A Survey of Open-source Chinese Word Segmentation Tools

|  |  |
| --- | --- |
| First Author  Affiliation / Address line 1  Affiliation / Address line 2  Affiliation / Address line 3  email@domain | Second Author  Affiliation / Address line 1  Affiliation / Address line 2  Affiliation / Address line 3  email@domain |

Abstract

This paper acts as a survey of existing open-source parsers for Chinese Word Segmentation (CWS). We applied several of these systems on Chinese Treebank v8.0 and Traditional Chinese Universal Dependencies Treebank v2.0. We give a quantitative evaluation of each parser. We present those results as well as an analysis of the types of errors persistent in the remaining percent margin and conclude that there exist certain types of context-resolvable errors that some kinds of yet untried post-processing systems might be well-suited to handle.

1. Introduction

Mandarin Chinese, whether penned or typed, does not delimit its words boundaries by spaces or any other boundary marker. Automatically deciphering those word boundaries and segmenting unrestricted text is therefore an essential preliminary task for any downstream language processing tasks.

issue of openness

showing that many of these open-source tools have reported results, but not all on the same corpus, this is an opportunity to test on the same thing

potentially make a taxonomy of errors

1. Background

“The key to accurate automatic word identification in Chinese lies in the successful resolution of […] ambiguities and a proper way to handle out-of-vocabulary words” (Xue, 2003). We will first give an overview of the main problems for segmenters, and then break down how types of systems which fall into a few important methodological categories overcome those obstacles.

* 1. Ambiguity

There are two main types of ambiguities when considering Chinese word segmentation. Lee & Huang (Chia-Ming Lee and Huang, 2013) explain each of them well:

“There are two types of ambiguities in CWS: overlapping and combinational ambiguities. They can be defined as follows, given a dictionary D and a string “abc”, if the set of substrings {ab,bc} ⊃ D, “abc” involves an overlapping ambiguity; given a dictionary D and a string {ab}, if the set of substrings {a, b, ab} ⊃ D, “ab” involves a combinational ambiguity.”

Luo et al. (Luo et al., 2002) note that combinational ambiguities (sometimes also called combinatorial or covering ambiguities) are much more difficult to resolve, as they rely on syntactic, pragmatic, and semantic information.

* 1. Out-of-Vocabulary (OOV)

Goh et al. (Goh et al., 2006) explained that most words never-before seen by a parser (i.e. not included in its lexicon) are proper nouns in the form of synthetic words. Synthetic words are those formed by combining two or more single-morpheme words (whether one-character, two-character, or transliterations) and whose meaning is derivable from its parts (Cheng et al., 2014).

In reality, because of the ever-expanding nature of language and lexicon, word lists will never be truly exhaustive. The inevitable use of proper names (often being comprised of lexical words), neologisms, and transliterations make it impossible for any segmenter to anticipate every word in running text. There must be some kind of work-around.

* 1. Types of Systems

Approaches to CWS can generally be categorized by looking to the following distinctions.

* + 1. Statistical vs Rule-based

First, in approaching Chinese word segmentation, there is the option to use statistical information or language-specific knowledge hand-coded into rules to inform the segmentation. With advances in computing power, statistical/probabilistic models quickly outpaced rule-based systems (Chang, 2007). There are instances of statistical systems incorporating linguistic rules (EX).

* + 1. Word-based vs Character based

Furthermore, within the statistical framework, a decision must also be made to use either characters or words as the basic unit. Word-based systems involve a range complexity from maximum matching to semi-Markov combined with Conditional Random Fields (CRFs). One major difficulty for word-based systems is that because a complete word-list is infeasible, OOV words will always be a hindrance. Chinese characters, however, are a relatively closed set (Chia-Ming Lee and Huang, 2013), which means that making inferences at that stage has the potential for higher accuracy, even for unknown words. At the character level, a common approach is to treat CWS as a sequence labeling task (Xue, 2003), enabling machine learning models such as Maximum Entropy (ME) and Conditional Random Fields (CRF) to be applied. Although it is possible for a segmenter to rely solely on weighted probabilities created by training on a corpus, many systems incorporate labels at the character-level of known information such as part-of-speech or character-position within a word. Systems can go even further to incorporate word-based and character-based information (Sun et al., 2009; Wang et al., 2014)

* + 1. All-at-once vs One-at-a-time

Segmentation of Chinese text is certainly a foundational task to many more advance language processing tasks; however, part-of-speech tagging can be concurrently attempted with and even inform segmentation. Segmenters that simultaneously perform this and other tasks are known as one-at-a-time or ensemble systems. Alternatively, many parsers function as stand-alone tools within a pipeline of language processing. (Sun et al., 2009) put forth a well-thought comparison and evaluation of these two approaches, further contextualizing their taxonomy within the character- or word-based framework. Although ensemble approaches consistently outperform pipeline models, they do have the downside of taking much longer to train as well as utilize. Ensemble models have the benefit of using tags at the character level to perform better on unknown words.

* 1. Domain Adaptation

When segmenters are trained on certain types or genres of text, it naturally follows that they perform better on segmenting similar texts. Technical jargon and scientific terms present a real problem for those kinds of systems. If the goal is to CWS technology to be advanced to such a robust degree that input genre is not a constraint, then domain adaptation must be taken into account.

Li and Xue (2016) annotated Chinese patent texts, discovering that even a small amount of target-domain training data is more effective in increasing performance compared to a large amount of out-of-domain data. Also addressing the issue of low-resource datasets, Xu et al. (2017) implemented a deep-stacking neural networks framework.

* 1. Bake-offs

Over the past 15 years, SIGHAN, sometimes in coordination with CIPS, has sponsored bake-offs in CWS. More recently, NLPCC has run shared-tasks for evaluation and comparison among systems. The datasets used in these shared tasks vary in genre, from newswire to microblog to wiki.

Place licence statement here for the camera-ready version. See Section 3.9 in the instructions for preparing a manuscript.

* + 1. SIGHAN Bake-offs

In 2003, the ACL-SIGHAN sponsored the 1st International Bake-off for CWS, sparking much progress in the field. The question of domain adaptation was made apparent as the evaluations showed that systems performed well on genres they were designed for, with no single segmenter performing best across domains (Sproat and Emerson, 2003). ]

In the 2nd bake-off in 2005, Stanford NLP brought to the table a CRF model that operates at the character level and incorporates linguistic features (Tseng et al., 2005).

These bake-offs consisted of almost entirely newswire datasets.

* + 1. NLPCC Shared-tasks

Different from the SIGHAN bake-offs, NLPCC sponsored a shared task in CWS with a focus on microblog data. In 2015, Boson NLP showed top performance with their ensemble approach (segmenting & tagging) based on a CRF model.

In 2016, Beijing University of Posts and Telecommunications came out on top by using a type of neural network called Long Short-Term Memory (LSTM) to included long-range dependencies in their inference.

* 1. Openly-available Systems

Many of the tools reported on in the literature are not freely available for

* + 1. FNLP

Fudan University

* + 1. Jieba1

This system employs a word-based statistical model, with a Hidden Markov Model for OOV detection.

* + 1. ZPar

Uses word-based perceptron.

* + 1. Nivre

Uses bidirectional RNN with CRF

* 1. Evaluation Metrics

The traditional metrics in evaluating the performance of segmenter are precision, recall, and F-measure. Recently, a psychometric-inspired metric has been proposed (Qian et al., 2016) and was even adopted in the NLPCC 2016 Shared-task. Due to increasing difficulty in comparing top systems at the highest level, the new metric is weighted at the word level according to difficulty of segmentation as determined by a committee of segmenters.

* 1. Our goal/this work in the context just given

We are interested in surveying several of these available open-source segmenters in order to put forth possible next steps that might advance the state of the art for collaborative approaches to CWS. Current top-performing systems in the literature are using character-based, ensemble, neural network approaches, and this agrees with the present evaluation of openly-available models.

1. Methodology

We used the double annotated gold standard files included with CTB v8.0 as our test set, training the models that require it on the rest of the corpus.

* 1. Preparing the data

We …

* 1. Evaluator

We created a simple evaluation tool for use in comparing these models. It works by first transforming both the gold standard segmentation and a segmenter’s predictions of the same text into a list of binary values. Specifically, at every bigram of characters moving left to right, whether or not they are separated or combined is notated as “s” or “c” and appended to a list. Take the following manually segmented sentence as an example:

“工程 // 近日 // 即将 // 完工.”

Gōngchéng jìnrì jíjiāng wángōng.

“The project will be completed within the next few days.”

The first bigram consists of “工” and “程”. The gold standard segmentation above would play at as follows:

[工程] => “c”

[程近] => “s”

[近日] => “c”

This process results in complete lists of “c”s and “s”s for segmenter and gold standard and finally compared to return accuracy, precision, and recall. An error is

1. Results

Something about how the systems did overall (in terms of what’s been tested on what corpora and jieba not being reported on at all). Graphs

|  |  |  |  |
| --- | --- | --- | --- |
| Parser | Accuracy | precision | recall |
| fnlp |  |  |  |
| jieba |  |  |  |
| Shao |  |  |  |
| stanford\_ctb |  |  |  |
| stanford\_PKU |  |  |  |
| zpar |  |  |  |

Table 1. Segmenters on CTB v.8.0

|  |  |  |  |
| --- | --- | --- | --- |
| Parser | Accuracy | precision | recall |
| fnlp |  |  |  |
| jieba |  |  |  |
| Shao |  |  |  |
| stanford\_ctb |  |  |  |
| stanford\_PKU |  |  |  |
| zpar |  |  |  |

Table 1. Segmenters on UD v.2.0

Explanation of charts.

1. Discussion

The advancements made it correctly segmenting unrestricted Chinese text are remarkable. High levels of accuracy positively affect many other areas of language processing. However, we have seen that within the (relatively small) margins of error that remain there is room for improvement.

You can imagine a hybrid system where the heavy lifting is done by tried and tested models and a post-processor simply handles a subtask to up the overall accuracy and performance. This may look something like DECCA’s automatic detection of errors or a constraint grammar wherein language-specific linguistic knowledge informs the resolution of certain kinds of errors (like those discussed previously.

the Ideal. With each of these, closed, etc

Publish the dataset when publish bake-off results.

1. Conclusion

Future work could investigate implementing a hybrid model which makes use of each method’s distinctive specialty.

Acknowledgements

ORCA grant from BYU.

Reference

HUAN G Chang. 2007. Chinese Word Segmentation : A Decade Revie w. :12.

Fei Cheng, Kevin Duh, and Yuji Matsumoto. 2014. Parsing chinese synthetic words with a characterbased dependency model. In *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), Reykjavik*.

Fei Cheng, Kevin Duh, and Yuji Matsumoto. Parsing Chinese Synthetic Words with a Character-based Dependency Model. :6.

Chia-Ming Lee and Chien-Kang Huang. 2013. Context-Based Chinese Word Segmentation using SVM Machine-Learning Algorithm without Dictionary Support - Semantic Scholar. *IJCNLP*.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2006. Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing. *Journal of Chinese Language and Computing*, 16(4):185–206.

Si Li and Nianwen Xue. 2016. Towards Accurate Word Segmentation for Chinese Patents. *arXiv:1611.10038 [cs]*, November. arXiv: 1611.10038.

Xiao Luo, Maosong Sun, and Benjamin K. Tsou. 2002. Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:2185–2194.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. , June.

Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun’ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. In page 56. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. :4.

Mengqiu Wang, Rob Voigt, and Christopher D. Manning. 2014. Two Knives Cut Better Than One: Chinese Word Segmentation with Dual Decomposition. In pages 193–198. Association for Computational Linguistics.

Jingjing Xu, Xu Sun, Sujian Li, Xiaoyan Cai, and Bingzhen Wei. 2017. Deep Stacking Networks for Low-Resource Chinese Word Segmentation with Transfer Learning. *arXiv:1711.01427 [cs]*, November. arXiv: 1711.01427.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.