Dual Subtitles as Parallel Corpora

Shikun Zhang, Wang Ling, Chris Dyer
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
shikunz@andrew.cmu.edu, lingwang@cs.cmu.edu, cdyer@cs.cmu.edu

Abstract

In this paper, we leverage the existence of dual subtitles as a source of parallel data. Dual subtitles present viewers with two languages simultaneously, and are generally aligned in the segment level, which removes the need to automatically perform this alignment. This is desirable as extracted parallel data does not contain alignment errors present in previous work that aligns different subtitle files for the same movie. We present a simple heuristic to detect and extract dual subtitles and show that more than 20 million sentence pairs can be extracted for the Mandarin-English language pair. We also show that extracting data from this source can be a viable solution for improving Machine Translation systems in the domain of subtitles.

Keywords: Parallel Data, Dual Subtitles, Machine Translation

1. Introduction

The availability of data determines the pace of natural language processing research. A large parallel corpus with good quality has tremendous value to the field of statistical machine translation (Brown et al., 1993). Many sources have been identified to extract such corpora. Early approaches have relied on parliament data (Koehn, 2005), where sessions transcribed and translated into multiple languages. However, a larger source of parallel data is the Web. Parallel webpages (Web pages that have multiple versions for different languages) have been shown to be a reliable source to extract bitext (Resnik and Smith, 2003; Munteanu and Marcu, 2005). Other approaches leverage the presence of translated material in more specific artifacts within the Web, such as social media (Ling et al., 2013b) and Wikipedia (Smith et al., 2010). In these domains, the large amount of data that is published justify the dedicated method to find translated material. Crowdsourcing translations have also been proposed as a method to obtain parallel data with a low budget (Ambati and Vogel, 2010; Zaidan and Callison-Burch, 2011; Post et al., 2012; Ambati et al., 2012). The goal in these methods is to solicit non-expert workers to translate documents, which are then used as parallel corpora.

Parallel data is not only useful for training and testing MT systems, many other NLP tasks have also benefited from these kinds of resource. Some examples include word sense disambiguation (Gale et al., 1992; Ng et al., 2003; Specia et al., 2005), paraphrasing (Bannard and Callison-burch, 2005; Ganitkevitch et al., 2012), text normalization (Ling et al., 2013a), annotation projection (Das and Petrov, 2011), language learning (Ling et al., 2011b), and other language-specific applications (Schwarck et al., 2010; Liu et al., 2011).

Our work focuses on the extraction of parallel data from subtitles from Movies and TV series, which have also been shown to be a viable source of parallel data (Xiao and Wang, 2009; Tiedemann, 2007a; Lavecchia et al., 2007; Tiedemann, 2007b; Tiedemann, 2008; Itamar and Itai, 2008; Tiedemann, 2012), where the content of the movie is translated into multiple languages. However, the most of the work above build corpra by combining several monolingual subtitles for the same video source, which are generally created by different translators. This generally leads to alignment challenges, which are not easily solved, and consequently, most of the parallel corpora that are extracted from subtitles are noisy. In our work, we exploit the existence of dual subtitles, which is a term we use to define subtitles that contain two languages. An example of this is shown in Figure 1..

Figure 1: Video with Dual Subtitles

In this type of subtitles, the viewer is presented with subtitles in two languages simultaneously, which are manually aligned by the author of the subtitles. This allows for a much more simple extraction process, and resulting corpora extracted from these subtitles are more clean and devoid of alignment errors. In this paper, we described our approach in building a Mandarin-English parallel corpus with 20 million sentence segment pairs. We also leverage the fact that the parallel data is self-contained, that is, both languages are present within the same file, which removes the need to group files that are translations of the same movie. Thus, we use a simple method that retrieves subtitle files from a pre-specified set of websites, detects the subtitle files that contain dual subtitles and extracts the parallel data automatically.

This paper is organized in the following fashion. Section 2.

Sequence	Time	Subtitle
629	00:48:01,140 → 00:48:05,110	托尼？你得上楼把情势控制住 Tony, you gotta get upstairs and get on top of this situation right now.
630	00:48:05,190 → 00:48:08,190	我一整天都在和国民兵通电话 Listen. I've been on the phone with the National Guard all day
631	00:48:09,530 → 00:48:12,360	说服他们不要开坦克车过来 trying to talk them out of rolling tanks up the PCH,
632	00:48:12,450 → 00:48:15,320	破你的大门拿走这些 knocking down your front door and taking these.

Table 1: Example of a subtitle file containing dual subtitles. The *Sequence*, *Time* and *Subtitle* columns represent the number of the segment within the movie, the timeframe that the subtitle is shown and the text that is within the subtitle, respectively.

provides the literature review on the extraction of parallel data from subtitles. Then, we describe our process to find and retrieve parallel data from subtitles in Section 3.. Experiments using this method are performed in Section 4.. Finally, we conclude and discuss future work in Section 5..

2. Related Work

Parallel data is essential for building most MT systems. It is required to train statistical MT models, optimize the model parameters and testing a system's translation quality. However, professional translations are generally expensive to obtain, so a large amount has been exerted on researching alternatives to obtain such datasets. In this section, we shall review the work that has been done to extract the existing parallel datasets.

2.1. Automatic Parallel Data Extraction

Automatic collection of parallel data is a well-studied problem. Approaches to retrieve parallel web documents automatically have received much interest from the MT community (Resnik and Smith, 2003; Fukushima et al., 2006; Li and Liu, 2008; Uszkoreit et al., 2010; Ture and Lin, 2012). These are generally focused on finding promising candidates using different kinds of inputs, such as URL similarity. This is done to reduce the potential number of candidates into a tractable amount. Then identifying truly parallel segments is performed by training classifiers with more expensive features.

Some more specific domains such as microblogs, such as Twitter, have also been targeted for parallel data extraction due to the extraordinary amounts of data that these contain. Some methods rely on Cross-Lingual Information Retrieval techniques (Jehl et al., 2012), while others attempt to find users that translate their messages (Ling et al., 2013b). Other work on finding specific domains in the Web include the extraction of data from Wikipedia (Smith et al., 2010), parenthetical translations (Lin et al., 2008) and anchor texts (Ling et al., 2011a).

Our method is similar to the methods above, in the sense that we find subtitle files that are posted on online forums, extracted them and find parallel sentences from those files. The advantage of this approach is that it does not require human input and can automatically find more parallel data as they are posted.

It is also possible to manually find sources for parallel data. One example is EUROPARL, where a large amount of parallel data is extracted from parliament procedures that are translated into multiple languages. Generally, these methods leverage the structure that these documents are stored in order to find documents that may be translations of each other.

Finally, crowd-sourcing techniques have been applied to the translation task, where non-expert effort is used to obtain translations (Zbib et al., 2012; Post et al., 2012). The main advantage of these methods is that it allows the selection of the data to be translated, while automatic extraction methods can only be used on translations that occur naturally. However, the number of translations that can be obtained daily is limited by on number of workers in the crowdsourcing platform that are can or are willing to perform the task.

2.2. Parallel Data Extraction from Subtitles

Previous work on extracting parallel data from subtitles generally use two monolingual subtitles in different languages from the same movie, and find parallel segments between the two subtitle documents. Aligning subtitle documents is a challenging task, and in (Lavecchia et al., 2007), many causes for misalignments are discussed. Firstly, some authors may add descriptions of movie scenes in the subtitles, causing extra segments to be present in the subtitle file, which are generally unaligned. Secondly, different authors may break segments in different times, which lead to segmentation mismatches. Finally, omitted segments by some authors are another source of misalignments.

The work in (Itamar and Itai, 2008) proposes an alignment algorithm that achieves a precision of 78% and recall of 74% on aligning parallel subtitle segments. While these results significantly outperform those obtained using the Gale and Church alignment algorithm (Gale and Church, 1993), we can still only approximately find three parallel sentences correctly in four subtitles and one in every four parallel sentences is discarded. Similar results are observed in (Tiedemann, 2007b; Tiedemann, 2008). It is clear that a considerable amount of effort was put into this problem, and we have yet to find an effective approach to solve this problem. In this work, we shall not address this problem. Instead, we will introduce another source of parallel data in the domain of movie subtitles, where it is highly likely that the sentences are well aligned and segmented, and that the extraction of parallel data can be performed with high precision and recall, with minimal effort. We will extract a parallel

Topic	Most probable words
1	time home back long coming party times question half answer buy million
2	ve made feel bad father head heard boy city listen hit mother fire
3	don work things care doesn problem won mind change remember worry sex
4	call car happen phone gonna side wanna back heart security safe cell
5	good uh idea family pretty friend business true reason tonight inside office
6	house mr president school today high men sister white speak hassan hot
7	find jack wrong left bit bauer play hand game fuck asked ctu law agent
8	big money thought told girl leave job dad deal late mom lost haven
9	place talking stop called case thinking hands truth pay telling child
10	life real show live guy rest cool force human couldn perfect state earth date

Table 2: Most probable words using LDA on the English side of the parallel data, run with 50 topics on one millions sentence pairs in the English side. The most representative 10 topics, identified manually, are shown.

dataset of 20 million sentence pairs for English-Mandarin, which is a highly demanded language pair, with a relatively low coverage in the Open Subtitles parallel corpora (Tiedemann, 2012). However, there are many technical aspects that must be addressed, namely, the detection and extraction of dual subtitles. We will show that using simple heuristics, it is possible to detect these accurately.

3. Building A Parallel Corpus from Dual Subtitles

While the previous work crawls parallel corpora from aligning monolingual subtitles, we attempt to find subtitles that contain two languages. Some example subtitle segments are shown in Table 1. Obviously, the extraction process is much simpler than previous work, since the problems related to the alignment of subtitles are solved by the authors of the subtitles. As the corresponding Mandarin and English subtitles are designed to show simultaneously in frames, this will allow us to extract cleaner corpora on the domain of movie and drama subtitles.

We will now describe our methodology to find 20 million parallel sentences in Mandarin-English.

3.1. Crawling Subtitles

Unfortunately, subtitle data cannot be found in a single website. These are generally posted as links in forums, such as Shooter forum[1]. Thus, we crawled multiple Mandarin forums and looked for posts that contain links to files that are in two commonly known subtitle extensions, Advanced SubStantion Alpha (ASS) and SubRip Text (SRT). Then, we convert each format into a list of subtitles with their respective filenames and timeframes.

3.2. Detecting Dual Subtitles

To find whether the subtitles are dual, we propose a simple language independent approach to do this. Dual subtitles generally contain two lines in one frame, one line for each language. On the other hand, monolingual subtitles usually only contain one line. If certain line happens to be long, we may have some monolingual subtitles more than one line. However, dual subtitles always have two lines in every frame, and that is unlikely for a monolingual subtitle file. Thus, we simply set a rule that a dual subtitle file is a file where all segments contain more than one line.

[1] http://www.shooter.cn/

3.3. Language Detection

For language detection, we implement a variant of the URL matching algorithm proposed in (Resnik and Smith, 2003), where the URL is used to find parallel websites. We keep the forum URL where each subtitle file was crawled from, and look into the html to find language indicators. In our case, we only keep a subtitle file, if it contains one element in {en, eng, english, 英, 英语} and one element in {chinese, mandarin, cn, zh, 中, 中文, 汉语}.

4. Experiments

Using the process described in Section 3., we obtained 20 million sentence pairs for the Mandarin-English language pair. In this section, we will answer two questions about this dataset. Firstly, how many sentence pairs extracted are indeed parallel. Secondly, what is the representation of this dataset. Finally, to what extent can this dataset improve machine translation.

4.1. Parallel Data Extraction

To compute the accuracy of our dataset, 200 extracted parallel sentences (chosen randomly) were annotated on whether they are parallel by an proficient bilingual speaker. Results indicate that all 200 samples were parallel. This is expected, since dual subtitles are aligned by the author prior to release.

There are cases where the translation is not literal. For instance, in Table 1, we observe see that in segment 629 the English sentence uses a comma after *Tony*, while the Mandarin translation uses a question mark after the translation 托尼. Likewise, segment 630 does not contain the translation for the word *Listen* in the Mandarin translation. A problem we frequently observe in the extracted data is that punctuation is frequently not used in the Mandarin translations and are replaced by whitespaces.

While, these imperfections are unfortunate, MT systems are generally tolerant to small errors in the parallel segments.

As a final step, we remove parallel sentences that are duplicates that are identical in both source and target sides.

	English-Mandarin	Mandarin-English
FBIS	6.99	3.48
Subtitles	**27.63**	**24.08**

Table 3: Translation quality evaluated with BLEU. The two rows correspond to the translation scores obtaining using different training corpus on the same tuning and testing sets.

4.2. Data Representation

We ran topic modeling on a sample of one million sentence pairs of this dataset using Mallet (McCallum, 2002) with 50 topics. The first ten topics are shown in Table 2. We can see that some topics contain very informal language, containing terms such as *sex*, and many conversational artifacts are present in the data, such as *uh*.

We also observe that some topics contain more informal words than others. For instance, topic 4 contains the popular informal expressions *wanna* and *gonna*, and seem to be representative of subtitles for movies that involve crime due to the terms *security*, *safe* and *cell*. On the other hand, topic 1 is represented by more formal terms, which leads us to believe that it represents a set of more formal subtitles.

4.3. Machine Translation Evaluation

As an extrinsic test, we shall test the quality of machine translation systems using our dataset. The goal is to show that the effectiveness of our corpora in improving the translation quality in the domain of movie subtitles.

4.3.1. Corpus

We consider an in-domain and an out-of-domain training corpus. The in-domain corpus is built by extracting 300K sentence pairs obtained from our extracted corpus. As out-of-domain corpus, we used the FBIS dataset which contains 300K high quality sentence pairs, which are extracted from broadcast news, which differs substantially from the subtitle domain.

As held-out data, we extract another 2000 sentence pairs from out subtitle corpus. We used 1000 sentences as development data and another 1000 sentences for testing.

Finally, we used 1 million Mandarin sentences from the subtitle corpus as monolingual data for language modeling.

4.3.2. Setup

Our setup follows a standard Moses (Koehn et al., 2007) pipeline. We build word alignment were generated using IBM Model 4, and then perform phrasal extraction (Ling et al., 2010). As for the reordering model, we use MSD reordering model and the distance-based reordering feature. As language model, we use a 5-gram smoothed (Kneser-Ney) model built using KenLM (Heafield, 2011). Finally, results are evaluated using using BLEU (Papineni et al., 2002).

For each translation direction, we then build two systems, one using the subtitles corpus and another one using the FBIS corpus as training corpora. Aside from the training corpus, the systems were left under the same conditions.

4.3.3. Results

Results are shown in Figure 3. As expected, we observe that using the subtitle corpus yields better scores in both directions, even using the same amount of parallel sentences. There are many factors for this large improvement. One reason is the mismatch between the news domain where the FBIS corpus was extracted from, which radically differs from most subtitles. One important factor is the presence of fundamental lexical gaps in the FBIS dataset, which are frequently used in movies. Some examples include informal terms, such as *wanna* and *gonna*, which are variations for the *want to* and *got to*. As these are left untranslated with the FBIS dataset these, and occur frequently, the gap between the translation quality of the two systems is naturally large. This also happens for Mandarin, where many expressions that are seldom found in the news domain are frequently mis translated.

Table 4 provides an examples of sentences that are incorrectly translated using the FBIS corpus. In most cases, due to the lack of domain specific knowledge the system built using the FBIS corpus tends to use literal translations. For instance, the expression 永别了 is equivalent to the English word *Farewell*, which is translated correctly using the subtitle corpus. However, it literally means *Forever*, which is the option that is frequently seen in the FBIS corpus, which is not an accurate translation in this context. The same happens with the proper name *Tommy*, which is also translated literally as soap. We can also see that in most cases, using a FBIS corpus yields translations that are radially different from the reference. As subtitles are generally short, in most cases, the translations obtained using the FBIS corpus tend to not match well with the reference, which is also a factor for the divergences in the BLEU scores between the systems.

Thus, we can see that the extracted corpus can be used to not only improve state-of-the-art MT systems on the translation of subtitles, but it can also be used to boost the quality of the MT systems on more informal domains.

5. Conclusion

In this work, we presented an alternative source for crawling parallel data from the subtitle domain. Previous methods attempt to align multiple subtitle files from different authors for the same movie, which result in alignment errors, and consequently, spurious parallel sentences.

We propose to crawl parallel data from dual subtitles, where two subtitles in different languages are presented simultaneously. As such, the effort to align sentences in the two languages is performed by the author of the subtitles, and the extraction process is simple and practically error free. Thus, minimal effort is needed to automatically align the subtitles.

Applied to Machine Translation, our extracted corpora shows that significant improvements can be obtained using such data compared to the usage of out-of-domain training corpora.

While it is not common practice to use dual subtitles in all countries, these are not only limited for the Mandarin-English language pair. These exist for many other language pairs, such as, Korean-English and Japanese-English. As

Original Mandarin	Reference	FBIS translations	Subtitle translation
永别了汤米	Good-bye, Tommy.	Forever a soup	Good-bye, Tommy.
你这个婊子养的骗子	You lying son of a bitch.	You this 婊raise the swindlers	You're a son of a bitch .
快点出去	Come on. Let's go!	faster out	Come on, let's go !

Table 4: Examples of translations performed using different systems from Mandarin to English. The *Original Mandarin* and *Reference* columns illustrate the original Mandarin sentence and its manual translation, while columns *FBIS translation* and *Subtitle translation* present the translations using the FBIS corpus and the Subtitle corpus.

future work, we shall proceed with the extraction of parallel data for other language pairs using this technique.

Acknowledgements

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533.
The PhD thesis of Wang Ling is supported by FCT Fundao para a Ciencia e a Tecnologia, under project SFRH/BD/51157/2010. This work was supported by national funds through FCT Fundacao para a Ciencia e a Tecnologia, under project PEst-OE/EEI/LA0021/2013.
The authors also wish to express their gratitude to the anonymous reviewers for their comments and insight.

6. References

Ambati, V. and Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ambati, V., Vogel, S., and Carbonell, J. (2012). Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1191–1194, New York, NY, USA. ACM.

Bannard, C. and Callison-burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *In ACL-2005*, pages 597–604.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fukushima, K., Taura, K., and Chikayama, T. (2006). A fast and accurate method for detecting English-Japanese parallel texts. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 60–67, Sydney, Australia, July. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, March.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods.

Ganitkevitch, J., Cao, Y., Weese, J., Post, M., and Callison-Burch, C. (2012). Joshua 4.0: Packing, pro, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 283–291, Montréal, Canada, June. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Itamar, E. and Itai, A. (2008). Using movie subtitles for creating a large-scale bilingual corpora. In *LREC*.

Jehl, L., Hiebel, F., and Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421, Montral, Canada, June. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-burch, C., Zens, R., Aachen, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C., and Bojar, O. (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, June. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Lavecchia, C., Smaili, K., Langlois, D., et al. (2007). Building parallel corpora from movies. In *The 4th International Workshop on Natural Language Processing and Cognitive Science-NLPCS 2007*.

Li, B. and Liu, J. (2008). Mining Chinese-English parallel corpora from the web. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*.

Lin, D., Zhao, S., Van Durme, B., and Paşca, M. (2008). Mining parenthetical translations from the web by word alignment. In *Proceedings of ACL-08: HLT*, pages 994–1002, Columbus, Ohio, June. Association for Computational Linguistics.

Ling, W., Lus, T., Graa, J., Coheur, L., and Trancoso, I. (2010). Towards a general and extensible phrase-extraction algorithm. In *IWSLT '10: International Work-*

shop on Spoken Language Translation, pages 313–320, Paris, France.

Ling, W., Calado, P., Martins, B., Trancoso, I., and Black, A. (2011a). Named entity translation using anchor texts. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA, December.

Ling, W., Trancoso, I., and Prada, R. (2011b). An agent based competitive translation game for second language learning. August.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2013a). Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84, Seattle, Washington, USA, October. Association for Computational Linguistics.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013b). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

Liu, F., Liu, F., and Liu, Y. (2011). Learning from chinese-english parallel data for chinese tense prediction. In *IJCNLP*, pages 1116–1124.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Munteanu, D. and Marcu, D. (2005). Improving machine translation performance by exploiting comparable corpora. *Computational Linguistics*, 31(4):477–504.

Ng, H. T., Wang, B., and Chan, Y. S. (2003). Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL03*, pages 455–462.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June. Association for Computational Linguistics.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

Schwarck, F., Fraser, A., and Schütze, H. (2010). Bitext-based resolution of german subject-object ambiguities. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 737–740, Stroudsburg, PA, USA. Association for Computational Linguistics.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Proc. NAACL*.

Specia, L., Graas, M. D., Nunes, V., and Stevenson, M. (2005). Exploiting parallel texts to produce a multilingual sense tagged corpus for word sense disambiguation. In *Proceedings of RANLP-05, Borovets*, pages 525–531.

Tiedemann, J. (2007a). Building a multilingual parallel subtitle corpus. *Proc. CLIN*, page 14.

Tiedemann, J. (2007b). Improved sentence alignment for movie subtitles. In *Proceedings of RANLP*, volume 7.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *LREC*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Ture, F. and Lin, J. (2012). Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 626–630, Montréal, Canada, June. Association for Computational Linguistics.

Uszkoreit, J., Ponte, J., Popat, A. C., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *COLING*, pages 1101–1109.

Xiao, H. and Wang, X. (2009). Constructing parallel corpus from movie subtitles. In *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, pages 329–336. Springer.

Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 105–110, New York, NY, USA. ACM.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwarz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proc. NAACL*.