Computer Engineering and Intelligent Systems www.iiste.org
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)

Sentence Alignment using MR and GA

Mohamed Abdel Fattah
FIE, Helwan University
Cairo, Egypt
mohafi2003@helwan.edu.eg

Abstract
In this paper, two new approaches to align English-Arabic sentences in bilingual parallel corpora based on mathematical regression (MR) and genetic algorithm (GA) classifiers are presented. A feature vector is extracted from the text pair under consideration. This vector contains text features such as length, punctuation score, and cognate score values. A set of manually prepared training data was assigned to train the mathematical regression and genetic algorithm models. Another set of data was used for testing. The results of (MR) and (GA) outperform the results of length based approach. Moreover these new approaches are valid for any languages pair and are quite flexible since the feature vector may contain more, less or different features, such as a lexical matching feature and Hanzi characters in Japanese-Chinese texts, than the ones used in the current research.
Key words: Sentence Alignment, English/ Arabic Parallel Corpus, Parallel Corpora, Machine Translation, Mathematical Regression, Genetic Algorithm.

1. Introduction

Recent years have seen a great interest in bilingual corpora that are composed of a source text along with a translation of that text in another language. Nowadays, bilingual corpora have become an essential resource for work in multilingual natural language processing systems (Fattah et al., 2007; Moore, 2002; Gey et al., 2002; Davis and Ren, 1998), including data-driven machine translation (Dolan et al., 2002), bilingual lexicography, automatic translation verification, automatic acquisition of knowledge about translation (Simard, 1999), and cross-language information retrieval (Chen and Gey, 2001; Oard, 1997). It is required that the bilingual corpora be aligned. Given a text and its translation, an alignment is a segmentation of the two texts such that the *n*th segment of one text is the translation of the *n*th segment of the other (as a special case, empty segments are allowed, and either corresponds to translator's omissions or additions) (Simard, 1999; Christopher and Kar, 2004). With aligned sentences, further analysis such as phrase and word alignment analysis (Fattah et al., 2006; Ker and Chang, 1997; Melamed, 1997), bilingual terminology and collocation extraction analysis can be performed (Dejean et al., 2002; Thomas and Kevin, 2005).

In the last few years, much work has been reported in sentence alignment using different techniques. Length-based approaches (length as a function of sentence characters (Gale and Church, 1993) or sentence words (Brown et al., 1991)) are based on the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. These approaches work quite well with a clean input, such as the Canadian Hansards corpus, whereas they do not work well with noisy document pairs (Thomas and Kevin, 2005). Cognate based approaches were also proposed and combined with the length-based approach to improve the alignment accuracy (Simard et al., 1992; Melamed, 1999; Danielsson and Muhlenbock, 2000; Ribeiro et al., 2001).

Sentence cognates such as digits, alphanumerical symbols, punctuation, and alphabetical words have been used. However all cognate based approaches are tailored to close Western language pairs. For disparate language pairs, such as Arabic and English, with a lack of a shared Roman alphabet, it is not possible to rely on the aforementioned cognates to achieve high-precision sentence alignment of noisy parallel corpora (however, cognates may be efficient when used with some other approaches). Some other sentence alignment approaches are text based approaches such as the hybrid dictionary approach (Collier et al., 1998), part-of-speech alignment (Chen and Chen, 1994), and the lexical method (Chen, 1993). While these methods require little or no prior knowledge of source and target languages and give good results, they are relatively complex and require significant amounts of parallel text and language resources.

Instead of a one-to-one hard matching of punctuation marks in parallel texts, as used in the cognate approach of Simard (Simard et al., 1992), Thomas (Thomas and Kevin, 2005) allowed no match and one-to-several matching of punctuation matches. However, neither Simard nor Thomas took into account the text length between two successive cognates (Simard's case) or punctuations (Thomas's case), which increased the system confusion and lead to an increase in execution time and a decrease in accuracy. We have avoided this drawback by taking the probability of text length between successive punctuation marks into account during the punctuation matching process, as will be shown in the following sections.

In this paper, non-traditional approaches for English-Arabic sentence alignment are presented. For sentence alignment, we may have a 1-0 match, where one English sentence does not match any of the Arabic sentences, a 0-1 match where one Arabic sentence does not match any English sentences. The other matches we focus on are 1-1, 1-2, 2-1, 2-2, 1-3 and 3-1.

There may be more categories in bi-texts, but they are rare. Therefore, only the previously mentioned categories are considered. If the system finds any other categories they will automatically be misclassified. As illustrated above, we have eight sentence alignment categories. As such, sentence alignment can be considered as a classification problem, which may be solved by using a mathematical regression or genetic algorithm classifiers.

The paper is organized as follows. Section 2, introduces English–Arabic text features. Section 3, illustrates the new approaches. Section 4, discusses English-Arabic corpus creation. Section 5, shows the experimental results. Finally, section 6 gives concluding remarks and discusses future work.

2. English–Arabic text features

As explained in (Fattah et al., 2007), the most important feature for text is the text length, since Gale & Church achieved good results using this feature.

The second text feature is punctuation marks. We can classify punctuation matching into the following categories:

A. 1-1 matching type, where one English punctuation mark matches one Arabic punctuation mark.

B. 1-0 matching type, where one English punctuation mark does not match any of the Arabic punctuation marks.

C. 0-1 matching type, where one Arabic punctuation mark does not match any of the English punctuation marks.

The probability that a sequence of punctuation marks $AP_i = Ap_1 Ap_2 ......Ap_i$ in an Arabic language text translates to a sequence of punctuation marks $EP_j = Ep_1 Ep_2 ......Ep_j$ in an English language text is $P(AP_i, EP_j)$. The system searches for the punctuation alignment that maximizes the probability over all possible alignments given a pair of punctuation sequences corresponding to a pair of parallel sentences from the following formula:

$$\arg\max_{AL} P(AL \mid AP_i, EP_j),$$

(1)

Since "*AL*" is a punctuation alignment. Assume that the probabilities of the individually aligned punctuation pairs are independent. The following formula may be considered:

$$P(AP_i, EP_j) = \prod_{AL} P(Ap_k, Ep_k).P(\delta_k \mid match),$$

(2)

Where $P(Ap_k, Ep_k)$ = the probability of matching Ap_k with Ep_k, and it may be calculated as follows:

$$P(Ap_k, Ep_k) = \frac{Number\ of\ punctuation\ pair\ (Ap_k, Ep_k)}{Total\ number\ of\ punctuation\ pairs\ in\ the\ manually\ aligned\ data}$$

(3)

$P(\delta_k \mid match)$ is the length-related probability distribution function. δ_k is a function of the text length (text length between punctuation marks Ep_k and Ep_{k-1}) of the source language and the text length (text length between punctuation marks Ap_k and Ap_{k-1}) of the target language.

$P(\delta_k \mid match)$ is derived straight as in (Fattah et al., 2007).

After specifying the punctuation alignment that maximizes the probability over all possible alignments given a pair of punctuation sequences (using a dynamic programming framework as in (Gale and Church, 1993)), the system calculates the punctuation compatibility factor for the text pair under consideration as follows:

$$\gamma = \frac{c}{max\,(m, n)}$$

Where γ = the punctuation compatibility factor,

c = the number of direct punctuation matches,

n = the number of Arabic punctuation marks,

m = the number of English punctuation marks.

The punctuation compatibility factor is considered as the second text pair feature.

The third text pair feature is the cognate. For disparate language pairs, such as Arabic and English, that lack a shared alphabet, it is not possible to rely only on cognates to achieve high-precision sentence alignment of noisy parallel corpora.

However many UN and scientific Arabic documents contain some English words and expressions. These words may be used as cognates. We define the cognate factor (cog) as the number of common items in the sentence pair. When a sentence pair has no cognate words, the cognate factor is 0.

3. The Proposed sentence alignment model

The classification framework of the proposed sentence alignment model has two modes of operation. First, is training mode where features are extracted from 7653 manually aligned English-Arabic sentence pairs and used to train a Mathematical Regression (MR) and Genetic Algorithm (GA). Second, is testing mode where features are extracted from the testing data and are aligned using the previously trained models. Alignment is done using a block of 3 sentences for each language. After aligning a source language sentence and target language sentence, the next 3 sentences are then looked at. We have used 18 input units and 8 output units for MR and GA. Each input unit represents one input feature as in (Fattah et al., 2007). The input feature vector X is as follows:

$$X = [\ \frac{L(S1a)}{L(S1e)},\ \frac{L(S1a)+L(S2a)}{L(S1e)},\ \frac{L(S1a)}{L(S1e)+L(Se2)},\ \frac{L(S1a)+L(S2a)}{L(S1e)+L(S2e)},\ \frac{L(S1a)+L(S2a)+L(S3a)}{L(S1e)},$$
$$\frac{L(S1a)}{L(S1e)+L(S2e)+L(S3e)},\ \gamma(S1a,S1e),\ \gamma(S1a,S2a,S1e),\ \gamma(S1a,S1e,S2e),\ \gamma(S1a,S2a,S1e,S2e)$$
$$,\ \gamma(S1a,S2a,S3a,S1e),\ \gamma(S1a,S1e,S2e,S3e),\ Cog(S1a,S1e),\ Cog(S1a,S2a,S1e),\ Cog(S1a,S1e,S2e)$$
$$,Cog(S1a,S2a,S1e,S2e),\ Cog(S1a,S2a,S3a,S1e),\ Cog(S1a,S1e,S2e,S3e)\ ]$$

Where:
$L(X)$ is the length in characters of sentence X; $\gamma(X,Y)$ is the punctuation compatibility factor between sentences $X\ and\ Y$; $Cog(X,Y)$ is the cognate factor between sentences $X\ and\ Y$.

The output is from 8 categories specified as follows. $S1a \to 0$ means that the first Arabic sentence has no English match.
$S1e \to 0$ means that the first English sentence has no Arabic match. Similarly, the remaining outputs are $S1a \to S1e$, $S1a+S2a \to S1e$, $S1a \to S1e+S2e$, $S1a+S2a \to S1e+S2e$, $S1a \to S1e+S2e+S3e$ and $S1a+S2a+S3a \to S1e$.

3.1. Mathematical Regression Model (MR)

Mathematical regression is a good model to estimate the text feature weights (Jann, 2005; Richard, 2006). In this model a mathematical function relates output to input. The feature parameters of the 7653 manually aligned English-Arabic sentence pairs are used as independent input variables and the corresponding dependent outputs are specified in the training phase. A relation between inputs and outputs is tried to model the system. Then testing data are introduced to the system model for evaluation of its efficiency. In matrix notation, regression can be represented as follow:

$$\begin{bmatrix} Y0 \\ Y1 \\ : \\ : \\ Ym \end{bmatrix} = \begin{bmatrix} X01 & X02 & X03 & .... & X018 \\ : & : & : & : & : \\ : & : & : & : & :: \\ : & : & : & : & : \\ Xm1 & Xm2 & Xm3 & .... & Xm18 \end{bmatrix} . \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ : \\ : \\ w_{18} \end{bmatrix}$$

(4)

Where

[Y] is the output vector.

[X] is the input matrix (feature parameters).

[W] is the linear statistical model of the system (the weights $w_1, w_2, \ldots\ldots w_{18}$ in equation 4).
m is almost the total number of sentence pairs in the training corpus.

For testing, 1200 English–Arabic sentence pairs were used as the testing data. Equation 4 was applied after using the defined weights from [W].

3.2. Genetic Algorithm Model (GA)

The basic purpose of genetic algorithms (GAs) is optimization. Since optimization problems arise frequently, this makes GAs quite useful for a great variety of tasks. As in all optimization problems, we are faced with the problem of maximizing/minimizing an objective function f(x) over a given space X of arbitrary dimension (Russell, & Norvig, 1995; Yeh, Ke, Yang & Meng, 2005). Therefore, GA can be used to specify the weight of each text feature.

For block of 3 sentences, a weighted score function, as shown in the following equation is exploited to integrate all the 18 feature scores, where w_i indicates the weight of f_i.

$$Score = w_1.Score_{f_1} + w_2.Score_{f_2} + ... + w_{18}.Score_{f_{18}} \quad (5)$$

The genetic algorithm (GA) is exploited to obtain an appropriate set of feature weights using the 7653 manually aligned English-Arabic sentence pairs. A chromosome is represented as the combination of all feature weights in the form of $(w_1, w_2, ..., w_{18})$. 1000 genomes for each generation were produced. Thousand genomes for each generation were produced. Evaluate fitness of each genome (we define fitness as the average accuracy obtained with the genome when the alignment process is applied on the training corpus), and retain the fittest 10 genomes to mate for new ones in the next generation. In this experiment, 100 generations are evaluated to obtain steady combinations of feature weights. A suitable combination of feature weights is found by applying GA. For testing, a set of 1200 sentence pairs was used. Eq. (5) is applied after using the defined weights from GA execution.

4. English-Arabic corpus

Although, there are very popular Arabic-English resources among the statistical machine translation community that may be found in some projects such as the “DARPA TIDES program, http://www.ldc.upenn.edu/Projects/TIDES/”, we have decided to construct our Arabic-English parallel corpus from the Internet to have significant parallel data from different domains. The approach of (Fattah et al., 2006) is used to construct the English-Arabic corpus.

5. Experimental results

5.1. Length based approach

Let’s consider Gale & Church’s length based approach as explained in (Fattah et al., 2007). We constructed a dynamic programming framework to conduct experiments using their length based approach as a baseline experiment to compare with our proposed system. First of all it was not clear to us, which variable should be considered as the text length, character or word? To answer this question, we had to do some statistical measurements on 1000 manually aligned English – Arabic sentence pairs, randomly selected from the previously mentioned corpus. We considered the relationship between English paragraph length and Arabic paragraph length as a function of the number of words. The results showed that, there is a good correlation (0.987) between English paragraph length and Arabic paragraph length. Moreover, the ratio and corresponding standard deviation were 0.9826 and 0.2046 respectively. We also considered the relationship between English paragraph length and Arabic paragraph length as a function of the number of characters.

The results showed a better correlation (0.992) between English paragraph length and Arabic paragraph length. Moreover, the ratio and corresponding standard deviation were 1.12 and 0.1806 respectively. In comparison to the previous results, the number of characters as the text length variable is better than words since the correlation is higher and the standard deviation is lower.

We applied the length based approach (using text length as a function of the number characters) experiment on a 1200 sentence pair sample, not taken from the training data. Table 1 shows the results. The first column in table 1 represents the category, the second column is the total number of sentence pairs related to this category, the third column is the number of sentence pairs that were misclassified and the fourth column is the percentage of this error. Although 1-0, 0-1 and 2-2 are rare cases, we have taken them into consideration to reduce error. Moreover, we did not consider other cases like (3-3, 3-4, etc.) since they are very rare cases and considering more cases requires more computations and processing time. When the system finds these cases, it misclassifies them.

5.2. Mathematical Regression approach

The system extracted features from 7653 manually aligned English-Arabic sentence pairs and used them to train a Mathematical Regression model. 1200 English–Arabic sentence pairs were used as the testing data. These sentences were used as inputs to the Mathematical Regression after the feature extraction step. Alignment was done using a block of 3 sentences for each language. After aligning a source language sentence and target language sentence, the next 3 sentences were then looked at as follows:

1- Extract features from the first three English sentences and do the same with the first three Arabic sentences.
2- Construct the feature vector X .
3- Use this feature vector as an input of the Mathematical Regression model.
4- According to the model output, construct the second feature vector. For instance, if the result of the model is $S1a \rightarrow 0$, then read the fourth Arabic sentence and use it with the second and third Arabic sentences with the first three English sentences to generate the feature vector X .
5- Continue using this approach until there are no more English-Arabic text pairs.

Table 2 shows the results when we applied this approach on the 1200 English – Arabic sentence pairs. It is clear from table 2 that the results have been improved in terms of accuracy over the Length based approach. Additionally, we applied the MR approach on the entire English-Arabic corpus containing 191,623 English sentences and 183,542 Arabic sentences. Then we randomly selected 500 sentence pairs from the sentence aligned output file and manually checked them. The system reported a total error rate of 6.1%.

We decreased the number of sentence pairs used for training the MR model to 4000 sentence pairs to investigate the effect of the training data size on the total system performance. These 4000 sentence pairs were randomly selected from the training data. The constructed model was then used to align the entire English-Arabic corpus. Then, we randomly selected 500 sentence pairs from the sentence aligned output file and manually checked them. The system reported a total error rate of 6.2%. The reduction of the training data set does not significantly change total system performance.

5.3. Genetic Algorithm approach

The system extracted features from 7653 manually aligned English-Arabic sentence pairs and used them to construct a Genetic Algorithm model. 1200 English–Arabic sentence pairs were used as the testing data. Using formula (5) and the 5 steps mentioned in the previous section (using GA instead of MR) the 1200 English–Arabic sentence pairs were aligned. Table 3 shows the results when we applied this approach. Additionally, we applied the GA approach on the entire English-Arabic corpus. Then, we randomly selected 500 sentence pairs from the sentence aligned output and manually checked them. The system reported a total error rate of 5.6%.

We decreased the number of sentence pairs used for training the GA to 4000 sentence pairs as in the previous section to investigate the effect of the training data size on total system performance. These 4000 sentence pairs were randomly selected from the training data. The constructed model was then used to align the entire English-Arabic corpus. Then, we randomly selected 500 sentence pairs from the sentence aligned output and manually checked them. The system reported a total error rate of 5.8%. The reduction of the training data set from 7653 to 4000 does not significantly change the total system performance.

Table 1
The results using length based approach

Category	Frequency	Error	% Error
1-1	1099	54	4.9%
1-0, 0-1	2	2	100%
1-2, 2-1	88	14	15.9%
2-2	2	1	50%
1-3, 3-1	9	6	66%
Total	1200	77	6.4%

Table 2
The results using the Mathematical Regression approach

Category	Frequency	Error	% Error
1-1	1099	49	4.4%
1-0, 0-1	2	2	100%

1-2, 2-1	88	14	15.9%
2-2	2	1	50%
1-3, 3-1	9	6	66.6%
Total	1200	72	6 %

Table 3
The results using Genetic Algorithm model

Category	Frequency	Error	% Error
1-1	1099	45	4.1%
1-0, 0-1	2	2	100%
1-2, 2-1	88	13	14.7%
2-2	2	1	50%
1-3, 3-1	9	6	66.6%
Total	1200	67	5.6%

5.4. Discussion

The Length based approach did not give bad results, as shown in table 1 (the total error was 6.4%). This is explained by the fact that there is a good correlation (0.992) and low standard deviation (0.1806) between English and Arabic text lengths. These factors lead to good results as shown in table 1. The Mathematical Regression and Genetic Algorithm approaches decreased the total error compared to the length based approach.

Using feature extraction criteria allows researchers to use different feature values when trying to solve natural language processing problems in general.

6. Conclusions

In this paper, we investigated the use of Mathematical Regression and Genetic Algorithm models for sentence alignment. We have applied our new approaches on a sample of an English – Arabic parallel corpus. Our approach results outperform the length based. The proposed approaches have improved total system performance in terms of effectiveness (accuracy). Our approaches have used the feature extraction criteria, which gives researchers an opportunity to use many varieties of these features based on the language pairs used and their text types (Hanzi characters in Japanese-Chinese texts may be used for instance).

References

Brown, P., Lai, J., Mercer, R., 1991. Aligning sentences in parallel corpora: In Proceedings of the 29th annual meeting of the association for computational linguistics, Berkeley, CA, USA.

Cain, B. J., 1990. Improved probabilistic neural networks and its performance relative to the other models: Proc. SPIE, Applications of Artificial Neural Networks, 1294, 354–365.

Chen, A., Gey, F., 2001. Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval: TREC 2001.

Chen, K.H., Chen, H.H., 1994. A Part-of-Speech-Based Alignment Algorithm: Proceedings of 15th International Conference on Computational Linguistics, Kyoto, 166-171.

Chen, S. F., 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information: Proceedings of ACL-93, Columbus OH, 9-16.

Christopher, C., Kar, Li., 2004. Building parallel corpora by automatic title alignment using length-based and text-based approaches: Information Processing and Management 40, 939–955.

Collier, N., Ono, K., Hirakawa, H., 1998. An Experiment in Hybrid Dictionary and Statistical Sentence Alignment: COLING-ACL, 268-274.

Danielsson, P., Mühlenbock, K., 2000. The Misconception of High-Frequency Words in Scandinavian Translation: AMTA, 158-168.

Davis, M., Ren, F., 1998. Automatic Japanese-Chinese Parallel Text Alignment: Proceedings of International Conference on Chinese Information Processing, 452-457.

Dejean, H., Gaussier, É., Sadat, F., 2002. Bilingual Terminology Extraction: An Approach based on a Multilingual thesaurus Applicable to Comparable Corpora: Proceedings of the 19th International Conference on Computational Linguistics COLING 2002, Taipei, Taiwan, 218-224.

Dolan, W. B., Pinkham, J., Richardson, S. D., 2002. MSR-MT, The Microsoft Research Machine Translation System: AMTA, 237-239.
Fattah, M., Ren, F., Kuroiwa, S., 2006. Stemming to Improve Translation Lexicon Creation form Bitexts: Information Processing & Management, 42 (4), 1003 – 1016.
Gale, W. A., Church, K. W., 1993. A program for aligning sentences in bilingual corpora: Computational Linguistics, 19, 75-102.
Ganchev, T., Tasoulis, D. K., Vrahatis, M. N., Fakotakis, N., 2003. Locally recurrent probabilistic neural networks for text independent speaker verification: Proc. of the EuroSpeech, 3, 1673–1676.
Gey, F.C., Chen, A., Buckland, M.K., Larson, R. R., 2002. Translingual vocabulary mappings for multilingual information access: SIGIR, 455-456.
Ker, S.J., Chang, J.S., 1997. A class-based approach to word alignment: Computational Linguistics, 23(2), 313-344.
Kraaij, W., Nie, J. Y., Simard, M., 2003. Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval: Computational Linguistics, 29(3), 381-419.
Melamed, I. D., 1997. A portable algorithm for mapping bitext correspondence: In The 35th Conference of the Association for Computational Linguistics (ACL 1997), Madrid, Spain.
Melamed, I. D., 1999. Bitext Maps and Alignment via Pattern Recognition: Computational Linguistics, March, 25(1), 107-130.
Moore, R.C., 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora: AMTA, 135-144.
Oard, D. W., 1997. Alternative approaches for cross-language text retrieval: In D. Hull, & D. Oard (Eds.), AAAI symposium in cross-language text and speech retrieval. American Association for Artificial Intelligence, March.
Pellom, B.L., Hansen, J.H.L., 1998. An Efficient Scoring Algorithm for Gaussian Mixture Model based Speaker Identification: IEEE Signal Processing Letters, 5(11), 281-284.
Resnik, P., Smith, N. A., 2003. The Web as a Parallel Corpus: Computational Linguistics, 29(3), 349-380.
Reynolds, D., 1995. Speaker identification and verification using Gaussian mixture speaker models: Speech Communication, 17, 91–108.
Ribeiro, A., Dias, G., Lopes, G., Mexia, J., 2001. Cognates Alignment: In Bente Maegaard (ed.), Proceedings of the Machine Translation Summit VIII (MT Summit VIII) – Machine Translation in the Information Age, Santiago de Compostela, Spain, 287–292.
Simard, M., 1999. Text-translation alignment: three languages are better than two: In Proceedings of EMNLP/VLC- 99, College Park, MD.
Simard, M., Foster, G., Isabelle, P., 1992. Using cognates to align sentences in bilingual corpora: Proceedings of TMI92, Montreal, Canada, 67-81.
Specht, D.F., 1990. Probabilistic neural networks: Neural Networks, 3(1), 109-118.
Thomas, C., Kevin, C., 2005. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria: Computational Linguistics and Chinese Language Processing , 10 (1), 95-122.
Fattah, M., Ren, F., Kuroiwa, S., 2007. Sentence Alignment using P-NNT and GMM: Computer Speech and Language, 21 (4), 594 – 608.