Experiments on Unsupervised Chinese Word Segmentation and Classification

Jin Hu Huang* David Powers

(Flinders University, Adelaide SA5042, Australia)
E-mail: {jin.huang, powers}@ist.flinders.edu.au

Abstract: There are several problems encountered for Chinese language processing as Chinese is written without word delimiters. The difficulty in defining a word makes it even harder. This paper explores the possibility of automatically segmenting Chinese character sequences into words and classifying these words through distributional analysis in contrast with the usual approaches that depends on dictionaries.
Key words: Unsupervised learning, Word segmentation, Word classification

Introduction

There is no explicit word boundary in Chinese text. Chinese orthography fails to represent word boundaries. The definition of a word is very important in Chinese language processing as these standards result in different segmentations and classifications. Chinese words are not inflected with respect to tense, case, person and number. As a result, Chinese word segmentation and classification is more difficult.

A given word, in a given syntactic context, has a grammatical role. A word such as "惩罚" in Chinese can be translated into punish/punishes/punished/punishing/punishment in English. We cannot tell the grammatical role solely based on the word: have to look beyond the word. In other words, we must observe how the given word functions in its given context.

Zellig Harris [1] proposed the substitutability of linguistic entity from the same class. Several researches [2,3,4] have worked on learning grammatical properties of words on English. Languages with a less overt morphology like Chinese may be simpler to analyze than English since with fewer tokens per type, there is less data on which to base a categorization decision. For example, a noun or pronoun can be substituted in a Chinese sentence without regard to its number or gender. The verb remains unchanged whether the noun is singular or plural as the substitution is made possible by the fact that Chinese verbs are not conjugated.

In our experiments we explored the possibility of segmenting Chinese character sequences into words and classifying them by their syntactic distribution.

1. Word Segmentation

Unlike English, Chinese words comprise several characters without delimiters, typically two, three, or four characters. Many characters can even stand alone as words in themselves. In order to classify the words, first we need segment the corpus into words. The problem of finding words in Chinese is analogous to the problem of identifying collocations in English, such as "put up with" or "object oriented".

* Supported by Flinders Postgraduate Research Scholarship

Several approaches have been developed for Chinese word segmentation. In general two main approaches are widely used: the statistical approach [5,6,7,8,9] and rule-based approach [10,11]. Some statistical approaches are based on the mutual information [5], which only captures the dependency among characters of a word. Some need large pre-annotated corpus for training [6,9], which is too expensive to prepare at present. Rule-based approaches require a pre-defined word list (dictionary, or lexicon). The coverage of the dictionary is critical for these approaches. Many researches use a combination of approaches [12]. Only these [5,7,8] are the most similar in aim to our work as they are also unsupervised and others are supervised and depend on hand segmented examples or dictionaries.

It has been long known that contextual information can be used for segmentation [13]. Dai, et al [14] used weighted document frequency as contextual information for Chinese word segmentation. Zhang, Gao and Zhou [15] used the context dependency for information retrieval – finding terms [words in a corpus or collocations]. Others used contextual entropy for unknown Chinese words identification [16] and automatic lexical acquisition [17]. Hutchens and Alder [18] and Kempe [19] used the contextual entropy to detect the separator in English and German corpus.

$$H(x_1, x_2) = -\sum_{x_3 \in \Sigma} p(x_3 \mid x_1, x_2) \log_2 p(x_3 \mid x_1, x_2) \quad (1)$$

$$H(x_2, x_3) = -\sum_{x_1 \in \Sigma} p(x_1 \mid x_2, x_3) \log_2 p(x_1 \mid x_2, x_3) \quad (2)$$

We use (1) and (2) to calculate the right and left contextual entropy. This entropy measures the uncertainty about the next symbol x_3 after having seen the left context x_1x_2. We call it contextual entropy. It will be low if one particular symbol is expected to occur with a high probability. Otherwise it will be high if the model has no “idea” what kind of symbol will follow the context. Across a word boundary there is a significant increase in the contextual entropy as we are not sure what kind of character will appear after a word boundary.

Most Chinese words are two characters. A bi-directional 2nd order Markov model is thus effective to detect the word boundaries for Chinese word segmentation. To find Chinese words, we look for character sequences that are stable in the corpus. The components of a word are strongly correlated but appear in various contexts in the corpus. Contextual entropy among components of a word is low. High entropy appears at word boundaries.

The process of segmenting the text is how to find the word boundaries. Across a word boundary there is a significant increase in contextual entropy and decrease in mutual information [20]. We apply the following algorithms to determine whether there is a word boundary between C and D for a string ABCD:

- If the entropy value, given the left context AB or the right context CD is at a crest, there is a boundary between AB and CD.
- If the sum of both entropy values give the left context AB or the right context CD is great than a certain threshold (optimally 9), there is a boundary between AB and CD.
- If the mutual information of BC is less than a certain threshold (optimally 3) compared with both mutual information of AB and CD, there is a boundary between AB and CD.

2. Word Classification

Part of speech is not well defined in Chinese and a dictionary can not include all words and all conceivable usages. Although Yu [21] built the grammatical knowledge-based dictionary it only includes 70,000 words at this stage. And building such a dictionary is time consuming without machine assistance. Brown [22] argued that letting the machine infer the classes rather than relying on dictionaries or other human-derived artifact may result in more robust systems.

Most approaches in the previous work in English [2,3,4] classify words instead of individual occurrences. They make the assumption that any given word will only belong to one category. Given the widespread part-of-speech ambiguity of words in Chinese this is problematic. Chang and Chen [23] make the same assumption. They present a method to classify Chinese words automatically into a predetermined number of categories. They try to find a class assignment for word that maximizes the probability of a corpus. A bigram approximation of this would state that the estimated probability of the text T of length L is modeled as the product over each word of the probability of *w* given the inferred class along with probability of given the previous.

$$P(T) = \sum_{i=1}^{L} p(W_i \mid C_i)\, p(C_{i-1} \mid C_i)$$

They optimize the class assignment of words so that probability the text is maximized using a simulated annealing approach.

We adopt the substitutional approach in the experiments. Words of the same class words are syntactically substitutable although they may not be appropriate semantically. Words in Chinese can frequently function in more than one lexical class. But in a given sentence, it is usually clear how a word is functioning. We construct the environments of each target word with respect to its left and right context. For a context "深化~改革", "足球 职称 征管 这项 渔村 训练 新闻 卫生 铁路 体制 体育 税制 水利 商业 企业 配套 农业 农村 内部 林业 科技 经济 金融 教育 教学 价格 机构 各项 高校 高教 党校 出版 城市 财政 财税 部队" can be substituted by each other in this context. Sun [24] suggests that the best watch window in collocation extraction is [-2, +1], [-3, +4], [-1, +2] for noun, verb and adjective respectively in Chinese. We use two words on each side of the concept as an environment for classification.

We use the following algorithms for word classification:

1. Segment raw text
2. Collect 5-grams for words
3. Produce context-word sets
4. Group words according to identical context and produce initial word classes
5. Merge word classes according to overlap
6. Replace the words with word classes in the contexts
7. Repeat 4,5,6 until no more classes can merge

3. Experiments & Future Work

We trained the bi-directional 2nd order Markov model on 220MB corpora mainly news from People Daily (91-95) using suffix array [25]. We stored both contextual entropy and mutual information for the bigrams with positive value. In order to evaluate our algorithm, we tested it on Penn Treebank Tagged Chinese Corpus including 325 articles from People Daily (94-98). We used recall and precision to measure our performance both on discovering word boundaries and words. A word is considered correctly segmented only if there is a word boundary in front of and at the end of the

word and these is no boundary among the word. We achieved 93.2% precision with 93.1% recall on discovering word boundaries and 81.2% precision with 81.1% on discovering words, although it should be noted that there is poor agreement on word segmentation amongst human annotators and at least three relative widespread conventions [26,27,28]. Our results should be lower than those judged by hand (which can bias judgements) and tested on non-standard corpora.

Peng and Dale [7] used successive EM phases to learn a probabilistic model of character sequences and pruned the model with a mutual information selection criterion. They achieved 75.1% precision with 74.0% recall on discovering words by repeatedly applying lexicon pruning to an improved EM training. Their results are tested on the same corpus as ours. Sproat [5] obtained 94% precision and 90% recall but only considered the correctness of two-character words. Fu and Wang[8] achieved 99.25% accuracy (recall) but did not provided precision.

Most errors caused by our approaches relate to numbers and names. As in the training corpus, numbers are written in full-width alphabetic number but in the test corpus numbers are written in Chinese character. We did not obtain enough statistics to predict the boundary. Chinese names and foreign names are difficult to process for all purely statistical approach. Both numbers and names appear quite frequently in the test corpus. The other major class of error is compound nouns. We segmented “开发区” as “开发/区”. But note that there is no standard definition for Chinese words. The segmentation in Penn Treebank Chinese tagged corpus is also sometimes debatable and the different conventions differ in treatment of complex nouns. An example in English of the arbitrariness is “cannot” versus “do not”.

We applied some rules for our model to complement our purely statistical model. We treated adjacent numbers and single word markers as a word. We obtained segmented corpus with about 96% precision. We applied suffix array again to collect 5-gram word strings on the segmented corpus. We merged two classes if two classes overlap 75%. If the adjacent words are both common to two contexts, we merge even if there is only 50% overlap. From experiments we get about 2000 classes and cannot merge again. Following shows some of our results.

In our approach we will get a large number of classes. One word maybe belongs to dozens of classes if it has a range of meanings and usages. How to merge the class is critical in our approach. Many sub-classes embody finer distinctions among them. Some rare words are difficult to classify because of lack of distributional evidence. Bad word segmentation also contributes some errors.

We are currently exploring how to merge the classes more efficiently and reduce the classes as well as the possibility of learning phrase structure through distributional information.

元^人次^人^平方米^亩^名^美元^件^家^公斤^个^吨 (classifier)
进一步^继续向前^继续^积极^不断^进一步^积极^大力 (Adv)
及其^及^和^各^等^的^ (Conj)
至于^因为^在^由^因^为^据^尽管如^鉴于^对^从 (Prepositions)
越^很^相当^十分^日益^更为^更加^比较 (Adv)
７^６^５^４^３０^３^２^１００^１０^１^一^五^四^三^七^六^九^二^八 (Number)
佟志广^邹家华^专家们^朱基^这就是^张震^张万年^张思卿^于永波^有人^叶选平^叶利钦^杨主席^杨尚昆^杨福昌^薛驹^宣言^徐惟诚^徐敦信^消息^西哈努克^吴作栋^吴仪^吴学谦^吴建民^吴邦国^文章^文件^温家宝^尉健行^王忠禹^王震^王兆国^王学贤^王汉斌^王丙乾^汪道涵^万里^通知^通报^田纪云^陶驷驹^她^他强调^他们^他^俗话^苏哈托^宋平^宋健^声明^沈国放 (Name)
组织^主管^种子^执法^植保^职能^政治^政府^政法^渔政^有关^邮电^医药^宣传^行政^刑侦^信访^物价^武装^文化^卫生^土地^统战^统计^体育^体改^司法^税务^水利^审计^涉农^商业^

人武^人事^侨务^气象^农资^农业^农牧^农经^民政^旅游^领导^林业^粮食^劳动 (Department)
祝贺^致意^支持^震惊^赞同^赞赏^赞成^忧虑^遗憾^欣慰^欣赏^谢意^慰问^同意^钦佩^满意^
理解^乐观^肯定^敬意^敬佩^欢迎^怀疑^关注^关心^关切^高兴^感谢^愤慨^反对 (V)
珠海^镇江^宜昌^徐州^新疆^香港^厦门^西安^武汉^无锡^温州^天津^台州^四川^沈阳^深圳^
绍兴^上海^莆田^宁波^南京^梅州^丽水^嘉兴^济南^吉安^湖南^杭州^海南^贵阳^桂林 (City)
组建^组成^展开^展出^运转^运营^运行^营业^议案^移交^邀请^选举^宣布^形成^问世^推出^
投产^通航^通车^停火^谈判^实行^实现^实施^施行^生效^设立^上任^上岗^确认^签字^签约^
签署^签订^启用^启动^批准^落成^立项^抗议^开诊^开业^开通^开始^开幕^开馆 (V & N)

Conclusion

This paper presents a distributional approach for word segmentation and classification. We use contextual entropy and mutual information for word segmentation. Mutual information captures the dependency inside the word. Contextual entropy captures the dependency with the contexts in which the word occurs. Then we classify words according to their occurrence statistics rather than the knowledge of stored in a dictionary. In addition our approach is unique in that words are allowed to belong to multiple independent classes.

Reference:

1. Zellig Harris, 1951, Structural linguistics, Chicago: University of Chicago Press.
2. Steven Paul Finch, 1993, Finding structure in language, PhD thesis, University of Edinburgh.
3. Eric Brill, 1993, A Corpus-Based approach to language learning, PhD thesis, University of Pennsylvania.
4. David Powers, 1997, Unsupervised learning of linguistic structure: an empirical evaluation, Int'l Journal of Corpus Linguistics 2#1:91-131.
5. Richard Sproat, Chilin Shih, 1990, A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4, 4.
6. Richard Sproat, Chilin Shih, William Gale and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics, 22(3).
7. Fuchun Peng and Dale Schuurmans, 2001, Self-supervised Chinese word segmentation, In F. Hoffman et al. (Eds.): Advances in Intelligent Data Analysis, Proceedings of the Fourth International Conference (IDA-01), Cascais, Portugal.
8. Fu Guohong and Wang Xiaolong, 1999, Unsupervised Chinese word segmentation and unkown word identification, Proceedings 5th NLPPRS, Beijing, China.
9. W J Teahan, Y Wen, R McNab and I Witten, 2000, A compression-based algorithm for Chinese word segmentation, Computational Linguistics, 26, 3.
10. C. L. Yeh and H. J. Lee, 1991, Rule-based word identification for mandarin Chinese sentences – a unification approach, Computer Processing of Chinese and Oriental Languages, Vol. 5, No. 2.
11. Swen Bing and Yu Shiwen, 1999, A graded Approach for the efficient resolution of Chinese word segmentation ambiguities, Proceedings 5th NLPPRS, Beijing, China.
12. J Y Nie, W Y Jin and M L Hannan, 1994, A hybrid approach to unknown word detection and segmentation of Chinese. ICCC'94. Singapore.
13. Zellig Harris, (1955) From phoneme to morpheme. Language, 31(2), 1955.
14. Yu Bin Dai, C. Kgoo and T. Loh, 1999, A new statistical formula for Chinese text segmentation incorporating contextual information, SIGIR'99 Berkley, CA USA

15. Jian Zhang, Jianfeng Gao, Ming Zhou, 2000, Extraction of Chinese compound words – an experimental study on a very large corpus. The second Chinese Language Processing Workshop attached to ACL2000, Hong Kong, October 8, 2000.
16. C. H Tung and H J Lee, 1994, Identification of unknown words from a corpus. Computer Processing of Chinese & Oriental Languages, Vol. 8 (supplement).
17. Chang, J. S, Lin Y. C. and Su, K. Y. A, 1995, Automatic construction of a Chinese electronic dictionary. Proceedings of the Third Workshop on Very Large Corpora.
18. Jason L. Hutchens and Michael Alder, 1998, Finding structure via compression. In D. Powers, ed., NeMLap3/CoNLL98, Sydney, Australia, 1998.
19. Andre Kempe, 1999, Experiments in unsupervised entropy-based corpus segmentation, Ninth Conference of the European Chapter of the Association for Computational Linguistics' 99 Workshop, 12th June 1999, Bergen, Norway.
20. D. Magerman and M. Marcus. 1990. Parsing a natural language using mutual information statistics, In Proceedings Eighth National Conference on Artificial Intelligence (AAAI 90).
21. Yu Shiwen, Zhu Yunfeng, Wangfei, Zhang Yunyun, 1998, The grammatical knowledge-base of contemporary Chinese, Hsinghua University and Guangxi Press. (Chinese)
22. Peter Brown, Vincent Dlla Pietra, Peter deSouza, Jenifer Lai and Robert Mercer, 1992, Class-based n-gram models of natural language, Computational Linguistics, 18(4).
23. Chang Chao-Huang and Chen Cheng-Der, 1994, A study on Corpus-based Classification of Chinese Words, International Conference on Chinese Computing '94, Singapore, June 1994.
24. Sun Honglin, 1998, Distributional property of collocations in the texts, Chinese Information Processing International Conference. (Chinese)
25. Mikio Yamamoto and Kenneth Church, 2000, Using suffix array to compute term frequency and document frequency for all substrings in a corpus, Computational Linguistics, Volume 27, Number 1.
26. GB/T 13175-92, 1993, Contemporary Chinese language word-segmentation specification for information processing, Technical report, Beijing.
27. Xia Fei, 1999, Segmentation guideline, Chinese Treebank project. Technical report, University of Pennsylvania.
28. Huang Chu-Ren, Chan Keh-jiann, Chang Lili and Chen Feng-yi, Segmentation standard for Chinese natural language processing. International Journal of Computational Linguistics and Chinese Language Processing, 2(2):47-62.

Authors:

Huang Jin Hu, Phd student, Main research interests are natural language processing, data mining, information retrieval.
David Powers, Associate Professor, Main research interests are natural language processing, information retrieval, artificial intelligence, machine learning.

无监督下的词切分和词分类的试验

黄金虎　大卫
南澳弗林德斯大学
{jin.huang, power} @ist.flinders.edu.au

摘要：中文书写没有分隔符给中文语言处理带来很大的困难，词难下定义给它带来更大的麻烦，本文与常规依靠词典的方法相反，探讨通过对词分布的分析来自动进行词切分和词分类的可能性。

关键词：无监督下学习；词切分；词分类