Building Parallel Corpora from Movies

Caroline Lavecchia, Kamel Smaïli, David Langlois

Building Parallel Corpora from Movies

Lavecchia Caroline[1], Smaïli Kamel[1], and Langlois David[12]

[1] LORIA, Campus Scientifique, BP239, 54506 Vandoeuvre-lès-Nancy, FRANCE
[2] IUFM of Lorraine
{lavecchi,smaili,langlois}@loria.fr

Abstract. This paper proposes to use DTW to construct parallel corpora from difficult data. Parallel corpora are considered as raw material for machine translation (MT), frequently, MT systems use European or Canadian parliament corpora. In order to achieve a realistic machine translation system, we decided to use movie subtitles. These data could be considered difficult because they contain unfamiliar expressions, abbreviations, hesitations, words which do not exist in classical dictionaries (as vulgar words), etc. The obtained parallel corpora can constitute a rich ressource to train decoding spontaneous speech translation system. From 40 movies, we align 43013 English subtitles with 42306 French subtitles. This leads to 37625 aligned pairs with a precision of 92,3%.

1 Introduction

Training machine translation systems require a huge quantity of bilingual aligned corpora. Even if this kind of corpora becomes increasingly available, there may be a coverage problem for a specific need. Building bilingual parallel corpora is an important issue in machine translation. Several French-English applications use either the Canadian Hansard corpus or corpora extracted from the proceedings of European Parliament (Koehn, 2005). One way to enrich the existing parallel corpora is to catch the important amount of free available movie subtitles. Several web-sites (*http://divxsubtitles.net*) provide files used for subtitling movies. This quantity of information may enhance the existing bilingual corpora and enlarges the nowadays-covered areas. Furthermore, subtitles corpora are very attractive due to the used spontaneous language which contains formal, informal and in some movies vulgar words. Our research group is involved in a speech-to-speech translation machine project dedicated to a large community. That is why subtitles corpora are very worthy.
The raw subtitle corpora can not be used without processing. In order to make these files convenient for use, it is first necessary to align bilingual versions of the same movie at paragraph, sentence or phrase level. Usually, subtitles are presented on two lines of 32 characters which is readable on six seconds in maximum (Vandeghinste and Sang, 2004), this technical constraint makes the alignment problem more difficult.

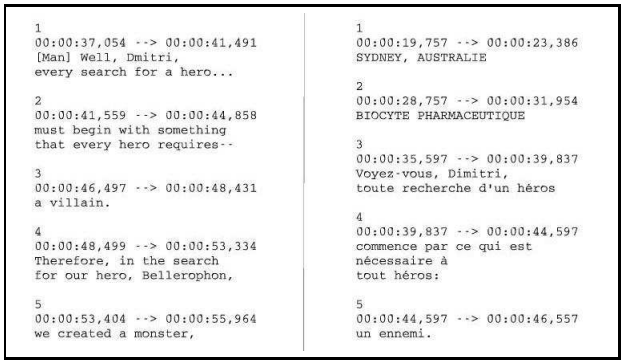
In this paper, we present a method which automatically aligns two subtitle files. This method is based on DTW (Dynamic Time Warping) algorithm. We pinpoint the specific features of subtitles and present a measure suitable to align efficiently.

2 An outline of the alignment problems

Our objective is to obtain as much pairs of aligned sentences from movie subtitles as possible. Two sentences are aligned if they are translations of one another. We get forty subtitle files in a text format in both English and French languages from the web-site *http://divxsubtitles.net*

2.1 Data description

A subtitle file is a set of phrases or words corresponding to: a set of dialogues, a description of an event or a translation of strings on screen (in general destinated to deaf people). A subtitle is a textual data usually displayed at the bottom of the screen. The text is written on original version or in a foreign language and corresponds to what is being said by an actor or what is being described. Fig. 1 shows a piece of subtitles extracted from the movie *Mission Impossible 2*.



```
1
00:00:37,054 --> 00:00:41,491
[Man] Well, Dmitri,
every search for a hero...

2
00:00:41,559 --> 00:00:44,858
must begin with something
that every hero requires--

3
00:00:46,497 --> 00:00:48,431
a villain.

4
00:00:48,499 --> 00:00:53,334
Therefore, in the search
for our hero, Bellerophon,

5
00:00:53,404 --> 00:00:55,964
we created a monster,
```

```
1
00:00:19,757 --> 00:00:23,386
SYDNEY, AUSTRALIE

2
00:00:28,757 --> 00:00:31,954
BIOCYTE PHARMACEUTIQUE

3
00:00:35,597 --> 00:00:39,837
Voyez-vous, Dimitri,
toute recherche d'un héros

4
00:00:39,837 --> 00:00:44,597
commence par ce qui est
nécessaire à
tout héros:

5
00:00:44,597 --> 00:00:46,557
un ennemi.
```

Fig. 1: Source and target movie subtitles

Each subtitle is characterized by an identifier, a time frame and finally a sequence of words. The time frame indicates the interval time the subtitle becomes visible on the screen. The sequence of words is the literal version of the dialogue or an event description. Subtitles as they are presented can not be used directly for alignment because the French and English subtitles do not match. In the example of Fig.1, the content of the first two subtitles mismatch, in fact the English subtitle begins with a dialogue when the French one does not. Because the movie is American, if any informative message is displayed on the screen, it is thus not necessary to repeat it into the English subtitle file. In the opposite in French the translation is necessary. This kind of difference occurs very frequently and produces gaps between the French and the English subtitles. In the next section, we detail the mismatch cases between the source and target subtitle files.

2.2 Source of subtitle delay

Several reasons are at the origin of delay between the source and the target subtitles, in the following we point out the most important of them.

Scene description insertion As pointed out before some scene movies are described by particular subtitles as illustrated by Fig. 1. The first two French subtitles situate physically the view, whereas this description is missing in the English version. Another example of mismatching is shown by figure 2. The English subtitles 13 to 15 describe

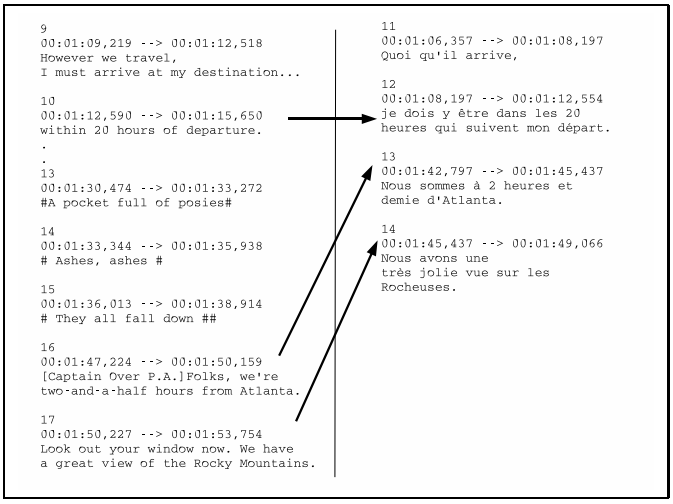


Fig. 2: Insertion of scene description

the scene, this description is skipped in the French version. This difference is due to the fact that subtitle files for a same movie are not necessarily written by the same person. One can decide to transcribe descriptions when another to let them down. Such descriptions in subtitles files are generally written in square brackets, between # or in upper case. Consequently, they are easily recognizable. To overcome this problem, we decided to remove all the identifiable descriptions from the text files. This solution is not sufficient to regulate and synchronize the source and target files.

Segmentation Unfortunately, even when descriptions are omitted in both languages, gaps between subtitles persist. In fact, a sentence in one language could be translated using several subtitles whereas in the other language it might be handled by only one subtitle. This will be entitled as a segmentation issue. A segmentation is the distribution of a sentence into one or several subtitles. For example, in Fig. 3, the English sentence *“However we travel, I must arrive at my destination within 20 hours of departure”*

is divided into two subtitles just like its corresponding French translation *"Quoi qu'il arrive, je dois y être dans les 20 heures qui suivent mon départ"*.

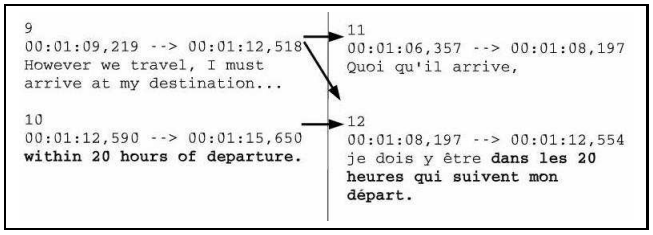


Fig. 3: Example of shifted segmentation

However, the segmentation is done differently in the two languages. Intuitively, the best way to proceed is to match the English subtitle 9 with the two French subtitles 11 and 12 and the English subtitle 10 with the French subtitle 12. Indeed, *"However we travel, I must arrive at my destination"* is the translation of *"Quoiqu'il arrive, je dois y être"* and *"within 20 hours of departure"* corresponds to *"dans les 20 heures qui suivent mon départ"*. Ideally, English subtitles 9 and 10 should be concatenated and matched with both French subtitles 11 and 12. Later, we will explain how to solve this problem.

Subtitle omission and insertion In addition to all the previous problems, some subtitles which transcript dialogues can occur in only one of the two versions. While it is simple to identify scene description insertions, it is difficult to decide automatically if a part of a dialogue has been omitted. In Fig. 4, we can distinguish several kinds of insertion.

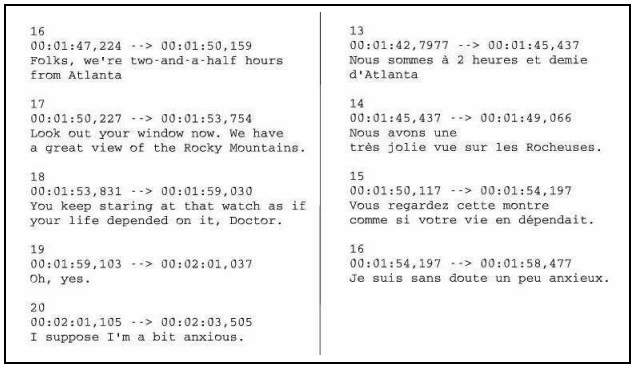


Fig. 4: An example of dialogue insertion

The English subtitle 17 which should match with French subtitle 14 contains an extra part: the phrase *Look at your window*. To overcome this problem, either we remove the entire pair $(17,14)$ and we loose information, or we keep it and we introduce noise. A third solution could be to remove the noise from the subtitle, but this way seems difficult because it needs a machine translation system. We can observe that English subtitle 19 has no corresponding in the French version. It does not match with any French subtitles. Removing it from the English script would be sufficient. The issue is how to automatically determine if a subtitle has or not an equivalent in the other language. We present in the next section the way we solved this problem.
To sum-up, we have seen that we can neither refer to subtitles identifiers (see Fig. 1) nor to time frames: sometimes the delay can reach 1.5 minutes. This delay in the movie is not regular, it grows up, it decreases, it rises again. It is difficult to find out any automatic rule to modelize this delay even if in certain research, authors refer to the use of frame time to align subtitles (Mangeot and Giguet, 2005). The only information in which we can focus on is the text. An alignment by hand is time and cost consuming, that is why we propose in the next section a method which automatically aligns subtitle pairs.

3 Alignment solutions

The major works aiming at solving the alignment of parallel corpora are based on dynamic programming. These works use a distance to evaluate the closeness between corpus segments. A segment can be a paragraph, a sentence or a phrase. The segmentation may be available or calculated automatically as in (Melamed, 1996). Several solutions and different options have been proposed, for more details we can refer to (Moore, 2002; Brown et al., 1991; Melamed, 1996; Vandeghinste and Sang, 2004; Gale and Church, 1991). One can find a comparative study about several of these methods in (Singh and Husain, 2005).

4 Dynamic Time Warping based on F-measure

Matching two subtitles can be considered as a classical problem of dynamic programming. As shown previously English and French subtitles are asynchronous. To align them, we utilize DTW based on F-measure. This measure is used to calculate the best path between two subtitle files. Intuitively, two subtitles are not considered as an aligned pair, if none or only few phrases of source and target match. This leads to guess that two subtitles do not match if their F-measure is weak.
In Fig. 5, each node (e, f) represents a potential matching point between English and French subtitle. A correct path begins by node $(0,0)$ and ends at node (E,F) where E is the number of English subtitles and F the number of French subtitles. From a node, the following shifts are possible:

- vertical progress from (e, f) to $(e, f+1)$: the subtitle e matches with two consecutive French subtitles (this case corresponds to the example given in 3)
- diagonal shift from (e, f) to $(e+1, f+1)$: the subtitle e matches with the subtitle f, then a shift towards $(e+1, f+1)$ is performed.

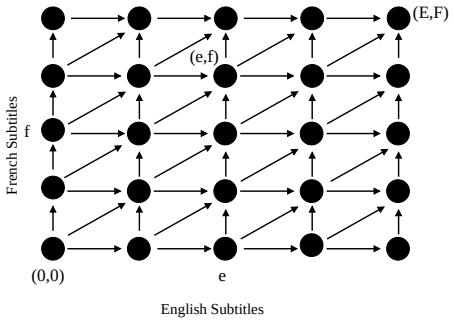


Fig. 5: Dynamic alignment for subtitles

– horizontal transition from (e, f) to $(e+1, f)$: the subtitle f matches with two consecutive English subtitles.

For each node (e, f), we define a matching score based on the F-measure (F_M) calculated as follows:

$$S(e,f) = max \begin{cases} S(e,f-1) + \beta_{F_m}(F_M(e,f) + \varepsilon) \\ S(e-1,f-1) + \alpha_{F_m}(F_M(e,f) + \varepsilon) \\ S(e-1,f) + \lambda_{F_m}(F_M(e,f) + \varepsilon) \end{cases}$$

α_{F_m}, β_{F_m} and λ_{F_m} are parameters chosen in order to find out the best alignment. These coefficients depend on the value of F_M (see section 5.1 for more details). One can notice that the previous formula uses a smoothed F-measure to prevent from a null value. F_M is calculated as follows:

$$F_M(e,f) = 2 \times \frac{R(e,f) \times P(e,f)}{R(e,f) + P(e,f)} \tag{1}$$

$$R(e,f) = \frac{match(e, tr(f))}{N(e)} \qquad P(e,f) = \frac{match(e, tr(f))}{N(f)} \tag{2}$$

$$match(e, tr(f)) = \sum_{i=1}^{n} \delta(e_i, tr(f_j)) \forall j \tag{3}$$

$tr(f)$ is a word-for-word translation of the French subtitle f. $tr(f)$ is obtained by using a French-English dictionary. $N(x)$ is the number of words in subtitle x. $match(e, tr(f))$ is the number of words which matches between the subtitles e and $tr(f)$ and the Kronecker $\delta(x,y)$ is a function which is 1 if x and y are equal and 0 otherwise. An example of matching is given in Fig. 6.

To make the matching more accurate, we decided to enhance the *match* function when an orthographic form occurs in both English and French subtitles. This makes proper names matching without introducing them into the dictionary.

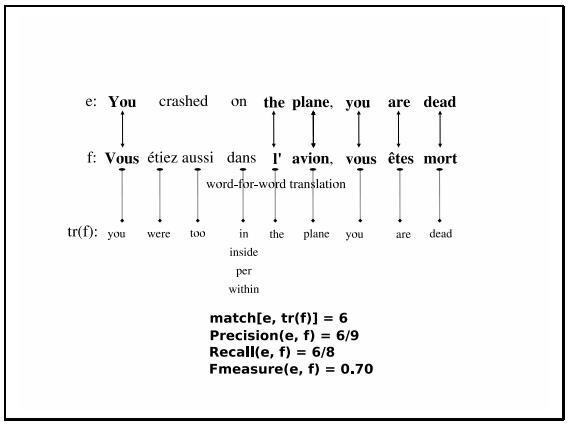


Fig. 6: Illustration of e and f matching

5 Evaluation

5.1 Test Corpora

Tests have been conducted on a corpus extracted from 40 movies. From each movie, we take out randomly around 35 English and their French corresponding subtitles. This leads to 1353 English subtitles (corpus T_E), and 1334 subtitles in French (corpus T_F). We aligned by hand the selected subtitles. This leads to 1364 ($\#A$) pairs of subtitles which constitute our reference corpus. We used a French-English dictionary extracted from the XDXF project[3]. It contains 41398 entries[4]. For the evaluation, we conducted the following procedure:

1. Removing from T_E and T_F subtitles describing events.
2. Alignment of English and French corpora.
3. Deletion of the unuseful subtitles: each matching pair for which the F-measure is zero is removed.
4. Comparison with the reference pairs.

A first test has been conducted to study the effect of α_{F_M}. We guess that if F_M is not null, we should give preference to the diagonal path.

In the following experiment, α_{F_M} varies from 1 (the diagonal is not favored) to 100 and β_{F_M} and λ_{F_M} are set to 1. Results in terms of recall, precision and F-measure are presented in Table 1.

$\#Tot.$ is the number of retrieved pairs. $\#C$ is the number of correct alignments. $\#I$ indicates the wrong identified pairs. With

$$Precision = \frac{\#C}{\#T} \quad Recall = \frac{\#C}{\#A} \tag{4}$$

The results showed that α_{F_M} parameter has a strong effect on the performance. We can notice that F_M increases with α_{F_M} until 7 and then the value becomes unstable. In

[3] http://xdxf.revdanica.com/

[4] Archive filename: comn_sdict05_French-English.tar.bz2

Table 1: Performance depending on α_{F_M} parameter

α_{F_M}	*#C*	*#I*	*#Tot.*	Rec.	Prec.	Fm.		α_{F_M}	*#C*	*#I*	*#Tot.*	Rec.	Prec.	Fm.
1	1063	842	1905	0.779	0.558	0.650		7	1119	97	1216	0.820	0.920	0.867
2	1124	213	1337	0.824	0.841	0.832		8	1118	96	1214	0.820	0.921	0.867
3	1124	114	1238	0.824	0.908	0.864		9	1119	94	1213	0.820	0.923	0.868
4	1121	99	1220	0.822	0.919	0.868		10	1118	94	1212	0.820	0.922	0.868
5	1121	98	1219	0.822	0.920	0.868		20	1116	93	1209	0.818	0.923	0.867
6	1120	97	1217	0.821	0.920	0.868		100	1114	92	1206	0.817	0.923	0.867

order to set the different parameters we have to remind our objective. In fact, we would like to collect as much aligned subtitles pairs as possible without introducing noise. Table 1 shows that this objective is reached when we maximize precision rather than F-measure. In fact, when precision increases, the number of False Positives[5] decreases. Considering this objective, we decided to set α_{F_M} to 9 in the following experiments. This value leads to 82% of recall and only 94 pairs mismatch. Analyzing results shows that the wrong identified pairs have sometimes a high F-measure. This is due to the weight of tool words (prepositions, conjunctions, ...). Such words are uniformly present in several subtitles which make the F-measure positive even if the French and English sentences do not match. This is particularly more critical when subtitles are short as illustrated on Table 2.

Table 2: Illustration of mismatching due to tool words

	E1 : Wallis **hold** on to this F1 : Wallace **tiens** moi cela	E1 : Wallis hold on **to** this F2 : Ulrich pense **à**
N(e)	5	5
N(f)	4	3
match	1	1
Prec.	1/4	1/3
Rec.	1/5	1/5
Fm.	0.22	0.23

Two potential pairs of alignment get the same F-measure if their constituent have the same length and the same number of matching words. The alignment (E1, F1) is considered correct whereas the second is wrong. Unfortunately, the F-measure refutes this fact. Indeed, the number of words matching in both pairs is the same but the correspondence in (E1, F2) concerns two small words (language tool word): *"à"* in French and *"to"* in English. It is obviously incongruous to let these small words having an important influence on the alignment decision. We can indicate that the proper name Wallace (Wallis) is missing from dictionary. A better dictionary coverage (including this proper name) will achieve a F-measure of 0.44 and allows the couple (E1, F1) to be a better alignment. To reduce the impact of tool words we modified the formula 5 as

[5] the number of incorrect alignments

follows:

$$match(e, tr(f)) = \sum_{i=1}^{n} \gamma \times \delta(e_i, tr(f_j)) \forall j \quad (5)$$

Where γ is smaller than one when e_i or f_j are tool words, otherwise γ is set to 1. Assigning less weights to tool words unfortunately does not improve results (Table 3). The more the weight decreases, the more F-measure, Recall and Precision fall. Naturally a subtitle is short (between 7 and 10 words) and furthermore it is formed by several tool words, it is henceforth difficult to do without this small words. By examining the subtitles pairs proposed by the automatic alignment (with $\alpha_{F_M} = 9$), we discover that 182 out of 1119 correct aligned pairs matched only because of tool words. By decreasing their weight in the match function, we decreased also the F-measure. This could explain also the last line of Table 3. When we omitted tool words (γ set to 0) we noticed that the number of proposed pairs felt considerably. We remind that in the procedure of alignment, we remove all the pairs (e, f) for which the F-measure is equal to 0. That is why all the pairs which matched only on tool words disappeared from the alignment, 289 subtitle pairs are concerned by this cut off.

Table 3: Impact of reducing the tool words' weight

γ	*#C*	*#I*	*#Tot*	Rec.	Prec.	Fm.		γ	*#C*	*#I*	*#Tot*.	Rec.	Prec.	Fm.
1.0	1119	94	1213	0.820	0.923	0.868		0.4	1056	171	1227	0.774	0.861	0.815
0.9	1097	134	1231	0.804	0.891	0.845		0.3	1044	189	1233	0.765	0.847	0.804
0.8	1097	134	1231	0.804	0.891	0.845		0.2	1040	192	1232	0.762	0.844	0.801
0.7	1097	134	1231	0.804	0.891	0.845		0.1	1039	194	1233	0.762	0.843	0.800
0.6	1097	133	1230	0.804	0.892	0.846		0.0	869	55	951	0.657	0.942	0.774
0.5	1097	133	1230	0.804	0.892	0.846								

By launching the developed alignment method on the total corpus (40 movies: 43013 English subtitles and 42306 French subtitles) we achieve 37625 aligned pairs.

6 Conclusion and perspectives

Working on parallel movie corpora constitutes a good challenge to go towards realistic translation machine applications. Indeed, movies corpora include so many common expressions, hesitations, coarse words,... Training decoding translation system on these corpora will lead to spontaneous speech translation machine systems. First results are very confident and can be used in order to constitute automatic aligned corpora. Tests have been conducted on a corpus of 40 movies, which correspond to 43013 English subtitles and 42306 French subtitles. By setting γ to 1 and α_{F_M} to 9, we obtained 37625 aligned pairs with a precision of $92,3\%$. This result is competitive in accordance to the state of art of noisy corpus alignment (Singh and Husain, 2005). However, we have to pursue our efforts in order to increase the precision which makes the parallel corpora noiseless. Several movies are available on the Internet, the result of the automatic alignment encourage us to boost our parallel corpus which is crucial for the decoding translation process. This work could be considered as a first stage towards a real time subtitling machine translation.

Bibliography

[Brown et al., 1991]Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176.

[Gale and Church, 1991]Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

[Koehn, 2005]Koehn, P. (2005). Europarl: A multilingual corpus for evaluation of machine translation. In *MT SUMMIT*, Thailand.

[Mangeot and Giguet, 2005]Mangeot, M. and Giguet, E. (2005). Multilingual aligned corpora from movie subtitles. Technical report, LISTIC.

[Melamed, 1996]Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In Brill, E. and Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–12. Association for Computational Linguistics, Somerset, New Jersey.

[Moore, 2002]Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the Association for Machine Translation in the Americas Conference*, pages 135–144.

[Singh and Husain, 2005]Singh, A. K. and Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*, pages 99–106.

[Vandeghinste and Sang, 2004]Vandeghinste, V. and Sang, E. K. (2004). Using a parallel transcript/subtitle corpus for sentence compression. In *LREC*, Lisbon, Portugal.