002). Compared to the text where we have all disfluencies, the translation score is improved by 3.4 BLEU points after removing all disfluencies. Manually removing the obvious filler words already improves the score by almost 0.9 BLEU points. A further improvement of 1.4 BLEU points is gained when other filler words and discourse markers are removed. We also conduct other experiments, where we remove rough copy tokens and non-copy tokens independently. That is, filler words are kept but each class of disfluency is deleted so that the impact can be compared. When we remove rough copy tokens and non-copy tokens independently of the filler words, the improvement in BLEU is smaller than when removing filler tokens. This is due to the fact that compared to other disfluencies, the number of filler tokens is huge, as shown in the Table 4. When we remove filler words and rough copy tokens at the same time, the BLEU score is improved by 0.74 points, compared to when only filler words are removed. We also conducted another experiment where we remove rough copy and non-copy tokens together. The improvement is bigger than the performance improvement we get when we remove a single class of disfluencies only - a strong indication that removing disfluencies does have synergy effects on the translation performance when both are cleaned up.

Finally, even though the reference might not be the best match for the reconstructed sentences, translating the reconstructed, grammatically better organized sentences gave us an improvement of around 0.3 BLEU points over the text where disfluencies were simply removed. This shows that generating a well-constructed sentence can improve the machine translation quality further.

5.3. Analysis

As shown in Table 5, removing speech disfluencies directly improves the translation performance. Especially removing filler words has a strong impact, as the number of tokens involved is the highest. Improvement gained by reconstructing sentences suggests that even if we detect all disfluent words, there is still room to improve the performance. This also suggests that spoken-style sentences, even when disfluencies are removed or do not occur at all, are harder to translate than the grammatically correct sentences produced by the annotators that of course come closer to written texts. These newly constructed sentences fit better with the translation models built using written-text style sentences.

We have been training a speech disfluency detection model using some parts of the data described in this paper, as shown in (Cho et al., 2013b; Cho et al., 2014). Using this data, we were able to improve the machine translation performance on both manual speech transcripts and automatic speech transcripts. It is our plan to extend the study in order to model well-structured sentences from spontaneous sentences.

6. Conclusion

In this paper, we introduced the KIT lecture corpus for spoken language processing and translation. The goal of building this corpus is to model spoken language phenomena in order to improve the performance of subsequent applications, such as machine translation systems. The largest part of our corpus covers diverse topics related to computer science, and contains various speaking styles from 15 different speakers.

This corpus, as mentioned earlier, is expected to reach a size of 130K. We also aim to work on other language pairs for the disfluency annotation task, so that we can compare the performance improvements. The targeted speech domain or style will be also extended to other genres of speech, such as meetings, where multiple speakers are involved. As the disfluency detection problem becomes more challenging with a growing number of speakers, we hope that such an expansion will provide a better insight into the difficulty of dealing with disfluencies for following applications.

7. Acknowledgements

'Research Group 3-01' received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. 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

Cho, E., Fügen, C., Herrmann, T., Kilgour, K., Mediani, M., Mohr, C., Niehues, J., Rottmann, K., Saam, C., Stüker, S., and Waibel, A. (2013a). A Real-World System for Simultaneous Translation of German Lectures.

In *Proc. of the International Conference on Speech and Language Processing (Interspeech)*, Lyon, France.

Cho, E., Ha, T.-L., and Waibel, A. (2013b). CRF-based Disfluency Detection using Semantic Features for German to English Spoken Language Translation. In *Proceedings of the International Workshop for Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Cho, E., Niehues, J., and Waibel, A. (2014). Tight Integration of Speech Disfluency Removal into SMT. In *Proceedings of the European Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Fitzgerald, E. and Jelinek, F. (2009). Linguistic Resources for Reconstructing Spontaneous Speech Text. In *Proc. of the Linguistic Resources and Evaluation Conference (LREC)*, Athens, Greece.

Fitzgerald, E., Hall, K., and Jelinek, F. (2009). Reconstructing False Start Errors in Spontaneous Speech Text. In *Proceedings of the 12th conference on European chapter of the Association for Computational Linguistics (ACL)*, Athens, Greece.

Fügen, C., Waibel, A., and Kolss, M. (2007). Simultaneous Translation of Lectures and Speeches. *Machine Translation*, 21:209–252.

Fügen, C. (2008). *A System for Simultaneous Translation of Lectures and Speeches*. Ph.D. thesis, Universiät Karlsruhe (TH), November.

Johnson, M. and Charniak, E. (2004). A TAG-based Noisy Channel Model of Speech Repairs. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL), Demonstration Session*, Prague, Czech Republic, June.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontneous Speech Corpus of Japanese. In *Proc. of the Linguistic Resources and Evaluation Conference (LREC)*.

Niehues, J., Herrmann, T., Vogel, S., and Waibel, A. (2011). Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, UK.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, Denver, Colorado, USA.

Stüker, S., Kraft, F., Mohr, C., Herrmann, T., Cho, E., and Waibel, A. (2012). The KIT Lecture Corpus for Speech Translation. In *Proc. of the Linguistic Resources and Evaluation Conference (LREC)*.

Venugopal, A., Zollman, A., and Waibel, A. (2005). Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *WPT-05*, Ann Arbor, MI.

Vogel, S. (2003). SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China.