High-quality bilingual subtitle document alignments with application to spontaneous speech translation

Andreas Tsiartas*, Prasanta Ghosh, Panayiotis Georgiou, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, United States

Received 24 July 2010; received in revised form 27 July 2011; accepted 27 October 2011

Available online 16 November 2011

Abstract

In this paper, we investigate the task of translating spontaneous speech transcriptions by employing aligned movie subtitles in training a statistical machine translator (SMT). In contrast to the lexical-based dynamic time warping (DTW) approaches to bilingual subtitle alignment, we align subtitle documents using time-stamps. We show that subtitle time-stamps in two languages are often approximately linearly related, which can be exploited for extracting high-quality bilingual subtitle pairs. On a small tagged data-set, we achieve a performance improvement of 0.21 F-score points compared to traditional DTW alignment approach and 0.39 F-score points compared to a simple line-fitting approach. In addition, we achieve a performance gain of 4.88 BLEU score points in spontaneous speech translation experiments using the aligned subtitle data obtained by the proposed alignment approach compared to that obtained by the DTW based alignment approach demonstrating the merit of the time-stamps based subtitle alignment scheme.

Keywords: Movie subtitle alignment; Spontaneous speech translation

1. Introduction

Speech-to-speech (S2S) systems are used to translate conversational speech among different languages. In S2S systems, a critical component is the statistical machine translator (SMT). Due to the broad range of topics, domains, and different speaking styles that need to be potentially handled, an enormous amount of bilingual corpora that adequately represent this variety is ideally required to train the SMT. Therefore the S2S research and development efforts have not only focused on manually collecting multilingual data but also on automatically acquiring data, for example, by mining bilingual corpora from the Internet matching the domain of interest.

It is advantageous for the SMT of an S2S system to be trained on bilingual transcriptions of spontaneous speech corpora because they match the spontaneous speech style of ultimate S2S usage. A source of bilingual corpora that has gained attention recently is movie subtitles. Aligned subtitle documents in two languages can be used in SMT training. In this work, our efforts focus on extracting high quality bilingual subtitles from movie subtitle documents.

* Corresponding author. Tel.: +1 9512885023.
E-mail addresses: tsiartas@usc.edu (A. Tsiartas), prasantg@usc.edu (P. Ghosh), georgiou@sipi.usc.edu (P. Georgiou), shri@sipi.usc.edu (S. Narayanan).

© 2011 Elsevier Ltd. All rights reserved.

This paper has been recommended for acceptance by ‘Guest Editors Speech–Speech Translation’.


0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
Corpora alignment research for training machine translators has been active since the early 90s. Past works have introduced a variety of methods for sentence alignment including the use of the number of tokens of each utterance (Brown et al., 1991), the length of sentences (Gale and Church, 1991), and the frequency, position and recency information under the dynamic time warping (DTW) framework (Fung and Mckeown, 1994).

Movie subtitle alignment as a source of training data in S2S systems is attractive due to the increasing number of available subtitle documents on the web and the conversational nature of speech reflected in the subtitle transcripts. Recently, there have been many attempts to align bilingual movie subtitle documents. For example, Mangeot and Giguet (2005) were one of the first to describe a methodology to align movie subtitle documents. Lavecchia et al. (2007) posed this problem as a sequence alignment problem such that the total sum of the aligned utterance-similarities is maximized. Tsiartas et al. (2009) proposed a distance metric under a DTW minimization framework for aligning subtitle documents using a bilingual dictionary and showed improvement in subtitle alignment performance in terms of F-score (Manning et al., 2009). Even though the DTW algorithm has been used extensively, there are inherent limitations due to the DTW assumptions. Notably, the DTW-based approaches have the disadvantage of not providing an alignment quality measure, resulting in the use of poor translation pairs depending on the performance of the alignment approach. Using such poor translation pairs results not only in degrading the performance but also in increasing the training and decoding time, an important factor in SMT design.

As a rule of thumb, increasing the amount of correct bilingual training data improves the SMT performance. Objective metrics for evaluating the performance of SMTs include the BLEU score (Papineni et al., 2002). Sarikaya et al. (2009) reported BLEU score improvements using subtitle data with only 49% accurate translations, demonstrating the usefulness of subtitle data. It should be noted that Sarikaya et al. included an additional step to their scheme by automatically matching the movies first, resulting in a potentially noisy step that can cause performance degradation. This step can be avoided since many subtitle websites offer deterministic categorization of subtitle documents with respect to the movie title. Importantly, their approach has not used any information from the sequential nature of bilingual subtitle documents alignment as done in DTW approaches.

Timing information has been considered in subtitle documents alignment. Tiedemann (2007b, 2008, 2007a) synchronized subtitle documents by using manual anchor points and anchor points obtained from cognate filters. In addition, an existing parallel corpus was used to learn word translations and estimate anchor points. Then, based on the estimated anchor points, subtitle documents were synchronized to obtain bilingual subtitle pairs. However, in many cases a parallel corpus is either not available or there is a domain mismatch, so in such cases anchor point estimation using parallel corpus is not a feasible option. Itamar and Itai (2008) introduced a cost function to align subtitle documents using subtitle durations and sentence lengths under the DTW framework to find the best alignments. However, this approach fails when the subtitle documents contain many-to-one and one-to-many subtitle pairs because they tend to skew the sentence length and subtitle timing duration. Even when there are only one-to-one subtitle pairs, it requires that the subtitles have approximately the same length which might not be true for all language pairs. Also, time shifts and offsets (Itamar and Itai, 2008) can distort the subtitle durations. Xiao and Wang (2009) proposed an approach that uses time differences, and the approach was applied only for subtitle documents having the same starting and ending time-stamps. They reported comparable performance to subtitle alignment works using lexical information. In addition, they reported performance gains by incorporating lexical information.

Time-stamps can be crucial and important in aligning subtitle document pairs. In this work, we aim to study the properties and the benefits of the timing information and matching bilingual subtitle pairs using time-stamps. We propose a two-pass method to align subtitle documents. The first pass uses the Relative Frequency Distance Metric (RFDM) (Tsiartas et al., 2009) under the DTW framework. Using the DTW approach and the lexical information, we identify bilingual subtitle pairs. It is crucial at this point to find pairs that are actual translations of each other and that have timing information describing the deterministic relation between the time-stamps. The identification and usage of these pairs is incorporated in the proposed approach. The second pass uses timing information to align subtitle documents. In particular, we assume that there exists an approximately linear mapping between the time-stamps of the bilingual subtitle documents that can align the bilingual subtitle pairs. This assumption is verified experimentally for most of the bilingual subtitle documents available in our bilingual subtitle sets. This approach results in high quality translation pairs and, in a small set with tagged mappings, significant improvement in the alignment accuracy is obtained compared to that in our prior work (Tsiartas et al., 2009). Also, the performance of this method is demonstrated by training and testing an SMT using downloaded subtitle documents from the web (http://www.opensubtitles.org) on a large scale.
This paper is structured as follows. In Section 2, we present the theory and implementation used in this work. In Section 3, we describe the experimental results and the evaluation methodology used in our approach. Finally, in Section 4, we summarize the results of this work.

2. Theory and methodology

We start by formulating the subtitle alignment problem under the DTW framework. Next, we formulate the timestamp-based subtitle alignment method. Finally, we describe the methodology used to align the subtitles under the proposed two-pass approach. The general diagram of the two-pass approach is shown in Fig. 1.

2.1. First step: DTW using lexical information

We follow the definition and approach as followed by Tsiartas et al. (2009). We define the utterance fragments with starting and ending time-stamps as subtitles and the sequence of subtitles of a movie as a subtitle document. The first part of the movie subtitle alignment problem is defined as follows:

Say the subtitles documents in two languages, \( L_1 \) and \( L_2 \) are to be aligned. We denote the \( i^{th} \) subtitle in the \( L_1 \) subtitle document as \( S_{L_1}^i \) and the \( j^{th} \) subtitle in the \( L_2 \) subtitle document as \( S_{L_2}^j \). Also, let \( N_1 \) and \( N_2 \) be the number of subtitles in the \( L_1 \) and \( L_2 \) subtitle documents respectively. We try to estimate the mappings \( m_{ij}^* \) that minimize the global distance as follows (Tsiartas et al., 2009):

\[
\{m_{ij}^*\} = \arg\min_{m_{ij}} \sum_{i,j} m_{ij} DM(S_{L_1}^i, S_{L_2}^j)
\] (1)

where \( m_{ij} = 1 \), if \( S_{L_1}^i \) aligns with \( S_{L_2}^j \) and \( m_{ij} = 0 \) otherwise and \( DM(S_{L_1}^i, S_{L_2}^j) \) is a distance measure between \( S_{L_1}^i \) and \( S_{L_2}^j \).

The above-mentioned optimization problem can be solved efficiently using the DTW algorithm under the following assumptions:
We denote the starting and ending time-stamps of the pairs, for example, subtitles which are bilingual translations of each other. Moreover, consider the two languages as follows:

The DTW block is shown in Fig. 1 in dashed rectangle (a). The details of the DTW algorithm used in this step is described in Appendix A. The inputs are two bilingual subtitle documents and the output is a list of aligned subtitles with their time-stamps. In the next section, we discuss the distance metric used by the DTW.

2.1.1. Distance metric

Following Tsiartas et al. (2009), we define the Relative Frequency Distance Metric (RFDM) between subtitles across the two languages as follows:

Consider the subtitle $S_{L1}^k$ and denote the words in that subtitle by $W_k$. Also, the words of the subtitle $S_{L2}^k$ are translated using a dictionary and the resulting bag of words of the translated subtitle is denoted by $B_j$. Note that both $B_j$ and $W_k$ contain words in the language $L_1$. First, we compute the unigrams distribution of the words in the $L_1$ subtitle document. Using the unigrams distribution of words in the $L_1$ subtitle document, the RFDM is defined as:

$$DM(S_{L1}^k, S_{L2}^j) = \left( \sum_{k \in W_k \cap B_j} p_k^{-1} \right)^{-1}$$

where $p_k$ is the relative frequency of the word $k$ in the $L_1$ subtitle document. RFDM has the property that it gives high-quality anchor points of subtitle pairs. The lower the RFDM score, the higher the similarity of the subtitles is. In particular, low RFDM occurs when there are infrequent words that match in both subtitles. For example, the sum of the inverse probability of infrequent words will be high and, thus, the inverse of the sum will be low. Hence, infrequent words in the text play the important role of aligning subtitle documents. Finally, RFDM is used as a distance metric to obtain the best mappings \{m^*_ij\}.

2.2. Second step: alignment using timing information

We select a subset of the best DTW output mappings \{m^*_ij\} and estimate a relation among the bilingual subtitles.

In this work, we argue that one can relate the time-stamps of most bilingual subtitles using a linear relation. We hypothesize that this linearity assumption stems from the fact that movies are played in different regions and versions with varying frame rates (slope) and varying offset times (intercept).

For this purpose, consider the scenario of aligning subtitle documents in two languages, say $L_1$ and $L_2$. Assume $L_1$ is the source language and $L_2$ is the target language. Also, assume that we know a-priori $M$ actual one-to-one matching pairs, for example, subtitles which are bilingual translations of each other. Moreover, consider the $i$th one-to-one pair.

We denote the starting and ending time-stamps of the $i$th subtitle in $L_1$ by $x_{1i}$ and $x_{2i}$ respectively. The starting and ending time-stamps of the matching subtitle in the $L_2$ subtitle document are denoted by $y_{1i}$ and $y_{2i}$. Hence, using the time-stamps of $M$ pairs, we define the set $P = \{(x_{1i}, y_{1i}), (x_{2j}, y_{2j}) : 1 \leq i \leq M\}$. In addition, we use the following definition:

**Definition 1.** The absolute error, $E$, of a set of $N$ pairs given a linear function $f(x) = mx + b$ is defined by:

$$E = \frac{1}{2N} \sum_{i=1}^{N} |m(x_{1i} - y_{1i}) + mx_{2i} - y_{2i} + 2b|$$

As discussed in the previous paragraph, the end goal is to approximate the relation of the starting and ending time-stamps of bilingual subtitles with an approximately linear function. Under the assumption of linear mapping, the time-stamps are related by $f \left( \frac{x_{1i} + x_{2i}}{2} \right) = \frac{y_{1i} + y_{2i}}{2}$, where $f$ is a linear function. Since in practice the relation is not exactly linear, due to factors like human error in tagging, we allow an absolute error bound for all the bilingual pairs.
Thus, we model the relation between time-stamps of subtitles in $L_1$ and $L_2$ with an $\alpha$, $\epsilon$-linear function of order $N$ which is defined next.

**Definition 2.** A function $f(x) = mx + b$ is called an $\alpha$, $\epsilon$-linear function of order $N$ if for a set of pairs $P = \{\{x_1, y_1\}, \{x_2, y_2\} : 1 \leq i \leq M\}$ there is a set $I \subseteq \{i : 1 \leq i \leq M\}$ of order $|I| = N$ pairs with $3 \leq N \leq M$ such that:

(i) $\frac{1}{\alpha} < \left| \frac{y_2 - y_1}{x_2 - x_1} \right| < \alpha, \; \forall i \in I$ and $\alpha > 1$

(ii) $\epsilon \leq \epsilon$, where $\epsilon$ is the absolute error of $I$ given the linear function $f(x)$.

**Definition 2** uses a linear function $f$ to relate a subset of the set of pairs, $P$, (the starting and ending time-stamps in the source language and the corresponding time-stamps in the target language) under two conditions. Initially, we have $M$ pairs (in practice returned by the DTW step). Then, a subset of $N$ out of $M$ pairs and a linear function $f$ based on the $\alpha$ and $\epsilon$ parameters are defined. The $\alpha$ parameter controls the allowed duration divergence of bilingual subtitles at subtitle level. The $\epsilon$ parameter establishes the connection between the linear function $f$ and the $N$ pairs by imposing a maximum absolute error between the linear function and the points.

In the ideal case, time-stamps are ideally scaled and shifted from source to target time-stamps, no noise is introduced and there are $N$ one-to-one pairs. Any two pairs selected will fall on a line with the same slope, $\alpha$, and $\epsilon \to 0$. Thus, if we could extract the $N$ noise-free one-to-one pairs, then, the relation would be simply a straight line connecting the middle points of the pairs. In other words, the lower the absolute error, the closer is the relation of the pairs to a line, thus, the more approximately linear their relation is. Hence, ideally, we want $\epsilon$ as small as possible.

On the other hand, in the practical case, humans will transcribe the movies separately. On top of the ideal time scaling and shifting, noise will be introduced to the time-stamp points. Hence, the absolute error is used to reflect the linearity of the pairs selected. Using the absolute error as a measure to reflect the linearity of the map offers a great advantage. The absolute error, $\epsilon$, is just an average of $N$ points, thus, $\epsilon$ is robust to $M$ and $N$ variations, making the absolute error comparable across aligning different bilingual subtitle documents. In addition, in practice, it is crucial to select $N$ reliable points to estimate the linear function, rather than considering all $M$ points. At the global level, a movie’s duration could be scaled by a few minutes or seconds. However, at the local level (subtitle level), this duration change is in the order of milliseconds and we expect the bilingual subtitles to have similar durations. For this purpose, $\alpha$ is used to filter bilingual subtitles with large duration divergence.

In summary, modelling the subtitles alignment problem using $\alpha$, $\epsilon$-linear functions offers various advantages compared to the DTW-based modeling approach (Tsiartas et al., 2009). First, $\alpha$ serves as a quality measure to accept or reject the pairs used to estimate the relation. Then, the absolute error, $\epsilon$, is employed to filter the sets of $N$ pairs that cannot describe a linear relation. Consequently, $\alpha$ and $\epsilon$ serve as measures for the quality of the alignments. In addition, alignment using $\alpha$, $\epsilon$-linear functions depends only on timing information rather than on the semantic closeness of the utterances which is more complicated to model.

Based on **Definition 2**, once the $\alpha$ is set, one can find no or infinitely many $m$’s and $b$’s that satisfy the three conditions. However, we seek $m^*$ and $b^*$ that minimize the squared-error of the pairs considered, so that the total squared error is minimum for the $N$ pairs. Such a function is defined next.

**Definition 3.** A function $f^*(x) = m^*x + b^*$ is called an optimal $\alpha$, $\epsilon$-linear function of order $N$ if for a set of pairs $P = \{\{x_1, y_1\}, \{x_2, y_2\} : 1 \leq i \leq M\}$ and $I \subseteq \{i : 1 \leq i \leq M\}$ of size $|I| = N$ the following are satisfied:

(i) The function $f^*$ is an $\alpha$, $\epsilon$-linear function of order $N$.

(ii) $f^*$ minimizes $MSE = \sum_{i \in I} \left( \frac{y_{1i} + y_{2i}}{2} - f^* \left( \frac{x_{1i} + x_{2i}}{2} \right) \right)^2$.

The optimal function parameters, $m^*$ and $b^*$ are estimated using the least squares line-fitting method. The difference between the least squares line-fitting and this method is that we are using a subset of high-quality mappings to estimate the line in order to control the quality of the linear relation. Thus, the relation is robust to errors either from bad estimates of the DTW step or from additional noise. For the sake of completeness, we show the formula for estimating the optimal estimates, $m^*$ and $b^*$ along with the proof in Appendix B.
2.2.1. Implementation

An overall diagram of the proposed implementation described in this section is shown in Fig. 1.

Select one-to-one mappings. As discussed in the previous section, the end goal of this approach is to estimate a relation between the subtitles in the $L_1$ and $L_2$ documents based only on the time-stamps under the assumption that they are approximately related by a linear function. Initially, we need to extract a set of reliable points that best describe the relation between the subtitles in $L_1$ and $L_2$ subtitle documents. For this purpose, we assume that the most reliable mappings are the $K\%$ one-to-one pairs with the lowest RFDM returned by the DTW approach. By one-to-one pairs, we mean the source subtitles each of which is related with exactly one subtitle in the target subtitle document.

This step is shown in the dashed rectangle (b) of Fig. 1. As shown in the diagram the input is the DTW-step output and the output is a list of ranked RFDM values.

Duration ratio bound. After keeping only the one-to-one mappings, $M$ mappings are left. At this point, our goal is to find one $\alpha$, $\epsilon$-linear function of order $N$ which could model the subtitles alignment problem using time-stamps. In practice, we optimize $\alpha$ on a development set and denote this value as $A$. Thus, $A$ acts as a bound to accept only the reliable mappings to be used in the $A$, $\epsilon$-linear function parameters estimation. To justify the usage of this bound, we study and present its relation with correct and incorrect mappings.

Fig. 2 shows the empirical distribution of the duration ratio, \[ \frac{|y_{2i} - y_{1i}|}{x_{2i} - x_{1i}} \], for correct mappings along with the empirical distribution for incorrect mappings. The distribution of correct mappings shows that the ratio of pairs duration is mostly in the range \( \frac{1}{2} < \frac{|y_{2i} - y_{1i}|}{x_{2i} - x_{1i}} < 2 \). Thus, it is reasonable in practice to impose this constraint on the duration ratio of the mappings to filter out the incorrect mappings.

Fig. 3 is a two-dimensional scatter-gram showing how the correct and incorrect mappings are distributed with respect to the log (RFDM) value and the duration ratio. As Fig. 3 suggests, mappings with low RFDM and duration ratio close to 1 justify the fact that they are important in selecting one-to-one mappings. Thus, DTW will return $K\%$ reliable mappings and $A$ will play the role of detecting the outlier points by imposing the constraint of the property (i) of the $\alpha$, $\epsilon$-linear functions (Definition 2). Hence, the thresholds $K$ and $A$ are important in filtering incorrect mappings while estimating the $A$, $\epsilon$-linear function parameters.

The duration ratio bound block is shown in the dashed rectangle (c) of Fig. 1. This block filters out mappings with duration ratio higher than $A$. The input to this block are the ranked RFDM mappings and the output is the subset of there mappings with duration ratio \( \frac{1}{A} < \frac{|y_{2i} - y_{1i}|}{x_{2i} - x_{1i}} < A \).
Line parameters estimation. As a consequence of the previous step, for a fixed $A = \alpha$, the $N$ pairs that satisfy $\frac{1}{A} < \frac{y_2 - y_1}{x_2 - x_1} < A$ are used to estimate the optimal slope, $m^*$, and intercept, $b^*$, of the $A, \varepsilon$-linear function (of order $N$ will be omitted but implied from this point onwards) using the results of Appendix B. Moreover, the absolute error is computed using the $N$ pairs and the function $f(x) = m^*x + b^*$.

The line parameters estimation block takes as input the mappings with duration ratio less than $A$ and outputs the optimal slope, $m^*$, intercept, $b^*$, the absolute error, $E$, of the $A, \varepsilon$-linear function and the filtered mappings. The line-parameters estimation block is shown in the dashed rectangle (d) of Fig. 1.

Absolute error threshold. Now, we need a measure to assess the level of linearity of the mapping. For this purpose, we define a fixed threshold, $E$. Due to the fact that $E$ is robust to $M$ and $N$ variations (as discussed in Section 2.2), $E$ is used as an upper bound to the check if the absolute error, $E$, is “low” enough. Hence, by assumption, we accept the $A, E$-linear modeling, if $E \leq E$. If this condition is not satisfied, the alignment cannot be modeled with an $A, E$-linear function of order $N$. In this case, one might choose another set of $N$ pairs or use only the DTW approach if there is no approximately linear relation between the time-stamps.

The absolute error threshold block is shown in the dashed rectangle (e) of Fig. 1. The input of this block are the $A, E$-linear function parameters and filtered mappings and the output is a decision if the $A, E$-linear function can model the subtitles relation. Also, this block output the $A, E$-linear function parameters.

Time-stamps mapping. With the $A, E$-linear function and the optimal slope, $m^*$ and intercept $b^*$ in place, we relate all starting time-stamps by translating the $L_1$ subtitle document time-stamps into the $L_2$ subtitle document time-stamps. In particular, assume $x_1$ is a starting time-stamp in the $L_1$ document. Then, the assigned starting time-stamp in the $L_2$ document is the point $y_1$ that minimizes the distance $D_1 = |y_1 - f(x_1)|$. Similarly, we relate all ending time-stamps in the $L_1$ document with ending time-stamps in the $L_2$ document. Assume, $x_2$ is an ending time-stamp in the $L_1$ document; then the assigned ending time-stamp in the $L_2$ document is the point $y_2$ that minimizes the distance $D_2 = |y_2 - f(x_2)|$. Also, we seek additional subtitle pairs by mapping $y_1$ with the starting time-stamp of $x_1$ that minimizes $D_3 = |x_1 - f^{-1}(y_1)|$ and by mapping $y_2$ with the ending time-stamp of $x_2$ that minimizes $D_4 = |x_2 - f^{-1}(y_2)|$. Note at this point that the pairs might not be one-to-one because the closest distance might suggest to merge two subtitle pairs. Next, we filter out mappings which do not satisfy $(D_1 < T$ and $D_2 < T$) or $(D_3 < T$ and $D_4 < T)$. $T$ is chosen empirically by maximizing the performance on a development set. The last step is important in checking for possible subtitle pairs that might not be modeled by the estimated relation.

The time-stamp mapping block is shown in the dashed rectangle (f) of Fig. 1. This block takes as input the $A, E$-linear slope, intercept and the subtitles documents, maps the subtitles based on the closest translated time-stamps, filters the mappings with distance greater than $T$ and, finally, the outputs a subset of the mappings by filtering non-matching subtitles based on the approach described above.

Mappings merging. Finally, we need a method to merge many-to-one, one-to-many, and many-to-many mappings because, in practice, there may not be a clear pair boundary between bilingual subtitles in $L_1$ and $L_2$ subtitle documents.

Fig. 3. This is the scatter-gram of the correct and incorrect mappings with respect to the log (RFDM) value and the duration ratio.
The goal is to identify many-to-one, one-to-many, and many-to-many mappings and merge them. Fig. 4 shows the fundamental rules used to merge two-to-one, and one-to-two mappings. For example, if subtitles a and b in the \( L_1 \) subtitle document are mapped to subtitle d in the \( L_2 \) subtitle document, we merge a and b subtitles and map them to d subtitle. This merging defines a two-to-one mapping. Similarly, the other rules define one-to-one and one-to-two mappings. To merge the subtitles in \( L_1 \) and \( L_2 \) subtitle documents, we apply recursively the rules shown in Fig. 4 for all subtitles in \( L_1 \) and \( L_2 \) documents until no subtitles can be merged. Fig. 5 shows an example of a three-to-three mapping merging. The above-mentioned basic rules are applied recursively until only the one-to-one rule can be applied. In this example, first we merge f and g subtitles in the \( L_1 \) subtitle document using the rule for merging two-to-one mappings. We continue in this fashion until f, g and h subtitles in the \( L_1 \) subtitle document are mapped to i, j and k subtitles in the \( L_2 \) subtitle document as shown in Fig. 5.

While Fig. 4 shows a closer look into how the mappings merging rules are applied, the integration of the mapping merging block into the algorithm is shown in the dashed rectangle (g) of Fig. 1. As shown in the diagram the input are the filtered aligned mappings and the output are the aligned merged mappings.

3. Experimental results

In this section, we describe the data collection and the experimental results. The experiments are divided into two sections: the pilot and the full-scale experiments. The pilot study using a small set of tagged bilingual mappings was used to understand the parameter trade-offs related to performance. Moreover, the pilot experiments section serves as a development set to optimize the parameters of the time-alignment approach. Finally, the full-scale experiments use the optimal parameters obtained by the pilot study and expand the experiments by aligning a large set of untagged bilingual subtitle document pairs. The aligned data are used to train a SMT system. Finally, the SMT performance is tested on the extracted bilingual sets and the BLEU score performance is reported.

3.1. Pilot experiments

3.1.1. Experimental setup

For the pilot experiments, we used the 42 Greek–English subtitle document pairs described in Tsiartas et al. (2009). In each subtitle document pair, a set of 40 consecutive English subtitles were paired with the corresponding Greek subtitles and we ended up with 1680 tagged pairs. The English subtitle documents have 1443 subtitles on average per movie with standard deviation 369. On the other hand, the Greek subtitle documents contain 1262 subtitles on average.
with standard deviation 334. The difference in the average number of subtitles indicates that subtitles in bilingual subtitle document pairs may not always have one-to-one correspondence. A typical example of an aligned bilingual subtitle is shown in Fig. 6, obtained from the movie “I am Legend”.

All subtitle documents are preprocessed and filtered from non-alphanumeric symbols similar to what one would do for cleaning text for statistical machine translation purposes. Then, the time-stamps and subtitle numbers are removed resulting in a list of Greek subtitles and a list of English subtitles per subtitle document. Each subtitle time-stamp is saved separately as well. For all Greek words available, a system was built to mine all the translations returned by the Google dictionary.1 Using the dictionary, each Greek subtitle is converted from Greek into a bag of words in English. Then, the RFDM is computed for all subtitle pairs. The best mappings are extracted using the DTW approach described in Appendix A. The parameters used in the DTW approach are the same parameters used by Tsiartas et al. (2009) since the data-sets are identical. Lastly, the method used to merge one-to-one, many-to-one, and one-to-many subtitle pairs is applied to also merge the subtitles of the DTW approach as described in Section 2.2.1.

The mappings obtained by the DTW approach are used to estimate the $A, E$-linear function and, in turn, use the function to align the subtitles. Initially, the pairs are ranked in ascending order of RFDM values. For various experimental values of $K\% = [0.01 \ 0.02 \ 0.03 \ 0.05 \ 0.07 \ 0.1 \ 0.15 \ 0.2 \ 0.4 \ 0.6]$, the $K\%$ lowest RFDM one-to-one mappings are extracted for each bilingual subtitle pair. For the $i$th bilingual subtitle document pair, keeping only one-to-one mappings results in $M_i$ mappings. Next, by varying $A = [1.05 \ 1.1 \ 1.15 \ 1.22 \ 1.3 \ 1.42 \ 1.5 \ 1.7 \ 1.9 \ 2.2]$, a subset of the one-to-one mappings of order $N_i$ is used to estimate the $A, E$-linear function of order $N_i$ for each bilingual subtitle document pair. Then, for different values of $E = [0.1 \ 0.2 \ 0.3 \ 0.4 \ 0.5 \ 0.6]$, the $A, E$-linear relation is accepted or rejected if $E_i \leq E$. The starting and ending time-stamps are mapped using the closest distance rule as described in Section 2.2.1. Finally, for different values of $T = [0.2 \ 0.5 \ 0.8 \ 1.5 \ 1.8 \ 2.5 \ 3 \ 5]$, outliers are filtered. The final mappings are obtained using the method to merge one-to-one, many-to-one, and one-to-many subtitle pairs as described in Section 2.2.1. For each combination of the parameters $K, A, E$, and $T$, we compute the balanced $F$-score (Manning et al., 2009, p. 156) averaged over all bilingual subtitle document pairs and the number of considered movies.

3.1.2. Results and discussion of pilot study

In this section, we aim to understand the trade-offs among the time-alignment approach parameters.

Fig. 7(a) shows the averaged $F$-score (vertical axis) and the corresponding number of movies (horizontal axis) for different $K, A, E$ and $T$ parameter values. Fig. 7(a) indicates that we can get an $F$-score close to 1 for some $K, A, E$, and

---

Fig. 7. (a) The averaged F-score of the time-alignment approach vs the number of movies for various $K$, $A$, $E$ and $T$ parameter values. (b) The averaged F-score using the DTW approach for the different number of movies considered when varying the $K$, $A$, $E$, and $T$ parameter values.

On the other hand, Fig. 7(b) indicates that the F-score of the DTW-based approach is much lower than that of the time-alignment case, considering the same number of movies. For example, when we consider the parameters aligning bilingual subtitle documents of 30 movies, the DTW-based approach F-score is less than 0.75 as opposed to the time alignment approach in which the F-score is close to 0.95. Furthermore, Fig. 7(a) suggests that there is a trade-off between the quality of the alignments (i.e. F-score) and the number of movies used. Thus, one should consider the amount of data and the quality of the bilingual subtitle pairs needed. Based on the quality and amount of data needed, the appropriate $K$, $A$, $E$, and $T$ values can be assigned. To understand the importance of the $\alpha$, $\epsilon$-linear functions and the associated parameters $K$, $A$, $E$, and $T$ in relating the time-stamps, we also computed the F-score using a linear relation estimated by the results in Appendix B using all the DTW output mappings. For this case, the resulting F-score was 0.56 which is even below the DTW-based approach.

Fig. 8 is a five-dimensional diagram representing the F-score as intensity against the values of $K$, $A$, $E$, and $T$ parameters. Similarly, intensity in Fig. 9 represents the number of movies aligned for each set of threshold values and, thus, is an indicator of the amount of parallel data extracted. An important parameter is the absolute error threshold, $E$, used to accept or reject the $A$, $E$-linear function alignment for the corresponding movie. Decreasing the absolute error threshold, $E$, the F-score increases but at the same time, as Fig. 9 suggests, the number of movies aligned decreases. In addition, the choice of the duration ratio threshold, $A$, becomes less important in filtering the incorrect DTW mappings when a low error threshold is used. This happens because the subtitle pairs, kept with low error, have approximately linearly related time-stamps obtained by the correct DTW mappings. In spite of giving high F-scores, the number of movies aligned is much less as $E$ decreases. On the other hand, as the threshold $E$ increases and as the threshold on duration ratio, $A$, approaches 1, the performance decreases but the number of movies modeled increases. The trade-off between $A$ and $E$ is important to consider in aligning subtitle documents. In practice, it is preferable to allow an absolute error threshold, $E$, greater than 0.4 and a duration ratio threshold, $A$, less than 1.6 since they maintain not only high F-scores but also more bilingual data compared to the case with low $E$ and high $A$. Intuitively, one can think that it is preferable to select accurate mappings at an earlier stage so that we can better estimate the $A$, $E$-linear function parameters. Allowing inaccurate mappings results in a higher absolute error, $E$ and, thus, subtitle document pairs are dropped by the $E$ threshold. Hence, the amount of bilingual data is reduced. If the quality of the alignment is more important than the size of the corpus, then a low $E$ and $A$ should be considered.
Fig. 8. This figure shows the $F$-score of the time alignment approach for various values of $K$, $A$, $E$, and $T$ parameters.

Fig. 9. The intensity in this figure shows the number of movies modeled by the time alignment approach for various values of $K$, $A$, $E$, and $T$ parameters.
Moreover, Fig. 8 suggests that increasing $K$, increases the $F$-score as well. However, the $F$-score increase rate is almost flat when $K > 0.1$. On the other hand, increasing $K$ above 0.2 reduces the number of movies aligned using $A$, $E$-linear functions and, in turn, decreases the bilingual data. The rationale behind this fact is that $K$ increases the number of DTW mappings used. Since we are choosing the mappings based on the RFDM score in increasing order, the more DTW mappings considered, the higher the RFDM score of the mappings considered in which we are less confident about their accuracy according to the RFDM score. Since the threshold $\frac{1}{A} < \frac{|y_2 - y_1|}{x_2 - x_1} < A$ might not always filter the misaligned mappings as Fig. 2 suggests, it will be preferable to choose the most reliable mappings with the lower RFDM score. Including possibly misaligned mappings, i.e. high RFDM score mappings, increases the error and, thus, reduces the number of subtitles accepted by the $E$ threshold. However, if $K\%$ is high and the $E$ is low, it suggests that the $K\%$ of the DTW mappings can be related with an almost linear relation and, thus, for those subtitle pairs the estimation of the $A$, $E$-linear function parameters is accurate resulting in a higher $F$-score. Hence, another trade-off to consider that affects the quality and the size of the extracted bilingual corpus is between the thresholds $K$, $E$ and $A$.

Finally, the value of absolute error of the starting and ending times differences threshold, $T$, takes place after accepting or rejecting the alignment of each bilingual subtitle movie. Fig. 9 shows the number of movies aligned is the same across all values of $T$ for a specific value of $K$, $E$, and $A$. Hence, $T$ does not affect the number of movies considered. However, Fig. 8 suggests that choosing a very low value of $T$ reduces the $F$-score. In this case, the $F$-score is reduced because recall is reduced and precision remains close to 1 as $T$ decreases below 1. On the other hand, as $T$ increases above 3, the precision decreases and the recall remains close to 1 resulting into a lower $F$-score. Fig. 8 suggests that $1 \leq T \leq 3$ maximizes the $F$-score.

The absolute error, $E$, plays an important role in deciding if the $A$, $E$-linear function can model the time-stamps’ relation. Thus, it is interesting to study the relationship between the absolute error, $E$, and the quality of the mappings. For this reason, we set $K=0.6$ and $A=1.5$ which are the optimal parameters for maximizing the $F$-score when 24 subtitle document pairs are selected. Using these parameters, we compute the absolute error, $E$, of the $A$, $E$-linear function. Fig. 10 suggests that there is a trade-off between the quality of the alignments and the absolute error, $E$. In practice, a low absolute error results in a higher $F$-score, precision, and recall. In particular, for absolute error, $E$, less than 0.2, we get almost perfect mappings with $F$-score close to 1 due to aligning movies with almost linearly related time-stamps. Fig. 10 also justifies the fact that reducing the error threshold, $E$, increases the $F$-score but, on the other hand, decreases the number of movies aligned because fewer subtitle document pairs will satisfy the $E$ threshold.

After analyzing the trade-offs between the various parameters, we choose two sets of parameters for the full-scale experiments. The first set of parameters is fixed to $K=0.6$, $A=1.5$, $E=0.6$ and $T=2$. This set is denoted by TA-1. For TA-1 pilot experiments, the $F$-score is 0.95, precision is 0.92, and recall is 0.98. The number of movies modeled by
TA-1 parameters is 24 movies. The corresponding DTW approach $F$-score for the 24 movies considered is 0.75. The second set of parameters produces alignments of less quality than TA-1 but more data. In particular, the second set of parameters is fixed to $K=0.15$, $A=1.1$, $E=0.5$, and $T=1.5$. This set is denoted by TA-2. For TA-2 pilot experiments, the $F$-score is 0.93, precision is 0.92, and recall 0.94. The number of movies aligned is 30. The corresponding DTW approach $F$-score for the 30 movies considered is 0.72.

3.2. Full-scale experiments

3.2.1. Experimental setup

For the full-scale experiments, we downloaded Spanish–English and French–English subtitle document pairs (http://www.opensubtitles.org/). For the Spanish–English subtitle document pairs, we collected 1758 Spanish subtitle documents and 1936 English subtitle documents. Note that these come from 699 unique movies. By combining all possible document pairs of movies, we end up with 4921 Spanish–English subtitle pairs including repeated subtitle documents for some movies.

On the other hand, for the French–English subtitle document pairs, we collected 1745 French and 2145 English movie subtitle documents out of 641 unique movies. By combining all possible document pairs, we end up with 5967 French–English subtitle document pairs including repeated subtitle documents for some movies.

For the above-mentioned subtitle documents, the non-alphanumeric symbols were filtered for all the subtitle documents. In addition, for all Spanish and French words available in the Spanish and French subtitle documents, we queried the Google dictionary and saved all the available English translations. Then, the bilingual subtitle documents pairs are aligned using the DTW procedure as described in Section 2.1 and the DTW mappings are obtained. Using the DTW mappings, the subtitle document pairs are aligned using the time-alignment algorithm described in Section 2.2. The time-alignment approach was run twice using the TA-1 and TA-2 parameters.

Since there are multiple subtitle document versions for each movie available, we can use the quality measures of the proposed approach to find the subtitle documents pair for each movie that maximizes the performance. Thus, among the multiple subtitle document pairs per movie, we select the subtitle document pair giving the lowest absolute error, $E$. Because the DTW baseline has no quality tests to accept or reject alignments, we randomly pick a subtitle document pair for each movie to align. Fig. 11 implies that there are approximately 95% of the movies having at least one bilingual subtitle document pair with absolute error $E < 1$. Hence, for the proposed approach, we align the subtitle pairs for each movie with the lowest error. The parameters used in Fig. 11 are $K=0.15$ and $A=1.1$ which are the parameters of TA-2. Using the $K$ and $A$ parameters of TA-1 yields similar results.

Finally, using parallel data from a corpus from aligned movie subtitles, we train the SMT models on each language pair separately. Experiments using the SMT trained on the TA-1, TA-2, and DTW corpora are denoted by TA-1, TA-2
Fig. 12. This figure compares the performance of the SMT models trained on the corpus created using the DTW-based approach and the models trained on the corpora extracted by the time-alignment approach with parameters TA-1 and TA-2 when the TRANSTAC development and test sets are considered. The experiments were repeated for various bilingual corpora sizes. The comparison is extended for the language pairs between English and Spanish, English–French, and vice versa.

and DTW respectively. Moreover, 2000 randomly picked utterances for tuning and 2000 randomly picked utterances for testing were used to evaluate the performance from the DARPA TRANSTAC English–Farsi data set. Only the English utterances were extracted and manually translated to Spanish and French for evaluating the performance. TRANSTAC is a protection domain corpus (e.g. dialogs encountered at military points). The randomly picked subset includes conversations of a spontaneous nature; for example, there are spontaneous discussions on various topics such as medical assistance related conversations, etc. Tuning and evaluation on this set is denoted by TRANSTAC. In addition, the development and test sets of the News Commentary corpus\(^2\) have been used to evaluate the experiments. We refer to the NEWS development and test set as NEWS-TEST.

The SMT requires language models of the target language to translate the source utterances. In each experiment, the training set of the target language is used to train the language models for each experiment as well. The trigram language models were built using the SRILM toolkit (Stolcke, 2002) and smoothed using the Kneser-Ney discount method (Kneser and Ney, 1995). We compared the performance of various combinations and sizes of the training sets using BLEU score (Papineni et al., 2002) on the TRANSTAC and NEWS test sets.

3.2.2. Results and discussion

Figs. 12 and 13 compare the performance of the SMT models obtained by training on the corpora extracted by the time-alignment approach and that extracted by the DTW-based approach in the TRANSTAC and NEWS-TEST domains. In addition, the comparison is extended into four language pairs, namely, English to Spanish, English to French, and vice versa.

\(^2\) Made available for the WMT10 workshop shared task http://www.statmt.org/wmt10/.
In Fig. 12, the goal is to compare the quality of the alignments in a spontaneous speaking style domain and, hence, the TRANSTAC domain is used for tuning and evaluating. The figure shows the performance gains of the models trained on the TA-1 and TA-2 corpora over the models trained on DTW-based approach corpus. In particular, the performance of TA-1 and TA-2 corpora is very close in terms of BLEU score; however, the parameters used in TA-2 could extract a larger bilingual corpus as shown in Fig. 12. In these experiments, the time alignment approach corpora consistently outperforms the DTW-based approach corpus across different language pairs and different bilingual corpus sizes by up to 2.53 BLEU score points for the English–Spanish experiments and by up to 4.88 BLEU score points for the English–French experiments. The improvement stems from the fact that the TA-1 and TA-2 corpora approaches have been shown in Section 3.1.2 to deliver $F$-scores close to 96% as opposed to the DTW-based approach corpus which is expected to deliver $F$-scores of 71% (Tsiartas et al., 2009). Thus, for a fixed amount of subtitle pairs, the $F$-score improvement of the alignment is translated into SMT performance boost showing the importance of the time-alignment based approach.

In Fig. 13, the goal is to compare the quality of the alignments in the broadcast news domain by using the NEWS test set. Similar to the TRANSTAC test set results, Fig. 13 indicates that SMT models trained on TA-1 and TA-2 outperform those trained on the corpus created using the DTW-based approach. We observe performance improvements of up to 1.2 BLEU score points for the English–Spanish experiments and by up to 2.65 BLEU score points for the English–French experiments. The performance improvement is consistent along all of the different bilingual corpus sizes. These experiments suggest that the time-alignment approach is superior to the DTW-based approach across different domains in terms of the SMT performance. We note that the $F$-score improvement delivered by the time-alignment approach is reflected even in domains not matching the subtitles speaking style such as in the NEWS-TEST domain.
For readers interested in the performance of subtitles for the TRANSTAC domain compared to additional corpora could refer to Appendix C.

4. Conclusions

In this work, we presented a novel method to align subtitle documents using timing information. We modeled the time-stamps relation using the so-called $\alpha$, $\epsilon$-linear functions. In Section 2.2, we presented some properties of the $\alpha$, $\epsilon$-linear functions. Then, we took advantage of the properties of the $\alpha$, $\epsilon$-linear functions to find an approximately linear relation between the time-stamps of two subtitles. In addition, we analyzed practical aspects of the $\alpha$, $\epsilon$-linear functions in modeling the subtitles alignment problem and showed the advantages over the standard DTW-based scheme.

Moreover, we have shown the importance of various parameters and the trade-offs associated with the quantity and quality of the data. The threshold $K$ plays the role of selecting the most reliable mappings from the DTW output with respect to the RFDM score. Additionally, we have shown the importance of the threshold $\Lambda$ in filtering unreliable mappings with low RFDM score. On the other hand, the threshold $E$ has been used to exclude subtitles with error $E \geq E$ in order to filter out subtitle pairs that cannot be modeled by the $\alpha$, $\epsilon$-linear function. In this work, the error was estimated using the absolute norm. For the sake of completeness, we have experimented with various norms, i.e. Euclidean norm, to compute the error, but no significant improvement was observed in terms of the $F$-score. Furthermore, a threshold on the start and end time differences has been imposed to filter out the outliers. Lastly, a method has been presented to join many-to-many subtitle pairs.

In the experiments section, we analyzed and compared the performance of the time-alignment approach and the DTW-based approach in a small set and showed improvement in terms of $F$-score. Moreover, we showed that the $F$-score improves by 39% (absolute) using the $\alpha$, $\epsilon$-linear function approach compared to a line fit obtained by linear regression on all DTW output pairs, indicating the importance of using the $\alpha$, $\epsilon$-linear function modeling. In addition, we have shown that our approach outperforms the DTW-based approach by up to 4.88 BLEU score points (approx. 25% relative improvement) on spontaneous speech Statistical Machine Translation (SMT) experiments. Overall the proposed approach can provide high quality bilingual translations by linearly relating the subtitle time-stamps and choosing parameters to set the quality and the quantity of the data extracted.

Appendix A. Dynamic time warping algorithm

The optimization problem described in (1) can be solved efficiently using DTW. The goal of the DTW algorithm is to find the best mappings $\{m_{ij}^*\}$ that minimize the global distance: $\sum_{i,j} m_{ij} DM(S_{i,1}^L, S_{j,2}^L)$. We use the notion of the

| Table A.1 |
| Steps in subtitle alignment using DTW approach. |

(i) Initialization:

$L(S_{i,1}^L, S_{j,2}^L) = DM(S_{i,1}^L, S_{j,2}^L)$
$L(S_{i,1}^L, S_{j,2}^L) = L(S_{i,1}^L, S_{j,2}^L) + DM(S_{i,1}^L, S_{j,2}^L)$
$\xi(S_{i,1}^L, S_{j,2}^L) = (S_{i,1}^L, S_{j,2}^L), j = 1, \ldots, N_2$
$L(S_{i,1}^L, S_{j,2}^L) = L(S_{i,1}^L, S_{j,2}^L) + DM(S_{i,1}^L, S_{j,2}^L)$
$\xi(S_{i,1}^L, S_{j,2}^L) = (S_{i,1}^L, S_{j,2}^L), i = 1, \ldots, N_1$

(ii) Iteration:

$L(S_{i,1}^L, S_{j,2}^L) = DM(S_{i,1}^L, S_{j,2}^L) + \min(L(S_{i,1}^L, S_{j,2}^L), L(S_{i,1}^L, S_{j,2}^L), L(S_{i,1}^L, S_{j,2}^L))$
$\xi(S_{i,1}^L, S_{j,2}^L) = \arg \min(L(S_{i,1}^L, S_{j,2}^L), L(S_{i,1}^L, S_{j,2}^L), L(S_{i,1}^L, S_{j,2}^L)), i = 2, \ldots, N_1, j = 2, \ldots, N_2$

(iii) Backtracking: Let there be $K$ mappings. Then

$m_{N_1,N_2} = m(S_{N_1}^L, S_{N_2}^L) = 1, m_{1,1} = m(S_{1}^L, S_{1}^L) = 1$
$\xi_K = \xi(S_{N_1}^L, S_{N_2}^L)$

\[ m(\xi_k) = 1 \]
\[ \xi_{k-1} = \xi(\xi_k), k = K, \ldots, 2. \]
cumulative distance for the best mappings starting from \((S_{11}^L, S_{12}^L)\) and ending at \((S_{j1}^L, S_{j2}^L)\), denoted by \(\mathcal{L}(S_{11}^L, S_{12}^L)\) (Tsiartas et al., 2009). We define \(\zeta(S_{11}^L, S_{12}^L)\) as the mapping prior to last in the best path terminating at \((S_{11}^L, S_{12}^L)\). Formally, the DTW algorithm is defined in Table A.1.

Appendix B. Optimal \(\alpha, \epsilon\)-linear function parameters

The optimal minimum mean square sense \(\alpha, \epsilon\)-linear function, \(f^*(x) = m^* + b^*\), of a set of points \(P = \{(x_{1i}, y_{1i}), (x_{2i}, y_{2i}) : 1 \leq i \leq M\}\) of order \(|P| = N, I \subseteq \{i : 1 \leq i \leq M\}\) with \(\frac{1}{\alpha} < \left| \frac{y_{2i} - y_{1i}}{x_{2i} - x_{1i}} \right| < \alpha \ \forall i \in I\) and \(\alpha > 1\) is given by the parameters:

\[
(i) \quad m^* = \frac{N(Y_1 + Y_2) - Z_1 - Z_2}{N \sum_{i \in I} (x_{1i} + x_{2i})^2 - \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2}
\]

where,

\[
Y_1 = \sum_{i \in I} y_{1i}(x_{1i} + x_{2i})
\]
\[
Y_2 = \sum_{i \in I} y_{2i}(x_{1i} + x_{2i})
\]

and

\[
Z_1 = \sum_{i \in I} y_{1i} \sum_{i \in I} (x_{1i} + x_{2i})
\]
\[
Z_2 = \sum_{i \in I} y_{2i} \sum_{i \in I} (x_{1i} + x_{2i})
\]

\[
(ii) \quad b^* = \frac{\left( \sum_{i \in I} y_{1i} + \sum_{i \in I} y_{1i} \right) \sum_{i \in I} (x_{1i} + x_{2i})^2 - (Y_1 + Y_2) \left( \sum_{i \in I} x_{1i} + \sum_{i \in I} x_{2i} \right)}{2 \left( N \sum_{i \in I} (x_{1i} + x_{2i})^2 - \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2 \right)}
\]

where \(Y_1\) and \(Y_2\) are defined in (i)

B.1. Proof of the optimal parameters using least squares

The goal is to find \(m^*\) and \(b^*\) such that:

\[
MSE = \sum_{i \in I} \left( \frac{y_{1i}}{2} + \frac{y_{2i}}{2} - f^* \left( \frac{x_{1i}}{2} + \frac{x_{2i}}{2} \right) \right)^2 \quad \because \text{MSE definition}
\]

\[
= \frac{1}{4} \sum_{i \in I} \left( y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right)^2 \quad \because \text{Linearity of } f
\]
(i) Taking the derivatives with respect to \( m^* \) and \( b^* \) and setting them to zero gives the equations needed to find \( m^* \) and \( b^* \) that minimize \( \text{MSE} \). Thus,

\[
\frac{d(\text{MSE})}{dm^*} = \frac{d}{dm^*} \left( \frac{1}{4} \sum_{i \in I} \left( y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right)^2 \right) = -\frac{1}{2} \sum_{i \in I} (x_{1i} + x_{2i}) \left( y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right) = -\frac{1}{2} \left( \sum_{i \in I} m^*(x_{1i} + x_{2i})^2 + y_2 - \sum_{i \in I} 2b^*(x_{1i} + x_{2i}) \right) = 0
\]

where \( Y_1 \) and \( Y_2 \) are defined in theorem Appendix B.

By rearranging terms, we get the following relation,

\[
Y_1 + Y_2 = \sum_{i \in I} m^*(x_{1i} + x_{2i})^2 + \sum_{i \in I} 2b^*(x_{1i} + x_{2i}) \quad (B.1)
\]

Next multiply both sides of (B.1) by \( N \) and we get the following equation of two unknowns, \( m^* \) and \( b^* \),

\[
NY_1 + NY_2 = N \sum_{i \in I} m^*(x_{1i} + x_{2i})^2 + N \sum_{i \in I} 2b^*(x_{1i} + x_{2i}) \quad (B.2)
\]

The next equation with unknowns \( m \) and \( b \) is obtained by differentiating \( \text{MSE} \) w.r.t. \( b \)

\[
\frac{d(\text{MSE})}{db^*} = \frac{d}{db^*} \left( \frac{1}{4} \sum_{i \in I} \left( y_{1i} - m^* x_{1i} - b^* + y_{2i} - m^* x_{2i} - b^* \right)^2 \right) = -\sum_{i \in I} (y_{1i} - m^* x_{1i} + y_{2i} - m^* x_{2i} - 2b^*) = -\sum_{i \in I} y_{1i} + m^* \sum_{i \in I} x_{1i} - \sum_{i \in I} y_{2i} + m^* \sum_{i \in I} x_{2i} + 2b^* \sum_{i \in I} 1 = -\sum_{i \in I} y_{1i} + m^* \sum_{i \in I} x_{1i} - \sum_{i \in I} y_{2i} + m^* \sum_{i \in I} x_{2i} + 2Nb^* = 0
\]

By rearranging terms, we get the following relation

\[
\sum_{i \in I} y_{1i} + \sum_{i \in I} y_{2i} = m^* \sum_{i \in I} x_{1i} + m^* \sum_{i \in I} x_{2i} + 2Nb^* \quad (B.3)
\]

Next multiply both sides of (B.3) by \( \sum_{i \in I}(x_{1i} + x_{2i}) \) and we end up with another equation with two unknowns, \( m^* \) and \( b^* \),

\[
Z_1 + Z_2 = m^* \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2 + 2Nb^* \sum_{i \in I} (x_{1i} + x_{2i}) \quad (B.4)
\]

where \( Z_1 \) and \( Z_2 \) are defined in Appendix B.
Now subtracting (B.4) from (B.2)

\[ N(Y_1 + Y_2) - Z_1 - Z_2 = m^\ast \left( N\sum_{i \in I} (x_{1i} + x_{2i})^2 - \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2 \right) \]

Solving w.r.t. \( m^\ast \)

\[ m^\ast = \frac{N(Y_1 + Y_2) - Z_1 - Z_2}{N\sum_{i \in I} (x_{1i} + x_{2i})^2 - \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2} \]

as required

(ii) We substitute \( m^\ast \) in (B.4), rearrange terms and we end up with

\[ b^\ast = \frac{\sum_{i \in I} y_{1i} + \sum_{i \in I} y_{2i} - m^\ast \sum_{i \in I} x_{1i} + \sum_{i \in I} x_{2i}}{2N} \]

\[ = \frac{\left( \sum_{i \in I} y_{1i} + \sum_{i \in I} y_{2i} \right) \sum_{i \in I} (x_{1i} + x_{2i})^2 - (Y_1 + Y_2) \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2}{2 \left( N\sum_{i \in I} (x_{1i} + x_{2i})^2 - \left( \sum_{i \in I} (x_{1i} + x_{2i}) \right)^2 \right)} \]

Fig C.1. This figure compares the performance of the SMT models trained on time-alignment (TA-2), NEWS and EUROP corpora when the TRANSTAC development and test sets are considered. The experiments were repeated for various bilingual corpora sizes. The comparison is extended for the language pairs between English and Spanish, English–French, and vice versa.
Appendix C. Additional experiments

Since the TRANSTAC bilingual corpus is not available in English–Spanish and English–French languages, we present here the comparison with Europarl version-5 (EUROP) (Koehn, 2005) and the NEWS corpora for readers interested in the performance of subtitles for this domain. As shown in Fig. C.1, SMT models trained on the subtitle corpus (TA-2) improve the performance by up to 14.34 BLEU score points on the TRANSTAC test set. Thus, on this data-set, it is beneficial to use subtitles for training the SMT models.

References