Automatic Construction of Discourse Corpora for Dialogue Translation

Longyue Wang*, **Xiaojun Zhang***, **Zhaopeng Tu**†, **Andy Way***, **Qun Liu***
* ADAPT Centre, School of Computing, Dublin City University, Ireland
† Noah Ark Lab, Huawei Technologies, China
{lwang, xzhang, away, qliu}@computing.dcu.ie, tu.zhaopeng@huawei.com

Abstract

In this paper, a novel approach is proposed to automatically construct parallel discourse corpus for dialogue machine translation. Firstly, the parallel subtitle data and its corresponding monolingual movie script data are crawled and collected from Internet. Then tags such as speaker and discourse boundary from the script data are projected to its subtitle data via an information retrieval approach in order to map monolingual discourse to bilingual texts. We not only evaluate the mapping results, but also integrate speaker information into the translation. Experiments show our proposed method can achieve 81.79% and 98.64% accuracy on speaker and dialogue boundary annotation, and speaker-based language model adaptation can obtain around 0.5 BLEU points improvement in translation qualities. Finally, we publicly release around 100K parallel discourse data with manual speaker and dialogue boundary annotation.

Keywords: Discourse Corpus, Dialogue, Machine Translation, Information Retrieval, Movie Script, Movie Subtitle

1. Introduction

Dialogue is an essential component of social behaviour to express human emotions, moods, attitudes and personality. To date, few researchers have investigated how to improve the machine translation (MT) of conversational material by exploiting their internal structure. This lack of research on the dialogue MT is a surprising fact, since dialogue exhibits more cohesiveness than single sentence and at least as much than textual discourse.

Although there are a number of papers on corpus construction for various natural language processing (NLP) tasks, dialogue corpora are still scarce for MT. Some work regarding bilingual subtitles as parallel corpora exists, but it lacks rich information between utterances (sentence-level corpus) (Lavecchia et al., 2007; Tiedemann, 2007a; Tiedemann, 2007b; Itamar and Itai, 2008; Tiedemann, 2008; Xiao and Wang, 2009; Tiedemann, 2012; Zhang et al., 2014). Other work focuses on mining the internal structure in dialogue data from movie scripts. However, these are monolingual data which cannot used for MT (Danescu-Niculescu-Mizil and Lee, 2011; Banchs, 2012; Walker et al., 2012; Schmitt et al., 2012). In general, the fact is that bilingual subtitles are ideal resources to extract parallel sentence-level utterances, and movie scripts contain rich information such as dialogue boundaries and speaker tags.

Inspired by these facts, our initial idea was to build dialogue discourse corpus by bridging the information in these two kinds of resources (i.e., scripts and subtitles). The corpus should be parallel, align at the segment-level as well as contain rich dialogue information. We propose a simple but effective approach to build our dialogue corpus. Firstly, we extract parallel sentences from bilingual subtitles, and mine dialogue information from monolingual movie scripts. Secondly, we project dialogue information from script utterances to its corresponding parallel subtitle sentences using an information retrieval (IR) approach. Finally, we apply this approach to build a Chinese–English dialogue corpus, and also manually annotate dialogue boundaries and speaker tags based on automatic results.

To validate the effect of the proposed approach, we carried out experiments on the generated corpus. Experimental results show that the automatic annotation approach can achieve around 82% and 98% on speaker and dialogue boundaries annotation, respectively. Furthermore, we explore the integration of speaker information into MT via domain-adaptation techniques. Results show that we can improve translation performance by around 0.5 BLEU points compared to baseline system.

Generally, the contributions of this paper include the following:

- We propose an automatic method to build a segment-level dialogue parallel corpus with useful information, for building large-scale dialogue MT systems;
- Through exploring dialogue information with MT, we show that speaker information is really helpful to dialogue MT systems;
- We also manually annotate about 100K sentences from our dialogue corpus. The gold standard dataset[1] can be further used to search for the coherence and consistency clues in discourse structure to implement a dialogue MT system.

The rest of the paper is organized as follows. In Section 2, we describe related work. Section 3 describes in detail our approaches to build a dialogue corpus as well as the structure of the generated database. The experimental results for both corpus annotation and translation are reported in Section 4. Finally, Section 5 presents our conclusions and future work plans.

2. Related Work

In the specific case of dialogue MT system, data acquisition can impose challenges including data scarcity, translation quality and scalability. The release of the Penn Discourse Treebank (PDTB)[2] (Prasad et al., 2008) helped bring about

[1] We release our DCU English-Chinese Dialogue Corpus in `http://computing.dcu.ie/~lwang/resource.html`.

[2] Available at `https://www.seas.upenn.edu/~pdtb`.

a new sense of maturity in discourse analysis, finally providing a high-quality large-scale resource for training discourse parsers for English. Based on PDTB, some have applied the insights to MT (Meyer and Popescu-Belis, 2012). A resource like the PDTB is extremely valuable, and it would be desirable to have a similar resource in dialogue or conversation as well.

There are two directions of work related to dialogue corpus construction. One is parallel corpora construction for dialogue or conversation MT (Lavecchia et al., 2007; Tiedemann, 2007a; Tiedemann, 2007b; Tiedemann, 2008; Itamar and Itai, 2008; Xiao and Wang, 2009; Tiedemann, 2012). Thanks to the effects of crowdsourcing and fan translation in audiovisual translation (O'Hagan, 2012), we can regard subtitles as parallel corpora. Zhang et al. (2014) leveraged the existence of bilingual subtitles as a source of parallel data for the Chinese-English language pair to improve the MT systems in the movie domain. However, their work only considers sentence-level data instead of extracting more useful information for dialogues. Besides, Japanese researchers constructed a speech dialogue corpus for a machine interpretation system (Aizawa et al., 2000; Matsubara et al., 2002; Ryu et al., 2003; Takezawa, 2003). They collected speech dialogue corpora for machine interpretation research via recording and transcribing Japanese/English interpreters' consecutive/simultaneous interpreting in the booth. The German VERBMOBIL speech-to-speech translation programme (Wahlster, 2013) also collected and transcribed task-oriented dialogue data. This related work focused on speech-to-speech translation including three modules of automatic speech recognition (ASR), MT and text-to-speech(TTS).

The other one is mining rich information from other resources such as movie scripts. Danescu-Niculescu-Mizil and Lee (2011) created a conversation corpus containing large metadata-rich collections of fictional conversations extracted from raw movie scripts. Both Banchs (2012) and CMU released dialogue corpora extracted from the Internet Movie Script Database (IMSDb).[3] Based on IMSDb, Walker et al. (2012) annotated 862 film scripts to learn and characterize the character style for an interactive story system, and Schmitt et al. (2012) annotated 347 dialogues to explore a spoken dialogue system. The resource of movie scripts, such as IMSDb, is good enough to generate conversational discourse for dialogue processing. However, monolingual movie scripts are not enough for MT which needs a large-scale bilingual dialogue corpus to train and tune translation models.

3. Building A Parallel Dialogue Corpus

As already stated, our presented parallel dialogue corpus is extracted from bilingual movie/episode subtitles and monolingual scripts. We extend previous work on movie scripts to scripts of TV series such as *Friends*. From IMSDb and SimplyScripts[4] and the like, we crawled movie/episode scripts data. In addition, we collected the English-Chinese bilingual subtitles from multiple audiovisual translation web resources such as Shooter[5] and Opensubtitles. [6] Based on the hypothesis that both a script and a subtitle exist for the same movie or episode, the method can be described in a pipeline as follows:

(1) given a monolingual movie/episode script, we identify dialogue boundaries and speaker tags using clues such as format and story structure tags in the script;

(2) for a bilingual subtitle, we align each sentence with its translation using clues such as format and time information;

(3) for each utterance in a processed script, we apply IR techniques to match it with the line(s) in its corresponding processed subtitle according to the shared language;

(4) for each matched term, we map the useful annotations such as speaker and dialogue boundaries from the script side to the matched line(s) in its subtitle side.

3.1. Script and Subtitle

Figure 1 depicts a browser snapshot illustrating an episode script layout of *Friends*. There are three kinds of information: speaker, shot/scene and action information in the script. The speaker element (red ellipses) contains the corresponding character who says the utterance(s). The shot/scene tags (e.g., "SCENE", "SHOT", "CUT INTO:" and "CUT TO:" etc.) can be regarded as the boundaries of dialogues. For instance, the tags "SCENE J" and "CUT TO:" refer to the beginning and end of a dialogue, respectively. The action (green frames) contains all additional information of a narrative nature and explains what is happening in the scene. In this work, we focus on the first two information type while ignoring final one.

Figure 2 is the corresponding bilingual subtitle of the script in Figure 1. Subtitles are often organized in two formats: Advanced SubStation Alpha (ASS) and SubRip Text (SRT). As most lines are one-to-one aligned on two language sides, it easy to process them into a parallel corpus. We also use line id and time line information to deal with one-to-many or mismatching cases.

Based on the above rules, we extract useful information from both scripts and subtitles. In order to obtain high-quality data, we also apply a series of techniques including language detection, simplified-traditional Chinese conversation, coding conversation and punctuation normalization. After processing scripts and subtitles, the next step is to match and project terms from script side to subtitle side.

3.2. Matching and Projection

Comparing examples in Figures 1 and 2, we found that the script and the subtitle share the same language (i.e., English). However, subtitle lines are not always the same as the utterances in a script for the actors may change their lines on site, either slightly or to a greater extent. For example, the first utterance in the script is *Later, when Monica's around, I want you to ask me about fire trunks* while the

[3] Available at `http://www.imsdb.com`.

[4] Available at `http://www.simplyscripts.com`.

[5] Available at `http://sub.makedie.me`.

[6] Available at `http://www.opensubtitles.org`.

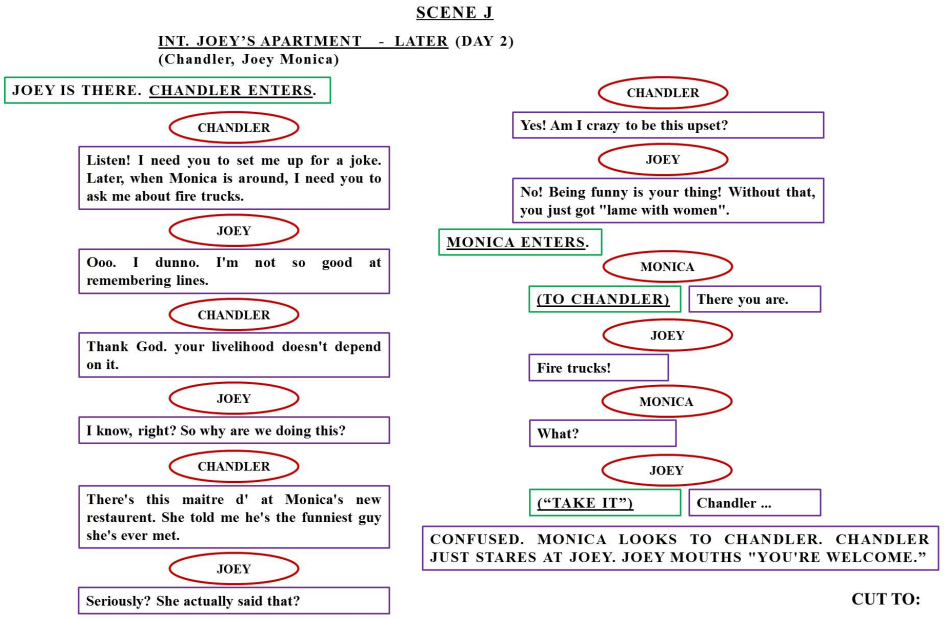

Figure 1: Examples of scripts of *Friends* in English

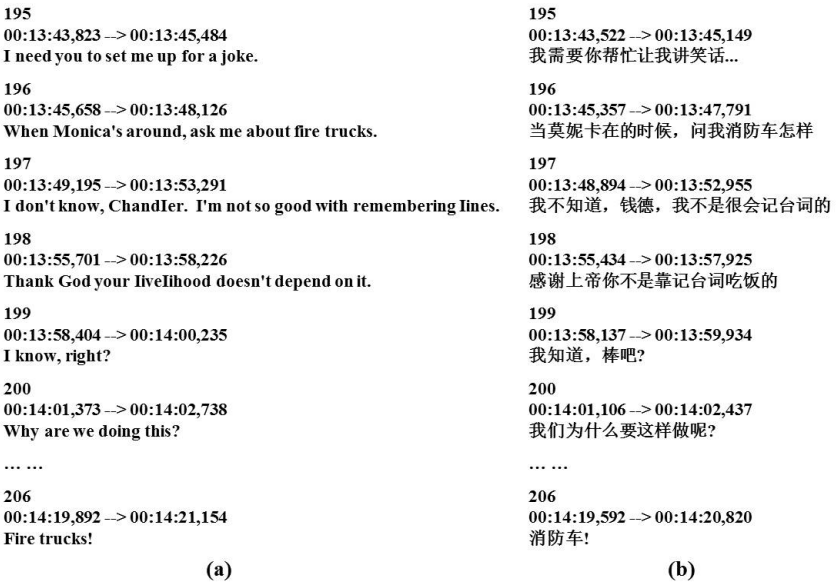

Figure 2: Examples of bilingual subtitles (SRT) of *Friends* in English and Chinese

corresponding line in the subtitle is *When Monica's around, ask me about fire trunks.*. Another phenomenon is that one utterance on script side may be split into several lines on subtitle side. This change is made to accommodate the size of the TV screen. It is a big challenge to deal with these changed, missing or duplicated terms during matching. All the above problems make the task a complex N-to-N matching where $N \geq 0$.

Therefore, we regard the matching and projection as an IR task (Wang et al., 2012a). The Vector Space Model (VSM) (Salton et al., 1975) is a state-of-the-art IR model in which each document is represented as a vector of identifiers (here we describe each identifier as a term). The ith utterance D_i in the script is represented as a vector $D_i = [w_{1,i}, w_{2,i}, ...w_{k,i}]$, in which k is the size of the term vocabulary. Many similarity functions can be employed to calculate the similarity between two utterance vectors (Cha, 2007). Here we apply the cosine distance:

$$sim(d_i, d_j) = \sum_{k=1}^{N} w_{i,k} \cdot w_{j,k} \sqrt{\sum_{k=1}^{N} w_{i,k}} \cdot \sqrt{\sum_{k=1}^{N} w_{j,k}} \quad (1)$$

where N is the number of terms in an utterance vector, and $w_{i,k}$ and $w_{j,k}$ represent the weight of the ith/jth term in the utterance D_i/D_j respectively. Technically, the distance between documents in VSM is calculated by comparing the

deviation of angles between vectors. A Boolean Retrieval Model sets a term weight to be either 0 or 1, while an alternative solution is calculating the term weights according to the appearance of a term within the document collection.
To calculate the term weights according to the appearance of a term within the document collection, we apply term frequency-inverse document frequency (*TF-IDF*) (Ramos, 2003) as one term-weighting model. The weight w of each term t is determined by its own term frequency $tf(t,d)$ in a document d and its inverse document frequency $idf(t,d,D)$ within the search collection. The definition of term weight $w_{t,d}$ is shown as in Eq. (2) and (3):

$$w_{t,d} = tf(t,d) \cdot idf(t,d,D) \quad (2)$$

$$idf(t,d,D) = \log\big(\frac{|D|}{|\{d \in D | t \in d\}|}\big) \quad (3)$$

where D is the total number of documents in the document collection.
In practice, we regard each utterance as a document and build the index for each movie script. Then we use each subtitle sentence as a query to search for target related utterances. In order to deal with inconsistency problems, we employ several strategies:

- For better indexing and searching, we split the sentences/utterances into the smallest units using a sentence splitter;
- Except for punctuation mark, we do not remove any stop words. Furthermore we low-case each word;
- For each original query, it can be split into n sub-queries. For each sub-query, we apply 1-best search. Then search results of the sub-queries are combined to vote for the best candidate for the original query.
- One query may be similar to several utterances in different lines of a script. The candidate closest to the last matched term is more likely to be the right answer. Thus we impose a window for short-distance searching.

After the script and subtitle are bridged, we project speaker tags and dialogue boundaries in scripts to their corresponding lines in subtitles. Finally, we preserve the results in XML format, which is illustrated in Figure 3.

4. Experiments and Results

For dialogue corpus construction, we apply our methods to a ten-season sitcom *Friends*. We extract and process both scripts and subtitles of *Friends* (described in Section 3.1) and then bridge them (described in Section 3.2) to build a dialogue corpus in the format of Figure 3. For data processing, we employ the sentence splitter and English tokenizer in the Moses toolkit and our in-house Chinese segmentor (Wang et al., 2012b). Furthermore, we employ Apache Lucene[7] for indexing and search tasks. Table 1 presents the main statistics of the resulting bilingual dialogue corpus. We obtained 5,428 bilingual dialogues with annotated speaker and dialogue boundary information.

[7]Available at `https://lucene.apache.org`.

Item	Size
Total number of scripts processed	236
Total number of dialogues	5,428
Total number of speakers	42
Total number of utterances	109,268
Average amount of dialogues per script	23
Average amount of speakers per dialogue	3.5
Average amount of utterances per dialogue	20

Table 1: Statistics of generated parallel dialogue corpus

To verify the validity of our methods (described in Section 3), we conduct an evaluation on the matching accuracy of speaker tags and dialogue boundaries in the generated corpus. To generate gold standard reference, we also manually annotate the dialogue information based on the generated parallel dialogue corpus. The agreements between automatic labels and manual labels is 81.79% on speaker and 98.64% on dialogue boundary, respectively. This indicates that the proposed automatic annotation strategy through mapping is reasonably trustworthy.
Furthermore, we conduct a simple experiment to explore the effects of speaker tags on dialogue MT. We first build a baseline MT engine using Moses (Koehn et al., 2007) on our generated parallel corpus (described in Table 1). We train a 5-gram language model (LM) using the SRI Language Toolkit (Stolcke, 2002) on the target side of parallel corpus. Besides, we use GIZA++ (Och and Ney, 2003) for word alignment and minimum error rate training (Och, 2003) to optimize feature weights. Based on the hypothesis that different types of speakers may have specific speaking styles, we employ a language model adaptation method to boost the MT system (Wang et al., 2014). Instead of building a LM on the whole data, we split the data into two separate parts according the speakers' sex and then build two separate LMs. As Moses supports multiple LM integration, we directly feed Moses two LMs. The translation results are listed in Table 2. For Chinese-to-English (i.e. "ZH-EN"), the baseline system achieves 20.32 and 16.33 in BLEU score on development and test data, respectively, while for English-to-Chinese (i.e. "EN-ZH"), the scores are 16.78 and 14.11 in BLEU score. The BLEU scores are relatively low because 1) we have only one reference, 2) the training corpus is small, and 3) dialogue MT is a challenging task. By using LM adaptation, we improve the performance on test data by +0.50 and +0.43 BLEU points on Chinese-to-English and English-to-Chinese tasks respectively.

Systems	Lang.	Dev Set	Test Set
ZH-EN	Baseline	20.32	16.33
	Speaker$_{LM}$	21.05	16.83 (+0.50)
EN-ZH	Baseline	16.78	14.11
	Speaker$_{LM}$	17.23	14.54 (+0.43)

Table 2: Translation results on speaker based language model adaption.

```
<dialogue id="4884" n_utterances="12">
    <context id="1" action= "JOEY IS THERE. CHANDLER ENTERS" >
        <utterance id="1" speaker="CHANDLER">
            <EN>I need you to set me up for a joke.</EN> <ZH>我需要你帮忙让我讲笑话...</ZH>
        </utterance>
        <utterance id="2" speaker="CHANDLER">
            <EN>When Monica's around, ask me about fire trucks.</EN> <ZH>当莫妮卡在的时候，问我消防车怎样</ZH>
        </utterance>
        <utterance id="3" speaker="JOEY">
            <EN>I don't know, ChandIer. I'm not so good with remembering Iines.</EN> <ZH>我不知道，钱德，我不是很会记台词的</ZH>
        </utterance>
        <utterance id="4" speaker="CHANDLER">
            <EN>Thank God your IiveIihood doesn't depend on it.</EN> <ZH>感谢上帝你不是靠记台词吃饭的</ZH>
        </utterance>
        <utterance id="5" speaker="JOEY">
            <EN>I know, right?</EN> <ZH>我知道，棒吧?</ZH>
        </utterance>
        <utterance id="5" speaker=" JOEY ">
            <EN>Why are we doing this?</EN> <ZH>我们为什么要这样做呢?</ZH>
        </utterance>
            … …
    </context>
    <context id="2" action= "MONICA ENTERS. TO CHANDLER" >
        … …
        <utterance id="12" speaker="JOEY">
            <EN>Fire trucks!</EN> <ZH>消防车!</ZH>
        </utterance>
    </context>
        … …
    <context id="3" action= "CONFUSED. MONICA LOOKS TO CHANDLER… …" >NULL</context>
</dialogue>
```

Figure 3: A sample of generated XML of dialogue in episode script

5. Conclusions and Future Work

We propose a novel approach to build a parallel dialogue discourse corpus from monolingual scripts and their corresponding bilingual subtitles. We identify the dialogue boundaries according to the scene or shot tags in the script to segment the monolingual dialogue, and then map the matched monolingual dialogues to the source part of the bilingual subtitles with the speaker and utterance elements in order to obtain the bilingual discourse dialogues. Finally we align the bilingual dialogue subtitle lines to produce suitable MT training material.

We expand the current dialogue generation resources from movie scripts to movie/episode scripts, and specify the current parallel corpus construction to bilingual dialogue corpus building based on bilingual subtitles. We pilot this approach on a 10-season sitcom *Friends* and automatically generated 5,428 bilingual parallel dialogue discourses. This is a quick way to generate a bilingual dialogue corpus.

To validate the effect of the proposed approach, we annotated the speaker and dialogue boundary elements manually in 4-season *Friends* data and compared the manual results with our automatic findings. Experimental results show that the automatic annotation approach can achieve around 81.79% and 98.64% on dialogue boundaries and speaker tags, respectively. Furthermore, we explore the integration of speaker tags into MT using domain-adaptation techniques. The experiments show that we can improve translation performance compared to a baseline system.

As far as future work is considered, we intend to explore automatic dialogue detection from bilingual subtitles. A reachable goal is to utilize the resulting bilingual dialogue corpus based on our approach to also summarize the discourse elements such as coherence and co-reference, speaker relationship and time information of the subtitle lines. Some supervised and semi-supervised methods and machine learning approaches can be used on these tasks.

6. Acknowledgements

This work is supported by the Science Foundation of Ireland (SFI) ADAPT project (Grant No.:13/RC/2106), and partly supported by the DCU-Huawei Joint Project (Grant No.:201504032-A, YB2015090061). It is partly supported by the Open Projects Program of National Laboratory of Pattern Recognition (Grant 201407353) and the Open Projects Program of Centre of Translation of GDUFS (Grant CTS201501).

7. Bibliographical References

Aizawa, Y., Matsubara, S., Kawaguchi, N., Toyama, K., and Inagaki, Y. (2000). Spoken language corpus for machine interpretation research. In *Proceedings of the 6th International Conference on Spoken Language Processing*, volume 3, pages 398–401, Beijing, China.

Banchs, R. E. (2012). Movie-dic: A movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 2, pages 203–207, Jeju Island, Korea.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1:300–307.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach

to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon.

Itamar, E. and Itai, A. (2008). Using movie subtitles for creating a large-scale bilingual corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 269–272, Marrakech, Morocco.

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 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Lavecchia, C., Smaili, K., and Langlois, D. (2007). Building parallel corpora from movies. In *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science*, pages 201–210, Funchal, Madeira, Portugal.

Matsubara, S., Takagi, A., Kawaguchi, N., and Inagaki, Y. (2002). Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 153–159, Las Palmas, Canary Islands - Spain.

Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation*, pages 129–138, Avignon, France.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Sapporo, Japan.

O'Hagan, M. (2012). From fan translation to crowdsourcing: Consequences of web 2.0 user empowerment in audiovisual translation. *Approaches to Translation Studies*, 36:25–41.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the 1st instructional conference on machine learning*, Piscataway, NJ USA.

Ryu, K., Matsubara, S., Kawaguchi, N., and Inagaki, Y. (2003). Bilingual speech dialogue corpus for simultaneous machine interpretation research.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3369–3373, Istanbul, Turkey.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 901–904, Colorado, USA.

Takezawa, T. (2003). Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 2757–2760, Geneva, Switzerland.

Tiedemann, J. (2007a). Building a multilingual parallel subtitle corpus. In *Proceedings of the 17th Conference on Computational Linguistics in the Netherlands*, pages 1–14, Leuven, Belgium.

Tiedemann, J. (2007b). Improved sentence alignment for movie subtitles. In *Proceedings of the 3rd Conference on Recent Advances in Natural Language Processing*, volume 7, pages 582–588.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1902–1906, Marrakech, Morocco.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey.

Wahlster, W. (2013). *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, Berlin.

Walker, M. A., Lin, G. I., and Sawyer, J. (2012). An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1373–1378, Istanbul, Turkey.

Wang, L., Wong, D. F., and Chao, L. S. (2012a). An improvement in cross-language document retrieval based on statistical models. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing*, pages 144–155, Chung-Li, Taiwan.

Wang, L., Wong, D. F., Chao, L. S., and Xing, J. (2012b). Crfs-based chinese word segmentation for micro-blog with small-scale data. In *Proceedings of the 2nd conference jointly organized by the Chinese Language Processing Society of China and the Association for Computational Linguistics Special Interest Group on Chinese Language Processing*, pages 51–57, Tianjin, China.

Wang, L., Lu, Y., Wong, D. F., Chao, L. S., Wang, Y., and Oliveira, F. (2014). Combining domain adaptation approaches for medical text translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 254–259, Baltimore, USA.

Xiao, H. and Wang, X. (2009). Constructing parallel corpus from movie subtitles. In *Proceedings of the*

22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy, pages 329–336, Hong Kong, China.

Zhang, S., Ling, W., and Dyer, C. (2014). Dual subtitles as parallel corpora. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1869–1874, Reykjavik, Iceland.