INTERSPEECH
2009 BRIGHTON

Context-driven automatic bilingual movie subtitle alignment

Andreas Tsiartas, Prasanta Ghosh, Panayiotis G. Georgiou, Shrikanth Narayanan

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

tsiartas@usc.edu, prasantg@usc.edu, georgiou@sipi.usc.edu, shri@sipi.usc.edu

Abstract

Movie subtitle alignment is a potentially useful approach for deriving automatically parallel bilingual/multilingual spoken language data for automatic speech translation. In this paper, we consider the movie subtitle alignment task. We propose a distance metric between utterances of different languages based on lexical features derived from bilingual dictionaries. We use the dynamic time warping algorithm to obtain the best alignment. The best F-score of ∼0.713 is obtained using the proposed approach.

Index Terms: subtitles alignment, dynamic time warping

1. Introduction

The statistical machine translator (SMT) is a critical part of cross-lingual speech-to-speech (S2S) systems. The training of an SMT requires a vast amount of parallel bilingual corpora; like most data-driven systems, the quantity and quality of training data reflecting the target domain's content and style dictate the coverage and performance of the resulting system. Hence, significant effort within the speech and language community has not only gone into new multilingual data collections, but devising automated data acquisition/selection schemes, such as from the Internet and other existing resources, that are appropriately matched to the target application domain needs. Furthermore, the design of S2S systems can benefit from SMT subcomponents built on bilingual transcriptions of spontaneous speech corpora because bilingual transcriptions of spontaneous speech corpora are more appropriate for training the SMT. One possible source of such corpora is movie subtitles readily available in multiple languages. The subtitles can be used in SMT training, if they can be aligned to form a parallel corpus. This is the topic addressed in the present paper.

Since the early 90's, corpora alignment has been a topic of research because of its usefulness in training machine translators. Brown et al. [1] extracted parallel utterances using information from the number of tokens in the utterances, without using any lexical cues. Following Brown's approach, Gale et al. [2] extracted parallel utterances using statistics of the length of utterances. Simand et al. [3] went one step further and they defined the cognateness measure, a measure stating how phonetically similar are two utterances, and they showed that this further improves alignment performance. Fung et al. [4] used the *dynamic time warping* (DTW) algorithm to align noisy parallel corpora using frequency, position and recency information.

Recently, movie subtitle alignment has become topic of significant research interest due to the increasing number of available subtitles on the web. Mangeot and Giguet [6] first described a methodology to align movie subtitles. Itamar et al. [10] introduced a cost function to align subtitles using non-contextual cues. They used DTW to find the best alignments. Tiedemann [8, 9, 7] synchronized subtitles by using manual anchor points and anchor points obtained from cognate filters. In addition, he used an existing parallel corpus to learn word translations to estimate anchor points. Then, based on estimated anchor points, he synchronized subtitles and obtained bilingual utterance 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. Lavecchia et al. [5] posed this problem as a sequence alignment problem such that the total sum of the aligned utterance-similarities is maximized. However, their proposed similarity metric suffers from stop words noise and affects the alignment performance. They also showed that their algorithm performance improves as they increase the weight of diagonal path. A high value of diagonal weight may override the importance of the similarity metric in DTW. Hence, such a strategy may not reflect the usefulness of the similarity metric.

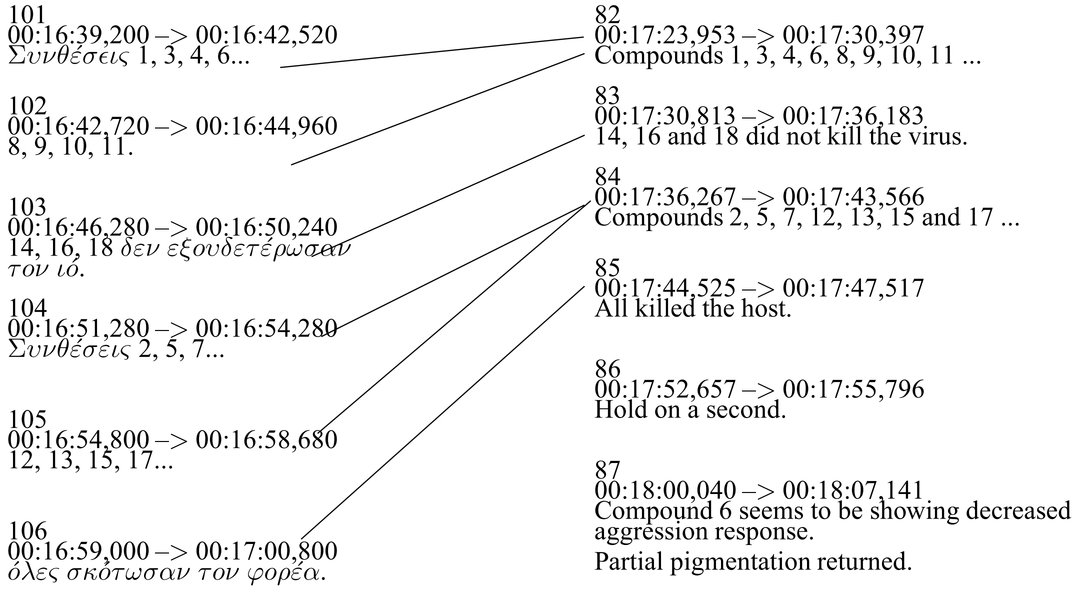
In this paper, we view this sequence alignment problem as a minimization problem. We propose a distance metric based on lexical cues that significantly improves alignment performance. In addition, we investigate the effect of stemming on the alignment quality. To the best of our knowledge, Lavecchia et al. [5] attempted to solve subtitle alignment using solely lexical cues. Thus, we compare the performance of our approach to that of [5]. We find that the performance using our approach yields a 17.5% relative improvement over that of [5].

This paper is structured as follows. In section 2, we describe the data that we collected. In section 3, we describe the proposed algorithm and the underlying formulation used in this work. In section 4, we describe the experimental setup and the evaluation methodology used in our approach. In section 5, we present the results of this work compared to prior efforts [5]. Finally, we summarize this work and propose some future directions.

2. Data set

To experiment with parallel subtitles alignment, we downloaded 42 Greek-English subtitles (http://www.opensubtitles.org/). For each subtitle, 40 random English utterances were manually aligned with the corresponding Greek utterances. In total, we obtained 1680 reference Greek-English utterance pairs. Some characteristics of this data set are: the utterance mappings are not always one-to-one, often one utterance in one language cor-

Work supported by NSF and NIH.



101
00:16:39,200 –> 00:16:42,520
Συνθέσεις 1, 3, 4, 6...

102
00:16:42,720 –> 00:16:44,960
8, 9, 10, 11.

103
00:16:46,280 –> 00:16:50,240
14, 16, 18 *δεν εξουδετέρωσαν τον ιό.*

104
00:16:51,280 –> 00:16:54,280
Συνθέσεις 2, 5, 7...

105
00:16:54,800 –> 00:16:58,680
12, 13, 15, 17...

106
00:16:59,000 –> 00:17:00,800
όλες σκότωσαν τον φορέα.

82
00:17:23,953 –> 00:17:30,397
Compounds 1, 3, 4, 6, 8, 9, 10, 11 ...

83
00:17:30,813 –> 00:17:36,183
14, 16 and 18 did not kill the virus.

84
00:17:36,267 –> 00:17:43,566
Compounds 2, 5, 7, 12, 13, 15 and 17 ...

85
00:17:44,525 –> 00:17:47,517
All killed the host.

86
00:17:52,657 –> 00:17:55,796
Hold on a second.

87
00:18:00,040 –> 00:18:07,141
Compound 6 seems to be showing decreased aggression response.
Partial pigmentation returned.

Figure 1: *An illustrative example of alignment between English and Greek subtitles from the movie "I am legend". Each utterance in the subtitles is preceded by the time stamps which are preceded by the utterance number*

responds to many utterances in the other language and vice versa. In addition, these two bilingual subtitles have different number of utterances, in some cases one language has utterances that do not exist in the other. An example of aligned utterances is shown in Fig. 1, obtained from the movie "I am Legend".

In this data, the average time length for 42 English subtitles is 7888 seconds with standard deviation 1584 seconds. For the Greek subtitles, the mean is 7898 seconds and the standard deviation is 1606 seconds. This shows that the subtitiles have about the same average duration. The average number of utterances per English subtitle is 1443 utterances with standard deviation 369. The average number of utterances per Greek subtitle is 1262 with standard deviation 334. This mismatch of average number of utterances is due to the fact that subtitles in two languages are created independently. Hence, sentences in two bilingual subtitles do not necessarily have one-to-one correspondence. Thus, in addition to one-to-one mappings, many-to-one and one-to-many mappings are required to align the subtitles.

3. Subtitle alignment approach

In this section, we formally state the problem definition of movie subtitle alignment. We also describe the alignment algorithm and the proposed distance metric.

3.1. Problem definition

The movie subtitle alignment problem is defined as follows:

Suppose we have to align subtitles in two languages, L_1 and L_2. Let us define $\mathcal{T}_k^{L_1} = \left\{S_i^{L_1} : i = k, \cdots, N_1\right\}$ and $\mathcal{T}_k^{L_2} = \left\{S_i^{L_2} : i = k, \cdots, N_2\right\}$, where $S_i^{L_1}$ is the i^{th} utterance of the subtitle in language L_1 and $S_j^{L_2}$ is the j^{th} utterance of the subtitle in language L_2. N_1 and N_2 are the number of utterances in subtitles in language L_1 and language L_2 respectively. The goal is to find the best mappings between utterances in two languages so that they reflect the best translations of the utterances in language L_1 to those in language L_2. Determining the best mappings can be expressed as the minimization of the distance, $\mathcal{D}$, between $\mathcal{T}_1^{L_1}$ and $\mathcal{T}_1^{L_2}$ as follows:

$$\left\{m_{ij}^{\star}\right\} = \arg\min_{m_{ij}} \mathcal{D}\left(\mathcal{T}_1^{L_1},\, \mathcal{T}_1^{L_2}\right) \tag{1}$$

where m_{ij} are the mappings. If $S_i^{L_1}$ matches with $S_j^{L_2}$, then m_{ij}=1, otherwise m_{ij}=0. The minimization problem has the following constraints:

1. Every utterance in $\mathcal{T}_1^{L_1}$ must have at least one mapping with an utterance in $\mathcal{T}_1^{L_2}$ and vice versa, which can be written as $\sum_i m_{ij} \geq 1$ and $\sum_j m_{ij} \geq 1$. In addition, we assume $m_{1,1} = 1$ and m_{N_1,N_2}=1. This means that the first and last utterances of both subtitles match.
2. The best mappings must not cross each other. If $S_i^{L_1}$ is aligned with $S_j^{L_2}$, then utterances in $\mathcal{T}_i^{L_1}$ can only be aligned with utterances in $\mathcal{T}_j^{L_2}$.

3.2. Proposed distance metric

Given $\mathcal{T}_1^{L_1}$ and $\mathcal{T}_1^{L_2}$, we want to find the best mappings $\left\{m_{ij}^{\star}\right\}$ such that it minimizes the global distance between two bilingual subtitles. We choose the distance metric $\mathcal{D}$ of (1) as follows:

$$\mathcal{D}\left(\mathcal{T}_1^{L_1}, \mathcal{T}_1^{L_1}\right) = \sum_{i,j} m_{ij} \mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) \tag{2}$$

where, $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$ is the distance metric between $S_i^{L_1}$ and $S_j^{L_2}$. Hence, the minimization problem of (1) becomes

$$\left\{m_{ij}^{\star}\right\} = \arg\min_{m_{ij}} \sum_{i,j} m_{ij} \mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) \tag{3}$$

The distance metric $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$ and the approach for solving the minimization problem using DTW are explained next.

3.2.1. *Relative-frequency based distance metric (RFDM)*

We propose the Relative-frequency based distance metric (RFDM) to capture the dissimilarity between two bilingual utterances, $S_i^{L_1}$ and $S_j^{L_2}$, based on their lexical context. The higher the value of the RFDM, the more dissimilar are the sentences. For computing the RFDM, we use a dictionary between language L_1 and language L_2. For each word in language L_2, we obtain its translations in language L_1 using the dictionary. Thus, we obtain a bag of words (BOW) in language L_1 for each utterance in the subtitle of language L_2. Common words between the BOW and an utterance in language L_1 are counted and used for calculating the RFDM.

Let the dictionary be denoted by $D_{L_2L_1}$, which takes one word from language L_2, w, and produces a set of words in language L_1, $D_{L_2L_1}(w)$. If the dictionary does not have the translation of the given word, it returns an empty set. Let $S_j^{L_2}$ have the following words $\left\{w_{j,1}^{L_2},\ w_{j,2}^{L_2},\ \cdots,\ w_{j,M_j^{L_2}}^{L_2}\right\}$. Using $D_{L_2L_1}$, we obtain a BOW $\mathcal{B}_j$ in language L_1; $\mathcal{B}_j = \left\{D_{L_2L_1}\left(w_{j,1}^{L_2}\right),\ D_{L_2L_1}\left(w_{j,2}^{L_2}\right),\ \cdots,\ D_{L_2L_1}\left(w_{j,M_j^{L_2}}^{L_2}\right)\right\}$. Similarly, let $S_i^{L_1}$ have the following words $\mathcal{W}_i = \left\{w_{i,1}^{L_1},\ w_{i,2}^{L_1},\ \cdots,\ w_{i,M_i^{L_1}}^{L_1}\right\}$. Suppose there are N_w unique words in this entire collection of words $\{\mathcal{B}_1,\ \mathcal{B}_1,\ \cdots, \mathcal{B}_{N_2}\}$ and w_k, appears C_{w_k} times in the entire collection $k = 1, ..., N_w$. The RFDM is computed as follows:

$$\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) = \left(\sum_{k=1}^{M_i^{L_1}} \frac{I_{w_{i,k}^{L_1}}}{C_{w_{i,k}^{L_1}}}\right)^{-1} \tag{4}$$

where

$$I_{w_{i,k}^{L_1}} = \begin{cases} 1 & \text{if } w_{i,k}^{L_1} \text{ is in both } \mathcal{W}_i \text{ and } \mathcal{B}_j \\ 0 & \text{otherwise.} \end{cases}$$

I_w is the indicator variable of word w and takes value 1, when w is in both $\mathcal{B}_j$ and $\mathcal{W}_i$. Words like 'the', 'is' are very frequent words and hence, their counts are high in general. Due to inverse of the counts in (4), frequent words contributes less (high distance) in the distance computation. Even one infrequent common word (i.e. C_{w_k} is small) reduces the distance to small value and thus, contributes more in distance computation.

3.3. Dynamic time warping algorithm

Having defined $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$, we now explain the DTW algorithm for solving (3). The goal of the DTW algorithm is to find the best mappings $\{m_{ij}^\star\}$ in between $\mathcal{T}_1^{L_1}$ and $\mathcal{T}_1^{L_2}$ to minimize $\sum_{i,j} m_{ij}\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$.

For the description of the DTW algorithm, let us define $\mathcal{L}(S_i^{L_1}, S_j^{L_2})$ as the total distance for the best mappings starting from $(S_1^{L_1}, S_1^{L_2})$ ending at $(S_i^{L_1}, S_j^{L_2})$. Let us also define $\zeta(S_i^{L_1}, S_j^{L_2})$ (called back-tracking variable) as a temporary variable, which stores the last mapping in the best path ending at $(S_i^{L_1}, S_j^{L_2})$.

The algorithm is described below:

1. *Initialization:*

$$\mathcal{L}(S_1^{L_1}, S_1^{L_2}) = \mathcal{DM}\left(S_1^{L_1}, S_1^{L_2}\right)$$

$$\begin{aligned} \mathcal{L}(S_1^{L_1}, S_j^{L_2}) &= \mathcal{L}(S_1^{L_1}, S_{j-1}^{L_2}) + \mathcal{DM}\left(S_1^{L_1}, S_j^{L_2}\right) \\ \zeta(S_1^{L_1}, S_j^{L_2}) &= \left(S_1^{L_1}, S_{j-1}^{L_2}\right),\ j = 1, ..., N_2 \end{aligned}$$

$$\begin{aligned} \mathcal{L}(S_i^{L_i}, S_1^{L_2}) &= \mathcal{L}(S_{i-1}^{L_1}, S_1^{L_2}) + \mathcal{DM}\left(S_i^{L_1}, S_1^{L_2}\right) \\ \zeta(S_i^{L_1}, S_1^{L_2}) &= \left(S_{i-1}^{L_1}, S_1^{L_2}\right),\ i = 1, ..., N_1 \end{aligned}$$

2. *Iteration:*

$$\begin{aligned} &\mathcal{L}(S_i^{L_1}, S_j^{L_2}) = \mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right) + \\ &\qquad \min\left\{\mathcal{L}(S_i^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_j^{L_2})\right\} \\ &\zeta(S_i^{L_1}, S_j^{L_2}) = \\ &\quad \arg\min\left\{\mathcal{L}(S_i^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_{j-1}^{L_2}), \mathcal{L}(S_{i-1}^{L_1}, S_j^{L_2})\right\} \\ &\qquad ,\ i = 2, ..., N_1,\ j = 2, ..., N_2 \end{aligned}$$

3. *Backtracking:* Let there be K mappings. Then

$$m_{N_1,N_2} = m(S_{N_1}^{L_1}, S_{N_2}^{L_2}) = 1,\ m_{1,1} = m(S_1^{L_1}, S_1^{L_2}) = 1$$
$$\xi_K = \zeta(S_{N_1}^{L_1}, S_{N_2}^{L_2})$$

$$\begin{cases} m(\xi_k) = 1 \\ \xi_{k-1} = \zeta(\xi_k),\ k = K, ..., 2. \end{cases}$$

4. Experimental procedure

For this experiment, we considered the dataset described in section 2. As a preprocessing step, we cleaned subtitles from noisy symbols (i.e. non-alphanumeric symbols) similar to what one would do for cleaning text for statistical machine translation purposes; we removed all the time stamps and the reference numbers (shown in Fig. 1). Finally, we removed all punctuation and capitalized all letters.

We mined English translations of Greek words by querying the Google dictionary ("http://www.google.com/dictionary"). Using translated words, we computed the RFDM. For comparison, we also computed the F-measure [5]. In addition, we inverted the F-measure (called IF-measure) and used it as $\mathcal{DM}\left(S_i^{L_1}, S_j^{L_2}\right)$ in (3) to obtain the best alignments.

The Google dictionary, in general, often does not return translations of words. For example, the English-Spanish Google dictionary returns translations of word 'say', but not for word 'says'. For mining English translations of Greek words, the Google dictionary returned translations of 4441 Greek words out of 33812 words queried. Due to this limitation, the F-measure often turned out to be zero and similarly, the IF-measure and the RFDM turned out to be ∞.

We used the manual mappings as references for evaluating the movie subtitle alignment task. We computed the precision, the recall, and reported the F-score [11], averaged over all movies. To check the effect of stemming on this problem, we repeated the above experiments with stemming [12]. Stemming was performed on the English words and on the English BOW obtained after translating Greek utterances.

5. Results

In table 1, we report the F-score of the movie subtitle alignment task using the F-measure, the IF-measure and the RFDM. Lavecchia et al. in [5] considered the movie subtitle alignment task as a maximization problem using the F-measure, unlike minimization of (3). From table 1, it can be seen that formulating the alignment problem as the minimization of the total distance is advantageous over formulating it as a maximization of the total similarity. The RFDM gives 8-11% absolute improvement in terms of the F-score compared to that of the F-measure.

Metric	No Stemming	Stemming
F-measure	0.609	0.607
IF-measure	0.681	0.688
RFDM	0.711	0.713

Table 1: The F-score of subtitle alignment using different metrics

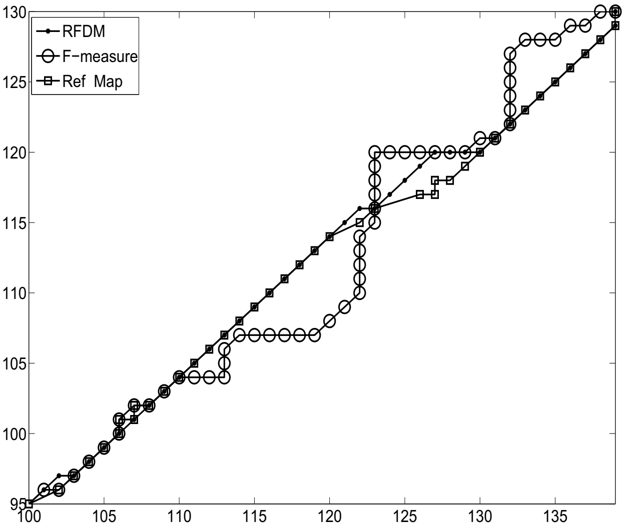


Figure 2: *A sample subtitle alignment using the RFDM and the F-measure*

For comparing the best mappings obtained by the F-measure and the RFDM, we picked a segment of the alignment obtained by these approaches in Fig. 2. The figure also shows the reference maps. It is clear that the maximization approach using the F-measure results in a path which is far off from the reference alignment, while the minimization approach is not. Among all the metrics used for minimization, the RFDM performs even better than the IF-measure. From table 1, it appears that the RFDM models the conceptual similarity between two utterances in two languages better than the F-measure.

While the maximization of the total F-measure using DTW requires fine tuning of the diagonal weight (Lavecchia et al. [5] reported best diagonal weight to be 9), the minimization approach works better than the F-measure with diagonal weight 1. Optimization of the diagonal weight of the minimization approach might have improved the results.

The F-scores obtained for metrics computed by stemming do not differ much from those without stemming. One possible reason could be that stemming introduced some noise, by mapping words of different meaning to the same root word. Due to such noise, any benefit from stemming might have been compensated.

6. Conclusion

In this paper, we proposed the relative-frequency based distance metric (RFDM) between two utterances in two languages under the dynamic time warping (DTW) framework. For movie subtitle alignment, we demonstrated $\sim$0.11 absolute improvement in terms of the F-score (17.5% relative improvement) compared to the maximization of the total similarity obtained by the DTW algorithm. We believe that in addition to lexical features, timing information can be used to improve the alignment performance. In future, we want to investigate how the dictionary sparsity can affect the alignment performance. We also want to investigate the effect of relaxing assumption 1 of section 3.1, thus, accounting for missing pieces of the dialog in the subtitles. Additionally, we want to investigate the quality of parallelism of the selected best alignments for removing noisy alignments due to very liberal interpretations. Finally, we intend to investigate the effects of subtitle transcripts on SMT performance.

7. Acknowledgements

Authors thank Margarita Siamptani for her efforts in providing the reference maps for all the movies.

8. References

[1] Brown P. F., Lai J. C., and Mercer R. L., “Aligning utterances in parallel corpora”, Meeting of the Association for computational linguistics, 1991.

[2] Gale W. A., Church K. W., “A program for aligning utterances in bilingual corpora”, Proc. of the 29 th Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991, pp 177-184.

[3] Simard M., Foster G. F., Isabelle P., “Using cognates to align sentences in bilingual corpora”, Proc. conference of the Centre for Advanced Studies on Collaborative research: distributed computing, Vol. 2, Toronto, Ontario, Canada, 1993, pp 1071-1082.

[4] Fung P., Mckeown K., “Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping”, Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, Maryland, 1994, pp 81-88.

[5] Lavecchia C., Smali K., Langlois D., “Building Parallel Corpora from Movies”, The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2007 (2007), Funchal, Madeira.

[6] Mangeot, M., and Giguet, E., “Multilingual aligned corpora from movie subtitles”, Technical report, LISTIC, 2005.

[7] Tiedemann J., “Building a multilingual parallel subtitle corpus”, In Proceedings of 17th CLIN, Leuven, Belgium, 2007.

[8] Tiedemann J., “Improved sentence alignment for movie subtitles”, In Proceedings of RANLP 2007, Borovets, Bulgaria, 2007, pp 582-588.

[9] Tiedemann J., “Synchronizing Translated Movie Subtitles”, Proceedings of the Sixth International Language Resources and Evaluation (LREC ’08), Marrakech, Morocco, 2008, pp 1902-1906.

[10] Itamar E. and Itai A., “Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora”, Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), 2008, pp 269-272.

[11] Yates R. B., Neto R. B., “Modern Information Retrieval”, New York, ACM Press, Addison-Wesley, 1st edition May 15, 1999.

[12] Porter M.F., “An Algorithm for Suffix Stripping”, Program, July 1980, vol 14, issue 3, pp 130-137.