Finding Alternative Translations in a Large Corpus of Movie Subtitles

Jörg Tiedemann
Department of Modern Languages
University of Helsinki
firstname.lastname at helsinki.fi

Abstract

OpenSubtitles.org provides a large collection of user contributed subtitles in various languages for movies and TV programs. Subtitle translations are valuable resources for cross-lingual studies and machine translation research. A less explored feature of the collection is the inclusion of alternative translations, which can be very useful for training paraphrase systems or collecting multi-reference test suites for machine translation. However, differences in translation may also be due to misspellings, incomplete or corrupt data files, or wrongly aligned subtitles. This paper reports our efforts in recognising and classifying alternative subtitle translations with language independent techniques. We use time-based alignment with lexical re-synchronisation techniques and BLEU score filters and sort alternative translations into categories using edit distance metrics and heuristic rules. Our approach produces large numbers of sentence-aligned translation alternatives for over 50 languages provided via the OPUS corpus collection.

Keywords: parallel corpora, alignment, machine translation

1. Introduction

Multilingual parallel corpora are useful resources for many purposes such as machine translation, cross-lingual studies, bilingual lexicography but also monolingual tasks such as word sense disambiguation and discovery (Dyvik, 2002) and the detection of idiomatic expressions (Villada Moirón and Tiedemann, 2006). In most cases, they contain exactly one version per language, referring to either the source text or its translation. However, there are cases in which one would like to consider alternative translations, for example, when evaluating machine translation using metrics such as BLEU (Papineni et al., 2002). In general it is not satisfactory (and reliable) to use these automatic evaluation methods with one reference translation only and scores obtained in this way may be quite misleading as many acceptable translation alternatives are not considered. Furthermore, reference-based parameter tuning is also effected by this limitation. Nevertheless, most research is reported with single-reference test/development sets as multiple translations are difficult to obtain.

In this paper, we investigate the use of user-contributed movie subtitles as a source of alternative translations. We base our study on the OpenSubtitles corpus included in OPUS (Tiedemann, 2012), which provides a large number of subtitles in various languages including alternative uploads for many movies and TV shows. In the following, we first discuss the issues of user-uploaded subtitles that can lead to translation differences and after that we propose our method for filtering the data set that makes it possible to identify various categories of subtitle alternatives.

As a side product we also obtain cross-lingual links between all subtitle files via the monolingually aligned subtitle alternatives that make it possible to create truly multilingual parallel corpora across all languages, which is not available in the previous version of the corpus. Yet another result is the identification of possible errors (spelling mistakes, encoding problems, etc.) in the collection. This information will be very useful to clean-up the corpus in future releases. More details and examples are given below.

2. Issues With User-Contributed Subtitles

Subtitles in OpenSubtitles are sorted by language, release year and movie ID. Language checking and other filtering and pre-processing techniques have been used to clean-up the data (Tiedemann, 2007a). Nevertheless, it still contains a lot of noise in terms of misspellings and character encoding issues. Figure 1 shows a few examples of misspellings in some Swedish subtitles.

Får jag fråga en sak?	F**a**r jag fr**a**ga en sak?
Ge mig väskan.	**C**e mig väskan.
Det här blir ditt rum .	Det här blir ditt rum**m** .

Figure 1: Typical spelling errors in subtitles.

Another difference can be due to the use of alternative punctuations that may also lead to differences in sentence segmentation and tokenisation. Certainly, these differences are not very interesting when looking for truly alternative translations. However, misspellings are important to identify for further cleaning of the data or for filtering out corrupted portions of the collection.

Another issue with OpenSubtitles is that TV shows do not have separate IDs for all their individual episodes. Hence, it is not possible to know from the ID whether a file is an alternative upload of subtitles for the same video or whether it refers to another episode of the same TV series. This fact complicates the extraction of candidates of translation alternatives. We handle this problem by a brute-force approach, aligning all possible combinations but selecting only those that pass a certain overlap threshold, which is explained in the next section.

Finally, it is common that user-contributed subtitles use slightly different timings when synchronising to the video. This issue is already adressed when aligning different translations with each other (Tiedemann, 2008) and we treat it in the same way when aligning alternative subtitles in the same language.

lang	punct	insert	spell	other
ar	119,612	25,668	85,325	34,508
bg	214,588	54,776	117,476	147,809
bn	4		9	
br			7	9
bs	54,346	9,416	478,163	101,664
ca	31	67	170	530
cs	646,407	124,977	1,327,214	571,956
da	61,261	17,529	94,216	34,649
de	68,730	40,124	113,470	24,228
el	757,978	88,862	1,017,116	272,948
en	18,195,828	9,329,184	16,006,229	4,788,733
eo	39		180	
es	1,771,127	558,997	2,442,295	2,031,533
et	35,160	10,029	18,988	32,390
eu	6	3	2	
fa	34,753	2,737	6,303	10,028
fi	49,707	7,787	50,718	18,258
fr	1,089,621	170,882	933,970	284,505
gl	414	15	244	50
he	200,041	95,069	55,980	135,558
hi	2	5	6	18
hr	291,406	66,687	1,378,230	546,773
hu	195,125	41,244	153,310	92,622
id	23,180	9,354	17,956	38,135
it	80,085	96,488	324,351	227,901
ja	3,027	1,568	589	3,981
ko	288	16	48	58
lt	453	757	1,859	464
lv	35	12	26	27
mk	2,484	819	3,306	3,750
ml	156	66	51	330
ms	363	296	303	1,524
nl	511,291	76,181	386,400	250,531
no	16,370	2,282	38,954	2,332
pl	1,336,093	297,981	903,037	590,141
pt	398,114	251,651	800,126	951,509
pt_br	987,801	628,544	1,315,918	2,088,517
ro	498,345	141,234	7,412,091	450,734
ru	35,807	12,289	27,458	56,933
si	432	10	188	14
sk	7,477	4,303	21,952	12,211
sl	303,897	55,096	257,415	167,579
sq	1,886	5,717	6,813	2,940
sr	562,258	120,725	3,221,672	1,166,972
sv	64,601	21,656	133,746	33,407
th	2,870	1,530	3,163	14,053
tl	9		67	
tr	423,006	105,329	629,594	289,557
uk	229	142	301	697
vi	1,137	2,116	1,232	4,100
zh	140,126	22,530	57,643	33,437
zh_tw	17,798	5,553	13,497	5,852

Table 1: Translation units in each category for languages in our collection. Language IDs follow the ISO-639-1 standard; extensions of local varieties are added such as pt_br for Brazilian Portuguese and zh_tw for Traditional Chinese.

3. Finding Subtitle Alternatives

In the first step of our procedure we look for subtitles referring to the same movie ID that have a time overlap of at least 50%. The time-overlap is a rather soft constraint that allows to filter out many candidates that have nothing to do with each other without restricting the selection too much as there are many wrongly synchronised candidates in the collection. We sort all possible candidates by their time overlap and consider a maximum of 20 subtitles per file that pass the overlap threshold to reduce the search space. For each candidate pair, we use the time-based alignment algorithm presented by Tiedemann (2007b) to match sentences with each other. Subtitle pairs that produce more empty alignments than non-empty ones are discarded.

More recently, we improved the pre-processing steps and properly link movies and TV episodes to IMDb identifiers using metadata from the provider. In that procedure, we also merge subtitles that are split into various parts due to the separate distributions of related video files. More details on the conversion steps are given in Lison and Tiedemann (2016). Having our collection sorted by movie identifiers, we can now proceed to align alternative candidates.

The first step in the process is a content-based filter. We compute BLEU scores between monolingually aligned subtitles and discard all file pairs that obtain a score below 50%. With this we get rid of most candidates that do not match for one reason or another. For the remaining pairs, we try to re-align the ones that have BLEU scores below 80% with synchronisation techniques proposed by Tiedemann (2008). In particular, we use lexical matches between tokens of five characters or more to find the best time adjustments that maximise the ratio between non-empty and empty alignments. After that we re-compute BLEU and keep the new alignment if the score improves. In order to clean-up alignment even more, we add another step that checks neighbouring sentences of the proposed alignments. In this step, we re-arrange alignments if those neighbours match sentences included in the current alignment unit. This greatly reduces the number of misaligned sentences.

After all these steps, we have a number of alternative subtitle pairs aligned at the sentence level. The next task is now to check to what degree they differ and to use that information to classify alternative translations into different categories.

4. Sorting Alternative Translations

For sorting translation alternatives, we rely on string comparison and some heuristics. We are interested in separating the following cases:

Insertions: Some sentences are identical except for some inserted text (words or phrases).

Punctuation differences: Sentence pairs that only differ in their use of punctuation and/or white-spaces.

Spelling differences: Minor differences in a few words in otherwise identical sentences.

Alternative translations: Sentence pairs that use paraphrased expressions or are substantially different from each other (in word order or any other way).

Misaligned sentences: Sentence pairs that should not be aligned because they do not refer to the same information, not even partially.

differences in punctuation/tokenisation	
Ask away ...	” Ask away ! ”
Please , stop crying .	Please stop crying .
Don ’t be a smart ass !	Don ’t be a smartass !
insertions / deletions	
But ... (screaming)	But ...
STORM : ln that case , why ?	- ln that case , why ?
Sasaki !	Dr. Sasaki !
My goodness .	Oh , my goodness .
variation in spelling (spelling errors)	
Only Magneto is capabl of something like this .	Only Magneto is capable of something like this .
I accuse those who are asleep ...	l accuse those whoo are asleep ..
And the visionary told them many other things .	And the visionary told the any other things .
However her heart swayed , Edith suffered .	However her heart swayed , Edih suffered .
misaligned sentences	
Go on .	Are you serious ?
Hit .	*Beat me .*
how much did you know of her ?	Haven ’t you learnt how to shoot ?
I really want to tell her about this	Kanzaki !
other (often paraphrased translations)	
” As mother wants it . ”	” As mother wishes . ”
” Did you mend my coat ? ”	Was it you who mended my coat ’ ? ”
What ’s the matter ?	What ’s wrong ?
Now this is the eight-inch pipe .	This is the 8-inch pipe .
” Let me have her to myself now that she is dying . ”	Now that she ’s dying , let me have her to myself .
” We don ’t have the sterilising oven going yet . ”	It may be full of germs , and our sterilizing oven isn ’t working yet .
” For a year thou wilt reap souls . ”	For a year thou wilt set souls free from the earthly realm .
I ’m gonna fire some of those people .	I ’m gonna fire some of them .

Figure 2: Alternative English subtitles sorted into different categories. The italic sentences in the *misaligned sentences* category are an example of a wrongly classified alignment.

To make these distinctions we compare each individual translation unit in our sentence-aligned collection of alternative subtitle pairs. We apply the following tests to the aligned strings for categorising the data:

Punctuation: First of all, we test whether they only differ in non-alphanumeric characters. If this is the case, then we put them into the *punctuation* class.

Spelling variations: Sometimes, encoding issues can cause tokenisation errors due to erroneously recognised characters. Usually, such errors are very consistent and involve only a few specific characters that are incorrectly converted. To handle these cases, we check whether the strings include different numbers of tokens but otherwise are very similar and of similar lengths. We do the latter by matching characters at identical positions and count the number of mismatches sorted by individual characters. We use the heuristics that sentences with more than 12 words may contain any number of mismatches between at most three different character pairs, sentences with more than 6 words but maximum 12 words may contain mismatches between two different character pairs and shorter sentences may have mismatches between at most one particular character pair. We run this procedure on strings with or without spaces to increase the recall. Sentence pairs in this category are marked as *spelling* errors.

Inserted content: The next step is to compute word-level edit distance (using insertion, deletion, substitution and match as the basic edit operations). We ignore punctuation in this procedure. If all words of one language are present in the other language string, then we put the translation unit into the *insertion* class.

Paraphrases: If there are no insertions and deletions but word substitutions according to the edit distance algorithm then we compute character-level distances for all non-matching words to see whether they are substantially different from each other or just minor spelling variants (indicating possible errors). We use two criteria to make decisions: (i) We apply thresholds over edit operations and length-normalised edit distance scores and (ii) we use thresholds over the maximum number of non-matching characters in a row and the positions of the edit operations. The following heuristics are used to mark strings as substantially different:

- The edit distance is more than one and the normalised edit distance is greater than 0.5.
- The strings are at least 5 characters long and the normalised edit distance is between 0.4 and 0.5.
- The normalised edit distance is between 0.3 and 0.4 and differences are at more than one non-contiguous edit position.
- There are more than three edit operations in a row.

If all non-matching word pairs are dissimilar according to one of these criteria, then we put the translation unit in the *other* class assuming that we have found a truly alternative

translation in our data set. In all other cases, we add the translation unit into the *spelling* class. If there is a mixture of word-level insertions, deletions and substitutions then we always put the sentence pair into the *other* class without checking the similarity of word substitutions.

Misalignments: Substantial differences may, of course, also be caused by erroneous sentence alignments. Therefore, we use additional heuristics to further filter the *other* class.

As a first rule, we mark an alignment as an error if one string is more than twice as long as the other string unless the time overlap is very high (over 90%) and the previous sentence pair is not marked as an erroneous link.

Furthermore, we assume that even paraphrased translation include matching tokens in most cases and that completely different sentences often refer to alignment errors, especially if they are surrounded by other misaligned sentences. Therefore, we count matching tokens in a given sentence pair (s_1, s_2) by a matching function $m(s_1, s_2)$. We use different weights for matching content words and function words (which we approximate by a string length factor) as well as for certain punctuations (question marks and exclamation marks count more than other types of punctuations). The final score is normalised by the number of tokens in s_1 and compared to a threshold that depends on the number of previously misaligned sentences (according to our heuristics). For the latter we define an exponentially decaying threshold of $1 - 0.9^{error\_count}$ with $error\_count$ giving the number of subsequent misaligned units before the current sentence pair. Alignments with a score below the threshold are then marked as *misaligned* unless they are at the beginning of the subtitle file and have a time-overlap of over 80% or unless there is no attested alignment error before and the time-overlap is more than 60%.

Global properties: Spelling errors are often caused by conversion problems and affect the entire subtitle file. Similarly, files with many alignment errors are often useless in most other places as well. For this reason, we also overwrite the categories *other* and *insert* for individual sentence pairs if the categories *spelling* and *misaligned* dominate in the entire subtitle pair (i.e. if there are more cases than *insert* and *other* together). Files with many character edit operations over only a very few types are also treated in the same way. This reduces the amount of sentences that are wrongly classified as *insert* or *other*, which helps to retrieve a cleaner set of truly paraphrased translation alternatives from the data.

With all these heuristics in place, we are now abel to run through our entire collection that covers 60 languages and language variants. For 52 of them we could extract alternative translations in at least one of the categories that we discuss in the paper. The overall statistics are given in Table 1 A few examples of each category are shown in Figure 2.

5. Applications

Alternative translations have various applications and the categorisation of our data makes it possible to separate different cases that are useful for different purposes.

Spelling variations are in most cases of the kind as shown in Figure 1. Together with the *punctuation* category, these translation units can be very useful for error correction and text normalisation. The *insertion* category can be useful for training systems that compress subtitles, simplify and summarise information, or extract extra-linguistic information such as background noise and speaker-labels, which is given in some subtitles like in the first two examples in Figure 1.

Probably the most interesting category is *other*, which contains mostly paraphrased translations. Even though this category is rather noisy due to the nature of the data and the limitations of our filtering approach, it still represents a valuable collection of true translation alternatives. The quality may vary quite a lot between different languages and further quality assessments need to be carried out to prove their use in downstream applications. One application is the use in machine translation evaluation. Automatic metrics such as BLEU are designed to include multiple reference translations in order to achieve good correlation with human judgements. We ran some initial SMT experiments with our subtitle corpus that demonstrate the effect of multi-reference test sets. The test sets cover ten blockbuster movies for which we can get a reasonable amount of alternative translations. Table 2 lists the results of these experiments showing the consistent increase of BLEU scores when using alternative translations from our data set.[1]

These experiments are by no means any prove of the quality of the alternative translations we extract but demonstrate that they match machine translation output, which indicates that they represent useful alternative reference translations for given input sentences.

Note that only a fraction of the sentences in our test sets have multiple reference translations. However, some of them have even more than two alternative translations. Figure 3 illustrates the proportions for each language pair showing that there are quite substantial differences between

Language pair	Single	Multiple
English-Arabic	8.88	9.28
English-Czech	21.92	22.12
English-Spanish	33.64	34.05
English-French	21.00	21.26
English-Portuguese (BR)	27.24	27.49
English-Russian	13.65	17.05
English-Turkish	15.68	16.14
English-Chinese	11.80	11.85
Arabic-English	25.10	25.34
Czech-English	28.90	29.67
Spanish-English	38.05	39.15
French-English	22.89	24.01
Dutch-French	18.29	18.46
Polish-English	25.82	26.62
Portuguese (BR)-English	31.95	33.34
Russian-English	24.62	25.49
Turkish-English	24.47	24.78
Chinese-English	17.59	18.20

Table 2: Comparing BLEU scores with single and multiple reference translations.

[1] Thanks to Pierre Lison for running the SMT experiments.

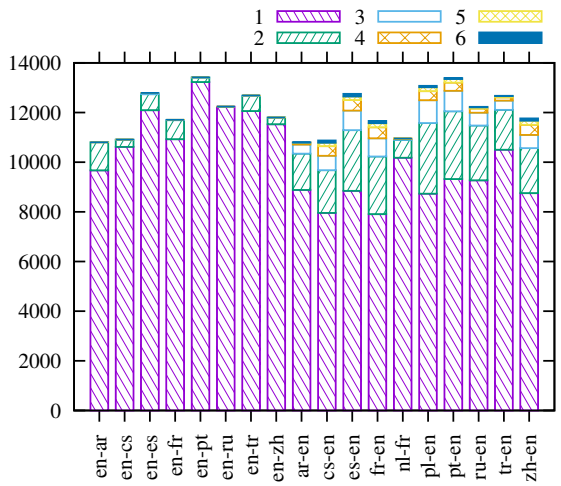

Figure 3: Alternative translations in SMT test sets: The number of sentences with up to 6 distinct reference translations extracted for the selected language pairs in our experiments.

the individual data sets.

6. Conclusions

In this paper, we present our work on aligning alternative versions of movie subtitles for extending the multilingual data set in OPUS, for cleaning and normalising subtitles, and for extracting paraphrases and multi-reference data sets for machine translation research. Time-based alignment, BLEU-based filters and string edit measures are effective tools for performing this task. All our data is publicly available (Tiedemann, 2016) and feedback is welcome to improve the quality of our collection.

7. Bibliographical References

Dyvik, H. (2002). Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, 16:311–326.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016)*, Portorož, Slovenia.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia (Pennsylvania, USA). ACL.

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

Tiedemann, J. (2007b). Improved sentence alignment for movie subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)*, Marrakesh, Marocco.

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

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy, April.

8. Language Resource References

Jörg Tiedemann. (2016). *OpenSubtitles2016 – Intra-Lingual Sentence Alignments*. University of Helsinki / Uppsala University, `http://opus.lingfil.uu.se/OpenSubtitles2016alt.php`, part of OPUS - The Open Parallel Corpus.