Acquis Communautaire Sentence Alignment using Support Vector Machines

Alexandru Ceauşu, Dan Ştefănescu, Dan Tufiş

Research Institute for Artificial Intelligence of the Romanian Academy
13, Calea 13 Septembrie, 050711, Bucharest
{alceausu, danstef, tufis}@racai.ro

Abstract

Sentence alignment is a task that requires not only accuracy, as possible errors can affect further processing, but also requires small computation resources and to be language pair independent. Although many implementations do not use translation equivalents because they are dependent on the language pair, this feature is a requirement for the accuracy increase. The paper presents a hybrid sentence aligner that has two alignment iterations. The first iteration is based mostly on sentences length, and the second is based on a translation equivalents table estimated from the results of the first iteration. The aligner uses a Support Vector Machine classifier to discriminate between positive and negative examples of sentence pairs.

1. Introduction

Sentence alignment is a prerequisite for any parallel corpora processing and has been proven that very good results can be obtained with practically no prior knowledge about the concerned languages. However, as the sentence alignment errors may be detrimental to further processing, ensuring higher sentence alignment accuracy is a continuous concern for many NLP practitioners.

Sentence alignment is not characterized only by accuracy. To be language pair independent is other important demand that a sentence aligner must meet. In addition, many implementations stress on their ability to use little computation resources.

The sentence aligner employs a Support Vector Machine classifier for the discrimination between "good" and "bad" sentence pairs. The aligner was tested on selected pairs of languages from the recently released 20-languages Acquis Communautaire parallel corpus (http://wt.jrc.it/lt/acquis/).

2. Related Work

One of the best-known algorithms for aligning parallel corpora (Gale and Church, 1991) is based on the lengths of sentences being reciprocal translations and a very popular implementation is the Vanilla aligner (http://nl.ijs.si/telri/Vanilla/) due to P. Danielsson and D. Ridings. Chen (1993) developed a method based on optimizing word translation probabilities that has better results than the sentence-length based approach, but it demands much more time to complete and requires more computing resources. Melamed (1996) also developed a method based on word translation equivalence and geometrical mapping.

Moore (2002) presents a hybrid approach that has three stages. In the first stage, the algorithm uses length-based methods for sentence alignment. In the second stage, a translation equivalence table is estimated from the aligned corpus resulted in the first stage. The method used for translation equivalents estimation is based on IBM model 1 (Brown, 1993). The final step uses a combination of length-based methods and word correspondence to find 1-1 sentence alignments. The aligner has an excellent precision for one-to-one alignments because it was meant for acquisition of very accurate training data for machine translation experiments. Another problem of this aligner is that it was tested on only 10,000 sentence pairs (it cannot process more than 100,000 sentence pairs).

3. Features Selection

In the process of features selection, any sentence pair can be characterized by a collection of scores for each feature. Therefore, the alignment problem can be reduced to a two-class classification task: discriminating between "good" and "bad" alignments. One of the best performing formalism for this task proves to be Vapnik's Support Vector Machine (Vapnik, 1995).

We used an out-of-the-box solution for Support Vector Machine (SVM) training and classification - LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm (Fan et al., 2005) with default parameters (C-SVC classification and radial basis kernel function).

The accuracy of the SVM model was evaluated (10-fold cross validation) on five manually aligned files from the Acquis Communautaire corpus for the language pairs English-French, English-Italian, and English-Romanian. For each language pair experiment we used approximately 1000 sentence pairs.

To train the model the SVM model, each sentence pair from the "gold standard" is characterized by a collection of scores on features like translation equivalence, word length correlation, word rank correlation, etc. The examples of "bad" alignment were generated automatically from the gold standard, replacing one sentence in a correctly aligned pair with another sentence in the three-sentence vicinity. The replaced sentence was randomly selected from either the previous or following sentences.

The SVM classifier performance increases considerably when it uses more highly discriminative features. The irrelevant features or those with less discriminative power negatively influence the SVM classifier accuracy. Therefore, in the process of features selection we evaluated a series of features, out of which the best performing are listed in the Table 1. The *non-word length correlation* in Table 1 refers to non-lexical tokens, language independent such as punctuation,

numbers and currency symbols. Among the features we dropped from further consideration were difference in length (in both words and characters) for candidate sentences and the difference of the relative positions of the candidate sentences.

The length difference between the alignment candidate sentences – even though it can successfully discriminate positive and negative examples with an accuracy of 85%– was discarded because it did not improved the model when the correlation of length difference was employed. The correlation of the length in characters of the sentences, although it has an accuracy of 96%, when it is used independently, it adds no additional information in combination with other features.

Surprisingly, given the expected monotonicity of aligned sentences numbers, the difference of the relative positions of the sentences was not a very discriminative feature. Its classification accuracy was only 62%.

	Precision
Translation equivalence	98.47
Word sentence length correlation	96.77
Character sentence length correlation	96.01
Word rank correlation	94.86
Non-word sentence length correlation	93.00

Table 1: 10-fold cross validation precision for each feature independently evaluated

Each feature presented in Table 1 has several components designed to account for the context in which the sentences are aligned. Also, equally important are the features attributes designed to help recognize the wrongly aligned sentence pairs.

3.1. Translation equivalence

Translation equivalence can be considered the most important feature for training the SVM model because the sentence pairs, when characterized only by this feature, can be classified as "good" or "bad" with an accuracy of 98.47%.

For the estimation of the translation equivalence, we use the well-known IBM model 1 (Brown et al. 1993). In the estimation, besides translation equivalence, we also use features dependent of the context of the alignment (Tufis et al. 2005a, b). The link locality feature accounts for the degree of the cohesion of links surrounding the candidate link. The link locality is computed for a window of words, the span of which is dependent on the length of the aligned sentences. Another feature we use in parameter estimation is the crossed links score that computes (for a window size also depending on the sentences lengths) the links that were crossed by the candidate link.

The parameter estimation phase of our aligner is an iterative process that uses different feature weights and thresholds for each of the iterations. The weights and thresholds are manually set in order to favour the alignment of anchor words in the early iterations.

For the purpose of sentence alignment, the recall of the word alignment is less important than the precision. Therefore, when building the translation equivalence dictionary we did not take into account the happax-legomena words. They were mapped to the "unknown" token. Unlike in the IBM model 1 implementation we did not consider the null alignments (words not translated in the other side of the bitext); we found that the null word-alignments do not help the sentence alignment process.

To ensure fast processing of mass documents our search for the best word alignment solution considered only the candidates the translation equivalence score of which was above an empirically established threshold (0.05). We sum the translation equivalence scores for the respective pairs and normalize it with the average length of the sentences in the analyzed pair. This figure is called the sentence-pair translation equivalence score (TES) and is one of the attributes of the translation equivalence feature. Other attributes are the translation equivalence scores of the preceding and succeeding pairs of sentences. Additionally, we used another attribute which proved to be very useful in detecting "bad" sentence alignments: the translation equivalence scores of the sentences in the candidate pair with the surrounding sentences. This attribute can improve the SVM classification based on the translation equivalence feature with up to 1%.

3.2. Word sentence length correlation

A correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value, the other has a low (high) value. A correlation coefficient of 0 means that there is no linear relationship between the variables.

	Length	Sentence text
Source	6	(a) fully compliant ;
	9	(b) compliant , but improvement desirable ;
	10	(c) not compliant , with minor deficiencies ;
	10	(d) not compliant , with serious deficiencies ;
	6	(e) not applicable ;
	6	(f) not confirmed .
	2	Article 11
	5	Answer of the appropriate authority
Target	6	(a) conformitate deplină ;
	12	(b) conform , dar este de dorit o ameliorare ;
	11	(c) nu este conform , prezintă deficiențe minore ;
	11	(d) nu este conform , prezintă grave deficiențe ;
	7	(e) nu este cazul ;
	7	(f) nu se confirmă .
	2	Articolul 11
	3	Răspunsul autorității competente

Table 2: Length correlation for a window of 8 sentences

In our case, the variables are the lengths (in words) of the sentences in source and target documents. We used correlation coefficient to measure the co-variance of these

lengths inside windows of different sizes. Correlation coefficients may be computed in many ways. We employed the following formula to compute the correlation coefficient used in our application:

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the means of the sample AVERAGE(X) and AVERAGE(Y).

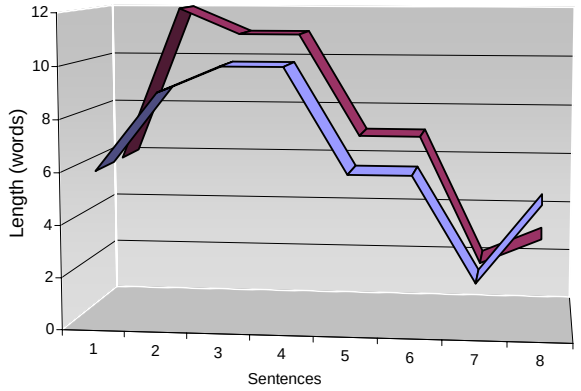

Figure 1: Length correlation for the sentences in Table 2

3.3. Word rank correlation

Word rank correlation is a feature that can help, by itself, the SVM model to label correctly as "good" or "bad" with an accuracy of 94.86 percent.

This feature can successfully replace the translation equivalence feature when a translation equivalence dictionary is not available.

Word rank correlation works on the common-sense belief that words with a high occurrence in the source corpus tend to be translated with words with high occurrence in the target corpus. The words of bitext to be aligned are sorted by their occurrence ranks. The top 25% of the words in each sorted list are considered representative for the bitext. Then, for each sentence in a candidate alignment pair we counted the representative words and these counts were normalized by the word lengths of the corresponding sentences. The absolute difference of these normalized counts represents the feature's score.

Other attributes of this feature are the word rank correlation scores similarly computed for the preceding and the succeeding pairs.

3.4. Non-word sentence length correlation

There are certain non-lexical tokens that are independent of the language pair such as the punctuation marks, numbers, currency symbols, etc. Counting these tokens (we call them non-words) and computing the correlation of their number for the sentence pairs, the SVM model, based on this feature, can distinguish a "good" pair from a "bad" one with an accuracy of 93 percent.

After considering each feature independently, we evaluated their combination. As can be seen in Table 2 non-word length and word rank correlation features add the same amount of information when word length correlation is used. Another observation that can be draw from table 2 is that language independent characters add much more information than word rank correlation when translation equivalence feature is used.

			I		II		
Translation equivalence				x	**x**	x	x
Word sentence length correlation	x	x	**x**	x	**x**	x	x
Non-word sentence length correlation		x	**x**		**x**		x
Word rank correlation	x		**x**			x	x
Precision (%)	97.87	97.87	**98.32**	98.72	**98.78**	98.51	98.75

Table 3: 10-fold cross validation precision of the SVM classifier using different combinations of features

4. The Sentence Aligner

We describe a hybrid sentence aligner that has 2 alignment phases: (a) length based sentence alignment; and (b) length and word-translation based sentence alignment.

The aligner does not have a-priori language specific information and its parameters are trained using just a small portion of human checked alignment data (1000 examples of correctly aligned pairs).

The first phase of the sentence aligner consists in training the SVM model on a Gold Standard that comprises 1000 samples. The features used in the first alignment phase (see table 2, column I) are the word sentence length, the non-word sentence length and the representative word rank correlation scores. This part of the alignment model does not significantly depend on the language pair of the bitext. The results shown in Table 3 for En-It, En-Fr and En-Ro bitexts were obtained using the SVM model learnt from the English-Romanian training data

The main feature used in the second phase is the translation equivalence. The sentence pairs which were classified as "good", with a score higher than 0.9, were used to estimate the translation equivalence table. A new SVM model is trained on the Gold Standard, this time using all the features four features. As one can see in Table 2, the feature combination with the best score is the one in column II.

The aligning process of the second phase has several stages and iterations. In the first stage, a list of sentence pair candidates for alignments is created and the SVM model is asked to return the probability estimates for these candidates being correct. The candidate pairs are formed in the following way: the i^{th} sentence in the source language is paired with the j^{th} presumably correspondent target sentence as well as with the neighboring sentences within a window the length of which is documents specific. The index j of the presumably correspondent target sentence is selected so that the pair <i,j> is the closest pair to the main diagonal of the length bitext representation. The window depends on the files length and on the files lengths difference, in terms of number of sentences.

In the second stage, an EM algorithm re-estimates the sentence-pairs probabilities in five iterations.

The third stage involves multiple iterations and thresholds. In one iteration step, the best-scored alignment is selected as a good alignment (only if above a pre-specified threshold) and the scores of the surrounding candidate pairs are modified as described below.

Let it be *(i, j)* the sentence pair considered a good alignment; then

- for the candidates *(i-1, j-1)* and *(i+1, j+1)* their respective scores are increased by a confidence bonus δ,
- for candidates *(i-2, j-2)* and *(i+2, j+2)* their respective scores are increased by δ/2,
- for candidate alignments which intersect the correct alignment*(i, j)*, their respective scores decreased by 0.1,
- for candidates *(i, j-1)*, *(i, j+1)*, *(i-1, j)*, *(i+1, j)* the respective scores are decreased by an amount in inverse ratio with their estimate probabilities; this will maintain the possibility for detections of 1-2 and 2-1 links; the correctness of these detections is directly influenced by the amount mentioned above,
- candidates *(i, n)* and *(m, j)* with $n \leq$ j-2, $n \geq$ j+2, $m \leq$ i-2, $m \geq$ i+2 are eliminated.

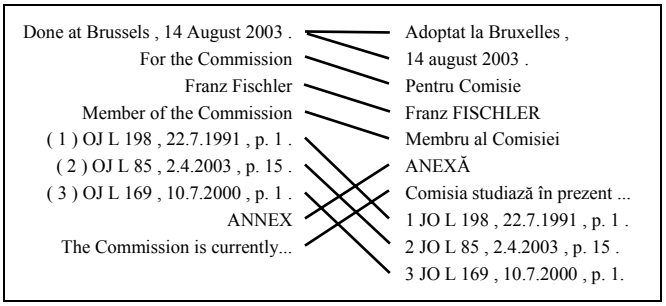

Figure 2: Example of the SVM sentence aligner's output for multiple and cross alignments

The sentence alignment is a multiple steps greedy process. Each step can be repeated several times. In the first step, the candidates are chosen only if their estimated probabilities are greater than a threshold of 0.99. The choice of a candidate as a good alignment pair influences the estimated probabilities of the neighboring candidates. This step is repeated, with the confidence bonus δ decreasing after each iteration, until no more candidates are chosen. The next steps are similar with the first step, the main difference being that the threshold of acceptance is lowered to 0.90, 0.85, 0.75, 0.65, 0.50, 0.35 and, finally, to 0.10. The δ confidence bonus decreases accordingly. For a threshold of 0.99, δ starts from 0.23, while for a threshold of 0.1 it starts from 0.1. We should note that the lowest value of the acceptance threshold is a matter of precision-recall compromise and depends on the intended application of the alignment results.

5. Evaluation

The evaluation of the aligner was made on 4 AcquisCom files (different from the ones used to evaluate the SVM model precision). Each language pair (English-French, English-Italian, and English-Romanian) has approximately 1000 sentence pairs.

The aligner can also use some specific features of the Acquis Communautaire parallel corpus. The most important is the fact that the corpus has the same numbers of articles in each language. In addition, the documents observe, irrespective of the language, a precise structuring (encoded as a unique DTD). Although not strictly needed, a preliminary alignment using the hard delimiters (in our case "articles") ensures a much faster processing. We did not use it for the evaluation presented in table 3.

	Precision	Recall	F-Measure
Moore En-It	**100**	97.76	98.86
SvmSent Align En-It	98.93	**98.99**	**98.96**
Moore En-Fr	**100**	98.62	99.30
SvmSent Align En-Fr	99.46	**99.60**	**99.53**
Moore En-Ro	**99.80**	93.93	96.78
SvmSent Align En-Ro	99.24	**99.04**	**99.14**

Table 3: The evaluation of SVM sentence aligner and Moore's Bilingual Sentence Aligner

As can be seen from table 3 our aligner can't best the precision of Moore's bilingual sentence aligner, but it has a very good recall indifferent of the language pairs.

6. References

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, vol. 19, 263-311, June 1993

Chen, S.F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 9-16

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005). Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University

Gale, W.A., Church, K.W. (1991). A program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 1991, 177–184

Melamed, I.D. (1996). A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report 96-22, University of Pennsylvania

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135-244

Tufiş, D., Ion, R., Ceauşu, Al., Stefănescu D. (2005a). Combined Aligners. In Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", June, 2005, *Ann Arbor, Michigan, June,* Association for Computational Linguistics, pp. 107-110

Tufiş, D., Ceauşu, Al., Ion, R., Ştefănescu, D. (2005b). An integrated platform for high-accuracy word alignment, JRC Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages, Arona, Italy

Vapnik, N. V. The Nature of Statistical Learning Theory. Springer, 1995.