Unsupervized Word Segmentation: the case for Mandarin Chinese

Pierre Magistry
Alpage, INRIA & Univ. Paris 7,
175 rue du Chevaleret,
75013 Paris, France
pierre.magistry@inria.fr

Benoît Sagot
Alpage, INRIA & Univ. Paris 7,
175 rue du Chevaleret,
75013 Paris, France
benoit.sagot@inria.fr

Abstract

In this paper, we present an unsupervized segmentation system tested on Mandarin Chinese. Following Harris's Hypothesis in Kempe (1999) and Tanaka-Ishii's (2005) reformulation, we base our work on the Variation of Branching Entropy. We improve on (Jin and Tanaka-Ishii, 2006) by adding normalization and viterbi-decoding. This enable us to remove most of the thresholds and parameters from their model and to reach near state-of-the-art results (Wang et al., 2011) with a simpler system. We provide evaluation on different corpora available from the Segmentation bake-off II (Emerson, 2005) and define a more precise topline for the task using cross-trained supervized system available off-the-shelf (Zhang and Clark, 2010; Zhao and Kit, 2008; Huang and Zhao, 2007)

1 Introduction

The Chinese script has no explicit "word" boundaries. Therefore, tokenization itself, although the very first step of many text processing systems, is a challenging task. Supervized segmentation systems exist but rely on manually segmented corpora, which are often specific to a genre or a domain and use many different segmentation guidelines. In order to deal with a larger variety of genres and domains, or to tackle more theoretic questions about linguistic units, unsupervized segmentation is still an important issue. After a short review of the corresponding literature in Section 2, we discuss the challenging issue of evaluating unsupervized word segmentation systems in Section 3. Section 4 and Section 5 present the core of our system. Finally, in Section 6, we detail and discuss our results.

2 State of the Art

Unsupervized word segmentation systems tend to make use of three different types of information: the cohesion of the resulting units (e.g., Mutual Information, as in (Sproat and Shih, 1990)), the degree of separation between the resulting units (e.g., Accessor Variety, see (Feng et al., 2004)) and the probability of a segmentation given a string (Goldwater et al., 2006; Mochihashi et al., 2009).

A recently published work by Wang et al. (2011) introduce ESA: "Evaluation, Selection, Adjustment." This method combines cohesion and separation measures in a "goodness" metric that is maximized during an iterative process. This work is the current state-of-the-art in unsupervized segmentation of Mandarin Chinese data.

The main drawbacks of ESA are the need to iterate the process on the corpus around 10 times to reach good performance levels and the need to set a parameter that balances the impact of the cohesion measure w.r.t. the separation measure. Empirically, a correlation is found between the parameter and the size of the corpus but this correlation depends on the script used in the corpus (it changes if Latin letters and Arabic numbers are taken into account during preprocessing or not). Moreover, computing this correlation and finding the best value for the parameter (i.e., what the authors call the *proper exponent*) requires a manually segmented training corpus. Therefore, this proper exponent may not be easily available in all situations. However, if we only consider their experiments using settings similar to ours, their results consistently lie around an f-score of 0.80.

An older approach, introduced by Jin and Tanaka-Ishii (2006), solely relies on a separation measure

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 383–387,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

that is directly inspired by a linguistic hypothesis formulated by Harris (1955). In Tanaka-Ishii (2005) (following Kempe (1999)) who use Branching Entropy (BE), this hypothesis goes as follows: if sequences produced by human language were random, we would expect the Branching Entropy of a sequence (estimated from the n-grams in a corpus) to decrease as we increase the length of the sequence. Therefore the variation of the branching entropy (VBE) should be negative. When we observe that it is not the case, Harris hypothesizes that we are at a linguistic boundary. Following this hypothesis, (Jin and Tanaka-Ishii, 2006) propose a system that segments when BE is rising or when it reach a certain maximum.

The main drawback of Jin and Tanaka-Ishii (2006) model is that segmentation decisions are taken very locally[1] and do not depend on neighboring cuts. Moreover, this system also also relies on parameters, namely the threshold on the VBE above which the system decides to segment (in their system, this is when VBE$\geq$ 0). In theory, we could expect a decreasing BE and look for a less decreasing value (or on the contrary, rising at least to some extent). A threshold of 0 can be seen as a default value. Finally, Jin and Tanaka-Ishii do not take in account that VBE of n-gram may not be directly comparable to the VBE of m-grams if $m \neq n$. A normalization is needed (as in (Cohen et al., 2002)).

Due to space constraints, we shall not describe here other systems than those by Wang et al. (2011) and Jin and Tanaka-Ishii (2006). A more comprehensive state of the art can be found in (Zhao and Kit, 2008) and (Wang et al., 2011).

In this paper we will show that we can correct the drawbacks of Jin and Tanaka-Ishii (2006) model and reach performances comparable to those of Wang et al. (2011) with as simpler system.

3 Evaluation

In this paper, in order to be comparable with Wang et al. (2011), we evaluate our system against the corpora from the Second International Chinese Word Segmentation Bakeoff (Emerson, 2005). These corpora cover 4 different segmentation guidelines from various origins: Academia Sinica (AS), City-University of Hong-Kong (CITYU), Microsoft Research (MSR) and Peking University (PKU).

[1] Jin (2007) uses self-training with MDL to address this issue.

Evaluating unsupervized systems is a challenge by itself. As an agreement on the exact definition of what a *word* is remains hard to reach, various segmentation guidelines have been proposed and followed for the annotation of different corpora. The evaluation of supervized systems can be achieved on any corpus using any guidelines: when trained on data that follows particular guidelines, the resulting system will follow as well as possible these guidelines, and can be evaluated on data annotated accordingly. However, for unsupervized systems, there is no reason why a system should be closer to one reference than another or even not to lie somewhere in between the different existing guidelines. Huang and Zhao (2007) propose to use cross-training of a supervized segmentation system in order to have an estimation of the consistency between different segmentation guidelines, and therefore an upper bound of what can be expected from an unsupervized system (Zhao and Kit, 2008). The average consistency is found to be as low as 0.85 (f-score). Therefore this figure can be considered as a sensible *topline* for unsupervized systems. The standard *baseline* which consists in segmenting each character leads to a baseline around 0.35 (f-score) — almost half of the tokens in a manually segmented corpus are unigrams.

Per word-length evaluation is also important as units of various lengths tend to have different distributions. We used ZPAR (Zhang and Clark, 2010) on the four corpora from the Second Bakeoff to reproduce Huang and Zhao's (2007) experiments, but also to measure cross-corpus consistency at a per-word-length level. Our overall results are comparable to what Huang and Zhao (2007) report. However, the consistency is quickly falling for longer words: on unigrams, f-scores range from 0.81 to 0.90 (the same as the overall results). We get slightly higher figures on bigrams (0.85–0.92) but much lower on trigrams with only 0.59–0.79. In a segmented Chinese text, most of the tokens are uni- and bigrams but most of the types are bi- and trigrams (as unigrams are often high frequency grammatical words and trigrams the result of more or less productive affixations). Therefore the results of evaluations only based on tokens do not suffer much from poor performances on trigrams even if a large part of the lexicon may be incorrectly processed.

因此基于符号的评估结果并未受三元语法模型较差表现的过多影响，即便大量的词典也未被正确处理

Another issue about the evaluation and comparison of unsupervized systems is to try and remain fair

in terms of preprocessing and prior knowledge given to the systems. For example, Wang et al. (2011) used different levels of preprocessing (which they call "settings"). In their settings 1 and 2, Wang et al. (2011) try not to rely on punctuation and character encoding information (such as distinguishing Latin and Chinese characters). However, they optimize their parameter for each setting. We therefore consider that their system does take into account the level of processing which is performed on Latin characters and Arabic numbers, and therefore "knows" whether to expect such characters or not. In setting 3 they add the knowledge of punctuation as clear boundaries and in setting 4 they preprocess Arabic and Latin and obtain better, more consistent and less questionable results.

As we are more interested in reducing the amount of human labor needed than in achieving by all means fully unsupervized learning, we do not refrain from performing basic and straightforward preprocessing such as detection of punctuation marks, Latin characters and Arabic numbers.[2] Therefore, our experiments rely on settings similar to their settings 3 and 4, and are evaluated against the same corpora.

4 Normalized Variation of Branching Entropy (nVBE)

Our system builds upon Harris's (1955) hypothesis and its reformulation by Kempe (1999) and Tanaka-Ishii (2005). Let us now define formally the notions underlying our system.

Given an n-gram $x_{0..n} = x_{0..1}\, x_{1..2} \ldots x_{n-1..n}$ with a left context $\chi_{\rightarrow}$, we define its *Right Branching Entropy* (RBE) as:

$$\begin{aligned} h_{\rightarrow}(x_{0..n}) &= H(\chi_{\rightarrow} \mid x_{0..n}) \\ &= -\sum_{x \in \chi_{\rightarrow}} P(x \mid x_{0..n}) \log P(x \mid x_{0..n}). \end{aligned}$$

The *Left Branching Entropy* (LBE) is defined in a symmetric way: if we note $\chi_{\leftarrow}$ the right context of $x_{0..n}$, its LBE is defined as:

$$h_{\leftarrow}(x_{0..n}) = H(\chi_{\leftarrow} \mid x_{0..n}).$$

The RBE (resp. LBE) can be considered as $x_{0..n}$'s *Branching Entropy* (BE) when reading from left to right (resp. right to left).

[2]Simple regular expressions could also be considered to deal with unambiguous cases of numbers and dates in Chinese script.

From $h_{\rightarrow}(x_{0..n})$ and $h_{\rightarrow}(x_{0..n-1})$ on the one hand, and from $h_{\leftarrow}(x_{0..n})$ and $h_{\leftarrow}(x_{1..n})$ we estimate the *Variation of Branching Entropy* (VBE) in both directions, defined as follows:

$$\begin{aligned} \delta h_{\rightarrow}(x_{0..n}) &= h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1}) \\ \delta h_{\leftarrow}(x_{0..n}) &= h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}). \end{aligned}$$

The VBEs are not directly comparable for strings of different lengths and need to be normalized. In this work, we recenter them around 0 with respect to the length of the string by substracting the mean of the VBEs of the strings of the same length. Writing $\tilde{\delta} h_{\rightarrow}(x)$ and $\tilde{\delta} h_{\leftarrow}(x)$. The normalized VBEs for the string x, or ***nVBEs***, are then defined as follow (we only defined $\tilde{\delta} h_{\leftarrow}(x)$ for clarity reasons): for each length k and each k-gram x such that $len(x) = k$, $\tilde{\delta} h_{\rightarrow}(x) = \delta h_{\rightarrow}(x) - \mu_{\rightarrow,k}$, where $\mu_{\rightarrow,k}$ is the mean of the values of $\delta h_{\rightarrow}(x)$ of all k-grams x.

Note that we use and normalize the variation of branching entropy and not the branching entropy itself. Doing so would break the Harris's hypothesis as we would not expect $\tilde{h}(x_{0..n}) < \tilde{h}(x_{0..n-1})$ in non-boundary situation anymore. Many studies use directly the branching entropy (normalized or not) and report results that are below state-of-the-art systems (Cohen et al., 2002).

5 Decoding algorithm

If we follow Harris's hypothesis and consider complex morphological word structures, we expect a large VBE at the boundaries of interesting units and more unstable variations inside "words." This expectation was confirmed by empirical data visualization. For different lengths of n-grams, we compared the distributions of the VBEs at different positions inside the n-gram and at its boundaries. By plotting density distributions for words vs. non-words, we observed that the VBE at both boundaries were the most discriminative value. Therefore, we decided to take in account the VBE only at the word-candidate boundaries (left and right) and not to consider the inner values. Two interesting consequences of this decision are: first, all $\tilde{\delta} h(x)$ can be precomputed as they do not depend on the context. Second, best segmentation can be computed using dynamic programming.

Since we consider the VBE only at words boundary, we can define for any n-gram w its *autonomy* as $a(x) = \tilde{\delta}_{\leftarrow} h(x) + \tilde{\delta} h_{\rightarrow}(x)$. The more an n-gram is autonomous, the more likely it is to be a word.

With this measure, we can redefine the sentence segmentation problem as the maximization of the autonomy measure of its words. For a character sequence s, if we call $Seg(s)$ the set of all the possible segmentations, then we are looking for:

$$\arg\max_{W \in Seg(s)} \sum_{w_i \in W} a(w_i) \cdot len(w_i),$$

where W is the segmentation corresponding to the sequence of words $w_0 w_1 \dots w_m$, and $len(w_i)$ is the length of a word w_i used here to be able to compare segmentations resulting in a different number of words. This best segmentation can be computed easily using dynamic programming.

6 Results and discussion

We tested our system against the data from the 4 corpora of the Second Bakeoff, in both settings 3 and 4, as described in Section 3. Overall results are given in Table 1 and per-word-length results in Table 2.

Our results (nVBE) show significant improvements over Jin's (2006) strategy (VBE $>$ 0) and are closely competing with ESA. But contrarily to ESA (Wang et al., 2011), it does not require multiple iterations on the corpus and it does not rely on any parameters. This shows that we can rely solely on a separation measure and get high segmentation scores. When maximized over a sentence, this measure captures at least in part what can be modeled by a cohesion measure without the need for fine-tuning the balance between the two.

The evolution of the results w.r.t. word length is consistent with the supervized cross-evaluation results of the various segmentation guidelines as performed in Section 3.

Due to space constraints, we cannot detail here a qualitative analysis of the results. We can simply mention that the errors we observed are consistent with previous systems based on Harris's hypothesis (see (Magistry and Sagot, 2011) and Jin (2007) for a longer discussion). Many errors are related to dates and Chinese numbers. This could and should be dealt with during preprocessing. Other errors often involve frequent grammatical morphemes or productive affixes. These errors are often interesting for linguists and could be studied as such and/or corrected in a post-processing stage that would introduce linguistic knowledge. Indeed, unlike content words, grammatical morphemes belongs to closed classes, therefore introducing this linguistic knowledge into the system may be of great help without requiring to much human effort. A sensible way to go in that direction would be to let unsupervized system deal with open classes and process closed classes with a symbolic or supervized module.

System	AS	CITYU	PKU	MSR
Setting 3				
ESA worst	0.729	0.795	0.781	0.768
ESA best	0.782	0.816	0.795	0.802
nVBE	0.758	0.775	0.781	0.798
Setting 4				
VBE $>$ 0	0.63	0.640	0.703	0.713
ESA worst	0.732	0.809	0.784	0.784
ESA best	0.786	0.829	0.800	0.818
nVBE	0.766	0.767	0.800	0.813

Table 1: Evaluation on the Second Bakeoff data with Wang et al.'s (2011) settings. "Worst" and "best" give the range of the reported results with differents values of the parameter in Wang et al.'s system. VBE $>$ 0 correspond to a cut whenever BE is raising. nVBE corresponds to our proposal, based on normalized VBE with maximization at word boundaries. Recall that the topline is around 0.85

Corpus	overall	unigrams	bigrams	trigrams
AS	0.766	0.741	0.828	0.494
CITYU	0.767	0.739	0.834	0.555
PKU	0.800	0.789	0.855	0.451
MSR	0.813	0.823	0.856	0.482

Table 2: Per word-length details of our results with our nVBE algorithm and setting 4. Recall that the toplines are respectively 0.85, 0.81, 0.85 and 0.59 (see Section 3)

One can also observe that our system performs better on PKU and MSR corpora. As PKU is the smallest corpus and AS the biggest, size alone cannot explain this result. However, PKU is more consistent in genre as it contains only articles from the People's Daily. On the other end, AS is a balanced corpus with a greater variety in many aspects. CITYU Corpus is almost as small as PKU but contains articles from newspapers of various Mandarin Chinese speaking communities where great variation is to be expected. This suggest that consistency of the input data is as important as the amount of data. This hypothesis has to be confirmed in futur studies. If it is, automatic clustering of the input data may be an important pre-processing step for this kind of systems.

References

Paul Cohen, Brent Heeringa, and Niall Adams. 2002. An unsupervised algorithm for segmenting categorical timeseries into episodes. *Pattern Detection and Discovery*, page 117–133.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weiming Zheng. 2004. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, page 673–680.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Changning. Huang and Hai Zhao. 2007. 中文分词十年回顾 (Chinese word segmentation: A decade review). *Journal of Chinese Information Processing*, 21(3):8–20.

Zhihui Jin and Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, page 428–435.

Zhihui Jin. 2007. *A Study On Unsupervised Segmentation Of Text Using Contextual Complexity*. Ph.D. thesis, University of Tokyo.

André Kempe. 1999. Experiments in unsupervised entropy-based corpus segmentation. In *Workshop of EACL in Computational Natural Language Learning*, page 7–13.

Pierre Magistry and Benoît Sagot. 2011. Segmentation et induction de lexique non-supervisées du mandarin. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France, June. ATALA.

Daichi Mochihashi, Takeshi. Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, page 100–108.

Richard W. Sproat and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

Kumiko Tanaka-Ishii. 2005. Entropy as an indicator of context boundaries: An experiment using a web search engine. In *IJCNLP*, page 93–105.

Hanshi Wang, Jian Zhu, Shiping Tang, and Xiaozhong Fan. 2011. A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3):421–454.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, page 843–852.

Hai Zhao and Chunyu Kit. 2008. An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework. In *The Third International Joint Conference on Natural Language Processing (IJCNLP2008), Hyderabad, India*.