Statistical Machine Translation of Subtitles: From OpenSubtitles to TED

Mathias Müller and Martin Volk

Institute of Computational Linguistics, Zurich, Switzerland

Abstract. In this paper, we describe how the differences between subtitle corpora, OpenSubtitles and TED, influence machine translation quality. In particular, we investigate whether statistical machine translation systems built on their basis can be used interchangeably. Our results show that OpenSubtiles and TED contain very different kinds of subtitles that warrant a subclassification of the genre. In addition, we have taken a closer look at the translation of questions as a sentence type with special word order. Interestingly, we found the BLEU scores for questions to be higher than for random sentences.

1 Introduction

The key ingredient for building successful Statistical Machine Translation (SMT) systems is a suitable and sufficiently large parallel corpus. For a number of language pairs, large subtitle corpora are available. The OPUS OpenSubtitles corpus (Tiedemann, 2009) contains fansubs for 54 languages. This collection amounts to several million parallel sentences for the most popular language pairs. On the other hand, there is the collection of subtitles from the TED talks (Cettolo et al., 2012), a series of high quality talks on "Technology, Entertainment, and Design". Although the TED collection is much smaller, it is interesting because of its wide coverage and complexity of topics.

Therefore, we set out to investigate the degree of similarity between the subtitles in the two corpora and to what extent this influences the quality of SMT systems trained on them. As a side issue we wanted to check the general usefulness of the OpenSubtitles corpus for training SMT systems which has been questioned repeatedly in the literature (see e.g. (Petukhova et al., 2012)). In order to gain deeper insight into the impact of the two corpora on specific linguistic phenomena, we evaluated their translation quality on questions.

In this paper we describe the differences between the two subtitle corpora and their impact on translation quality. We trained a number of Moses systems for that purpose, using parallel subtitles in English, French and German. In all, we trained twelve systems under the exact same conditions (preprocessing, Moses command line options). Each system was then tested on three test sets including a set containing only questions.

I. Gurevych, C. Biemann, and T. Zesch (Eds.): GSCL 2013, LNAI 8105, pp. 132–138, 2013.

2 OpenSubtitles vs. TED

Both OpenSubtitles and TED are collections of parallel sentences derived from subtitles. So, we are dealing with sentence-aligned corpora and not subtitle-aligned corpora as, for instance, Volk et al. (2010). Contrary to our intuition, Volk et al. (2010) found that subtitle-aligned corpora are as good for building SMT systems for subtitles as sentence-aligned corpora. The average sentence length is around 6 words per sentence for OpenSubtitles and around 17 for TED, irrespective of the language. It is surprising that the figures across languages are so similar. We believe that this is an artifact of the automatic alignment. Only those sentences that are of similar length were aligned. However, these numbers clearly indicate that OpenSubtitles and TED subtitles are different from one another. Some randomly chosen lines may illustrate this point.

Example 1. **Subtitle examples from OpenSubtitles and TED**[1]
OpenSubs: You miss me, today?
OpenSubs: Faut y penser avant.

TED: The first law is two-colorability: You can color any crease pattern with just two colors without ever having the same color meeting.
TED: Et avec ce qu'il trouve sur place, il entre et fait son petit studio qui sert de base de travail.

The differences in length can be explained if we consider the circumstances: OpenSubtitles are taken from regular movies and series, where sentences tend to be short and where subtitles are shortened to fit on the screen in the available time. TED talks on the other hand treat rather complex subjects that in turn demand more complex sentences. Any differences may have an influence on MT performance when crosstesting between corpora, more precisely on BLEU scores in our case. For our experiments, we used OpenSubtitles, as prepared in the OPUS project, for language pairs between EN, FR and DE in all possible combinations. The OpenSubtitles collections for these languages are large, ranging from 2.8 million sentence pairs for DE-FR up to 20 million sentence pairs for EN-FR. These corpora should be sufficient, given that Hardmeier and Volk (2009) argue that 1 million sentence pairs is suitable for subtitle SMT. Still we need to remember that subtitles in OpenSubtitles are of an unknown quality. Some are controlled and consistent, but others contain spelling errors or strange wordings. All texts are already sentence-aligned and formatted in a Moses-friendly way. As for the limitations of sentence-alignment techniques, see Tiedemann (2009).

The TED collection (Cettolo et al., 2012) is smaller. These subtitles are crawled from `www.ted.com`, a platform offering talks that were recorded at TED conferences or similar events. These videos sometimes come with subtitles translated into 30 or more languages by volunteers from within the TED community. The translations are generally high-quality because TED requires translators

[1] Obviously unrelated sentences, not translations that correspond.

to peer-review their work and prove their proficiency. These corpora, too, have been preprocessed and sentence-aligned by (Cettolo et al., 2012) as to allow using them in Moses with ease. Again, we only used the parts in EN, FR and DE.

In total, we built 12 SMT systems between EN, FR and DE on the basis of 6 different corpora from TED and OpenSubtitles. We used the same material for both directions, for example the same corpus to translate EN-FR and FR-EN. Table 1 shows the corpora's sizes as the number of lines (roughly equal to the number of sentences) and the number of words using a naive tokenization.

Table 1. Corpus sizes

Corpus	Language	Number of lines	Number of words
OpenSubtitles EN-DE (DE-EN)	EN	4,654,635	26,266,191
	DE	"	27,189,072
OpenSubtitles EN-FR (FR-EN)	EN	19,858,798	119,682,551
	FR	"	115,456,439
OpenSubtitles DE-FR (FR-DE)	DE	2,862,370	16,946,049
	FR	"	16,818,332
TED EN-DE (DE-EN)	EN	63,865	1,029,090
	DE	"	1,034,657
TED EN-FR (FR-EN)	EN	114,582	1,916,788
	FR	"	2,000,958
TED DE-FR (FR-DE)	DE	62,148	967,935
	FR	"	1,056,758

3 Building SMT Systems

Starting out with the corpora described in the section above, we built phrase-based Moses systems (Koehn et al., 2007) and tested their performance. Moses is used widely and is the state-of-the-art tool for statistical machine translation. We divided each corpus into training set (97 percent), tuning set (1 percent) and test set (2 percent) and assigned parallel lines randomly to the sets[2]. All of the data passed through the usual stages of preprocessing as we cleaned, lowercased and tokenized it using scripts offered by the Moses toolkit. For word-alignment we used GIZA++ (Och and Ney, 2003) which is implemented as part of Moses.

In order to build the language model we employed SRILM (Stolcke et al., 2011) together with Kneser-Ney discounting for smoothing, and interpolation as a back-off model for probabilities. These two options are an official recommendation by the Moses developers[3]. In general, we used the standard methods and options where possible and consistently applied the same rules to all systems.

[2] While creating the test set via randomly allocating the lines is statistically sound, it might be more natural to test the subtitles of whole movies or series.

[3] See `http://www.statmt.org/moses/?n=FactoredTraining/BuildingLanguageModel` for further information.

The baseline systems were then tuned with MERT (also part of Moses), optimizing them with respect to the tuning set. After assembling complete Moses systems, we tested each on three different test sets to obtain BLEU scores.

4 First Results

We conducted several experiments on the TED and OpenSubtitles collections. In all cases, the performance of the MT systems was measured with *multibleu*, a script distributed with Moses. Table 2 reports the BLEU scores of our systems. The most straightforward test set for each system is the 2 percent of the original corpus set aside at the beginning, the "native test set". In contrast, the "foreign test set" is the native set's equivalent from the other collection. In other words, TED systems are subjected to foreign test data taken from OpenSubtitles and vice versa.

Table 2. Performance results in BLEU scores

Language pair	System	Test set	
		native	foreign
DE-EN	OpenSubtitles	27.92	20.56
	TED	25.06	14.29
DE-FR	OpenSubtitles	17.18	14.65
	TED	17.64	9.69
EN-DE	OpenSubtitles	19.55	16.93
	TED	24.38	12.59
EN-FR	OpenSubtitles	22.86	23.56
	TED	31.87	14.89
FR-DE	OpenSubtitles	13.42	10.72
	TED	13.12	8.34
FR-EN	OpenSubtitles	23.52	**24.92**
	TED	**33.37**	16.87

First, let us consider the difference in performance between OpenSubtitles and TED systems when confronted with their native test sets. Out of six OpenSubtitles systems, the highest scores are achieved with DE-EN, scoring almost 28. Only TED systems surpass 30, and only when translating between EN and FR. Thus, good performance results wherever EN is involved, irrespective of its being the source or target language. On the other hand, combinations of DE and FR lead to the lowest scores. To understand this, we have to bear in mind that most movies and all TED talks are in English in the first place. Often, the English transcription is done first and translators base their work on English subtitles. Therefore, combinations with English (EN-FR, FR-EN, DE-EN, EN-DE) can be expected to be translations of one another, whereas combinations between DE and FR (DE-FR, FR-DE) are not directly related.

With regard to the foreign sets, the performance ranking changes somewhat. OpenSubtitles translating from FR to EN now takes the lead, resulting in a BLEU score of approximately 25. In general, the TED systems are affected more severely when confronted with foreign test lines, their scores plummeting to 50 percent of the former value in some cases.This indicates that the OpenSubtitles systems are more apt at translating TED than the other way round. Also, it implies that TED systems are more overfitted and OpenSubtitles systems more universal if our goal is to translate subtitles in general.

5 Investigating MT Quality for Questions

As a case study we have investigated the MT quality of questions. Questions are special because they have word order that is different from assertive clauses, and they use question words and special auxiliary verbs. For the sake of simplicity, a question is a line ending with a question mark. Here are some typical questions.

Example 2. **Question examples from OpenSubtitles and TED**
OpenSubs: Is the needle in his femoral artery, Mr. Palmer?
OpenSubs: Für dich sind wir nur Leichen, oder?

TED: And we asked ourselves, why couldn't it be exhibitionistic, like the Met, or like some of the other buildings at Lincoln Center?
TED: Was ist die Botschaft, was ist das Vokabular und die Grammatik, die von diesem Gebäude ausgesandt wird, und was sagt es uns über uns selbst?

A line that ends in a question mark in some cases might not be a question. For example, the English TED sentence in the example above is an assertion or an indirect question. But such cases are rare and are ignored here. Some of the lines ending with a question mark in one language do not have an equivalent counterpart, i.e. in the translation there is no question mark at the end. We disregarded them for our tests.

Our questions test set contains only questions from the native test set. We used the "question set" both for a quantitative (performance measures in BLEU scores) and qualitative analysis (manual error categorisation). With respect to questions, OpenSubtitles systems performed slightly better compared to the native test set, their scores climbing one or two BLEU points (see table 3). The scores of TED systems adapted slightly. Given the differences in performance between the native and question set, questions surprisingly score higher than the average subtitle of any type. We speculate that this might be due to the fact that questions are shorter than the average subtitle. The latter is 33.7 characters long – calculated over all the lines we took from OpenSubtitles. The average question line taken from OpenSubtitles counts no more than 27.2 characters. We get similar values for the TED corpora.

In order to evaluate the performance of translating questions qualitatively, we have looked at up to 100 translated questions from EN-DE and FR-DE, both

Table 3. BLEU scores for native and questions test sets

System	Test set	
	native	questions
OpenSubtitles DE-EN	27.92	31.11
TED DE-EN	25.06	27.09
OpenSubtitles DE-FR	17.18	19.61
TED DE-FR	17.64	17.14
OpenSubtitles EN-DE	19.55	21.73
TED EN-DE	24.38	27.99
OpenSubtitles EN-FR	22.86	23.28
TED EN-FR	31.87	29.7
OpenSubtitles FR-DE	13.42	15.37
TED FR-DE	13.12	15.37
OpenSubtitles FR-EN	23.52	23.55
TED FR-EN	33.37	30.61

OpenSubs and TED. In particular, we have paid attention to the types of errors that occur. Our categories are fragmentation (translation unit span too narrow), omission, lack of agreement, difficulty with ambiguous terms, reordering, ulexis issues and addition (of a phrase). The following errors were repeatedly made. Systems translating from FR to DE frequently omitted an infinitive, whereas this never happened when translating from EN to DE. Also, only EN-DE systems treated many auxiliary verbs as full verbs. Out-of-vocabulary problems are a more serious issue with FR-DE, presumably because verb-pronoun compounds like "atterrissez-vous" or "a-t-il" are common in French. The majority of lexis errors is concerned with a hyphenated French word like those. Deliberately tokenizing these forms as part of the preprocessing would alleviate this effect.

6 Conclusion

Parallel corpora of subtitles are a valuable source for machine translation and are frequently used. We compared corpora from the TED and OpenSubtitles collections, and we suggest that "subtitles" is in fact too broad a category. Four rows of test sets revealed that the systems can hardly be used interchangeably, since sentence length, broad applicability and subtitle quality mark stark differences between subtitles from OpenSubtitles and from TED. They may be so different that they might best be treated as different genres indeed. We isolated questions and found slightly better BLEU scores for them as compared to randomly selected sentences.

In future studies, it might prove fruitful to incorporate data from both collections into one system and assign weights to each in order to counteract the different sizes of the training corpora. One way to achieve this is to combine the phrase tables resulting from building translation models (see Sennrich (2012)).

References

Cettolo, M., Girardi, C., Federico, M.: Wit^3: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), Trento, Italy, pp. 261–268 (2012)

Hardmeier, C., Volk, M.: Using linguistic annotations in statistical machine translation of film subtitles. In: Nodalida (2009)

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., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: ACL (2007)

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

Petukhova, V., Agerri, R., Fishel, M., Penkale, S., del Pozo, A., Maucec, M.S., Way, A., Georgakopoulou, P., Volk, M.: SUMAT: Data collection and parallel corpus compilation for machine translation of subtitles. In: LREC, pp. 21–28 (2012)

Sennrich, R.: Perplexity minimization for translation model domain adaptation in statistical machine translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Avignon, France (2012)

Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: Update and outlook. In: Proceedings IEEE Automatic Speech Recognition and Understanding Workshop (2011)

Tiedemann, J.: News from opus - a collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, vol. V, pp. 237–248. John Benjamins, Amsterdam (2009)

Volk, M., Sennrich, R., Hardmeier, C., Tidström, F.: Machine translation of TV subtitles for large scale production. In: Proceedings of the Second Joint EM+/CNGL Workshop on Bringing MT to the User: Research on Integrating MT in the Translation Industry, Denver, pp. 53–62 (2010)