CS 410 Fall 2021 Technology Review
Chinese Word Segmentation Toolkit - PKUSEG

University of Illinois, Urbana-Champaign
Chia-Chi, Chen

Introduction

The enriched information in human language, especially written words, makes it critical for people to mine and analyze text data. The basic unit in Chinese language is a word, which consists of one or more characters. Due to the fact that Chinese sentences are simply words concatenated with one another, it is very difficult to decompose one sentence back to individual words. Therefore, distinguishing word boundaries to perform word segmentations becomes the primary and most important step for processing Chinese text documents since the quality of word segmentation directly influences the performance of the following tasks. In this review, we will discuss a newly released toolkit PKUSEG, which targets multi-domain word segmentation in Chinese, explain some of the techniques and models behind, and compare it with the most widely used Chinese language processing toolkit jieba.

Segmentation Models

Traditional segmentation models mainly rely on the Hidden Markov Model (HMM), which is a generative model that maximizes the joint likelihood for the observed training set. There are two main issues for such a generative model. First, to define joint probability, you will need to enumerate all possible observation sequences, which in reality is not practical. Second, generative models must make very strict independent assumptions on the observations. To solve these problems, an alternative discriminative model Conditional Random Field (CRF) was proposed. CRF models the conditional probability, which is the probability of possible label sequences given an observed sequence. The conditional probability can depend on non-independent features of observations, without forcing the model to account for the distribution of those dependencies. The probability of a transition between labels may depend not only on the current observation, but also on the past and future observations as well, which allows the model to have more relaxed restrictions on the dependence of observations. PKUSEG takes the advantage of CRF by maximizing the log likelihood of the tags of the reference sequences. To further address the problem of having a ton of parameters for training, the adaptive online gradient descent based on the feature frequency information (ADF) was adapted. As a result, the learning rates are combined into a vector which enables them to be automatically adjusted according to the frequency of the parameter, forcing the features with higher frequencies to be more adequate.

toolkit	PKUSEG	jieba
model	Conditional Random Field (CRF)	Hidden Markov Model (HMM)

Domain Specific Training

Most word segmentation toolkits only provide a single coarse-grained segmentation model, which is mostly trained on news domain datasets. In real-world applications, the domain of text varies which makes it difficult to achieve high performance without domain-specific segmentation rules. PKUSEG tackles this problem in two steps. In the first step, a pre-trained model is chosen as a base model and then fine-tunes it with a mixed combination of datasets consisting of multiple domains. This large-scale, multi-domain dataset combination helps to obtain a more comprehensive coarse-grained segmentation model. On top of that, the second step is to apply domain-specific data to this coarse-grained model individually in order to further improve the model into four fine-grained domains, including news, medicine, tourism, and web. With this additional customization toward different domains, PKUSEG achieve a better performance on domain-specific datasets compare to other toolkits.

Result Comparison

There are four datasets used for comparison. MSRA and PKU are two datasets provided by the Second International Chinese Word Segmentation Bakeoff, which are both from the news domain. CTB8 is a hybrid dataset consisting of text from news and web domain. Lastly, WEIBO is a web domain dataset. To measure the performance between different toolkits, we must ensure that the comparison is fair. Since only PKUSEG provide domain-specific segmentation models, here, we compare the results base on the coarse-grained segmentation model each toolkit provides. We can see that PKUSEG outperforms jieba on all four datasets.

	MSRA	CTB8	PKU	WEIBO
PKUSEG	87.29	91.77	92.68	93.43
jieba	81.45	79.58	81.83	83.56

To show the improvements of using domain-specific segmentation models, we can compare the results between coarse-grained and domain-specific models from PKUSEG. We can tell that domain-specific models actually do a better job on these domain-specific tasks compared with the above results.

	Precision	Recall	F-score
MSRA	96.94	96.81	96.88
WEIBO	93.78	94.65	94.21

Other Features

PKUSEG provides their pre-trained coarse-grained segmentation model along with domain-specific models for you to easily use. If you know the exact domain of your dataset, using a domain-specific model would achieve a better result. PKUSEG also allows users to train a new customized model with their own dataset. This is very useful for users who are segmenting a specific dataset where its domain is not included in the four listed categories. To improve the recognition of new words, users can also import their own word dictionary into the model to cover the words that do not occur in the original dictionary provided by PKUSEG. In addition to segmentation, PKUSEG can also label POS for words in a sentence. These features make PKUSEG more flexible on a more generalized task.

Conclusion

PKUSEG is a multi-domain Chinese word segmentation toolkit which provides high accuracy pre-trained models for users to easily play with. It also allows users to perform customization on top of the coarse-grained model with their own datasets. General coarse-grained segmentation might be a good starting point, but it is not enough for managing the explosive amount of text data. Domain-specific model is an unstoppable trend that helps deal with the varieties of text data and further improves the accuracy of the word segmentation task.

References

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, Xu Sun. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. arXiv:1906.11455v2

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. pp. 282-289. Morgan Kaufmann.