Quality Analysis of Patent Parallel Corpus by the Scale

Isamu OKADA
Faculty of Business Administration,
Soka University
1-236, Tangi, Hachioji City, Tokyo,
192-8577, JAPAN
okada@soka.ac.jp

Shinichiro MIYAZAWA (Shumei U.),
Kazunari ISHIDA
(Tokyo U. of Agriculture),
Nobuhiko SHIMIZU (Shumei U.),
and **Toshizumi OHTA**
(U. of Electro-Communications)

Abstract

Large-scale parallel corpus is extremely important for translation memory, example-based machine translation, and the support system to create English sentences. Organized collection or establishment of large-scale corpus is currently ongoing; however it is a difficult project in terms of copyrights as well as economic efficiency. To investigate general tendency of large-scale corpus helps to improve economical efficiency of parallel corpus collection as well as system establishment. In this study, therefore, the relationship between the scale of parallel corpus and the degree of correspondence is clarified, using parallel corpus for patents.

1. Introduction

Large-scale parallel corpus is extremely important for translation memory, example-based machine translation, and the support system to create English sentences. It is generally assumed that the more the parallel corpus is, the hit ratio will improve, although noise (example sentences that cannot practically be referred) increase is possible.

Since collection of massive parallel corpus is also an issue for copyrights, it is a fairly considerable task. It is considered that accuracy of parallel corpus generally describes an S-shaped curve for a hit ratio, depending on the scale of corpus. This is because a certain scale is considered to be necessary since the degree of correspondence does not go up when the scale is small.

Furthermore, since the hit ratio has an upper limit, it won't go up once it hits the limit, even if the scale is large to some extent. To assume general tendency of large-scale corpus helps to improve economical efficiency of parallel corpus collection as well as system establishment.

Although attempts to evaluate large-scale corpus have hardly been made regardless of its importance, Fujii, etc. evaluated large-scale corpus by an in-person questionnaire (Fujii and Ishikawa. 2001). On the other hand, we analyzed the quality of corpus with a quantitative approach. In this study, the relationship between the scale of parallel corpus and the degree of correspondence is statistically clarified, using parallel corpus by Japio (Japan Patent Information Organization).

2. Data: Patent Corpus and Patent Dictionary

To investigate the relationship between the scale of parallel corpus and the degree of correnspondence, we use raw data about patent corpus. The patent corpus has become increasingly important in terms of searching (M. Iwayama, A. Fujii, A. Takano and N. Kando. 2001). The co-author is a member of a joint study group between AAMT (Asia-Pacific Association for Machine Translation) and Japio named the AAMT/Japio Special Interest Group on Patent Translation, which began the attempt to utilize patent corpus into machine translation. Under these circumstances, there is an urgent need to clarify the nature of patent corpus.

Japio provided us the one-year corpus for the 2003 issue of the unexamined patent publication bulletin (348,061 cases). In addition, dictionaries for specific areas were also provided as sample data from Japio's dictionary (patent dictionary). The sample data in the patent dictionary totals 12,695 words, and the following include areas and respective word numbers. IPC represents International Patent Classification.

Category J: C11 (gene) - 4,789 words (corresponding IPC: C12N)
Category J: P03 (measurement, etc.) - 3,124 words (corresponding IPC: G05B, G05D, G05F, G05G, G05X, G06C, G06D, G06E, G06G, G06J, G06K, G06M, G06N, G06T, G06X, G07B, G07C, G07D, G07F, G07G, G07X, G08B, G08C, G08G)
Category J: P05 (electric digital data processing) - 4,782 words (corresponding IPC: G06F)

3. Basic Policy and Definition

Quality classified by scale for parallel corpus is analyzed with the following policy using patent sentences provided in both Japanese and English pairs. First, the sentences are divided three parts; corpus sentences, test sentences, and the remainder. A corpus translation system is constructed by the corpus pairs. Next, Japanese sentences of the test part are translated by the corpus system to translated English sentences. The English sentences of the test part are regarded as correct sentences. Finally, translated sentences and correct sentences are compared and calculated the degree of correspondence. Therefore, we can evaluate efficiency of the corpus system.

To formalize the policy strictly, we can define several terms. Let S=(J,E) be the pairs (# is 348,061) of Japanese and English patent sentences of provided by the Japio. $J = \{J_1,J_2,\ldots,J_n\}$ be a set of Japanese sentences consists of a Japanese sentence J_i , and $E=\{E_1,E_2,\ldots,E_n\}$ be a set of English ones is in the same way. A pair $S_i=(J_i,E_i)$, both sentences have a same suffix, is Japanese and English sentences which are translated each other.

Second, we divide S into a corpus part and a test part. For a suffix set, $N= \{1,2,\ldots,n\}$, based on a selection rule, R, N is divided into C and T, where $C \cup T \subseteq N$, $C \cap T = \Phi$. Let $S_C=\{(J_i,E_i)|i \in C\}$ be a set of corpus sentences and $S_T=\{(J_i,E_i)|i \in T\}$ be a set of test sentences.

Third, we construct a corpus system with S_C and let CP(・) be a translated sentence with the system. For J_t, all Japanese sentences of the test part S_T, $E'_t=CP(J_t)$, translated English sentences are made by the system. Let E_t be a correct English sentence of J_t. We can calculate a corresponding degree of E'_t and E_t using a corresponding function H(・) , that is $H(E_t,E'_t)$.

Finally, For all test sentences J_T, a set of corresponding degrees, $H(E_T,E'_T)$, is calculated. Let Hs(R) be an efficiency for a selection rule R, which is statistic quantity integrated $H(E_T,E'_T)$.

4. Experiment Policy

In this paper, an efficiency of the corpus system is defined as the degree of correspondence of terms which are in the patent dictionary (Japio's dictionary) for simplicity. This is because we have to focus on the technical terms in order to characterize the patent corpus. That is why we define a corresponding function as follows;

$$H(A,B) = |K(A) \cap K(B)| / |K(A) \cup K(B)|$$

where let K(A) be a set of terms which are consists of sentence A and in the patent dictionary.

A efficiency, Hs(R), which is a measurement, is defined as a couple of two values; one is an average of max corresponding degree of each test sentence, $Max(H_t)$, and the another is an average of the number of corresponding sentences of each test sentence, $Num(H_t,x)$. Let

$$Hs(R)=(Ave(Max(H)), Ave(Num(H,x)))$$

where $Max(H_t)=Max\{H(E_t,E'_t)\}$

and $Num(H_t,x) = |\{H|H(E_t,E'_t) > x\}|$.

Based on the basic polity, we structure the following experiment algorithm.

1) the bilingual part (summary part) is extracted from the provided corpus (one-year patent publication bulletin). Specifically, only the title part in a document, [Title of Invention] (the tag part of <B542> in English), and [Summary] (the tag part of <SDOAB> in English) remain. The summary consists of [Problem to Be Solved] and [Solution].

2) Since each document consists of multiple sentences, it is separated by each sentence. This increases the amount of data to about five times more; approximately 1,600 thousand sentences.

3) The following processing is made for each sentence:

(1) Only nouns are extracted with morphological analysis.

(2) Nouns not listed in the dictionary (Japio's patent dictonary) are deleted as noise. Therefore a sentence which has no terms in the dictionary is omitted from targets of this experiment. For example, the number of English sentences in "G08C" covered this time is 834. However, if sentences with zero words are excluded, it will be 652 sentences.

4) A selection rule $R=(R_T,R_C)$ which divide into a corpus part and a test part is formulated.

(1) 500 sentences of the test sentences are selected in the order of data from up-to-date data (R_T=LAT).

(2) 500 sentences of the test sentences are selected at random (R_T=RND).

The following algorithm processes after the test part is selected.

(3) p sentences of the corpus sentences are selected at random (R_C=(RND,p)).

(4) p sentences of the corpus sentences are selected in the order of data from up-to-date data (R_C=(LAT,p)).

(5) p sentences of the corpus sentences are selected with constant time density (R_C=(WIN,p)). Constant time density is

defined that the numbers of sentences extracted from a same preparation time are constant. There are 119 published times in the published bulletins in our data.

(6) Let p=a*2^n (a=100、n=0～6) and p be the number of all sentences. That is, p are 8 ways, p= 100、200、400、800、1,600、3,200、6,400 , and all.

The selection rule is determined above. For example, R=(LAT,(RND,100)) means a rule of (1) as test part select rule and (3) as corpus part select rule with p=100. Notice when p be all, sets of the corpus sentences are correspond to each other because of C ∪ T=N, namely, R(・,(RND,p)) = R(・,(LAT,p)) = R(・,(WIN,p)).

The total number of selection rules are 44(=2*(3*7+1)). We process the following algorithm every rule.

5) Every test sentence can be processed as follows:

(1) A Japanese sentence of the test sentence is broken down with morphological analysis and extracts norms which are in the patent dictionary to translate corresponding English words. Let a hit sentence be what shares the same technical term of the English words among the corpus sentences

(2) A corresponding degree between the test sentence and the hit sentence is calculated.

(3) Using the degree, Maximizing corresponding degree, $Max(H_t)$, and the number of corresponding sentences, $Num(H_t,x)$ are calculated in a test sentence. For simplicity, we assume x=0, namely, the number of corresponding sentences is the number of hit sentences.

6) An efficiency of a selection rule, Hs(R), is calculated based on $Max(H_t)$ and $Num(H_t,x)$ for all test sentences. Besides, to analyze visually, we show that curves with the degree of corresponding at the vertical axis and the number of cumulative matched combinations lined up in the order of the degree of corresponding at the horizontal axis are graphed for selection rules.

5. Experiments

If all the provided data is processed with our computer equipments, it would take approximately two and a half months to calculate all data, e.g., analysis of all data is impossible in terms of time. Therefore, we narrowed down to small areas including areas of "measurement, etc" and "gene."

According to the area of "measurement, etc." named as “G08C” in the IPC, the total number of subject sentences is 1152. The 500 sentences of them are needed as test sentences, i.e., the sizes of corpus sentences are p= 100, 200, 400 and 642.

On the other hand, in the area of "gene" named as “C12N” in the IPC, the total number of subject words is 176,946, and that of subject sentences is 5759. The 500 sentences of them are needed as test sentences, i.e., the sizes of corpus sentences are p= 100, 200, 400, 800, 1600, 3200 and 5259, namely, p is 7 ways of corpus size.

6. Results and Analysis

Our experiments of the two areas have little difference. Therefore, we show the results of the area of “gene” only.

6.1 Comparison of "C12N" by Selection Rule

Figure 1 and 2 show the efficiencies of the corpus selection rules. We can analyze the following points with these figures.

(1) There is no difference among LAT, RND and WIN as rules to select corpus sentences.

(2) The linear relationship of the average maximum corresponding degree can be observed by log at the horizontal axis and normal at the vertical axis. This investment suggest an insight that it is possible to forecast economical efficiency. Since it is possible to forecast the relationship between the scale of corpus and the corresponding degree, the quality of corpus (correspondence) can be forecasted from the cost (the number of corpus collected).

(3) Since the average in the maximum degree of correspondence has linearity in regards to the scale and degree of similarity, a regression line can be obtained. When a regression formula is calculated for a policy "LL" (a method to select test sentences is LAT and a method to select a corpus sentence group is LAT), a regression formula with 99.6% determination coefficient can be calculated as follows:

[Degree of Correspondence] = 5.214 x $\log_{10}$[Number of Data] - 1.396 (R^2=99.6%)

Using this regression formula, it is theoretically possible to obtain the scale of corpus necessary to find any degree of correspondence. For example, we can understand that approximately 4,000 thousand sentences in the number of data are necessary to make the corresponding degree 33%. On the contrary, it is possible to estimate a degree of corresponding when the scale of corpus is known. For example, when corpus is developed with approximately 1,600 thousand sentences in the number of data

(equivalent to patent information for one year), it is possible to estimate that the degree of correspondence would be approximately 31%.

(4) WIN, LAT and RND are also similar in terms of the trend in the number of corresponding sentences. Since linear relationship is observed with log-log at both axes and it is possible to determine the extent of degree of similarity that can be obtained with the large scale in the number of data, it seems possible to estimate economical efficiency.

6.2 Comparison by Scale for "C12N"

Figure 3 shows a cumulative graph for "C12N."

(1) As a basic trend, the distribution closely relates to the damping function.

(2) Comparison was also made by scale (7 types), however there is almost no difference. This is due to the small difference in the scale, therefore it is hard to mention without making the difference of scale large.

(3) In the case of comparison by selection method (6 types) with the same scale, there was almost no difference in the pattern.

7. Conclusion

We consider that one method to analyze the quality of massive corpus has been established in this study. Issues to be solved in the future include the following:

1) Resolution of differences due to the sentence length

There is a tendency that a shorter sentence has a higher degree of similarity. A device is necessary for improvement, such as change to an "absolute number of matched words" only upon calculation of a degree of correspondence.

2) Resolution of explosion of combined computational complexity

To avoid explosion of combined computational complexity, we only handled G08C and C12N. The number of words for G08C is 39,931, and the "possible sentences to be covered" total 652. On the other hand, the number of "words" for C12N is 176,946, and the "possible sentences to be covered" total 5,259. Although assumption in the case of larger scale is possible with regression analysis to some extent, the ability to directly handle massive patent information is necessary for accurate analysis, and for this purpose it is necessary to develop a method to avoid explosion of computation. Although this is possible by algorithm improvement to some extent, parallel processing is necessary in principle, such as utilization of cluster computers.

3) Resolution of the small amount in the dictionary

(1) The patent dictionary provided this time has 4,789 words for C11 (gene), 3,124 words for P03 (measurement, etc.), and 4,782 words for P05 (electric digital data processing), totaling 12,695 words, which we feel small for analysis. More patent dictionaries are required to improve data reliability as well as accuracy of analysis.

(2) Although we currently limit to patent dictionaries, it is necessary to include not only technical dictionaries but also general dictionaries to clarify general quality of massive corpus. These seem to be effective data upon utilization of corpus and dictionaries for machine translation of patents.

4) Relationship between Improvement of Hit Ratio and Noise Increase

It is necessary to study the relationship between improvement of hit ratio and noise increase as well.

5) Universal Quality of Large-Scale Corpus

We need to analyze corpus for other than patent to clarify its general trend.

We hope to solve the above issues and work on improvement of algorithms as well as diversification of analysis methods.

8. Acknowledgements

We express our gratitude to the member of AAMT/Japio Special Interest Group on Patent Translation for providing patent data and supporting this study.

References

M. Iwayama, A. Fujii, A. Takano and N. Kando. 2001. Patent Retrieval Challenge in NTCIR-3. *IPSJ SIG Technical Reports*, FI-63: pp.49-56 (in Japanese).

A. Fujii and T. Ishikawa. 2001. Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, pages.196-203.

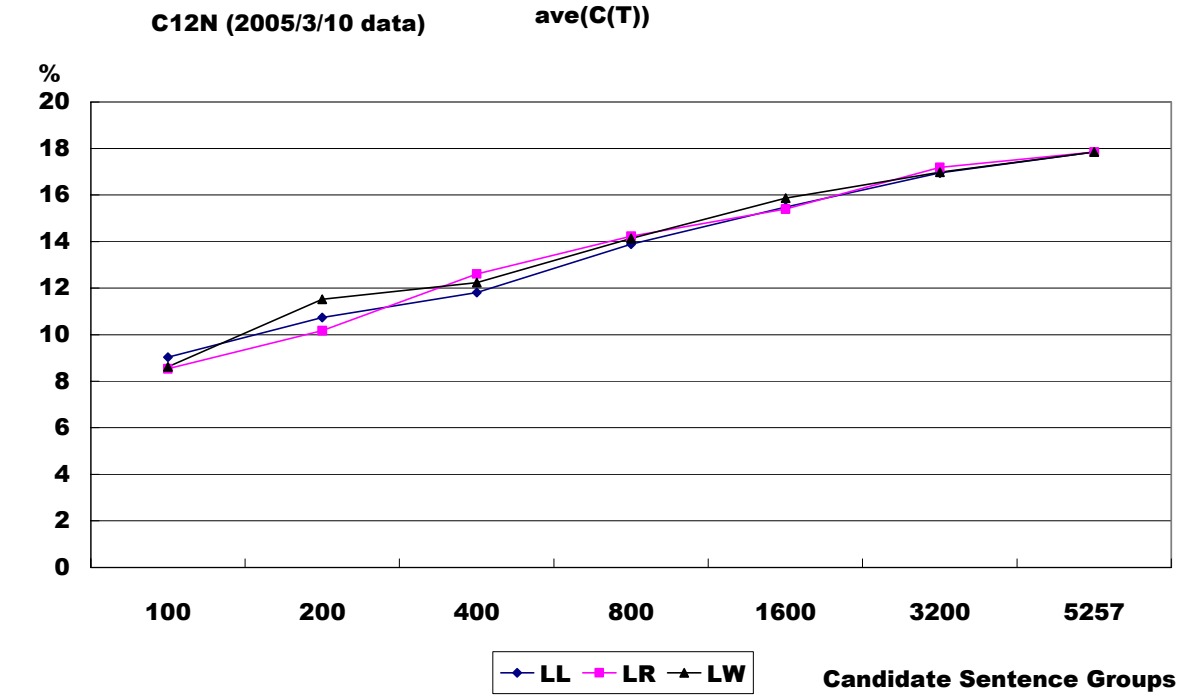

Figure 1: Average in the Maximum Degree of Correspondence in R_T=LAT

The blue line(LL) means R_T=LAT and R_C=(LAT,p), the red line(LR) means R_T=LAT and R_C=(RND,p), and purple (LW) means R_T=LAT and R_C=(WIN,p). The horizontal axis shows the size of p. The vertical axis shows the average in the maximum degree of correspondence, Ave(Max(H)).

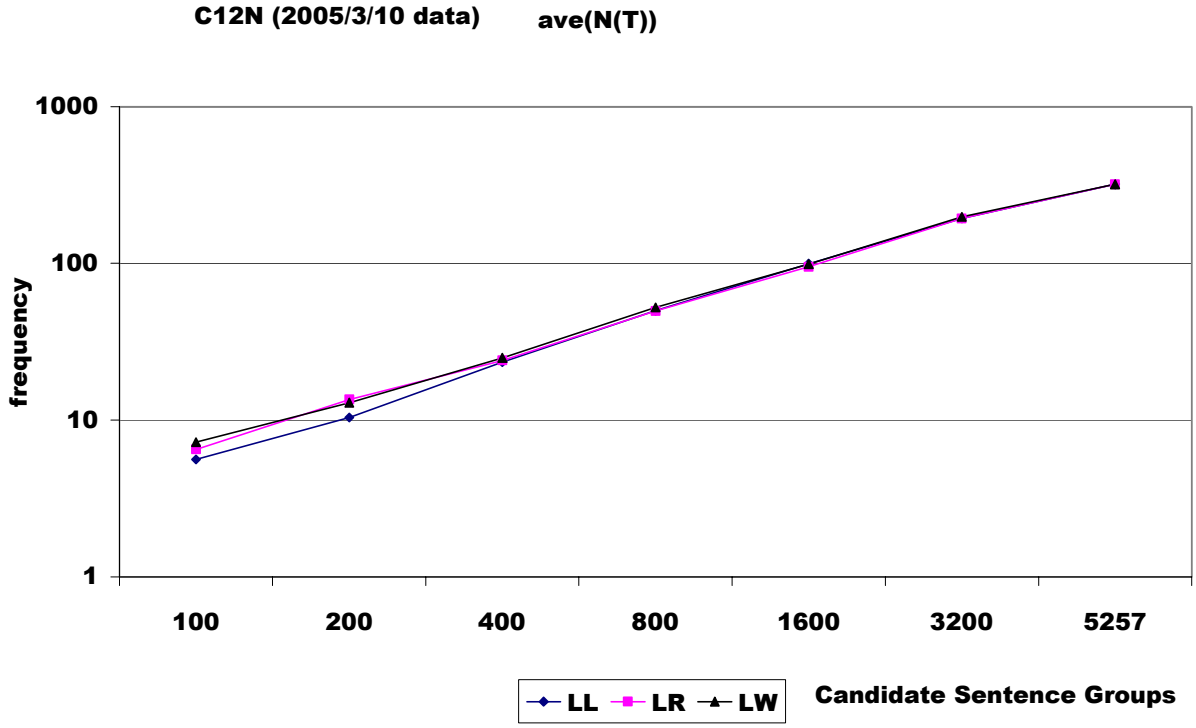

Figure 2: Average in the Number of Corresponding Sentences in R_T=LAT.

The blue line(LL) means R_T=LAT and R_C=(LAT,p), the red line(LR) means R_T=LAT and R_C=(RND,p), and purple (LW) means R_T=LAT and R_C=(WIN,p). The horizontal axis shows the size of p. The vertical axis shows the average in the total number of corresponding sentences, Ave(Num(H,0)).

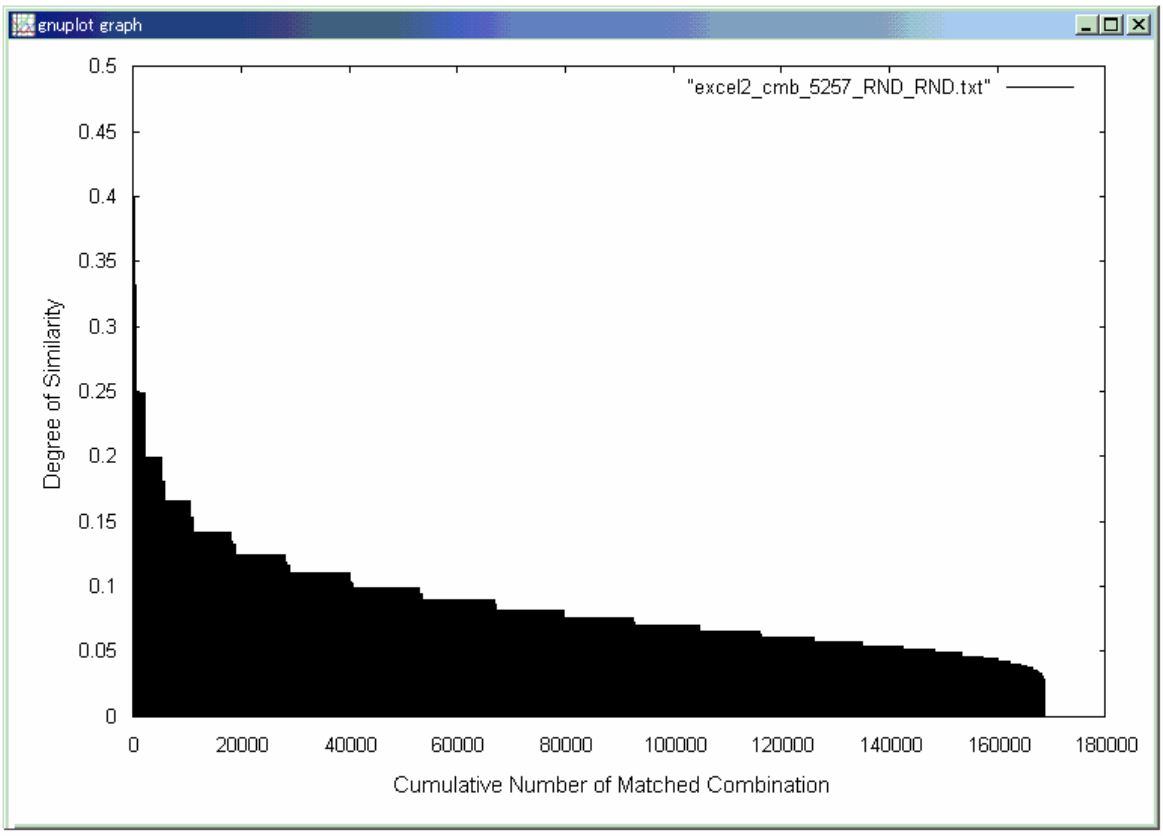

Figure 3: Distribution of Degree of Correspondence for 5,257 Sentences in R_T=RND and R_C=(RND,5257)

The x-axis is the cumulative number of matched combination and the y-axis is the degree of similarity in the order of larger degree of correspondence.