Word segmentation based on proportional analogy and majority voting

Zongrong Zheng and Yves Lepage

Graduate School of Information, Production and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
zzr0427@toki.waseda.jp, yves.lepage@waseda.jp

Abstract

This paper proposes a new method for word segmentation based on proportional analogy and majority voting. Firstly, we use an analogy-based method to propose segmentation hypotheses. Secondly, we use majority voting to make the final decision on where to segment. As an important and original feature, our method does not need any given lexical knowledge or pre-processing training phase. Preliminary results show that this simple approach compares well with segmentation methods for Chinese reported in previous studies. We also present results on English.

1. Introduction

Words are usually considered a basic unit. In some languages like Chinese, texts are continuous sequences of characters without spaces between words. In those cases, it is generally agreed that word segmentation should be the first pre-processing step in any natural language processing (NLP) system. The performance of the best Chinese segmenters has reached 96% in F-score, as reported in the second SIGHAN Chinese segmentation bakeoff was 95% (Emerson, 2005) and nowadays reaches 96% (Chen et al., 2015). These best existing methods rely on massive training data. Existing methods can be roughly classified as either dictionary-based or statistical-based methods.

Dictionary-based methods usually rely on large-scale lexicons and are built upon a few basic "mechanical" segmentation methods based on string matching. Without a large, comprehensive dictionary, the success of such methods degrades.

Statistical-based methods consider the segmentation problem as a classification problem on characters and usually involve elaborated learning models trained on large-scale corpora. The problem is to have the machine learning procedure extract the relevant information from the training data so as to reproduce the segmentation standard given in the training data (Kit and Liu, 2005).

All of these methods require prior lexical knowledge or a training phase. How to efficiently achieve human performance in word segmentation on any language without the knowledge of wordhood is still a challenge (Huang et al., 2007). By contrast to the above-mentioned approaches, we propose a method that directly makes use of the training data while segmenting, without using an explicit lexicon or extracting knowledge in a learning phase beforehand.

The notion on which our proposal relies, proportional analogy, is introduced in Section 2. The principle of the method is described in Section 3. Section 4. presents the implementation details. Section 5. details the evaluation of the method in Chinese and English and gives a comparison with state-of-the-art methods.

2. Proportional analogy

Proportional analogy has been proposed for various natural language processing tasks, like machine translation (Lepage and Denoual, 2005) or the computation of semantic relations (Turney and Littman, 2005). A proportional analogy is a relationship between four objects, noted $A : B :: C : D$ in its general form. On numbers, analogies like: $5 : 15 :: 10 : 30$ are nowadays commonly written as:

$$\frac{5}{15} = \frac{10}{30}$$

By using words, sequences of words or sentences instead of numbers, we get proportional analogies between words, sequences of words or sentences. For instance, the following example is a true analogy between sequences of words:

$$I\ walked : to\ walk :: I\ laughed : to\ laugh$$

We use the algorithm given in (Lepage, 1998) (Lepage, 2004) for the resolution of analogical equations. This algorithm is based on the characterization of proportional analogies shown in Formula (1).

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \quad (1)$$

Here, a is a character, whatever the writing system, and A, B, C and D are strings of characters. $|A|_a$ stands for the number of occurrences of character a in the string of characters A and $d(A,B)$ stands for the edit distance between strings A and B, with insertions and deletions as only edit operations. The input of this algorithm is three strings of characters, words, sequences of words or sentences. Its output is a string of characters in analogy with the input. The following applies this algorithm to a sequence of Chinese characters:

会前 : 社会前 :: 会前进的 : x ⇒ x = 社会前进的

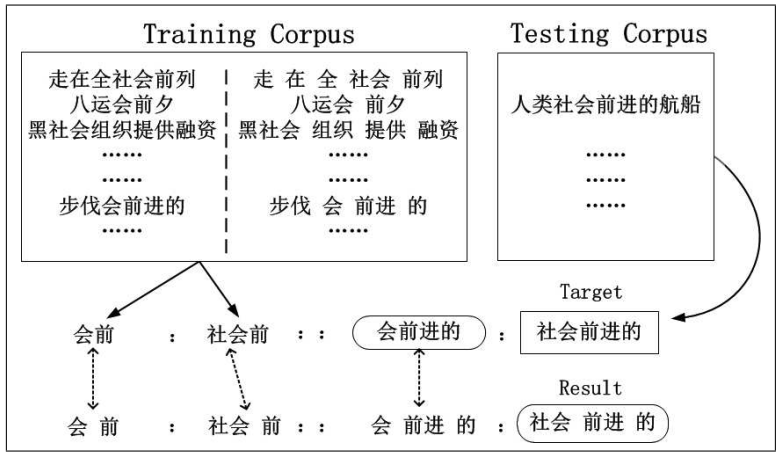

Figure 1: Sketch of the Chinese word segmentation method based on proportional analogy

3. Proposed method for word segmentation using proportional analogy

We propose a new word segmentation method based on proportional analogies. Crucially, we no longer need any explicit lexical knowledge (lexicon) nor pre-processing phrase (training). The following gives the basic idea of the method, directly inspired by the example-based machine translation method introduced in (Lepage and Denoual, 2005).

Let us suppose that we have a corpus of sentences in their usual unsegmented form and their segmented form. We call it the training corpus. A line in such a training corpus may look like:

Thisisaformofcopyright. | *This is a form of copyright .*

Let D be an unsegmented input sentence to be segmented into segmented sentence $\tilde{D}$.

(i) We build all analogical equations $A_i : B_j :: x : D$ with the input sentence D and with all pairs of sub-strings (A_i, B_j) from the unsegmented part of the training corpus. According to Formula 1, not all analogical equations have a solution. In order to get more analogical solutions and reduce time in solving analogical equations, we only consider sub-strings A_i and B_i which are at a distance less than a given threshold from D;

(ii) We gather all the solutions x of the previous analogical equations and only keep the solutions, named $C_{i,j}$, which belong to the training corpus. As it is easy to map from unsegmented part to segmented part for any sub-strings in training corpus, for each $C_{i,j}$, A_i and B_i, we retrieve their corresponding segmented forms $\widetilde{C_{i,j}}$, $\tilde{A}_i$ and $\tilde{B}_i$ in the segmented part of the training corpus;

(iii) We then build all possible analogical equations

$$\widetilde{A_i} : \widetilde{B_i} :: \widetilde{C_{i,j}} : y$$

We output the solutions $y = \widetilde{D_{i,j}}$ of all these analogical equations. They are hypotheses of segmentation for D. We record the number of times each hypothesis was output. Recall that different analogical equations may generate identical solutions.

Figure 1 gives a sketch of the method described above.

4. A word segmentation system using proportional analogy

We now describe the details of our implementation of the analogy-based word segmentation method. The key point is to generate as precise proportional analogies as possible. As the solutions of these proportional analogies may not all be correct, we consider them as hypotheses of segmentation. According to Formula 1, the longer the sentences are, the more difficult the constraints are to satisfy due to data sparseness. It means that longer sentences more easily miss analogical solutions and so miss hypotheses of segmentation more easily. Splitting sentences is necessary. We split sentences into n-grams, i.e., sub-strings of fixed length n. Our system is thus divided into two parts: generating hypotheses of segmentation for n-grams and re-combining segmentation hypotheses to generate a complete segmented result for the entire input sentences.

4.1. Generating segmented hypotheses for n-grams

We adopt the method proposed in Section 4. to generate the segmented result of n-grams in practice in our system. The work flow of generating segmentation hypotheses for n-grams is shown in Figure 2.

According to Formula 1, A and B should share characters with D to get a solution from equation $A_i : B_j :: x : D$. It means that A and B should be similar to D up to a certain extent. We use TRE agrep[1], an approximate regex matching library, to retrieve sub-strings which are similar to the input D from the training corpus. We use edit distance, with only insertions and deletions as edit operations, to quantify how similar two strings are to one another. Any two of these similar substrings form an analogical equation

[1] http://laurikari.net/tre/

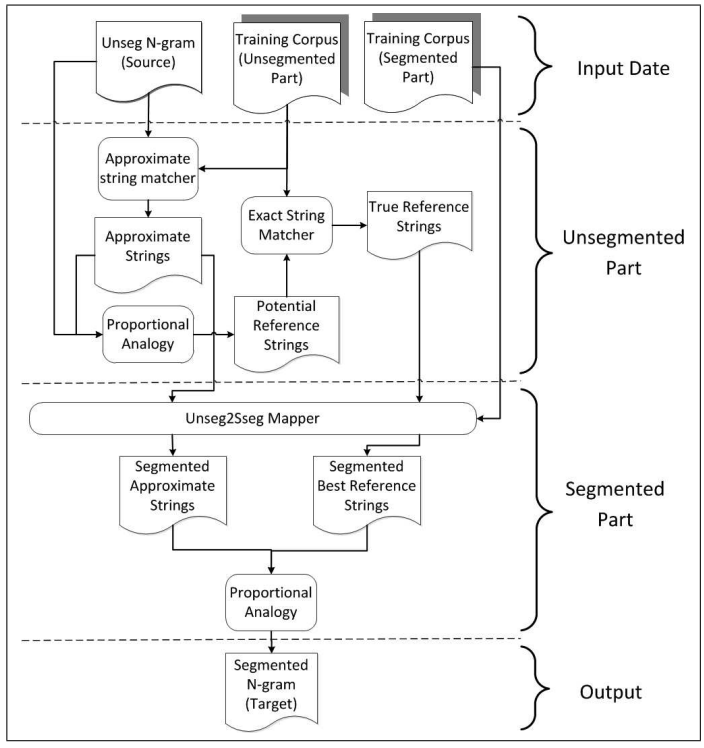

Figure 2: Work flow for the generation of segmentation hypothesis for n-grams

with the input D. In general, not all solutions of the equations occur in the training corpus. Consequently, only the solutions which occur in the segmented part of the training corpus are considered as segmentation hypotheses. Notice that different analogical equations may generate identical solutions.

The same segmentation hypotheses can be generated several times by different analogical equations. We record this number of occurrences. It is natural to think that the larger the number of occurrence is, the more likely the segmentation hypothesis is.

4.2. Re-combining n-gram segmentation hypotheses

We use majority voting rules to recombine the segmentation hypotheses of n-grams. A segmentation hypothesis can be represented as a sequence of characters and delimiters

$$c_1 D_1 c_2 D_2 ... c_{n-1} D_{n-1} c_n,$$

with its number of occurrences m. In this form, D_i is either a space or not a space. We let all segmentation hypotheses vote for D_i.

When D_i is a space, it means that this segmentation hypothesis votes m times for segmentation. When D_i is not a space, it votes m times against segmentation. Figure 3 is an example to illustrate the use of majority voting in our system. We sum up the votes in favor and against segmentation and output the final results according to the vote results.

5. Experiments

5.1. Evaluation

To evaluate the effectiveness of our proposed method, we conduct experiments on Chinese and English. The segmentation accuracy is evaluated by test recall (R), test precision (P) and balanced F-score, i.e., the harmonic mean of recall and precision (see Equations 2, 3 and 4). All evaluation results in this paper were obtained using the official scoring script of the second SIGHAN international Chinese word segmentation bakeoff (Emerson, 2005) downloaded from the official website[2].

$$\mathrm{R} = \frac{\text{number of correctly segmented words}}{\text{total number of words in gold standard segmentation}} \quad (2)$$

$$\mathrm{P} = \frac{\text{number of correctly segmented words}}{\text{total number of words in segmentation result}} \quad (3)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (4)$$

Our experiments follow the closed track. It means that no extra resource other than the training corpus is used.

5.2. Experiments on Chinese

On Chinese, we perform our experiments on the widely used Chinese word-segmented corpus, PKU (Peking University), used in the second SIGHAN international Chinese word segmentation bakeoff. The training set and test set are publicly available from the official website.

[2]http://www.sighan.org/bakeoff2005/

	c_1	D_1	c_2	D_2	c_3	D_3	c_4	D_4	c_5	D_5	c_6	D_6	c_7	D_7	c_8	D_8	c_9
# of occurrence	人		类		社		会		前		进		的		航		船
9	人	9	类	9	社	9	会	9	前								
3			类	3	社	3	会	3	前	3	进						
1					社	1	会	1	前	1	进	1	的				
11					社	11	会	11	前	11	进	11	的				
37					社	37	会	37	前	37	进	37	的				
4									前	4	进	4	的	4	航	4	船
1									前	1	进	1	的	1	航	1	船
votes for seg (␣)		**0**		**12**		**0**		**49**		**1**		**53**		**4**		**0**	
votes against seg		**9**		**0**		**61**		12		**56**		1		1		**5**	
segmentation result	人		类	␣	社		会	␣	前		进	␣	的	␣	航		船

Figure 3: Recombination of segmentation hypotheses of n-grams using majority voting

5.2.1. Influence of n-gram length and distance threshold

In this experiment, we measure the influence of the length of n-grams and the edit distance threshold. In general, the longer the length of a segmentation hypothesis, the more reliable this hypothesis. But as discussed in Section 4., longer n-grams miss hypotheses of segmentation more easily. So the length of n-grams will influence the segmentation results.

In addition, the larger the edit distance threshold used, the more similar sub-strings may be retrieved. To measure this, we conduct experiments using different lengths of n-grams and different edit distance thresholds.

According to our majority voting method, we would consider a position is not segmented if no segmentation hypothesis votes for it. The results in Table 1 shows that this data sparseness problem is more serious when we use larger lengths of n-grams.

Length of n-grams	Edit dist. threshold	Word count	P	R	F
6	3	79828	85.5	65.4	74.1
5	3	95079	90.0	82.0	85.8
4	2	99103	90.8	86.2	88.4
3	2	103186	90.9	89.9	90.4

Table 1: Performance of our method with different n-gram lengths and edit distance thresholds

5.2.2. Comparison with state-of-the-art systems

Based on the results obtained above, we set the length of n-grams to 3 and the edit distance threshold to 2. Table 2 shows our empirical results on the same data set as others systems reported in the SIGHAN 2005 bakeoff. Our system achieves significantly better results than the baseline. Very importantly, the R_{oov} score shows that our method exhibit a reasonable ability to deal with out-of-vocabulary (OOV) word and to guess segmentation for them. However, the R_{iv} score shows that our method performs slightly worse on in-vocabulary (IV) word recognition. Compared with the best result reported in SIGHAN 2005 (Tseng et al., 2005), our result still shows room for improvement. But as a simple method which does not need lexical knowledge not pre-compilation of knowledge extracted from the training data, our method showed a reasonably high potential for Chinese word segmentation.

Models	P	R	F	R_{oov}	R_{iv}
Baseline	84.3	90.7	87.4	6.9	95.8
Ours	90.9	89.9	90.4	60.7	91.6
Best05	95.4	94.6	95.0	78.7	95.6
(Ma and Hinrichs, 2015))	95.5	94.6	95.1	76.0	
(Chen et al., 2015)	96.3	95.9	96.3		

Table 2: Performance of our method on the Chinese SIGHAN 2005 PKU data set compared to a baseline and the best result (Best05) in SIGHAN 2005 bakeoff. All results are closed set

5.3. Experiments on English

We perform experiments on English data to show the robustness and the validity of the method. English word segmentation, on the contrary to Chinese, is not a necessary task, but it is a testbed in reproducing by machine the human intuition of wordhood (De Marcken, 1996). We use the English part of the European Parliament Proceedings Parallel Corpus (Koehn, 2005) (Europarl)[3] to build a training set and a test set. We randomly select 20,000 sentences from the English part of the Europarl corpus as a training set and randomly select another distinct 1,000 sentences as a test set. We delete all spaces between words to generate unsegmented data. The OOV rate in the test set relatively to the training set is 0.019.

[3]http://www.statmt.org/europarl/

Length of n-grams	P	R	F	R_{oov}	R_{iv}
10	95.6	95.5	95.5	44.1	96.0
11	94.5	94.0	94.3	40.7	95.0
12	93.5	91.4	92.4	35.6	92.0
13	90.7	86.4	88.5	23.7	87.0

Table 3: Performance of our method on English with different lengths of n-grams

Test set N^o	OOV rate	Size of training set	P	R	F
1	1.8	20,000	95.0	94.7	94.8
		40,000	96.7	96.3	96.5
2	2.2	20,000	95.6	95.5	95.5
		40,000	96.3	95.9	96.1
3	2.6	20,000	94.5	94.0	94.3
		40,000	96.0	95.4	95.7

Table 4: Performance of our method on English with different sized training set

5.3.1. Influence of n-gram length and distance threshold

In a first experiment, we examine the influence of the n-gram length. We set the edit distance to 3. Table 3 shows that the length of n-grams also influences results in English, but, compared to Chinese, a much higher value of n is needed. Although not really directly comparable, it is observed that the method can get higher scores in F-measure in English than in Chinese.

5.3.2. Influence of the size of the training data

To evaluate our method further, we perform three more experiments with a larger training corpus, which contains 40,000 sentences of English monolingual corpora. We also randomly select two more testing sets which contain 1,000 sentences each. They exhibit different out-of-vocabulary rates. We keep the length of n-grams to 10 and the edit distance threshold to 3. The results are shown in Table 4. Naturally, the method achieves better scores when using a larger training set. As for the F-measure scores obtained (higher than 94%), these results demonstrate that the method performs well in identifying English words. Of course, when the OOV rate increases, the F-measure scores decrease slightly (but not systematically as the F-score on the third row in Table 4 shows).

6. Conclusion

We presented a method for word segmentation based on proportional analogy to output segmentation hypotheses for n-grams and based on majority voting to make the final decision on where to segment when recombining the n-grams into complete sentences.

As an important and original feature, the proposed method does not need any prior lexical knowledge nor any pre-compilation of any knowledge extracted from the training data. Only two parameters have to be determined in advance: the length of the n-grams used and a distance threshold.

The method achieves a desirable accuracy both on Chinese and English. In Chinese, despite its simplicity, the method is well above the recognized baseline. The method shows a reasonable performance in word identification, as measured by its recall on out-of-vocabulary words.

7. References

Chen, Xinchi, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang, 2015. Long short-term memory neural networks for Chinese word segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

De Marcken, Carl, 1996. Linguistic structure as composition and perturbation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Emerson, Thomas, 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 133. Jeju Island, Korea.

Huang, Chu-Ren, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot, 2007. Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics.

Kit, Chunyu and Xiaoyue Liu, 2005. An example-based Chinese word segmentation system for CWSB-2. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Koehn, Philipp, 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT summit X)*, volume 5.

Lepage, Yves, 1998. Solving analogies on words: an algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics.

Lepage, Yves, 2004. Analogy and formal languages. *Electronic notes in theoretical computer science*, 53:180–191.

Lepage, Yves and Etienne Denoual, 2005. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282.

Ma, Jianqiang and Erhard Hinrichs, 2015. Accurate linear-time chinese word segmentation via embedding matching. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning, 2005. A conditional random field word segmenter for SIGHAN bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171.

Turney, Peter D and Michael L Littman, 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.