Abstract. Chinese text segmentation is a well-known and difficult problem. On one side, there is not a simple notion of “word” in Chinese language making really hard to implement rule-based systems to segment written texts, thus lexicons and statistical information are usually employed to achieve such a task. On the other side, any piece of Chinese text usually includes segments present neither in the lexicons nor in the training data. Even worse, such unseen sequences can be segmented into a number of totally unrelated words making later processing phases difficult. For instance, using a lexicon-based system the sequence 巴罗佐 (Bāluōzuǒ, Barroso, current president-designate of the European Commission) can be segmented into 巴 (bā, to hope, to wish) and 罗佐 (luōzuǒ, an undefined word) changing completely the meaning of the sentence. A new and extremely simple algorithm specially suited to work over short Chinese documents is introduced. This new algorithm performs text “self-segmentation” producing results comparable to those achieved by native speakers without using either lexicons or any statistical information beyond the obtained from the input text. Furthermore, it is really robust for finding new “words”, especially proper nouns, and it is well suited to build lexicons from scratch. Some preliminary results are provided in addition to examples of its employment.

Introduction

Chinese texts are written without any kind of “separator” apart from punctuation marks. This, of course, is not a problem for any human skilled in Chinese reading but remains a challenge for NLP systems. Thus, text segmentation, that is, the process of inserting “spaces” between Chinese “words” is a common previous stage for several tasks, for instance machine translation or information retrieval. However, such segmentation problem is not an easy one since there is not a simple notion of “word” in Chinese language. This makes really hard the development of rule-based systems for Chinese segmentation and therefore two main approaches are commonly followed, namely, lexicon- and statistical-based methods.

Lexicon-based systems rely mainly on pre-built sets of Chinese entries and employ that information along with some heuristics in order to segment the input texts, e.g. Chen and Liu (1992). Main limitation of such systems lies on the fact that a lexicon containing any conceivable Chinese segment is unfeasible since new “words” (proper nouns, brands or neologisms) are continuously coined. On the other side, statistical-based systems “distill” information from training data to help finding the most likely “break points”, e.g. Sproat and Shih (1990), Maosong et al. (1998) or Fu and Luke (2003). Hybrid approaches combining machine-learning and lexicon knowledge have also been proposed, for instance by Sproat et al. (1996).

So, Chinese text segmentation is by no means a closed problem and there is still room for improvement and new proposals. In this paper the so-called self-segmentation algorithm is introduced. This new method can also be considered a hybrid one since it uses both statistical information and lexicon knowledge; however, the lexicon is not built in advance but after processing the input data and the statistical information is not previously obtained by training on external corpora but just from the input text. Because of this, while “classical” Chinese segmenters can work on a few characters the self-segmentation algorithm can only be fully exploited by applying it to short documents (i.e. a few hundred characters). Thus, main application of this new algorithm is not text segmentation but “from-scratch” lexicon building since it performs quite robustly and finds equally easily both common Chinese words and rare entries such as phonetic spellings of non-Chinese proper nouns.

The Chinese Self-Segmentation Algorithm

As it has been said, the proposed algorithm does not employ external lexicons or corpus information beyond the obtained from the input data. The operation is quite simple. First, raw input text must be split into a sequence of chunks using as break points roman characters, numbers and punctuation marks (see Fig. 1).
This preprocessed text is submitted to the self-segmentation algorithm which processes it in sequential stages. First, it finds every \( n \)-gram (\( n>2 \)) appearing in the text twice or more times storing both the \( n \)-gram and its count. Secondly, for each chunk from the input data it finds every \( n \)-gram appearing within the chunk, the so-called “candidate” \( n \)-grams. Once all the candidates for a chunk have been found they are “tested” in decreasing length and appearance order (e.g. 巴罗佐 is checked before 巴罗). The test for each \( n \)-gram is very simple, if counts for both constituent \((n-1)\)-grams are strictly bigger than the \( n \)-gram count then such \( n \)-gram candidate is “unacceptable” otherwise the \((n-1)\)-grams are the “unacceptable” ones. When all the candidates have been tested and a list of acceptable \( n \)-grams has been found the algorithm goes through that list in decreasing length and count order introducing spaces before and after each acceptable \( n \)-gram found within the chunk (see Fig. 2).

During this later stage every segment found within the chunks is added to a “pre-lexicon” if the segment appears twice or more times in the whole segmented text. This pre-lexicon is used afterwards to segment again each chunk. This process is analogous to the one applied with acceptable \( n \)-grams, that is, traversing the lexicon in decreasing length and count order. This new segmentation phase can also introduce new entries to the lexicon. Thus, the final output of this algorithm includes both a segmented Chinese document and a tentative lexicon. Some preliminary results are provided below. It must be said, however, that by its own nature this algorithm does not always provide valid segments according to Wu’s Monotonicity Principle (Wu, 1998) and that the built lexicons, although quite accurate, usually include not only “words” but also complex sentences. A way to overcome these problems is discussed later while the self-segmentation algorithm is provided in the Appendix.

Preliminary results

Thorough testing of this algorithm over standardized corpora is needed but preliminary results are highly encouraging. Applied over five web pages\(^1\) from the Chinese version of BBC News website it has been found that (in average): (1) around 70% of the entries within the tentative lexicons appear in the CEDICT\(^2\) lexicon, (2) around 83% of the lexicon entries actually appearing within the news items also appear in the CEDICT, and (3) at least 37% of the lexicon entries from the news items are correctly segmented since their translations occur literally in the English versions of the original web pages. Table 1 provides more details about these claims while Table 2 shows some entries from the tentative lexicons built for those short documents and Figure 3 includes some selected sentences (headlines, leads and captions) comparing their self-segmentation with results produced by native speakers.

Table 1. Whole text from five web pages was self-segmented. These are the number of entries from the web pages that appear in CEDICT dictionary, the number of entries which are actually from the news items and appear in CEDICT and the number of literal translations of Chinese segments into English words that appear in the English version of the same news items.

---

\(^1\) Arafat in France: http://news.bbc.co.uk/chinese/simp/hi/newsid_3960000/newsid_3963900/3963989.stm  
British Gov. and Linux: http://news.bbc.co.uk/chinese/simp/hi/newsid_3960000/newsid_3961800/3961891.stm  
Bush Website: http://news.bbc.co.uk/chinese/simp/hi/newsid_3960000/newsid_3963800/3963879.stm  
Buttiglione: http://news.bbc.co.uk/chinese/simp/hi/newsid_3959000/newsid_3959100/3959187.stm  
EU New Constitution: http://news.bbc.co.uk/chinese/simp/hi/newsid_3960000/newsid_3964000/3964071.stm

\(^2\) CEDICT is a public-domain Chinese-English dictionary started by Paul Denisowski and now available at http://www.mandarintools.com/cedict.html
Table 2. Tentative lexicons built by self-segmenting previously mentioned web pages. Only entries appearing twice or more times in the news items and with their corresponding translations occurring in English version are shown. Entries which are not available in CEDICT are shown in bold.

Fig. 3. Some headlines, leads and captions self-segmented within previous web pages. The algorithm results are compared with human-made segmentations. Matches between both ones are noteworthy, specially taking into account that the text used by the algorithm is so scarce (around 800 characters each document).

Discussion
As it can be seen from above preliminary results the segmentation achieved by this new algorithm is similar to that produced by human beings. In contrast to state-of-the-art segmenters, which rely on lexicons and statistical information extracted from external corpora, the proposed algorithm simply uses measures obtained from the input text and not only does not employ external lexicons but can be used to build Chinese lexicons from scratch. Nevertheless, by using short Chinese documents some sentences are not totally segmented and thus some entries from the built lexicons are not “words” but complex statements.

In order to apply this algorithm to develop an accurate Chinese lexicon such issues can be overcome by finding (e.g. using search engines) unseen documents containing one or more “promising”\(^3\) entries from the tentative lexicon. This new documents can be self-segmented obtaining reinforcement for some entries within the lexicon in addition to new entries. Table 3 shows some reinforced and new lexicon entries obtained from a USENET post reached by using the keyword 阿拉法特 (Ālāfātè, Arafat).

Table 3. A new lexicon built from a post published in talk.politics.china on May 28\(^{th}\) 2002. Only entries appearing twice or more times in the post are shown. New entries to CEDICT are shown in bold while “reinforced” entries (i.e. entries appearing both in original and new document) appear with grey background.

Conclusion
An extremely simple algorithm to perform self-segmentation of Chinese documents has been described. This algorithm does not use either external lexicons or statistical information apart from the input data. Nevertheless, it provides segmentation results comparable to those achieved by native speakers and, in addition to this, rather accurate lexicon entries. Such algorithm can be applied to build Chinese lexicons from scratch beginning with some seed documents and relying on search engines to find new unseen but related documents which will provide reinforcement for some of the entries besides increasing the lexicon size with new entries. The algorithm could also be applied by other segmentation techniques as a preprocessing stage since it is quite robust when facing proper nouns.

Acknowledgments
The author wishes to thank Luo Li and Ho-Meng Chang Shu by their help with the original Chinese texts.

References
Chen, K.J, and Liu, S.H. (1992) Word Identification for Mandarin Chinese Sentences. In Proc. of COLING-92, pp. 101-107.
Fu, G., and Luke, K.K. (2003) A two-stage statistical word segmentation system for Chinese. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 156-159.
Maosong, S., Dayang, S., and Tsou, B.K. (1998) Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. Proceedings of the 36\(^{th}\) Annual Meeting of the ACL, pp. 1265-1271.
Sproat, R., and Shih, C. (1990) A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4, pp. 336-351.

---

\(^3\) The longest and the most frequent the entry the most promising it is.
Appendix

Algorithm selfSegmentation (chineseChunks)
Input: Chinese text split into the list chineseChunks using roman characters, numbers and punctuation marks as break-points
1. from n ← 2 to MAX_NGRAM_SIZE do
2.   for each n-gram ngram found in chineseChunks do
3.     weights(ngram) ← count(chineseChunks, ngram)
4.   loop
5. loop
6. for each chunk chunk in chineseChunks do
7. if length of chunk > 2
8.   from n ← 2 to length of chunk do
9.     for each n-gram ngram found in chunk do
10.    count ← weights(ngram)
11.    if count > 1
12.      candidates(n)(ngram) ← count
13.    end if
14.  loop
15.  loop
16. if there exist candidates
17.   from n ← size of candidates+1 to 2 do
18.     for each candidate ncandidate in candidates(n) do
19.       for each candidate n1candidate in candidates(n-1) do
20.         if n1candidate found in ncandidate
21.             if weights(n1candidate) > weights(ncandidate)
22.                ncandidate is unacceptable
23.             else
24.                n1candidate is unacceptable
25.           end if
26.       end if
27.     end if
28.   loop
29. loop
30. from n ← size of candidates+1 to 2 do
31.   for each candidate ncandidate in candidates(n) do
32.     if ncandidate is not unacceptable
33.       sortedInsert (acceptables, ncandidate)
34.     end if
35.   loop
36. loop
37. segmentedChunk ← chunk
38. for each candidate candidate in acceptables do
39.    segmentedChunk ← replace(candidate, <BLANK>candidate<BLANK>, segmentedChunk)
40. loop
41. chunk ← segmentedChunk
42. lexiconEntries ← explode(<BLANK>, chunk)
43. for each entry entry in lexiconEntries do
44.    lexicon(entry) ← lexicon(entry)+1
45. loop
46. end if
47. end if
for each entry entry in lexicon do
    if lexicon(entry) = 1
        drop lexicon(entry)
    end if
loop
sort lexicon by descending length and count order
for each chunk chunk in chineseChunks do
    segmentedChunk ← chunk
    for each entry entry in lexicon do
        segmentedChunk ← replace(entry, <BLANK>entry<BLANK>, segmentedChunk)
    loop
    chunk ← segmentedChunk
    lexiconEntries ← explode(<BLANK>, chunk)
    for each entry entry in lexiconEntries do
        lexicon(entry) ← lexicon(entry)+1
    loop
loop
for each entry entry in lexicon do
    if lexicon(entry) = 1
        drop lexicon(entry)
    end if
loop
sort lexicon by descending length and count order
return lexicon
Input data: 英国外相斯特劳对BBC说，英国可能将会在2006年举行公民投票。

Split data:
1. 英国外相斯特劳对
2. 说
3. 英国可能将会在
4. 年举行公民投票

Input data: 英国外相斯特劳对

Candidate n-grams:

| n-gram | Weight in whole document |
|--------|--------------------------|
| 斯特劳 | 2                        |
| 英国   | 3                        |
| 斯特   | 2                        |
| 特劳   | 2                        |

Is 斯特劳 acceptable?

| n-gram | 斯特 | weight |
|--------|------|--------|
| 1st constituent | 斯特 | weight |
| 2nd constituent | 特劳 | weight |

Yes, 斯特劳 is acceptable

Acceptable n-grams:

| 斯特劳 | 2 |
| 英国   | 3 |

Segmentation steps:

0. 英国外相斯特劳对
1. 英国外相 斯特劳 对
2. 英国 外相 斯特劳 对

Output: 英国 · 外相 · 斯特劳 · 对

Naïve literal translation: English · Foreign Minister · Straw · to

| News item | Total entries found in CEDICT | Entries from news items found in CEDICT | Translations of segments found in English version |
|-----------|-------------------------------|------------------------------------------|-----------------------------------------------|
| Arafat in France | 54/77 (70%) | 40/48 (83%) | 15/48 (31%) |
| British Government Linux | 45/64 (70%) | 34/37 (92%) | 17/37 (46%) |
| Bush Website | 42/63 (67%) | 28/33 (85%) | 14/33 (42%) |
| Buttiglione | 81/114 (71%) | 67/85 (79%) | 21/85 (25%) |
| EU New Constitution | 57/82 (70%) | 40/49 (82%) | 17/49 (35%) |
| Average | 279/400 (70%) | 209/252 (83%) | 98/252 (37%) |
### Arafat Emergency Treatment France

| Arabic   | English           |
|----------|-------------------|
| أرافات | Arafat (Yasser)   |
| فرنسا  | France            |

| Palestinian | English |
|-------------|---------|
| بلالستن   | Palestine |
| رام الله   | Ramallah |
| اسرائيل   | Israel   |

| Chinese   | English       |
|-----------|---------------|
| 治疗      | to treat / to cure / (medical) treatment / cure |
| 医院      | hospital      |
| 医生      | doctor        |
| 星期      | week          |
| 巴黎      | Paris         |
| 血液      | blood         |
| 否认      | to declare to be untrue / to deny |
| 紧急      | urgent        |
| 总理      | premier / prime minister |
| 领导      | lead / leading / to lead / leadership |

### British Government and Linux

| Chinese   | English       |
|-----------|---------------|
| 政府       | government    |
| 英国       | England / Britain / English |
| 软件       | (computer) software |
| 使用       | to use / to employ / to apply / to make use of |
| 微软       | Microsoft     |
| 中国       | China / Chinese |
| 公司       | (business) company / company / firm / corporation / incorporated |
| 产品       | goods / merchandise / product |
| 机构       | organization / agency / institution |
| 减少       | to lessen / to decrease / to reduce / to lower |
| 系统       | System        |
| 硬盘       | Windows (operating system) |
| 免费       | free (of charge) |
| 国家       | country / nation |
| 安装       | install / erect / fix / mount / installation |
| 开发       | exploit (a resource) / open up (for development) / to develop |
| 依赖       | to depend on / to be dependent on |
| 国家       | country / state / nation |

### EU New Constitution

| Chinese   | English       |
|-----------|---------------|
| 成员国    | member country |
| 立陶宛    | Lithuania     |
| 联盟      | European Union |
| 英国       | England / Britain / English |
| 宪法       | constitution (of a country) |
| 成员       | member        |
| 欧洲       | Europe / Europen |
| 罗马       | Rome / Roman  |
| 领袖       | leader        |
| 条约       | treaty / pact |
| 议会       | parliament / legislative assembly |
| 主席       | chairperson / premier / chairman |
| 举行       | to hold (a meeting, ceremony, etc.) |
| 联盟       | alliance / union / coalition |
| 仪式       | ceremony      |
| 国家       | country / state / nation |
| 新         | new / newly   |

### Barroso Butiglione Problem

| Italian   | English       |
|-----------|---------------|
| 正在       | Barroso (Jose Manuel) |
| 意大利    | Italy / Italian |
| 议员       | member of a legislative body / representative |
| 提名       | nominate      |
| 主席       | chairperson / premier / chairman |
| 名单       | list (of names) |
| 撤回       | withdraw      |
| 否决       | veto          |
| 接受       | to accept / to receive |
| 人选       | person(s) selected (for a job, etc.) |
| 投票       | to vote / vote |
| 政治       | politics / political |
| 原机       | crisis        |
| 部长       | head of a (government, etc) department / section chief / section head / secretary / minister |
| 派出       | to withdraw / to abort / to quit |
| 争议       | controversy / dispute |
| 同意       | to agree / to consent / to approve |

### Bush Campaign Website Blocked

| Chinese   | English       |
|-----------|---------------|
| 发言人     | spokes person |
| 竞选       | run for (electoral) office / campaign |
| 布什       | Bush (George) |
| 网站       | network station / node / website |
| 境外       | outside (a country's) borders |
| 封锁       | blockade / seal off |
| 美国       | America / American / United States / USA |
| 官方       | official / (by the) government |
| 总统       | president (of a country) |
| 原因       | cause / origin / root cause / reason |
| 发表       | to issue (a statement) / to publish / to issue / to put out |
| 安全       | safe / secure / safety / security |
| 网络       | computer, telecom, etc.) network |
| 攻击       | to attack / to accuse / to charge / (military) attack |
| 导致       | to lead to / to create / to cause / to bring about |
| 许可       | to allow / to permit |
| 露出       | to leak out / to divulge / to reveal |
Alaouite visits France for emergency medical treatment.

Arafat receives urgent medical treatment in France.

Arafat's campaign chief透露 that Bush's official campaign website was blocked to overseas visitors due to "security" reasons.

Overseas visitors saw "You are not authorized to view this page".

The president designate of the European Commission Barroso withdrew the list of nominated EU commissioners, Barroso's decision was taken on the eve before the voting about this issue to be held today in the European Parliament.

Buttiglione refused to quit from the nomination.

Leaders of the 25 EU countries signed the new Constitution in Rome.

A report from the British Office of Government Commerce considers that: using LINUX products can largely reduce costs, therefore it's "feasible" the use of LINUX products by British Government institutions.

A report from the British Office of Government Commerce considers that: using LINUX products can largely reduce costs, therefore it's "feasible" the use of LINUX products by British Government institutions.
| Term            | Translation            |
|-----------------|------------------------|
| 巴勒斯坦       | Palestine              |
| 领导人          | leader                 |
| 犹太           | Judea                  |
| 签字           | to sign (a signature)  |
| 武器           | weapon / arms          |
| 犹太           | Judea                  |
| 签字           | to sign (a signature)  |
| 武器           | weapon / arms          |
| 领导人          | leader                 |
| 犹太           | Judea                  |
| 签字           | to sign (a signature)  |
| 武器           | weapon / arms          |
| 领导人          | leader                 |
| 犹太           | Judea                  |
| 签字           | to sign (a signature)  |
| 武器           | weapon / arms          |
| 领导人          | leader                 |
| 犹太           | Judea                  |
| 签字           | to sign (a signature)  |
| 武器           | weapon / arms          |