Sentence Alignment of a Japanese-Italian Parallel Corpus. Towards a web-based Interface

Patrizia Zotti　　**Riccardo Apolloni**　　**Yuji Matsumoto**

Graduate School of Information Science

Nara Institute of Science and Technology

{zottip, r-apolloni, matsu}@is.naist.jp

1 Introduction

To align an OCR-generated dataset of excerpts from Japanese novels and their translations into Italian, we have adopted the length-based alignment algorithm developed by Church and Gale [3]. We have developed a web-based interface to run the alignment program interactively, while in this paper we focus only on the alignment results we have conducted.

2 Related Studies

Sentence alignment models may be based on sentence length (Church and Gale [3]; Brown *et al.*[1]), word correspondence (Kay [4]; Kay and Roscheisen [5]; Brown et al. [2], Melamed [6]), or on composite methods (Simard and Plamondon [8]; Moore 2002 [7]).

3 Our Approach

The Church and Gale method is mainly based on three assumptions. (1) It is reasonable to predict the length of a target text from the length of the source text, namely the relationship between aligned sentences is more or less constant. More formally we could say that the lengths of the source and target sentences are related. For Japanese and Italian we have found a ratio≈3: if the Japanese sentence is made of 25 characters, it is reasonable to assume that the length of the Italian sentence is about 75 characters. (2) If we accept to bet 100 yen on an Italian sentence of 75 characters against a Japanese sentence of 25 characters, how much are we willing to bet on 150 Italian characters against 50 Japanese characters? According to Church and Gale we should bet something less than the previous 100 yen. More formally we can say that deviations from the mean value increase with the length of the sentence under examination. (3) The third hypothesis is that if we would bet in a "scientific" other than "prudent" way, we could base our decisions on a Gaussian distribution hypothesis. Therefore, if we know the mean ratio (≈3) and also an average deviation of the lengths of the Italian sentences aligned to Japanese sentences of 50 and 75 characters, could adjust *ad hoc* his bets.

Although Church and Gale claim the method is fairly language-independent, it was originally developed to process texts in European languages. Therefore, we deemed it necessary to fit the parameters (mean, variance and distance) with Japanese-Italian specific values. To estimate the parameters we have processed the PEI corpus [9], a parallel dataset spanning 7000 semi-automatically aligned and human checked Japanese-Italian pairs.

The dataset is made of texts belonging to three domains: parliamentary proceeding (2000 sentences), news articles (2000 sentences) and literary works (3000 sentences).

The news data consists of 2000 Japanese sentences from the Yomiuri Shinbun translated into Italian. The translations are quite literal and one Japanese sentence always corresponds only to one Italian sentence. The parliamentary

proceedings data consists of 2000 pairs. Here translations are less literal and the language is quite formal. The literary works dataset is the most interesting but troublesome at the same time. The original source and target data differ in size since translations are often free: sentences may be merged, split, dropped or added. There is also a fair amount of dialogues that imposes a reflection on how to deal with them in automatically aligning new data.

3.1 Sentences Boundaries

We adopt as boundaries for Japanese sentences the period (。), the question mark (？) the exclamation mark (！), and any pair of quotation marks (「」『』). Italian sentences boundaries are identified with the period (.), the question mark (?), the exclamation mark (!), the colon (:) the semi-colon (;), and any pair of quotation marks (« »).

Examining the manually aligned parallel data we have observed that there is a fair amount of dialogues within which it is often difficult, if not futile, to identify minimum units semantically independent even if separated by the mentioned delimiters. We have therefore decided not to divide the content of the quotation marks, and to consider any text between a pair of quotation marks as one single unit, whether the above-mentioned boundaries are present within the dialogue or not. The use of the colon and the semi-colon as sentence boundaries may seem unusual if not wrong. However, we have observed that – for stylistic reasons, and to adapt the text to Italian reader – a sequence of short Japanese sentences is quite often translated in Italian with one single sentence including as many semi-colon or colon [1] as many corresponding Japanese sentences as in Table 1.

[1] From a stylistic point of view, the semi-colon indicates a pause halfway between the period and the comma. Its use often depends on a personal stylistic choice, though it is mainly used between complex clauses and between complex enumerations, to suggest a pause in formal terms but not in terms of content. The colon informs that what follows shows or explain what has been said before. It is also used to introduce a direct speech.

Table 1 – Excerpts from the parallel dataset

Source	Target
帰って荷物を置いた後、ひさしぶりの麦茶を飲む間もなく、おなかが痛い、と眉を寄せてトイレへ駆け込んだ。そのままテレビのドラマが終わるまで出て来なかった。	Avevamo appena poggiato le valigie nell'ingresso e non vedevamo l'ora di sorseggiare il primo tè verde dopo parecchi giorni, quando mamma si produsse in una smorfia e corse in bagno lamentandosi di un male alla pancia; vi rimase chiusa fino al termine della fiction che stavano dando alla TV.
父がそのかたわらを通り過ぎて玄へ向かい、靴を履き始めたので「お父さん、どこへ行くの？」	Poi, a un certo punto, vidi mio padre che mi passava accanto diretto verso l'ingresso. «Papà, dove vai?» gli chiesi notando che si stava infilando le scarpe.

3.2 Parameters Estimation

Before commenting on the data, we need to specify that we use a variable what Church and Gale call "distance", namely the difference between the actual length of the Italian sentence and its expected length compared to the respective variance. This allows us to reduce the calculations we should do on cumulative probabilities to the case of a normalized Gaussian (mean≈0, variance≈1).

The manually aligned dataset spans about 860k Japanese characters against 300k Italian characters with a mean ratio≈2,8. Max, mean and variance length of Italian sentences are equal to 712, 123,4 and 6723,4, against max, mean and variance length of Japanese sentences equal to 248, 43,3 and 774,5 (Fig. 1).

The best regression (a0≈-477 a1≈18) is shown in Fig. 2, where the vertical axis shows the relative deviation from the expected value and the horizontal axis shows a measure of the sentence length.

Fig. 3 shows the distribution of distances: the curve follows quite closely a normalized Gaussian distribution. By varying the slope from 18 to 10 (but not the mean) the correspondence is almost perfect (Fig. 4).

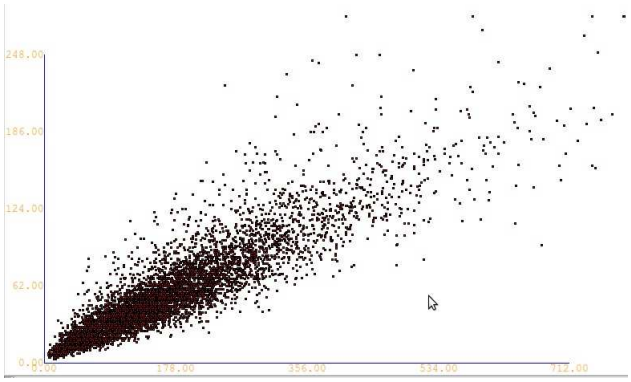

Fig. 1 – The horizontal axis shows the length in characters of Italian sentences, while the vertical axis shows the corresponding length of Japanese sentences.

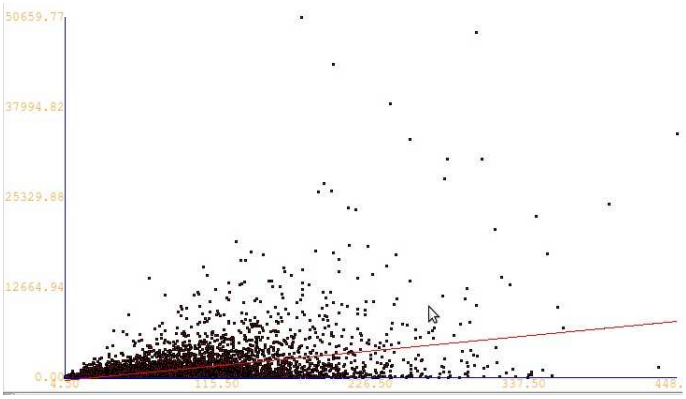

Fig. 2 – Regression. The vertical axis shows the relative deviation from the expected value. The horizontal axis shows a measure of the sentence length.

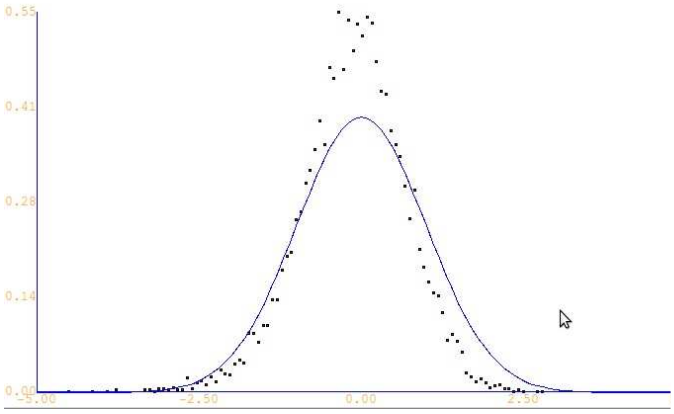

Fig. 3 – Distance distribution (mean≈2.8, slope≈18)

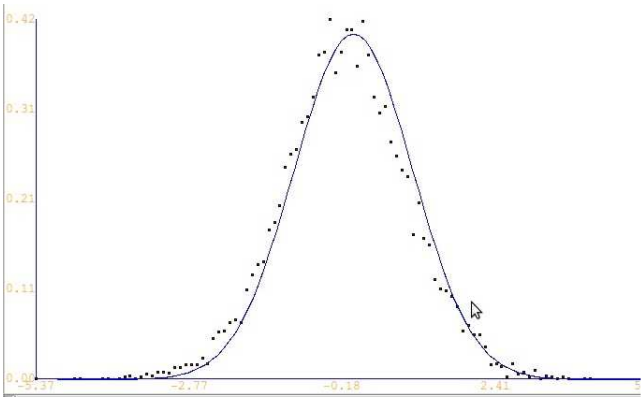

Fig. 4 – Distance distribution (mean≈2.8, slope≈10).

3.3 Matching Probability Estimation

We consider a sentence each portion of text ending with a sentence boundary as specified in 3.1. By processing the manually checked aligned pairs, we have found the matching probabilities shown in Table 2. We have then estimated the probability score for each match by combining the frequency of the matches found in the manually aligned dataset with the Gaussian distribution of lengths, as shown in Table 3. We have added to the algorithm the matches 3:1 and 1:3, not included in the Church and Gale model. We have included the matches 1:0 and 0:1 because, even if we have not found them in our dataset, from a conceptual point of view they may occur. Moreover, we have considered the probabilities 1:2, 2:1, 1:3 and 3:1 in a separate way, while Church and Gale make use of the aggregate data.

Table 2 – Matching probability found in the dataset

Match	Frequency	Probability
1:1	5923	0.8460
1:0	0	
0:1	0	
1:2	401	0.0573
2:1	375	0.0536
2:2	156	0.0223
3:1	49	0.0070
1:3	38	0.0054
3:2	19	0.0027
2:3	19	0.0027

Table 3 – Parameters used in the algorithm

Parameters
Mean≈2.85
Sigma2slope≈12
MATCHES["1:1"]=0.800
MATCHES["1:0"]=0.002
MATCHES["0:1"]=0.002
MATCHES["1:2"]=0.050
MATCHES["2:1"]=0.050
MATCHES["2:2"]=0.020
MATCHES["3:1"]=0.006
MATCHES["1:3"]=0.006

4 Evaluation and Discussion

We have evaluated the algorithm on a dataset spanning 132 lines (any sequence of characters ending with a linefeed) in Japanese and 108 in Italian. By running the split-line function we got 234 Italian sentences against 218 sentences. By running the aligning algorithm we got an output of 192 pairs of which 23 were wrong, with score 0.88.

Upon observation of the pairs obtained as output we have decided to run the algorithm by excluding the colon as sentence delimiter. We got 220 Italian sentences against 218 Japanese ones, and 188 pairs. The score rose to 0.95.

Since our data belong to the domain of literature, we have to deal with many stylistic issues and personal choices made by authors and translators. Before deciding to delete the colon as sentence delimiter we need to evaluate the algorithm on a bigger amount of data written by different authors and translators. We also need to process more data to cope with the big amount of short and long dialogues.

5 Conclusion

In this paper we have described an implementation of the Church and Gale length-based algorithm on the case of Japanese and Italian texts. By fitting the original parameters with Japanese-Italian specific values by processing a dataset spanning 7000 aligned pairs, we first got a score of 0.88 in the first experiment and of 0.95 in the second one.

The software we have developed will become freely available under the GNU licence.

References

[1] Brown *et al.* (1991). Aligning Sentences in Parallel Corpora. Proceedings of 29th Annual Meeting of the Association for Computational Linguistics, 169-176. Berkeley, CA.

[2] Brown *et al.* (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19/2, 263-311.

[3] Gale, William A., Church, Kenneth W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19/1, 75-102.

[4] Kay, Martin (1991). Text-Translation Alignment. ACH/ALLC 1991. Making Connections Conference Handbook. Tempe, Arizona.

[5] Kay, Martin, Roscheisen Martin (1993). Text-Translation Alignment. *Computational Linguistics* 19/1, 121-142.

[6] Melamed, I. Dan (1996). A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report, University of Pennsylvania.

[7] Moore, Robert C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. *Proceedings of AMRA*, 135-144.

[8] Simard, Michel, Plamondon, Pierre (1988). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation* 13/1, 59-80/

[9] Zotti, Patrizia (2013). Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e Applicazioni [How to Build a Japanese-Italian Parallel Corpus. Methodologies and Applications], in M. Casari, P. Scrolavezza (eds), *Giappone, storie plurali*, 351-63. I libri di Emil-Odoya Edizioni. Bologna.