Detecting Machine-translated Subtitles in Large Parallel Corpora

Pierre Lison, A. Seza Doğruöz
Norwegian Computing Center, Independent Researcher
Oslo, Norway, Turkey
plison@nr.no, a.s.dogruoz@gmail.com

Abstract

Parallel corpora extracted from online repositories of movie and TV subtitles are employed in a wide range of NLP applications, from language modelling to machine translation and dialogue systems. However, the subtitles uploaded in such repositories exhibit varying levels of quality. A particularly difficult problem stems from the fact that a substantial number of these subtitles are not written by human subtitlers but are simply generated through the use of online translation engines. This paper investigates whether these machine-generated subtitles can be detected automatically using a combination of linguistic and extra-linguistic features. We show that a feedforward neural network trained on a small dataset of subtitles can detect machine-generated subtitles with a F_1-score of 0.64. Furthermore, applying this detection model on an unlabelled sample of subtitles allows us to provide a statistical estimate for the proportion of subtitles that are machine-translated (or are at least of very low quality) in the full corpus.

Keywords: Parallel corpora, Machine Translation, Quality Estimation

1. Introduction

The availability of movie and TV subtitles for a large number of languages and linguistic genres makes them particularly useful for the construction of parallel corpora. Currently, the largest collection of subtitles is the OpenSubtitles corpus with 3.35 billion sentences covering 60 languages (Lison et al., 2018). In addition to their textual content, subtitles are also associated with precise timestamps indicating when each subtitle block should be displayed. These timestamps allow subtitles to be efficiently aligned across languages based on time overlaps (Tiedemann, 2007; Tiedemann, 2008). These time-based alignments can in turn be used to extract multilingual parallel corpora.[1]. In addition to the OpenSubtitles corpus, other corpora based on subtitles include the SUMAT data collection (Petukhova et al., 2012), the collection of dual subtitles from (Zhang et al., 2014), the Tehran English-Persian parallel corpus (Pilevar et al., 2011) and the Japanese-English subtitle corpus (Pryzant et al., 2017).

From a linguistic perspective, parallel corpora derived from subtitles are appealing due to their coverage of a broad range of conversational genres and speaker styles. Subtitles are also widely used in practical NLP applications, notably for neural and statistical machine translation (Belinkov and Glass, 2016; van der Wees et al., 2016; Wang et al., 2017), but also conversational modelling (Lison and Bibauw, 2017; Krause et al., 2017), semantic role labelling (Akbik et al., 2016) and distributional semantics (Lison and Kutuzov, 2017; Speer and Lowry-Duda, 2017).

Despite their benefits, corpora extracted from subtitle repositories also have some shortcomings. The most important issue is the varying quality of the subtitles in terms of linguistic fluency, faithfulness to the dialogues in the source material (movie or TV episode), and adherence to formatting guidelines. Subtitles made available in online repositories such as OpenSubtitles[2] are typically created by movie and TV fans rather than translation and subtitling professionals. An important portion of subtitles are not even produced by human translators at all (professional or not) but are merely generated using online machine translation engines based on other existing subtitles. The linguistic quality of these machine-generated subtitles is typically quite low, as they are typically left unedited and contain numerous grammatical and translation errors.

This paper presents a machine learning model for detecting such machine-translated subtitles based on a combination of linguistic and extra-linguistic features. Despite the difficulty of the detection task, the model achieves a reasonable performance and can be used to either filter out low-quality subtitles from the corpus or assign them with a document weight that can be passed to downstream applications.

The rest of the paper is organised as follows. The next section reviews related work on detecting machine-translated texts. Section 3 presents the dataset employed in this paper and provides several examples of translation errors observed in machine-translated subtitles. Section 4 defines the linguistic and extra-linguistic features that can be employed for detecting such subtitles. Section 5 details the empirical evaluation and error analysis of the approach. Section 6 shows how the detection model can be used to estimate the number of machine-translated subtitles in the full corpus and ultimately enhance the overall corpus quality. Finally, Section 7 concludes the paper.

2. Background

The comparison between machine-translated, human-translated and "original" (non-translated) texts has been the subject of numerous studies in translation studies and machine translation research. Translated texts can often be distinguished from non-translated texts due to interferences from the source language (where some aspects of the source language "spill" onto the translation output) combined with artifacts of the translation process that are independent of the source language (Koppel and Ordan, 2011). In particular, human-translated texts often make use of a more "standard" language than original texts (Toury, 1995), allowing them to be detected automatically (Kurokawa et al., 2009).

[1] `http://opus.nlpl.eu/OpenSubtitles2018.php`

[2] `http://www.opensubtitles.org`

Language	**Number**				
English	669	Swedish	46	Chinese (simplified)	10
Indonesian	580	Danish	45	Tamil	10
Spanish	519	Russian	43	Norwegian	10
Portuguese (Brazilian)	462	Serbian	41	Catalan	7
Romanian	327	Slovenian	40	Thai	6
Hebrew	326	Malay	37	Chinese (traditional)	6
Turkish	269	Albanian	35	Esperanto	5
Bulgarian	220	Dutch	31	Bengali	4
Arabic	193	Vietnamese	29	Basque	4
Polish	127	Ukrainian	26	Finnish	4
Persian	101	Japanese	23	Lithuanian	3
Portuguese	100	Hungarian	22	Korean	3
Italian	98	Estonian	18	Galician	2
Croatian	97	Slovak	17	Macedonian	2
German	92	Sinhalese	15	Malayalam	1
French	82	Hindi	14	Tagalog	1
Czech	79	Bosnian	11	Urdu	1
Greek	76	Telugu	10		
				Total:	**4 999**

Table 1: Number of subtitles explicitly marked with a "machine-generated" flag in the OpenSubtitles corpus, distributed by subtitling language.

This standardisation make them well-suited for language modelling (Lembersky et al., 2012). The term "*translationese*" is often used to refer to these peculiarities of translated documents compared to non-translated ones.

In comparison with human-translated texts, machine-translated documents are of course subject to various type of translation errors (Vilar et al., 2006; Stymne and Ahrenberg, 2012) that degrade the quality of the resulting texts. Arase and Zhou (2013) presented a data-driven approach aimed at detecting low-quality translations in web texts, using monolingual corpora only as input. Their features specifically focused on "phrase salads" in which the phrases of sentences are correct in isolation but become inaccurate when put together as a complete sentence. Aharoni et al. (2014) described a related approach and found a correlation between the performance of the machine learning model and the human evaluation of translation quality.

The two aforementioned approaches focused on specific language pairs for which large quantities of in-domain data is either already available or can be generated. In contrast, the detection model presented in this paper aims to be applicable to any language pair, without relying on the occurrence of translation errors specific to a given source or target language. Indeed, as explained in the next section, machine-generated subtitles can be found in virtually every language present in the corpus. Furthermore, these subtitles do not include any information about the subtitle it was translated from, nor even the source language. The detection model must therefore scale to a broad spectrum of possible language pairs while relying on a relatively small number of parameters (due to the modest amount of machine-generated subtitles available for training).

It should also be noted that machine-generated subtitles have been present in subtitle repositories since the early 2000s. As a consequence, the translations are a result of a broad mixture of translation tools, from early rule-based MT systems to modern APIs for statistical and neural machine translation. This leads to large disparities in the translation quality and typical error patterns observed in these subtitles. This stands in contrast with the aforementioned approaches which only relied on translations generated from specific, well-optimised statistical machine translation systems to train and evaluate their models.

3. Data

3.1. Subtitle corpus

The data employed in this paper comes from the latest version of the OpenSubtitles corpus released as part of the OPUS corpus repository (Tiedemann, 2012; Lison et al., 2018). The latest release comprises 3.73 million subtitles[3] in 60 languages. Each subtitle is converted into Unicode, segmented into sentences and tokenised according to the procedure outlined in (Lison and Tiedemann, 2016). For each language pair, subtitles associated with the same movie or TV episode (identified through their IMDB identifier[4]) are aligned at the sentence level, based on the respective timestamps of the two subtitles (Tiedemann, 2008). This alignment procedure leads to a total of 1 782 bitexts (language pairs must share at least one common movie or TV episode in order to form a bitext).

In addition to the tokenised sentences, each subtitle is also enriched with meta-data information regarding the movie or TV episode (release year, genre, original language) and the subtitle itself (upload date, user ratings, etc.). Unfortunately, we do not have any direct information about who

[3] In this paper, we use the term "subtitle" to refer to the whole file that contains the transcriptions for a given movie or TV episode. Each subtitle is itself composed of many (up to several thousands) subtitle blocks, where each block contains at most two lines of text and is associated with a start time and end time.

[4] `http://www.imdb.com`

created a given subtitle or for which purpose it was created. Some subtitles are created from scratch by fans, while others are "ripped" from official DVD releases or TV streams (which can sometimes be inferred from the presence of OCR errors in the subtitles). Yet another subset of subtitles are translations from other existing subtitles. For instance, a movie fan might wish to create a Spanish subtitle for a Japanese movie, but, not being fluent in Japanese, might opt for translating from an existing English subtitle instead of creating the subtitle from scratch. The translation quality of these subtitles is uneven at best, especially when translated with the help of online translation engines and left unedited. This is especially the case for subtitles uploaded before 2010, at a period where machine translation engines were of a much lower quality than today.

To address these quality issues, the administrators of the OpenSubtitles website have asked their users to mark machine-generated subtitles with an explicit flag when uploading new subtitles. However, only a small fraction of the machine-generated subtitles have so far been annotated with this flag (4 999 subtitles in total) as users are reluctant to declare that their uploaded subtitles are of lower quality. Table 1 illustrates the distribution of these subtitles by language.

3.2. Translation issues

One reason for this particularly low quality of machine-generated subtitles stems from the fact that, with the possible exception of documentaries, subtitles are conversational in nature and typically contain many short-sentences whose interpretation is tightly coupled with the preceding context. This content is ignored by machine translation engines as they operate at the sentence level.

This leads to problematic translations such as in the example below, extracted from an English subtitle. The subtitle is made for a 1945 Danish movie but the subtitle is apparently translated from an existing French subtitle.

(1) * *And Michael? It must come back, you hear?*
(**French**): Et Michael? Il doit revenir, vous entendez?
'And Michael? He must come back, you understand?'

We observe from Example (1) that the 3^{rd} person pronoun 'il' is mistranslated into 'it', while the preceding utterance makes it clear that the pronoun refers to a person.

Other well-documented translation errors include inaccurate lexical choices, wrong word order or mismatched inflectional endings. Here are two other examples of failed translations from the same subtitle, including both wrong lexical choices and grammatical errors:

(2) * *Come, you will see well.*
(**French**): Venez, vous verrez bien.
'Come, you'll see.'

(3) * *How are you take you?*
(**French**): Comment vas-tu t'y prendre?
'How will you go about it?'

Here is yet another example of failed translation, this time in a Dutch subtitle machine-translated from English:

(4) * *Hij is gonna verkopen ons allen langs de rivier.*
(**English**): 'He's gonna sell us all down the river'

Several translation mistakes are at play in Example (4). First of all, the English expression 'sell X down the river' is translated literally. Second, the word 'gonna' is not translated at all and simply repeated in the Dutch output. Finally, Dutch word order – which is verb-final in subordinate clauses – is not respected.

Another common error when translated into prop-drop languages (Doğruöz, 2014) relates to the use of redundant subject pronouns . The example below illustrates a redundant subject pronoun in Turkish:

(5) * *Ben* *telefonumu* *aldı*
I telephone-poss.1sg-acc take-past I
Ben *döndü* *ve* *bu* *iki* *gövde* *vardı.*
turn-past and these two body.
'I took my phone, I turned and there were these two bodies.'

Example (5) illustrates two translation issues. First, the two verbs ('take' and 'turn') lack person agreement markers. In addition, the second subject pronoun is redundant since it was already used in the first sentence and does not deliver new or contrastive information.

4. Approach

The detection of machine-translated subtitles is a challenging task, as we have no direct information about the actual source subtitle (or even the source language) that was used as translation input. Furthermore, the machine-translated subtitles are spread over a wide range of languages, as illustrated in Table 1. The features of the detection model must therefore be as language-independent as possible.

The features employed in the presented approach can be divided in two groups:

- *Target-side features*, extracted from the subtitle itself.
- *Subtitle pair features*, extracted by determining the most likely source subtitle and extracting similarity features between the source and target sentences.

4.1. Target-side features

Target-side features are defined on the sole basis of the subtitle itself. One important observation is that machine-generated subtitles typically contain a slightly larger proportion of rare/unknown tokens than their human-generated counterparts. Indeed, source-side tokens that the MT engine is unable to translate will often be repeated in the target sentence, as in the following example (where the contracted word '*tryin'*' is seemingly not understood by the MT engine and left untranslated in French):

(6) * *Regarde comme il est tryin' pour prendre sa température.*
(**English**): Looks like he's tryin' to take her temperature.

In order to detect such rare or unknown tokens, we relied on statistical language models to (1) determine the number of tokens unknown to the language model and (2) compute the log-probabilities over the bigrams extracted from a given subtitle. The language models are derived from the Google Web 1T 5-gram corpus (Brants and Franz, 2006) when available and are estimated from the OpenSubtitles corpora otherwise (excluding the subtitles used in the evaluation). The number of unknown tokens (such as “tryin’ in French) and the number of bigrams with very low log-probabilities are then integrated as features to the machine learning model. To account for the fact that distinct languages will have distinct distributions for these log-probabilities (due to e.g. differences in the vocabulary size of the various language models), the thresholds are empirically determined as percentiles of these language-specific distributions.

Subtitles are also associated with meta-data such as the release year, movie genre, release type (e.g. DVD) and original language of the movie or TV episode. These variables are also included as features in the machine learning model using one-hot encodings. Finally, a small number of subtitles include explicit clues in the beginning or end of the subtitles indicating that a machine translation engine was used. The occurrence of these cues (notably the presence of the words “Google” or “auto-translated”) are also integrated as target-side features.

4.2. Subtitle pair features

The comparison between the source-side and target-side sentences can also yield useful information.

Identification of source subtitle

The first step is to identify the source subtitle that may have served as input to the machine translation engine. To determine this source, we first determine a list of potential candidates, namely subtitles associated with the same movie or TV episode but written in another language.

To find the most likely source subtitle among this list of potential candidates, a good criteria is to look at the timestamps (start and end times of subtitle blocks, in milliseconds) that are used in the subtitle. Indeed, subtitles translated from other subtitles will often have identical or near-identical timestamps, as there is no reason for the user to modify these timings. More precisely, assume a subtitle s_t written in language $l(s_t)$ and associated with the movie or TV episode with IMDB identifier $I(s_t)$. We wish to identify the source subtitle s_s from the same IMDB $I(s_s) = I(s_t)$ but written in language $l(s_s) \neq l(s_t)$ and that stands closest to s_t in terms of timestamps associated with each subtitle block. One way to measure this proximity is to extract the set of all timestamps $T(s_s)$ for subtitle s_s and the set of all timestamps $T(s_t)$ for subtitle s_t, and compute the Jaccard coefficient between the two sets:

$$J\left(T(s_s), T(s_t)\right) = \frac{|T(s_s) \cap T(s_t)|}{|T(s_s) \cup T(s_t)|} \tag{7}$$

We can then rank the list of subtitle candidates s_s for a given target subtitle s_t according to this Jaccard coefficient. To limit the number of candidates to consider, we constrain the possible source languages $l(s_s)$ to be either:

- A large “pivot language”, such as English, Spanish, Russian, French, or Arabic ;
- The original language of the movie or TV episode.

The vast majority of machine-translated subtitles are indeed translations from these restricted set (mostly due to the wider availability of subtitles in these languages).

Surface-level features

Once the most likely source subtitle is determined, one can align the sentences from the two subtitles using the time-based method described in (Tiedemann, 2008) and extract features from the aligned sentence pairs.

One simple set of features is defined by the ratio between the number of tokens (and characters) in the source and target sentences. Indeed, machine-generated subtitles will often consist in literal translations of the source-side sentences, and will typically have have an average ratio close to one. On the other hand, subtitles created by human users will often show more variation in their transcription of the original dialogues, with some parts being left out, rephrased or selectively presented. This higher degree of variation will in turn lead to larger differences in the ratios of tokens (and ratios of characters) between the source and target sentences. These length ratios are, however, language-specific, as the average number of tokens may vary from language to language (as modelled in machine translation through word penalties). These differences are taken into account by rescaling the ratios by language.

Syntactic features

We can observe empirically that machine-translated subtitles are also more likely to follow the syntactic structure of the source subtitle than their human-generated counterparts. This is again due to the fact that machine-translated subtitles have more literal alignments than subtitles created by human users.

To capture this similarity, we extract the sequence of POS tags and dependency relations of the source and target subtitles through UDPipe (Straka and Straková, 2017) and extract k-gram precision scores from them:

$$\text{precision}_k = \frac{|k\text{-grams in both source and target}|}{|k\text{-grams in source}|} \tag{8}$$

The precision scores for each pair of (source,target) subtitles are then employed as features.

5. Evaluation

The features described in the previous section can be used to learn a classifier that detects whether a given a subtitle is likely to be machine-translated.

5.1. Experimental design

The dataset used for the experiments consists of a sample of 54 999 subtitles from the OpenSubtitles corpus, divided in two classes. The first class consists of the 4 999 subtitles explicitly marked as machine-generated in their meta-data (see Table 1). The second class comprises 50 000 subtitles that are (presumed to be) human-generated. As there is no absolute guarantee that a subtitle is not machine-generated,

Model	Hyper-parameters	Precision	Recall	F_1 score	Accuracy
Keyword baseline	"Google" at start/end of subtitle	1.000	0.017	0.030	0.910
Jaccard baseline	Jaccard coefficient $\geq$ 0.99	0.360	0.248	0.294	0.841
Logistic regression	Regularisation = l_2, C =1	0.266	0.757	0.394	0.787
	Regularisation = l_2, C =10	0.267	0.758	0.395	0.787
	Regularisation = l_1, C =1	0.263	0.756	0.390	0.784
	Regularisation = l_1, C =10	0.262	0.756	0.389	0.783
Support Vector Machines	Kernel = linear, C =1	0.268	0.751	0.395	0.790
	Kernel = linear, C =10	0.244	0.750	0.356	0.744
	Kernel = RBF, C =1	0.372	0.803	0.508	0.858
	Kernel = polynomial, C =1	0.340	0.708	0.460	0.848
K-nearest neighbours	Nb. neighbours = 1	0.610	0.514	0.558	0.925
	Nb. neighbours = 5	0.436	0.684	0.532	0.890
	Nb. neighbours = 10	0.359	0.757	0.486	0.854
Decision tree	Min. samples per leaf = 1	0.436	0.431	0.434	0.897
	Min. samples per leaf = 2	0.428	0.453	0.440	0.895
	Min. samples per leaf = 5	0.399	0.521	0.452	0.884
Random Forest	Nb. estimators = 10	0.718	0.409	0.521	0.931
	Nb. estimators = 50	0.758	0.449	0.564	0.937
	Nb. estimators = 100	0.772	0.448	0.567	**0.937**
Gradient Boosting	Nb. estimators = 10	0.710	0.412	0.521	0.931
	Nb. estimators = 50	0.753	0.449	0.563	0.936
	Nb. estimators = 100	0.762	0.444	0.561	0.936
Neural network (MLP)	1 hidden layer with dim. 10	0.377	0.808	0.513	0.860
	1 hidden layer with dim. 50	0.506	0.697	0.585	0.909
	1 hidden layer with dim. 100	0.580	0.661	0.617	0.925
	1 hidden layer with dim. 200	0.622	0.657	**0.638**	0.932
	2 hidden layers with dim. (10, 10)	0.374	0.812	0.512	0.858
	2 hidden layers with dim. (50, 10)	0.504	0.685	0.580	0.909

Table 2: Experimental results for the task of detecting machine-generated subtitles in a dataset of 54 999 subtitles, of which 9 % are explicitly marked as machine-generated.

we selected the subtitles that had the highest average user ratings (as users are more likely to give a high user rating to a high-quality, human-translated subtitle than a machine-generated one). The 50 000 subtitles were sampled according to the same language distribution than the 4 999 subtitles to avoid statistical biases between the two classes.

All features were scaled by removing the mean and scaling to unit variance. In addition, we found that transforming the feature values to follow a uniform distribution using quantiles information ("quantile transform") improved the performance of most classifiers. Features whose values may depend on language-specific properties (such as the average number of tokens per sentence) were scaled on a language by language basis. Class reweighting was used to compensate for the class imbalance in the dataset.

The performance of these classifiers is evaluated through 10-fold stratified cross-validation on the dataset of 54 999 subtitles, with the precision, recall, F_1-score and accuracy as performance metrics.

5.2. Models

Two simple, rule-based baselines are employed:

1. The first baseline looks at the occurrence of the token "Google" in the first and last sentences of the subtitle (which are typically indicative of a machine-translation, such as in "Tradução by Google"). This baseline has perfect precision, but only covers a small fraction of the machine-translated subtitles.

2. The other baseline looks at whether the Jaccard coefficient from Equation (7) is $\geq$ 0.99, indicating that the timestamps are identical or near-identical to another subtitle for the same movie or TV episode. This baseline has a higher recall but a lower precision, as many subtitles will share the same timings without being translations from one another (this is notably the case for subtitles extracted from DVD releases).

The following machine-learning models were estimated based on the features in Section 4 :

1. Logistic regression (with L_1 or L_2 regularisation)

2. SVMs (with linear, RBF or polynomial kernels)

3. K-nearest neighbours

4. Decision trees (with Gini as split criterion)

5. Random forests and gradient boosting trees

6. Feed-forward neural networks with one or two hidden layers. The networks use rectified linear units as activation layer and Adam as optimisation algorithm.

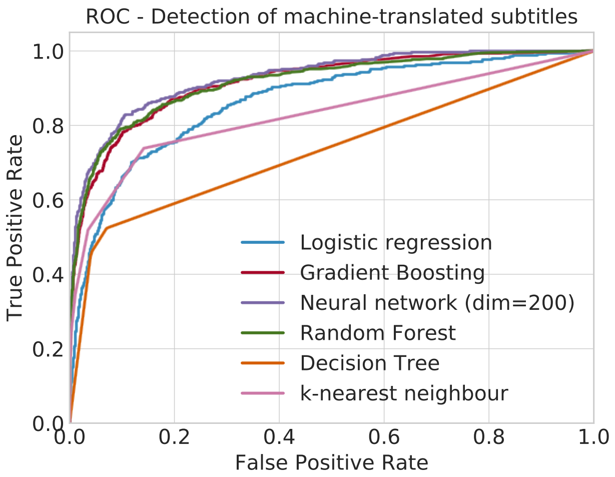

Figure 1: ROC curve for 6 machine learning models on the task of detecting machine-translated subtitles, based on the dataset of 54 999 subtitles (of which 10 % are known to be machine translated).

5.3. Results and error analysis

The results are shown in Table 2. The best performing models are feed-forward neural networks with one hidden layer, with a F_1 score of 0.638. Random forests achieve a slightly higher accuracy on this dataset, but accuracy is a less relevant metric than F_1 given the class imbalance of this task. The performance gain of neural networks seems to indicate the existence of non-linear interactions between the features that cannot be accounted for by “shallow” models such as logistic regression. All feature families described in Section 4 seem to be useful for the task (based on a small-scale feature ablation study). The most discriminative features for the task are the Jaccard coefficient, the occurrence of the “Google” keyword, and the number of unknown tokens according to the language model.

Figure 1 shows the ROC (Receiver Operating Characteristics) curve for each family of machine-learning models with the exception of SVMs which do not directly provide probabilistic estimates. The curve plots the true positive rate (equivalent to the recall) against the false positive rate when the discrimination threshold is varied.

The results demonstrate nevertheless the difficulty of the task. We conducted an error analysis of the classification results, and found most errors to be imputable to two factors. The first factor is that the “machine-translated” flags associated with the 4 999 subtitles are not always accurate. We observed a number of subtitles that were flagged as machine-translated that were surprisingly well written and lacked any obvious translation errors. In other words, their inclusion in the set of machine-translated subtitles is most likely due to a human classification error. Unfortunately, a manual cleanup of this dataset would require finding annotators capable of assessing the fluency of subtitles in most of the languages listed in Table 1, which would constitute a major undertaking.

Furthermore, the set of 50 000 subtitles assumed to be human-generated also has some shortcomings. One important problem, described in (Lison and Tiedemann, 2016) stems from the fact that many subtitles are extracted from video streams through Optical Character Recognition (OCR) and include therefore optical recognition errors, such as the letter ‘i’ being mistaken for the letter ‘l’. These spelling errors are a source of confusion for the language model used to identify unknown tokens and determine bigram log-probabilities. We also observed subtitles including a mixture of machine-generated and human-edited sentences, often combined with numerous spelling and grammatical errors. This leads to a relatively large number of false positives. It should nevertheless be pointed out that these false positives also reflect subtitles of low-quality that one might wish to prune out of the corpus .

6. Discussion

6.1. Estimates on full corpus

The detection models presented in Section 5 can be employed to extrapolate the total number of machine-translated subtitles – or at least on the number of subtitles of suspiciously low quality – in the full corpus. We selected a random sample of 30 000 subtitles from the OpenSubtitles corpus (excluding the subtitles used in the evaluation). We then extracted the features from Section 4 and applied the most accurate detection model (the feedforward neural network with one hidden layer of 200 dimensions) on these features. As the output probabilities of the neural network are not calibrated, we perform probability calibration using Platt’s sigmoid model (Guo et al., 2017).

The resulting distribution of probabilities (using Kernel Density Estimation) is illustrated in Figure 2. We can observe from the figure that most of the probability mass lies within the lower half of the distribution, but a small proportion of subtitles has a high probability of being machine-translated according to the detection model.

Base on this empirical distribution, we can proceed to estimate the number of machine-translated subtitles on the full OpenSubtitles corpus through a Poisson Binomial Distribution, which corresponds to the sum of independent Bernoulli trials that are not identically distributed (in this case, the probabilities of being machine-translated). The mean of this distribution is set to 327 K (out of 3.735 million subtitles) with a standard deviation $\sigma = 335.8$. In other words, the proportion of machine-translated subtitles (and other subtitles of similarly low quality) amounts to about 9 $\%$ of the total corpus.

6.2. Corpus filtering

The detection models presented in Section 5 can be used to detect at least a substantial portion of the machine-translated subtitles in the OpenSubtitles corpus. As illustrated by the ROC curve in Figure 1, the neural model is notably able to detect 51 % of the machine-subtitles with a false positive rate of just 1 %. Given the sensitivity of the model to the number of unknown tokens and the bigram log-probabilities, the detected subtitles are presumably also the ones with the lowest quality in terms of linguistic flu-

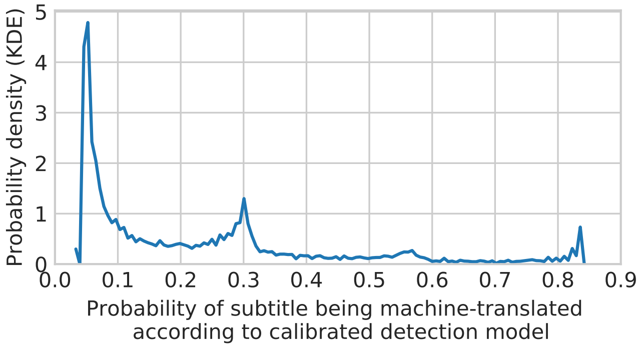

Figure 2: Distribution of probability values given by the calibrated neural network on the set of 30 000 subtitles of unknown class. Kernel Density Estimation is employed for the probability density function.

ency (and thus the ones causing the most important degradation to the quality of the resulting corpus).

The predictions from the detection model can be exploited in several ways. The most straightforward is to directly filter out these (presumed) machine-translated subtitles from the corpus. Alternatively, one can integrate the outputs of the prediction as a distinct feature in the statistical rescoring model of (Lison et al., 2018), which associates each sentence alignment with a numerical score. The latter approach has the advantage of allowing for various filtering levels, from conservative (keeping all subtitles in the corpus) to aggressive (removing all suspicious subtitles), without committing to a specific threshold. Finally, one can also transform the prediction into weights associated with each subtitle. Such weights can be used in various downstream applications, for instance when training machine translation models (Matsoukas et al., 2009)

Although the evaluation presented in this paper focused on subtitles, it should be pointed out that most features employed in the detection models (with the exception of metadata features) are genre-independent and can be extracted on other types of parallel or comparable corpora.

7. Conclusion

Parallel corpus extracted from movie and TV subtitles can be particularly noisy and include a large number of low-quality subtitles. One important cause of this low-quality is the presence of subtitles translated from other subtitles through online machine translation tools. Detecting and pruning out (or downsampling) these subtitles is therefore expected to enhance the overall quality of such corpora.

The present paper described a data-driven approach to the detection of machine-translated documents based on a combination of linguistic and extra-linguistic features. Experimental results show that a detection model based on a feed-forward neural network with one hidden layer is able to achieve reasonable performance on this task. In contrast with previous work, the machine learning models are not optimised for a specific language pair or translation model and can be directly applied to any multilingual corpus. The detection model can be used to filter out machine-translated documents from the corpus or assign them to a lower weight in downstream applications.

Future work will investigate how to further improve our understanding of the relations between subtitles and uncover the “history” behind each subtitle. Subtitles are indeed connected to each other in a myriad of ways:

- Some subtitles are translations of subtitles in other languages, as addressed in this paper. These translations may be done by (professional or amateur) human translators, machine translation tools, or a combination of both (machine-assisted translation).
- A second group consist of subtitles that are part of the same release (for instance, subtitles included in the same DVD). Such subtitles are often created by the same translation/subtitling company and are therefore relatively close at a structural level, although they are typically not translations of one another.
- Finally, many subtitles are corrections of previous subtitles in the same language (for instance to correct spelling, grammatical or formatting errors).

The relations above are important for the construction of parallel corpora from subtitles, as they provide key insights on the relative quality and proximity of each pair of subtitle. In the longer term, we wish to integrate these inferred relations into the ranking model employed for the document-level alignment process (Lison and Tiedemann, 2016).

References

Aharoni, R., Koppel, M., and Goldberg, Y. (2014). Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 289–295. ACL.

Akbik, A., Guan, X., and Li, Y. (2016). Multilingual aliasing for auto-generating proposition banks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 3466–3474.

Arase, Y. and Zhou, M. (2013). Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1597–1607. The Association for Computer Linguistics.

Belinkov, Y. and Glass, J. (2016). Large-scale machine translation between Arabic and Hebrew: Available corpora and initial results. *arXiv preprint arXiv:1609.07701*.

Brants, T. and Franz, A. (2006). Web 1T 5-gram corpus version 1. Technical report, Google Research.

Doğruöz, A. S. (2014). On the borrowability of subject pronoun constructions in Turkish-Dutch contact. *Constructions and Frames*, 6(2):143–169.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting* *of the Association for Computational Linguistics (ACL 2011)*, pages 1318–1326. ACL.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. L. (2017). Edina: Building an open domain socialbot with self-dialogues. *CoRR*, abs/1709.09816.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*.

Lembersky, G., Ordan, N., and Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Lison, P. and Bibauw, S. (2017). Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2017)*, pages 384–394, Saarbrücken, Germany. ACL.

Lison, P. and Kutuzov, A. (2017). Redefining context windows for word embedding models: An experimental study. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (Nodalida 2017)*, pages 284–288, Göteborg, Sweden. Linköping University Electronic Press.

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)*.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan. (accepted).

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 708–717. Association for Computational Linguistics.

Petukhova, V., Agerri, R., Fishel, M., Penkale, S., del Pozo, A., Maucec, M. S., Way, A., Georgakopoulou, P., and Volk, M. (2012). SUMAT: Data collection and parallel corpus compilation for machine translation of subtitles. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Pilevar, M. T., Faili, H., and Pilevar, A. H. (2011). Tep: Tehran English-Persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79. Springer.

Pryzant, R., Chung, Y., Jurafsky, D., and Britz, D. (2017). JESC: japanese-english subtitle corpus. *CoRR*, abs/1710.10639.

Speer, R. and Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. ACL.

Stymne, S. and Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Tiedemann, J. (2007). 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)*, pages 1902–1906, Marrakesh, Marocco.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Benjamins translation library. John Benjamins Publishing Company.

van der Wees, M., Bisazza, A., and Monz, C. (2016). Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 2571–2581.

Vilar, D., Xu, J., D'haro, L., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Wang, L., Tu, Z., Zhang, X., Liu, S., Li, H., Way, A., and Liu, Q. (2017). A novel and robust approach for pro-drop language translation. *Machine Translation*, pages 1–23.

Zhang, S., Ling, W., and Dyer, C. (2014). Dual subtitles as parallel corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.