Statistical Model of Length Based Methods in Developing English-Myanmar Parallel Corpus

Kyi Kyi Han, Hla Hla Htay
University of Computer Studies, Yangon
kyikyihan10@gmail.com, hlahlahtay123@gmail.com

Abstract

The paper is about developing an aligned English-Myanmar parallel corpus. This paper will describe method for aligning sentences based on simple statistical model of length. Parallel corpus is pairs of translation alignment of the bitext. We have used Gale & Church method to get translation alignments. Parallel corpus helps in making statistical bilingual dictionary, in supporting statistical machine translation and in supporting in training data for word sense disambiguation. The program is implemented in Java language.

Keywords : character, parallel corpus, translation, word

1. Introduction

Corpus is the body of text collections which are useful for Natural Language Processing (NLP). Parallel corpus helps in making statistical bilingual dictionary, in supporting statistical machine translation and in supporting as training data for word sense disambiguation and translation disambiguation. One useful step is to identify correspondences between sentences in one language and in other language. This paper will describe a method and a program for aligning sentences based on a simple statistical model of character lengths.

2. Related Work

Brown et al. [7] performed bitext alignment with a statistical technique to build parallel corpus. They designed an algorithm based on sentence length in terms of the number of words contained in sentence. No other lexical details of the sentence are utilized and they have achieved accuracy in excess of 99% in a randomly selected set of 1000 sentence pairs.

Bing Zhao et.al [2] presented bilingual sentence alignment technique which used five candidate scores based on perplexity and sentence length. Then a linear regression model based on those candidates is proposed and trained to predict sentence pairs' alignment quality scores solicited from human subjects.

Alexandru Ceausu et.al [1] described a hybrid sentence aligner that has four stages: (I) length and geometric based sentence alignment; (II) estimation of the translation model; (III) length, geometric and word translation based sentence alignment; (IV) recovery of the non 1-1 sentence alignments. Their alignment tool used a Support Vector Machine classifier and does not have language specific information and its parameters are trained using just a small portion of human checked alignment data.

Gale and Church's approach [4] has been widely adopted for the alignment of European languages and has subsequently been improved with complementary techniques. Their method is based on a simple statistical model of sentence length measured in terms of characters. This method uses the fact that longer sentences tend to be translated into longer sentences in the target language and shorter sentences tend to be translated into shorter sentences.

Fung and Yee [6] describe an approach for finding translations from non-parallel yet comparable texts. They make use of word frequency to find the translation pair. This approach makes use of the observation that words that appear in the context of words that are translations of each other are similar. It makes use of an already existing bilingual lexicon to find the meaning of these context words.

Sentence aligning program based on Gale and Church algorithm has been very successful in aligning some European language pairs. Nowadays, English language is very important and very useful around the world. English and Myanmar are not similar language but this alignment is based on statistical model of character length. Dictionaries are also updated every five years long.

Sentence alignment used models that just compared the lengths of units of text in the parallel corpora. While it seems strange to ignore the richer information available in the text, it turns out that such an approach can be quite effective, and its efficiency allows rapid alignment of large quantities of text. The rationale of based methods is that short sentences will be translated as short sentences and long sentences as long sentences. Length usually is defined as the number of characters.

3. Length Based Statistical Alignment Model

As was shown in the sentence alignment survey in Section 2, the sentence length ratio is also a very good indication of the alignment of a sentence pair for languages from a similar family such as French and English.

The method of Gale and Church depends simply on the length of source and translation sentences measured in characters. Gale and Church algorithm uses sentence length to evaluate how likely an alignment of some number of sentences in is with some number of sentences in possible alignments in the study were limited to {1:1, 1:0, 0:1, 2:1, 1:2, 2:2}. This made it possible to easily find the most probable text alignment by using a dynamic programming algorithm, which tries to find the minimum possible distance between the two texts, or in other words, the best possible alignment.

First of all, most of the time sentences do not align one-to-one. Sometimes a sentence may be translated in 2-3 sentences in the other language or some part of a text may be deleted or some additional sentences may be added to the text which has no matches in the corresponding text.

Gale and Church claim that their algorithm is language independent. Gale et al. algorithm is based on a simple and straightforward presumption: the lengths of sentence translations are proportional. In other words, longer sentences are translated into longer sentences and shorter sentences are translated into shorter sentences. For a given pair of candidate sentence alignments, it is needed to compute 2 steps for finding-of candidate sentence

1. Local distance
2. Two-side-distance

Then sentence alignments pairs are chosen dynamically. See section 3.3 and figure 1 and 2.

3.1. Find local distance

a) The sentence lengths of candidate sentence alignments,
b) The a-priori likelihood of the various possible cases of alignment (Prob (match)),
c) The mean ratio of sentence lengths (c) and variance of the character numbers in the target language per character in the source language (s^2).
The mean ratio c and the variance s^2 were also determined empirically by Gale and Church from parallel corpus data. Gale et al. suggested that $c \approx 1$ and $s^2 \approx 6.8$ can be used as language independent parameters for most European languages.

With all these parameters determined, the distance between the given candidates can be measured using the formulae in Figure1.

$$\delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}}$$

$$Prob\ (\delta) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\delta} e^{\frac{-z^2}{2}}\, dz$$

$$Prob\ (\delta\ |match) = 2(1 - Prob(|\delta|))$$

$$Prob(match|\delta) = Prob(\delta|match)Prob(match)$$

$$Distance\ (l_1, l_2) = logProb(match|\delta)$$

Figure 1. The formulae for computing distance

3.2. Find two-side-distance

The distance function, two-side-distance d, is defined in a general way to allow for insertions, deletion, substitution, expansion, contraction and merger. The function takes four arguments: x_1, x_2, y_1, y_2.Let

1. $d(x_1, y_1; 0,0)$be the cost of substituting x_1 with y_1,

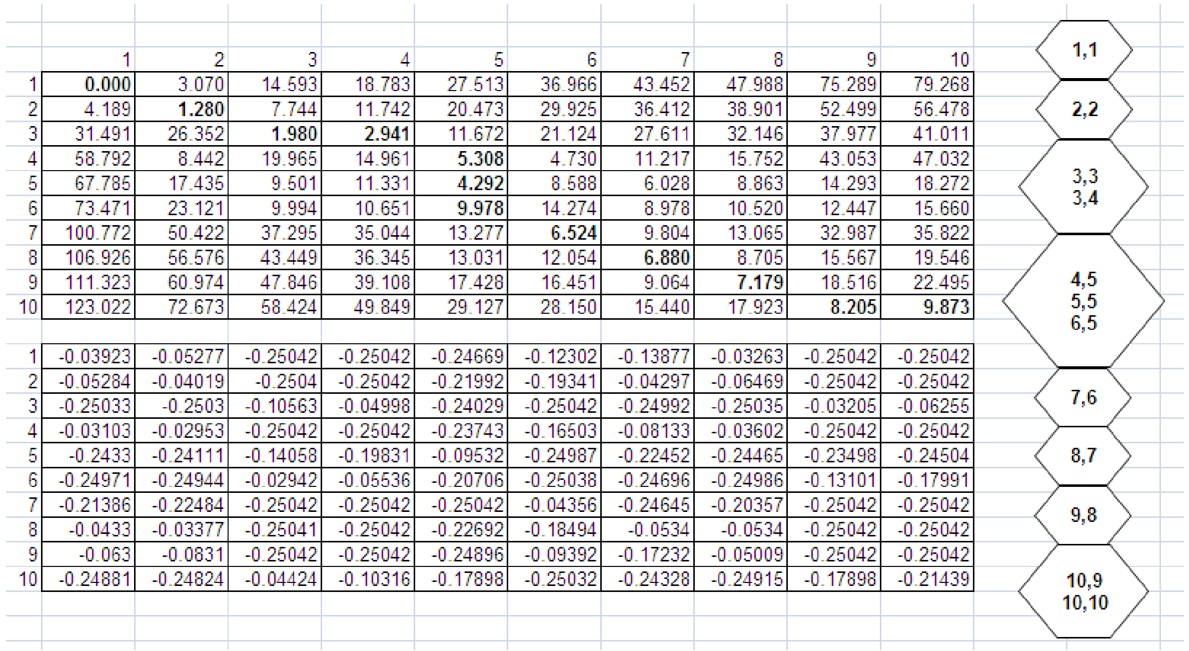
2. $d(x_1, 0; 0,0)$be the cost of deleting x_1 with y_1,
3. $d(0, y_1; 0,0)$be the cost of inserttion of y_1,
4. $d(x_1, y_1; x_2, 0)$be the cost of contracting x_1 and x_2 to y_1 and x_2 to y_1,
5. $d(x_1, y_1; 0, y_2)$be the cost of expanding x_1 to y_1 and y_2,
6. $d(x_1, y_1; x_2, y_2)$ be the cost of merging x_1 and x_2 and matching with y_1 and y_2 .

$$D(i,j) = min \begin{cases} D(i,j\text{-}1) + d(0,t_j{:}0,0) \\ D(i\text{-}1,j) + d(s_i,0{:}0,0) \\ D(i\text{-}1,j\text{-}1) + d(s_i,t_j{:}0,0) \\ D(i\text{-}1,j\text{-}2) + d(s_i,t_j{:}0,t_{j\text{-}1}) \\ D(i,j\text{-}1) + d(s_i,t_j{:}s_{i\text{-}1},0) \\ D(i,j\text{-}1) + d(s_i,t_j{:}s_{i\text{-}1},t_{j\text{-}1}) \end{cases}$$

Figure 2. Dynamic Programming Algorithm

3.2 Dynamic Programming Algorithm

The algorithm is summarized in the following recursive equation. Let s_i, i=1...I, be the sentence.



	1	2	3	4	5	6	7	8	9	10
1	**0.000**	3.070	14.593	18.783	27.513	36.966	43.452	47.988	75.289	79.268
2	4.189	**1.280**	7.744	11.742	20.473	29.925	36.412	38.901	52.499	56.478
3	31.491	26.352	**1.980**	**2.941**	11.672	21.124	27.611	32.146	37.977	41.011
4	58.792	8.442	19.965	14.961	**5.308**	4.730	11.217	15.752	43.053	47.032
5	67.785	17.435	9.501	11.331	**4.292**	8.588	6.028	8.863	14.293	18.272
6	73.471	23.121	9.994	10.651	**9.978**	14.274	8.978	10.520	12.447	15.660
7	100.772	50.422	37.295	35.044	13.277	**6.524**	9.804	13.065	32.987	35.822
8	106.926	56.576	43.449	36.345	13.031	12.054	**6.880**	8.705	15.567	19.546
9	111.323	60.974	47.846	39.108	17.428	16.451	9.064	**7.179**	18.516	22.495
10	123.022	72.673	58.424	49.849	29.127	28.150	15.440	17.923	**8.205**	**9.873**

1	-0.03923	-0.05277	-0.25042	-0.25042	-0.24669	-0.12302	-0.13877	-0.03263	-0.25042	-0.25042
2	-0.05284	-0.04019	-0.2504	-0.25042	-0.21992	-0.19341	-0.04297	-0.06469	-0.25042	-0.25042
3	-0.25033	-0.2503	-0.10563	-0.04998	-0.24029	-0.25042	-0.24992	-0.25035	-0.03205	-0.06255
4	-0.03103	-0.02953	-0.25042	-0.25042	-0.23743	-0.16503	-0.08133	-0.03602	-0.25042	-0.25042
5	-0.2433	-0.24111	-0.14058	-0.19831	-0.09532	-0.24987	-0.22452	-0.24465	-0.23498	-0.24504
6	-0.24971	-0.24944	-0.02942	-0.05536	-0.20706	-0.25038	-0.24696	-0.24986	-0.13101	-0.17991
7	-0.21386	-0.22484	-0.25042	-0.25042	-0.25042	-0.04356	-0.24645	-0.20357	-0.25042	-0.25042
8	-0.0433	-0.03377	-0.25041	-0.25042	-0.22692	-0.18494	-0.0534	-0.0534	-0.25042	-0.25042
9	-0.063	-0.0831	-0.25042	-0.25042	-0.24896	-0.09392	-0.17232	-0.05009	-0.25042	-0.25042
10	-0.24881	-0.24824	-0.04424	-0.10316	-0.17898	-0.25032	-0.24328	-0.24915	-0.17898	-0.21439

Figure 3. Calculated Cost and Distance

The algorithm is summarized in the following recursion equation. Let $_{si,}$ i = 1 ... I, be the sentences of one language, and t_j, j = 1.. J, be the translations of those sentences in the other language. Let d be the distance function described in the previous section, and let D(i,j) be the minimum distance between sentences $s_1,...s_i$ and their translations $t_1,...t_j$, under the maximum likelihood alignment. D(i,j) is computed by minimizing over six cases (substitution, deletion, insertion, contraction, expansion, and merger) which, in effect, impose a set of slope constraints. That is, D(i,j) is defined by the following recurrence with the initial condition D(i,j) = 0. The example for cost and distance which is dynamically chosen can be seen in figure 3.

4. Aligning English – Myanmar Text

The purpose of sentence alignment is to identify correspondences between sentences in one language and sentences in another language. Note that alignment may not be one to one, although one to one alignments are more frequent.

4.2. Preprocessing

Before the aligning, it is needed to segment the bitext in to sentence.

Plain English and Myanmar texts containing one sentence per line are input to the alignment program. The output will be two separate aligned files with line to line correspondence.

Firstly, Myanmar text we use are from online newspaper. All of the Myanmar text fonts are Win--Inwa. Therefore, we use the fonts convertor Win-2-Zawgyi to convert to Zawgyi fonts. And then, we saved that English and Myanmar text files with the same file name with different extension (.eng and .mm). If you load the English text file, the Myanmar text file is also automatically loaded because they have the same name.

Second step is the sentence segmentation. In *Myanmar* script, we have “။”as unique sentence boundary marker. Therefore segmenting paragraphs into sentences is trivial. In case of *English language*, however, detecting sentence boundary is not entirely trivial. Even though there are explicit sentence boundary markers such as the period (.), the question mark (?) and the exclamation mark (!), the same symbols can be used for other purposes. We use the Open NLP Tools for English sentences segmentation. You can download from the internet: (http://opennlp.sourceforge.net/models/english/sentdetect/).

4.3. Experimental Results

We have experimented in two ways:

- with language independent setting where c=1 and s^2=6.8
- with variation on c and s^2 where one of which is fixed and the other is increasing or decreasing 1.

for sentence length in terms of the number of *words* or *character* contained in sentence. The reported result can be seen section 4.3.1 and 4.3.2. Among all the variations, we have shown only the best one from both character and word based. We have taken a sample of 10 files for evaluation. It is found that character based is best with language independent setting and word based is best with c=1 and s2=6.8. See more detail in Table 1 and Table 2. It is also notice that changing the variance (s^2) does not affect the accuracies but mean (c)

Table 1. Character-based with c=1, s2= 6.8

File Name	System Reported	Correctly Aligned	(%) of accuracy
16050911	8	6	75
06050922	8	5	62.5
06120833	24	17	70.83
15120844	23	16	69.57
19120855	14	9	64.29
11120866	7	5	71.43
21120877	8	6	75
23120888	9	7	77.78
27120899	7	7	100
05040900	7	5	71.43
Total	**115**	**83**	**72.17**

Table 2. Word-based with c=1, s2= 6.8

File Name	System Reported	Correctly Aligned	(%) of accuracy
16050911	8	6	75
06050922	9	5	55.56
06120833	27	17	70.83
15120844	23	16	69.57
19120855	14	9	64.29
11120866	5	2	40
21120877	9	9	100
23120888	9	7	77.78
27120899	8	8	100
05040900	6	4	66.67
Total	**123**	**97**	**78.86**

5. Conclusion

In this paper is constructing an aligned English-Myanmar parallel corpus. This paper will describe method for aligning sentences based on simple statistical model lengths. Parallel corpus is pairs of translation alignment of the bitext. We have used Gale & Church method to get translation alignments. Parallel corpus is a scarce resource in making statistical bilingual dictionary, in supporting statistical machine translation and in supporting in training data for word sense disambiguation. We utilized Gale & Church method in terms of length of character and word contained in the sentence. It is found that changing the variance (s^2) does not affect the accuracies but mean (c). When the mean is changed, it is the best with c=1 with the character length based and it is the best with c=1 with the word based. In the future, we will use the available bilingual dictionary to improve the alignment accuracies. The program is implemented in Java language with user friendly interface.

References

[1] Alexandru Ceausu, Dan Stefanescu, Dan Tufi, "Acquis Communautaire sentence alignment using Support Vector Machines".

[2] Bing Zhao et.al, "Efficient Optimization for Bilingual Sentence Alignment Based on Linear Regression", HLT-NAACL 2003 Workshop: Building and Using Parallel Text,Data Driven Machine Translation and Beyond , pp. 115-118, Edmonton, May-June 2003

[3] Dan Melamed, "Empirical Methods for Exploiting Parallel Texts", MIT press, Cambridge, New York City, 2001.

[4] Gale, W.A. and K.W. Church, "A program for aligning sentences in bilingual corpora". In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkley, pp. 177–184, 1991.

[5] J. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries", In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington D.C., pp. 16-19, 1997.

[6] P. Fung and L. Yee. "An IR approach for translating new words from nonparallel, comparable texts", In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, pp 414–420, 1998.

[7] Peter F. Brown, Jennifer C. Lai, Robert L. Mercer, "Aligning sentences in parallel corpora, Association for Computational Linguistics", Morristown, NJ, USA, California, Pp: 169 - 176, 1991.

[8] Philip Resnik, "Mining the Web for Bilingual Text", In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), University of Maryland, College Park, Maryland, June 1999.

[9] Piao, Scott Songlin, "Sentence and Word Alignment Between Chinese and English", Ph.D thesis, 2000.