PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, Xu Sun

MOE Key Lab of Computational Linguistics, School of EECS, Peking University

{1500012842,jingjingxu,zhangyi16,renxc,xusun}@pku.edu.cn

Abstract

Chinese word segmentation (CWS) is a fundamental step of Chinese natural language processing. In this paper, we build a new toolkit, named **PKUSEG**, for multi-domain word segmentation. Unlike existing single-model toolkits, PKUSEG targets at multi-domain word segmentation and provides separate models for different domains, such as web, medicine, and tourism. The new toolkit also supports POS tagging and model training to adapt to various application scenarios. Experiments show that PKUSEG achieves high performance on multiple domains. The toolkit is now freely and publicly available for the usage of research and industry.[1]

1 Introduction

Chinese word segmentation is a fundamental task of Chinese processing. Since words define the basic semantic unit of Chinese, the quality of segmentation directly influences the performance of downstream tasks. In recent years, Chinese word segmentation has undergone great development. The best-performing systems are mostly based on conditional random fields (CRF) (Lafferty, 2001; Sun et al., 2012). However, despite the promising results, these approaches heavily rely on feature engineering. To tackle this problem, many researches (Chen et al., 2015; Cai and Zhao, 2016; Liu et al., 2016; Xu and Sun, 2016) explore neural networks to automatically learn better representations.

Recently, there arise several public segmentation toolkits, such as jieba, HanLP, and so on. For efficiency, they are built on traditional segmentation models, like perceptron (Zhang and Clark, 2007) or CRF, rather than time-consuming neural networks. These toolkits only provide a single coarse-grained segmentation model, mostly trained on news domain data. In real-world applications, the domain of text varies and the text from different domains has different domain-specific segmentation rules. This increases the difficulty of segmentation and drops the performance of existing toolkits on text from various domains.

To address this challenge, we propose a multi-domain segmentation toolkit, PKUSEG, based on the work of Sun et al. (2012): We adopt a fast and high-precision model CRF as implementation. PKUSEG includes multiple pre-trained domain-specific segmentation models. Since some domains may be of low resources, we use pre-training technique to improve the quality of segmentation. We first pre-train a coarse-grained model on a mixed corpus, including millions of data from news and web domains. Then, we fine-tune the coarse-grained model on specific domain data to get fine-grained models. In addition to provided segmentation models, PKUSEG also allows users to train a new model on their own domain data. Furthermore, POS tagging is supported in PKUSEG to adapt to various scenarios. Experimental results show that PKUSEG achieves high performance on multi-domain datasets.

In summary, PKUSEG has the following characteristics:

- **Good out-of-the-box performance.** The default word segmentation model provided by PKUSEG is trained on a large-scale, curated, multi-domain dataset, which shows stable and high performance across various domains.

- **Domain-specific pre-trained models.** PKUSEG also comes with multiple pre-trained models that are fine-tuned on texts of different domains, which further elevates domain-specific performance, suitable for analyzing in-domain texts.

---

[1] https://github.com/lancopku/pkuseg-python/

- **Easy transfer learning.** For advanced users, PKUSEG supports transfer learning based on the default multi-domain model. Users can fine-tune the model on their custom segmented texts.

- **POS tagging.** PKUSEG also provides users with POS tagging interfaces for further lexical analysis.

2 Implementation

This section gives the detailed description of toolkit implementation.

2.1 Conditional Random Field

Despite the better performance, we do not use neural networks as implementation due to the time-consuming training process. Instead, we use a well-performing and fast-training model, CRF, as implementation, considering the trade-off between time cost and high accuracy. We optimize the weights of CRF by maximizing the log likelihood of the tags of the reference sequence. When calculating the log likelihood, the log likelihood function can be calculated by the recursive algorithm in linear time. During inference, the Viterbi (Forney, 1973) algorithm is adopted. The goal is to find the sequence of tags by dynamic programming.

2.2 ADF Algorithm

For CRF with many high-dimensional features, the amount of parameters is very large, leading to expensive training cost. To address this problem, we use adaptive online gradient descent based on feature frequency information (ADF) (Sun et al., 2012) for training. The ADF algorithm does not use a single learning rate for all parameters like stochastic gradient descent (SGD), instead it turns the learning rate into a vector with the same dimension as the parameters. The learning rate of each parameter is automatically adjusted according to the frequency of the parameter. The idea is that the feature with higher frequency will be more adequate.

2.3 Pre-training

To handle the problem of low-resource for some domains, we adopt pre-training technique in PKUSEG following the work of Xu and Sun (2017). We mix news and web data together as pre-training data. News data comes from dataset PKU provided by the Second International Chinese Word Segmentation Bakeoff. Web data comes from dataset Weibo provided by NLPCC-ICCPOL 2016 Shared Task (Qiu et al., 2016). A hybrid dataset CTB is also involved into pre-training. In the process of fine-tuning, models are initialized with the pre-trained model and trained on domain-specific data. So far PKUSEG supports four fine-grained domains, including news, medicine, tourism, and web. Considering the covered domains are limited, we also provide a pre-trained model for generalization.

2.4 A Large-Scale Vocabulary

One major difficulty of multi-domain segmentation is spare domain-specific words. It is hard to cover all of these words on the training set. Therefore, to increase the coverage rate of PKUSEG, we automatically build a large-scale domain vocabulary. The word resource is crawled from Sogou website[2] and extracted from the training data of PKU, MSRA, Weibo, and CTB. In total, we extract almost 850K words. The distribution of words is shown in Table 1.

Domain	Vocabulary Size
Medicine	447K
Location	117K
Name	105K
Idiom	50K
Organization	31K
Training Words	100K
total	850K

Table 1: The distribution of words in the extracted vocabulary.

3 Usage

PKUSEG has high precision performance along with user-friendly interfaces. It is developed based on standard python3 libraries. PKUSEG supports common running platforms, such as Windows, Linux, and MacOS.

3.1 Installation

PKUSEG offers two user-friendly installation methods. Users can easily install it with PyPI and the corresponding models will be downloaded at the same time. A typical command is:

---

[2]https://pinyin.sogou.com/dict/

```
pip3 install pkuseg
```

Users also can install PKUSEG from GitHub. After downloading the project code from GitHub, users can run the following command to install PKUSEG:

```
python setup.py build_ext -i
```

Noting that the downloaded project from GitHub does not include pre-trained models, users need to additionally download them from GitHub or train a new model.

3.2 Segmentation

The followings are the detailed introduction of segmentation interfaces.

Domain-specific Segmentation

If a user is aware of the domain of the text to be segmented, then he/she can use the domain-specific model. An example code of specifying a model is shown in Figure 1. If a model is toolkit-provided, users can directly use the domain name to call it, e.g, "medicine", "touirsm", "web", and "news". The model is automatically loaded based on parameter "model_name". If the model is user-trained, "model_name" refers to the model path.

```
import pkuseg
seg = pkuseg.pkuseg(model_name='medicine')
#The domain-specific model will be downloaded
automatically.
text = seg.cut('我爱北京天安门')
print(text)
```

Figure 1: An example code of specifying the model of "medicine" domain.

Coarse-grained Segmentation

Although PKUSEG is designed to satisfy the situation where users know the domain of the text to be segmented, we also provide a coarse-grained model in case that the user can not distinguish the target domain. The coarse-grained model works under the default mode. Figure 2 shows an example code using the default mode.

User-defined Dictionary

To better recognize new words, users can add a dictionary to cover the words that do not occur in the dictionary of PKUSEG. The provided dictionary file should follow the following format. Each

```
import pkuseg
seg = pkuseg.pkuseg()
#The default model will be loaded.
text = seg.cut('我爱北京天安门')
print(text)
```

Figure 2: An example code of segmentation with the default model.

row has a single word and the dictionary file is encoded with the UTF-8 format. Figure 3 shows the usage of a user-defined dictionary.

```
import pkuseg
seg = pkuseg.pkuseg(user_dict='my_dict.txt')
text = seg.cut('我爱北京天安门')
print(text)
```

Figure 3: An example code of using a user-defined dictionary.

Model Training

PKUSEG also allows users to train a new model from scratch with their own training data. Figure 4 is an example code for showing how to train a new model.

```
import pkuseg
pkuseg.train(trainFile='user_train.txt',testFile='user
_test.txt', savedir='mypath')
```

Figure 4: An example code of training a new model with user-provided data.

Segmentation with POS Tagging

In addition to segmentation, PKUSEG can also label POS tags for words in a sentence. The usage of POS tagging interfaces is shown in Figure 5.

4 Experiment

This section evaluates the performance of PKUSEG.

4.1 Dataset

MSRA & PKU. MSRA and PKU are from news domain and provided by the Second International Chinese Word Segmentation Bakeoff.

```
import pkuseg
seg = pkuseg.pkuseg(postag=True)
text = seg.cut('我爱北京天安门')
print(text)
```

Figure 5: An example code of segmenting and POS tagging.

CTB8. Chinese Tree Bank is a hybrid domain dataset.[3] It consists of approximately 1.5 million words from Chinese newswire, government documents, magazine articles, various broadcast news and broadcast conversation programs, web newsgroups and weblogs.

Weibo. This dataset comes from the NLPCC-ICCPOL 2016 Shared Task. Different with the popular used newswire datasets, this dataset consists of many informal micro-texts.

Medicine & News & Tourism. The corpus is originally constructed in Likun Qiu (2015) by annotating multi-domain texts.

4.2 Out-of-domain Results

To show the effect of domain knowledge on segmentation performance, we train a model on CTB8 dataset and report its performance on different datasets. Here we choose CTB8 as an example because CTB8 is a hybrid dataset. The results are shown in Table 2. Here are results without using vocabularies. We can see that the performance drops obviously on out-of-domain datasets. Since different domain has its unique segmentation standard, it is not suitable to provide one single model for various domain data. The result demonstrates the necessity of fine-grained segmentation toolkits.

CTB8 Training	Testing	F1
In-domain	CTB8	95.69
Out-of-domain	MSRA (News)	83.67
	PKU (News)	89.67
	Weibo (Web)	91.19
Average	All Average	90.06
	OOD Average	88.18

Table 2: Results of PKUSEG with a model trained on CTB. "All Average" is the average F1 score of all datasets. "OOD Average (Out-of-domain Average)" is the average results of datsets except for CTB.

---

[3] https://catalog.ldc.upenn.edu/LDC2013T21

4.3 Pre-training Results

We combine existing large-scale datasets together, including PKU (news), Weibo (web), and CTB8 (hybrid), and use them as pre-training data to obtain a coarse-grained model. Then the coarse-grained model is used to fine-tune domain-specific models. Table 3 shows the effect of pre-training. The pre-training models perform much better in terms of average score, especially on datasets with lower resource (e.g., tourism).

	w/o Pre-train	w. Pre-train
Medicine	95.61	95.10
Web	94.75	95.49
News	97.58	97.80
Tourism	96.36	97.10
Average	96.08	96.37

Table 3: F1 scores of PKUSEG with and without pre-training. Here are results without using dictionaries. Web data comes from the Weibo dataset.

4.4 Default Performance

Considering the fact that many users tend to use the default mode to test performance, with the default model and vocabulary of PKUSEG. We also report experimental results on the default mode. The results are shown in Table 4. As we can see, the performance of the default model performs worse than that of domain-specific models. Therefore, we recommend users to use domain-specific models, rather than the default model, if the user can distinguish the domain of text.

PKUSEG	F1
MSRA	87.29
CTB8	91.77
PKU	92.68
Weibo	93.43

Table 4: Performance of PKUSEG on different domains with the default mode.

To learn more about the practical application of PKUSEG, we also show some segmentation examples that randomly crawled from articles which covers the domains of medicine, travel, web text, and news. The segmentation results are shown in Table 5. PKUSEG has high accuracy when dealing with words that need professional domain knowledge.

Medicine
医联 平台 ： 包括 挂号 预约 查看 院内 信息 化 验单 等 ， 目前 出现 与 微信 、 支付宝 结合 的 趋势 。
Medical Association platform includes registration appointment, in-hospital information management, etc. There is a trend of integration with WeChat and Alipay.

Travel
在 这里 可以 俯瞰 维多利亚港 的 香港岛 ， 九 龙 半岛 两岸 ， 美景 无敌。
It overlooks Victoria Harbour and the two sides of the Kowloon Peninsula. The view is so beautiful.

Web
【 这是 我 的 世界 ， 你 还 未 见 过 】 欢迎 来 参 加 我 的 演唱会 听点 音乐
This is my world that you have not seen before. Welcome to participate in my concert to listen to music.

News
他 不 忘 讽刺 加州 ： " 加州 已 在 失控 的 高 铁 项目 上 浪费 了 数十亿美元 ， 完全 没有 完 成 的 希望 。
He did not forget to satirize California, "California has been wasting billions of dollars on the uncontrolled high-speed rail projects, which is of no hope being completed at all".
乌克兰 政府 正式 通过 最新 《 宪法 修正 案 》 ， 正式 确定 乌克兰 将 加入 北约 作 为 重要 国家 方针 ， 该 法 强调 ， " 这项 法 律 将 于 发布 次日 起 生效 " 。
The Ukrainian government officially adopted the latest Constitutional Amendment, confirming that Ukraine will regard joining the NATO as an important national policy. The law emphasizes that it will take effect from the next day.

Table 5: Examples of segmentation on various domains with the domain-specific models.

5 Conclusion and Future Work

In this paper, we propose a new toolkit PKUSEG for multi-domain Chinese word segmentation. PKUSEG provides simple and user-friendly interfaces for users. Experiments on widely-used datasets demonstrate that PKUSEG performs well with high accuracy. So far PKUSEG supports domains like medicine, tourism, web, and news. In the future, we plan to release more domain-specific models and improve the efficiency of PKUSEG further.

References

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1197–1206.

G. D. Forney. 1973. The viterbi algorithm. *Proc. of the IEEE*, 61:268 – 278.

John Lafferty. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289. Morgan Kaufmann.

Houfeng Wang Likun Qiu, Linlin Shi. 2015. Construction of multi-domain chinese dependency treebanks and analysis of influencing factors on dependency parsing. *Journal of Chinese Information Processing*, 29(5):69.

Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. Exploring segment representations for neural segmentation models. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2880–2886.

Xipeng Qiu, Peng Qian, and Zhan Shi. 2016. Overview of the nlpcc-iccpol 2016 shared task: Chinese word segmentation for micro-blog texts. In *NLPCC/ICCPOL*, volume 10102 of *Lecture Notes in Computer Science*, pages 901–906. Springer.

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 253–262. The Association for Computer Linguistics.

Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursive neural network for chinese word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 567–572.

Jingjing Xu and Xu Sun. 2017. Transfer learning for low-resource chinese word segmentation with a novel neural network. *CoRR*, abs/1702.04488.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In *ACL*. The Association for Computational Linguistics.