

Constructing Parallel Corpus from Movie Subtitles

Han Xiao¹ and Xiaojie Wang²

¹ School of Information Engineering, Beijing University of Post and Telecommunications
artex.xh@gmail.com

² CISTR, Beijing University of Post and Telecommunications
xjwang@bupt.edu.cn

Abstract. This paper describes a methodology for constructing aligned German-Chinese corpora from movie subtitles. The corpora will be used to train a special machine translation system with intention to automatically translate the subtitles between German and Chinese. Since the common length-based algorithm for alignment shows weakness on short spoken sentences, especially on those from different language families, this paper studies to use dynamic programming based on time-shift information in subtitles, and extends it with statistical lexical cues to align the subtitle. In our experiment with around 4,000 Chinese and German sentences, the proposed alignment approach yields 83.8% precision. Furthermore, it is unrelated to languages, and leads to a general method of parallel corpora building between different language families.

Keywords: sentence alignment, parallel corpora, lexical cues.

1 Introduction

Text alignment is an important task in Natural Language Processing (NLP). It can be used to support many other NLP tasks. Lots of researches have been done on bilingual alignment [1,2,3,13], and some specific kinds of corpora have gained more and more focus. One of the typical examples is the movie subtitles, it is free available and has rich semantic. [4] showed that a Machine Translation(MT) system gave a slightly better result when training on subtitles compared to Europarl (Europarl Parallel Corpus). [4] also argued that the text genre of "file subtitles" is well suited for MT, in particular for statistical MT.

A subtitle file in a movie is a textual data corresponding to: a set of dialogues, a description of an event or sounds, noises and music (that is often called as hearing impaired subtitle). The popular subtitle formats are based on the time [6]. They are characterized by an identifier, a time frame and finally a sequence of words. Usually each text piece consists of one or two short sentences shown on screen with an average seconds [7]. [8] mentioned that the readers for subtitles have only a limited time to perceive and understand a given subtitle so the internal understanding complexity is small. The linguistic subtitle structure is closer to oral language with great variability. As a result the language of subtitling covers a broad variety of any conceivable topic, even with exaggerated modern youth language. Still, [5] pointed out widespread unknown words in subtitles. They comprise proper names of people and products,

rare-word forms and foreign words. One must notice the fact that different language versions of subtitles for a same movie are not necessarily written by the same person. Still, the amount of compression and re-phrasing is different for various languages, and it also depends on cultural differences and subtitle traditions. The Fig. 1 shows a short example of German subtitles and their Chinese correspondences in SubRip format.

92 00:07:44,330 --> 00:07:47,390 我的小学老师说过我有两个脑袋	94 00:07:26,327 --> 00:07:29,922 Meine Grundschullehrerin sagte, ich sei mit einem zweifachen Hirn,
93 00:07:47,460 --> 00:07:49,430 却只有半颗心	95 00:07:30,007 --> 00:07:32,441 aber mit nur einem halben Herzen geboren.
94 00:07:49,500 --> 00:07:50,990 真的？ 是呀	96 00:07:33,127 --> 00:07:35,357 Wow! Die klingt ja reizend!
95 00:07:51,070 --> 00:07:53,300 哇，她似乎挺可爱的	97 00:07:36,327 --> 00:07:38,158 Die Wahrheit ist, dass ich...
96 00:07:54,440 --> 00:07:57,890 其实我并不喜欢人们	98 00:07:38,247 --> 00:07:40,317 Ich mag Menschen nicht besonders.

Fig. 1. A piece of Chinese and German subtitles extracted from the movie “Beautiful Mind”. The Chinese subtitle 96 is divided into two subtitles 97 and 98 in its corresponding German translation, and Chinese subtitle 94 does not occur in German version.

Sentence alignment is an almost obligatory first step for making use of German-Chinese subtitles. It consists of finding a mapping between source and target language sentences allowing for deletions, insertions and some n:m alignments, but it restricts crossing dependencies. Most works on automatic alignment of film subtitles are still in its infancy. [6] handled the alignments on time with a variation by empirically fixing the global time-shifting at 500 milliseconds. [11] showed that the alignment approach based on time overlap combined with cognate recognition is clearly superior to pure length-based alignment. The results of [11] are of 82.5% correct alignments for Dutch-English and 78.1% correct alignments for Dutch-German. The approach of [11] is entirely based on time information, thus, it often requires the subtitles to be equally synchronized to the original movie. A method named Dynamic Time Warping (DTW) for aligning the French-English subtitles was obtained from Internet [10]. The approach of [10] requires a special bilingual dictionary to compute subtitle correspondences, and [10] reports 94% correct alignments when turning recall down to 66%. However, the corresponding lexical resources in [10] cannot be easily accessed.

In this paper, we propose a new language-independent approach considered both time frames and lexical content of subtitles, which automatically aligns the subtitle pairs. The purpose is to extract the 1:1 alignment from German and Chinese subtitles for our subtitle translation. Section 2 dedicates to the method presented by us for subtitle alignments, and our improvement by using lexical anchors. The evaluations of result with comparison to traditional length-based approach are discussed in Section 3.

2 Alignment Solution

Before alignment can be applied, the subtitle corpus needs to undergo a few preprocessing steps. Each German subtitle file has been tokenized and corresponding Chinese subtitle has been segmented, which are crucial for the success of selecting corresponded word. We intend to combine stemming, which has been demonstrated to improve statistical word alignment [9]. For German, a light stemmer (removing inflections only for noun and adjectives) presents some advantages. Despite its inflexional complexities, German has a quite simple suffix structure, so that, if one ignores the almost intractable problems of compound words, separable verb prefixes, and prefixed and infixed "ge", an algorithmic stemmer can be made quite short and effective. The umlaut in German is a regular feature of plural formation, so its removal is a natural feature of stemming, but this leads to certain false confluences (for example, schön, beautiful; schon, already). In future work, we would like to improve the stemmer especially for the irregular morphological variations used by verbs and nouns. We expect that the effect of combining these morphological operations will reduce the sparse data problem and speed up the computation of correspondence pairs.

2.1 Dynamic Time Warping

A considerable number of papers [1-3] have examined the aligning sentences in parallel texts between various languages. These works define a distance based on length or lexical content, which involves the use of dynamic programming. Since the time information is explicitly given in subtitle file, intuitively, corresponding segments in different translations should be shown at roughly the same time. However, this case does not occur very often. Every subtitle file is built independently from others even for the original video track. This results in growing time gaps between corresponding beans. The time span is never identical at the millisecond level.

In order to handle this problem, we apply dynamic programming to calculate the best path between two subtitle files. This algorithm uses the interval of the start time from two subtitles to evaluate how likely an alignment between them. Two subtitles are not considered as an aligned pair if their start times are far away from each other. To make it easily find the most probable subtitle alignment, the possible alignments in our approach are empirically limited to {1:1, 1:0, 0:1, 2:1, 1:2, 2:2}. Initially, the German and Chinese subtitles are asynchronous. The cost of all possible alignments, which are measured by time differences, has been considered from the beginning to the end of the subtitle file.

Let $T(i, j)$ be the lowest cost alignment between subtitle $1, \dots, i$ and $1, \dots, j$ where i is the index of Chinese and j is the index of German subtitle. Previously the $T(0, 0)$ is set to 0. Then one can define and recursively calculate $T(i, j)$ as follows:

$$T(i, j) = \min \begin{cases} T(i, j-1) + \lambda_{01} \text{cost}(i, j+1) \\ T(i-1, j) + \lambda_{10} \text{cost}(i+1, j) \\ T(i-1, j-1) + \lambda_{11} \text{cost}(i, j) \\ T(i-1, j-2) + \lambda_{12} \text{cost}(i, j-1) \\ T(i-2, j-1) + \lambda_{21} \text{cost}(i-1, j) \\ T(i-2, j-2) + \lambda_{22} \text{cost}(i-1, j-1) \end{cases}, \quad (1)$$

where λ is the inverse of priori probability in order to give a lower cost to more frequency match types.

This leaves determining the cost function, $\text{cost}(i, j)$ as follow:

$$\text{cost}(i, j) = [\delta_{\text{cur}}(i, j) - \delta_{\text{prev}}(i, j)]^2, \quad (2)$$

$$\delta_{\text{cur}}(i, j) = i_{\text{start}} - j_{\text{start}}, \quad (3)$$

where $\delta_{\text{cur}}(i, j)$ calculates the start time delay between i and j . One may notice that the previous delay may cause growing delay for the following subtitles. In order to solve this problem, the previous delay δ_{prev} must be subtracted from the current delay. According to the different align mode, δ_{prev} is selected as follows:

$$\delta_{\text{prev}}(i, j) = \begin{cases} \Delta(i, j-1), \text{ align } 0:1 \\ \Delta(i-1, j), \text{ align } 1:0 \\ \Delta(i-1, j-1), \text{ align } 1:1 \\ \Delta(i-1, j-2), \text{ align } 1:2 \\ \Delta(i-2, j-1), \text{ align } 2:1 \\ \Delta(i-2, j-2), \text{ align } 2:2 \end{cases} \quad (4)$$

So, in essence, Δ is a matrix that gives the delay caused by each possible alignment. For each step, once the align mode of i and j is determined by (1), $\Delta(i, j)$ must be set to $\delta_{\text{cur}}(i, j)$ of the selected mode. That is to say, the matrix Δ is built dynamically in the procedure.

2.2 Lexical Cues Extension

Since the previous algorithm is based on the purely time information, it ignores the richer information available in the text. To obtain an improvement in alignment accuracy, the lexical content must be considered. Intuitively, the lexical content could be used to find reliable anchor points. Previous work [2,9,11] focused on using bilingual dictionary to find anchor points in the subtitle pairs, or using string similarity measures [11] such as the longest common subsequences to decide the most relevant candidate pairs. Here, we apply the approach based on measures of association on roughly parallel texts, which has been processed by dynamic programming in Section 3, to derive the bilingual dictionary automatically. Then we find the anchor points by means of this dictionary. It's assumed that the aligned region is the terms from 1:1, 2:1, 1:2 and 2:1 matches. Though the validity of the co-occurrence clue is obvious for

parallel corpora, it also holds for comparable corpora and unrelated corpora [12]. The χ^2 test used by [13] is efficient way of identifying word correspondence, it has a simple form as follow:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}, \quad (5)$$

where N is the number of all roughly alignments except 1:0 and 0:1 mode. For a selected German word G and a Chinese word C, the two-by-two contingency tables are built for them. O_{11} counts the number aligned pairs which co-occurrence G and C, O_{12} is the count of pairs that have G in German subtitles, but lost C in the correspondence Chinese subtitle. O_{21} is the count of pairs in which Chinese subtitle have C, but the aligned German subtitle misses G. O_{22} counts the number of pairs that have neither G nor C.

For each Chinese word and word in current subtitles, we use (1) to calculate their χ^2 score. Using the confidence level of $\alpha = 0.05$ and the critical value $\chi^2 = 3.845$, we can decide whether C and G are good candidate for translation pair or not. Note that all entries belonging to the words are not found in the German stop list and Chinese stop list. Since the size of each subtitle file is limited, the word C may have several correspondent words G with the same score. For them, the pairs which score the highest are remained. Table 1 shows the result for 6 Chinese words and their correspondent German translations in 685 results totally.

Table 1. 6 Results for Chinese words and their corresponding German translations. Bold words are accepted translations.

Chinese word	Expected translation	Candidate German translations automatically generated			
啤酒	Bier	bier			
五角大楼	Pentagon	pentagon			
恭喜	gluckwunsch	gluckwunsch			
安排	arrangieren	arrangi	gebet	treffenmit	wiederholt
女孩子	Mädel	atomphysik	erzahlt	heisse	madel
全世界	Welt	erklart	global	kommunismus	sowjetsist

As the Table 1 illustrated, in many cases our program predicts the expected word with other typical associates mixed. Since a number of reliable word correspondences have been found, we can use them to predict anchor points by some simple heuristics. We count a score that indicates the number of words to match between the Chinese subtitle C and German subtitle G as follows

$$(C, G) = \frac{1}{n} \sum_{j=1}^n \delta(\text{tr}(C_j), G_j) \forall j, \quad (6)$$

$$\delta(x, y) = \begin{cases} 1, & y \in x \\ 0, & y \notin x. \end{cases} \quad (7)$$

where $tr(C_i)$ is the translation of word C_i in Chinese subtitle C based on the previous bilingual dictionary. Since our dictionary may provide several candidate German translations for a Chinese word, and $tr(C_i)$ can be a word or a collection. Therefore a Kronecker $\delta(x, y)$ is given to check the matches. Assuming that an alignment should mainly consist of match translations, we can use a threshold σ for this score to decide whether an alignment is likely to be correct or not, thus, to be an anchor point.

3 Evaluation

We examine the pure-length based approach [13] and the DTW based on time-delay with its lexical extension that we proposed in this paper. The evaluation has been conducted on a corpus extracted from randomly selected 10 movies. For each movie, we take out randomly around 400 Chinese and their German corresponding subtitles, which result in 4,128 aligned sentence pairs. Our selection is based on the principle that the sentence pairs are at initial of the movie and consecutive within each movie. All of t sentence pairs these pairs are manually aligned. When conducting some previous evaluations, most of them limited their test to few dozens for each movie, which obviously facilitates the task. We separated 1,000 sentence pairs from all this manually aligned pairs as a training set. We then used relative frequencies of each occurring alignment mode in training set to estimate the parameter λ . For efficiency reasons we round them into integers as shown in Table 2.

Table 2. Adjusted priors for various alignment types

Parameters	Value
$\lambda_{0:1}$	30
$\lambda_{1:0}$	56
$\lambda_{1:1}$	1
$\lambda_{1:2}$	23
$\lambda_{2:1}$	11
$\lambda_{2:2}$	79

Therefore it will cause the algorithm to give 1:1 match for a priority, which is most common. These parameters will be shared in the three approaches we evaluated. For the length-based approach, the number of German characters generated by each Chinese character is also calculated in our training set, the mean $c = 2.667$, with a standard deviation $\sigma = 1.040$.

The result of time-based dynamic programming and its lexical extension are listed in Table 3. The threshold σ as mentioned above was set to 0.05 previously. To be able to compare our results with other work, the evaluations are presented in terms of recall and precision. However, in some cases the count of exact matches is somewhat arbitrary. We count partially correct alignments, which have some overlap with

Table 3. Performance of different alignment approaches

Approach	Recall	Precision	F-measure	Add partially correct		
				Recall	Precision	F-measure
Len	30.8%	30.1%	30.4%	38.2%	37.3%	37.7%
Len+ Lexi	37.5%	50.6%	43.1%	45.8%	61.8%	52.6%
Time	73.4%	67.8%	70.5%	83.0%	76.7%	79.7%
Time+ Lexi	66.4%	72.4%	69.3%	76.8%	83.8%	80.1%

correct alignments in both Chinese and German beads. In order to make it easier to compare and uniform, the partial correct is defined as 50% correct, and we added it accordingly to the recall and precision.

The pure length-based approach showed their weakness being compared to other approaches on sentence alignment of subtitles. The possible reason could be the inaccurate Gaussian assumption of $l_2 - l_1$ in this specific domain, where l_2 and l_1 are the length of potential aligned sentences in German and Chinese. Since the linguistic structure of subtitle is closer to the oral language with great variability, this leads the translators of this sort of material to use informal and incompatible translations. The translations may drop some elements for cultural reasons. The correlation coefficient of German and Chinese sentence length in subtitles is 0.813, which indicates the sentence lengths between these two languages are not perfectly correlated. It results the poor performance of length based approach.

The score showed that the dynamic programming with lexical extension yields better precision, which could be expected due to the anchor points, since the lexical extension actually finds the translation pairs from the coarse aligned sentences. These reliable word alignments limit the search space for local minima. While we only allow the 1:1 alignments to be anchors, the original 1:2, 2:1 and 2:2 alignments will be divided into several pairs, which results more retrieved pairs than actual ones and leads the low recall. Our purpose is not to align all subtitles, but just to produce an aligned the subset of the corpus for further research. The developed alignment method on the total Chinese and German subtitle corpus retrieved only 1:1 and 1:2 alignments, for which there is a correct rate of 88.4%.

4 Conclusions

A sentence alignment method for movie subtitles is proposed in this paper. The proposed approach is based on the time-shift information in subtitles and it uses dynamic programming to minimize the global delay. The statistical lexical cues are also introduced to find word correspondence. As we have shown in the evaluation in Section 3, this additional technique yields better performance, it enhances about 7% of precision. Future work will be based on IBM Word Alignment Model to retrieve translation pairs instead of using co-occurrence statistic. The subtitle alignment is a novel and broad domain, and it may give a true picture of the translation quality and a useful system. The results of this paper may boost the research towards a practical MT system between German and Chinese.

References

1. Brown, P., Lai, J.C., Mercer, R.: Aligning Sentences in Parallel Corpora. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics, Berkeley, California, pp. 169–176 (1991)
2. Wu, D.K.: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In: Proceedings of the 32th Annual Conference of the Association for Computational Linguistics, Las Cruces, New Mexico, pp. 80–87 (1994)
3. Shemtov, H.: Text Aligment in a Tool for Translating Revised Documents. In: Proceedings of the 6th Conference on European Chapter of the Association for Computational Linguistics, Utrecht, The Netherlands, pp. 449–453 (1993)
4. Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., O'Hagan, M.: Improving the Quality of Automated DVD Subtitles via Example-based Machine Translation. In: Proceedings of Translating and the Computer, Aslib, London, vol. 28 (2006)
5. Martin, V.: The Automatic Translation of Film Subtitles. A Machine Translation Success Story? In: Resourceful Language Technology: Festschrift in Honor of Anna, vol. 7. Uppsala University (2008)
6. Mathieu, M., Emmanuel, G.: Multilingual Aligned Corpora from Movie Subtitles. Rapport interne LISTIC, p. 6 (2005)
7. Vandeghinste, V., Sang, E.K.: Using a Parallel Transcript/Subtitle Corpus for Sentence Compression. In: LREC, Lisbon, Portugal (2004)
8. Popowich, F., McFetridge, P., Turcato, D., Toole, J.: Machine translation of Closed Captions. *Machine Translation* 15, 311–341 (2000)
9. Och, F., Ney, H.: Improved Statistical Alignment Models. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440–447 (2000)
10. Lavecchia, C., Smaïli, K., Langlois, D.: Building Parallel Corpora from Movies. In: 5th International Workshop on Natural Language Processing and Cognitive Science, Funchal, Portugal (2007)
11. Tiedemann, J.: Improved Sentence Alignment for Movie Subtitles. In: Proceedings of the 12th Recent Advances in Natural Language Processing, Borovets, Bulgaria, pp. 582–588 (2007)
12. Reinhard, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, pp. 519–526 (1999)
13. Gale, W.A., Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19(1), 75–102 (1993)