Text Mining for Chinese Literature: An Analysis of Jin Yong’ s Wuxia Novels

Cao Mengqi, Xu Feng, Xu Shuni
School of Artificial Intelligence
Nanjing University, China
{mg20370004,dz20370012,mg20370045}@smail.nju.edu.cn

Abstract

With the explosion of text data, researchers pay more attention to extracting useful information from large amount of text data by using data mining methods. In our final project, we aim to mine useful information from Jin Yong’s Wuxia novels. We conduct multiple text mining tasks on word embeddings, including ML-based classification, clustering, etc. Our research consists of three parts of work. First, we collect 15 Jin Yong’s Wuxia novels and perform tokenization using multiple existing Chinese tokenizers. Second, we generate the word embeddings of the tokenized data. Finally, we perform data mining techniques on the word embeddings for the purpose of showing the capability of word embeddings to capture the semantic features in selected domain and Chinese literature. Based the result of our experiments, we successfully discovered some interesting information in Jinyong’s Wuxia novels. We also suggest that existing methods may not be enough to mine Chinese literature efficiently, and there remains opportunities and challenges on extracting interesting features and learning an effective representations from Chinese literature.

1 Introduction

Text mining, the practice of using computational and statistical analysis on large collections of digitized text, is becoming an increasingly important way of extracting patterns from written texts. This technique gives us information we could never access by simply reading the texts, which is rapidly penetrating the industry, right from academia and healthcare to businesses and social media platforms.

We take Chinese literature text mining as an example and at first introduce Jin Yong, one of the most famous Chinese novelist. He was a Chinese Wuxia (martial arts and chivalry) novelist and his Wuxia novels have a widespread following in Chinese communities worldwide. His 15 works written between 1955 and 1972 earned him a reputation as one of the greatest and most popular wuxia writers ever.

Novel writers, like Jin Yong, build their texts out of many central components, including subject, form, and specific word choices. Therefore, while performing text mining techniques on such long text data, it's difficult to extract useful features and conduct structural knowledge representation.

2 The aim of the study

This is an empirical survey of Chinese literature text mining on Jin Yong' s Wuxia novels. Our aims are stated as follows:

Firstly, we aim to obtain implicit semantic information from original unstructured novels. We are looking forward to extracting semantic patterns from each raw novel texts and obtaining structural information. We claim that such implicit information can be widely used in knowledge graphs and book recommendation systems.

Secondly, we aim to conduct various text mining techniques on these novels, including classification and clustering in both statistical learning and machine learning ways. Considering that our studies are text mining tasks, we can re-formulate text mining problems on our dataset, perform prevalent text mining techniques and compare the performances of existing algorithms.

Finally, we aim to raise important issues on Chinese literature text mining. The significant differences between Chinese and English mean different processing techniques. Moreover, compared with other Chinese text mining tasks, Chinese Wuxia literature text mining has its own noteworthy issues and challenges, which we would try to explore.

3 Related Works

3.1 Chinese Word Segmentation

Chinese word segmentation has been studied with considerable efforts in the NLP community. The most popular classical word segmentation methods is based on sequence labeling (Xue, 2003). Recently, researchers have incorporated neural network based approaches (Meng et al., 2019) to reduce efforts of human feature engineering. There are many toolkits available online for Chinese word segmentation, including Jieba (Sun, 2012), PKUSeg (Luo et al., 2019), Jiagu (Ownthink, 2019) and so on.

3.2 Word Embedding

Different from methods based on Vector Space Model (Salton et al., 1975), word embeddings are dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per distributional hypothesis. In our studies, we have considered 15 novels as a whole corpus and trained a Word2Vec (Mikolov et al., 2013) language model.

3.3 Topic Model

We have learned Latent Sentiment Analysis (Deerwester et al., 1990) and Probabilistic Latent Sentiment Analysis (Hofmann, 2013) in class. In our studies, we have implemented another topic model, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Both LSA and LDA have same input which is Bag of Words in matrix format. LSA focus on reducing matrix dimension while LDA solves topic modeling problems.

3.4 Text Classification Algorithms

Text classification aims to differentiate between sentences (or its segments) of different status.In general, text classification is comprised of three major steps: 1) Divide text into segments using segmentation methods. 2) Extract features from each segment (or words). 3) Build a model to classify each segment (or world).

3.5 Text Clustering Algorithms

(Cadez et al., 2000; Gaffney & Smyth, 1999) proposed to group similar text segements into clusters by using a regression mixture model and the EM algorithm. (Lee et al., 2007) proposed to partition text data into segments and to build groups of close segments using the Hausdorff Distance. (Li et al., 2010) further proposed an incremental clustering algorithm, aiming to reduce the computational cost and storage of received data.

4 Dataset

Chinese title	Date of first publication	Character count
鹿鼎记	24 October 1969	1,230,000
天龙八部	3 September 1963	1,211,000
神雕侠侣	20 May 1959	979,000
笑傲江湖	20 April 1967	979,000
倚天屠龙记	6 July 1961	956,000
射雕英雄传	1 January 1957	918,000
书剑恩仇录	8 February 1955	513,000
碧血剑	1 January 1956	488,000
飞狐外传	11 January 1960	439,000
侠客行	11 June 1966	364,000
连城诀	12 January 1964	229,000
雪山飞狐	9 February 1959	130,000
白马啸西风	16 October 1961	67,000
鸳鸯刀	1 May 1961	34,000
越女剑	1 January 1970	16,000

Tab. 4 shows the statistics of Jin Yong' s 15 novels. We take all novels as a whole and perform various text mining techniques on it.

5 The Methodology

5.1 Text Classification

5.1.1 Various Classification Models

For text classification tasks, we analyze several categories of classification algorithms on the latent embedding.

1. Parametric models.

   - Tree models. In general, Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision Tree models are especially suitable for categorical features. We use the Gini index as the metric of choice of node.
   - Ensemble models (Gradient Boost Decison Trees, AdaBoost and Random Forest models). In general, ensemble models combine decisions from multiple sub-models to enhance the overall performance. Adaboost models adjust the weight of the samples to avoid bad performance on some specific sample. Random Forest models generate a set of different features to generate difference predictive models, so that each model are not biased toward some specific features. Gradient Boost Decision Trees is a practical implementation of an ensemble of cart trees.
   - Linear model(Linear Regression). Linear Regression models predict the class of the data by taking a linear combination of the features of the class.

2. Non-parametric models.

   - Nearest Neighbors (K Nearest Neighbors). K Nearest Neighbors is one of the most common distance based classification methods. In KNN, training samples around the given data are used to predict the class label.
   - Naive Bayes(Gaussian Naive Bayes). Gaussian Naive Bayes model is especially suitable for discrete features.

5.1.2 LDA Topic Model

Given a passage of text, we aim to automatically analyse it and determine which Wuxia novel it talks about. Such text samples may be extracted from book reviews on the Internet, scripts for TV adaptions, or original novel text.

Apart from word-embedding-based method, we implement text classification in an old-school statistical way. We choose the most widely used classical topic model named Latent Dirichlet Allocation (LDA). Before the NN based topic model, LSA and LDA (and their variants) are best approaches to deal with text mining problems. The crucial assumption of topic models is that each document is a mixture of various topics and each topic is a mixture of various words. Due to the limited words of our report, the mathematical details of LDA is omitted.

5.2 Text Clustering

5.2.1 Relation Extraction

Given a passage of novel text, we aim to automatically analyze it and determine the relationship between words in novels. Relationships like parents and children, teachers and students, husband and wife, present in a relation graph in order to help people who have not read the novels to understand the relationship between characters.

After word segmentation, each word represents an entity. Entities like persons and organizations, form the most basic unit of the information. Occurrences of entities in a text are often linked through well-defined relations. For example, occurrences of person and organization in a text may be linked through relations such as employed at.(Pawar et al., 2017) The task of relation extraction is to identify such relations automatically.

A reasonable assumptions is that entity pairs with the same semantic relationship have similar contextual information. The corresponding context information of each entity pair can be used to represent the semantic relationship of the entity pair, then all entity pairs can be clustered.(Thy Tran et al., 2020) Aiming to words frequently appearing in novel text, there is a simple relations clustering to identify relations automatically.

6 Experiments

6.1 ML-based Text Classification

We performe classification algorithms on 14 books with three different type of nouns, including name of people, name of faction, name of the kongfu. The nouns are mapped to a latent embedding by Word2Vec model implemented in gensim[1]. These nouns and their ground truth labels are acquired from Jinyongwang[2]. For each type of nouns, the training set and testing set are splitted by a fixed ratio of 0.8. Tab. 6.1 shows the performance of the classification algorithms mentioned in Sec. 5.1.1, where SVM stands for Support Vector Machine, MultinomialNB stands for Multinomial Naive Bayes model, KNN stands for K Nearest Neighbors, DT stands for Decision Trees, RF stands for Random Foreset, LR stands for Linear Regression, GBDT stands for Gradient Boost Decision Tree and Adaboost stands for itself. All the models achieve a relatively high score. It is worth noting that even the non-parametric model KNN achieves a average score among all the models, where the parametric model SVM performs most poorly. GBDT, which has the highest complecity, performs the best on the classification task.

6.2 LDA-based Text Classification

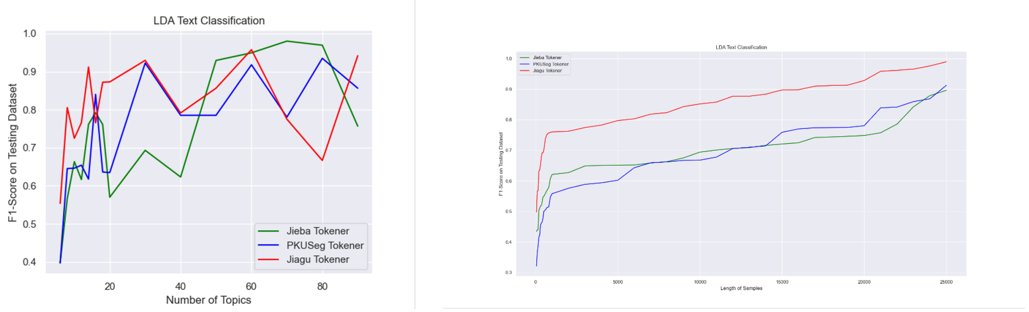
After tokenization and stopword filtering, we randomly extract fixed-length consecutive words for testing and keep the remaining texts as training samples. For each novel, the number of testing samples depends on the length of novels. We have implemented LDAs with various test sample lengths (from 20 to 25000) and hidden topic numbers (from 5 to 90). In order to verify how Chinese tokenizer affects performances, we have compared our LDA performances with three prevalent Chinese tokenizers, which are Jieba, PKUSeg and Jiagu. The performance eveluation report of those three Chinese tokenizers is available on here, where the results show that the superiority of Jiagu tokenizer.

[1] https://pypi.org/project/gensim/

[2] www.jingyongwang.com

	SVM	MultinomialNB	KNN	DT	RF	LR	GBDT	Adaboost
倚天屠龙记	96.39	97.42	96.91	93.97	96.44	97.42	96.49	95.36
笑傲江湖	93.89	96.95	94.66	92.37	95.88	96.95	96.95	90.08
天龙八部	98.02	97.03	97.52	96.44	99.06	98.51	98.02	95.54
鹿鼎记	96.64	98.66	97.32	96.98	96.85	97.99	98.72	95.97
书剑恩仇录	97.46	96.61	97.46	93.90	97.29	97.46	93.90	94.92
侠客行	89.23	92.31	89.23	92.92	90.77	93.85	90.77	89.23
射雕英雄传	89.32	93.20	92.23	86.80	94.85	93.20	91.46	91.26
神雕侠侣	98.92	98.92	96.77	97.63	98.60	98.92	97.85	97.85
碧血剑	89.32	92.23	88.35	83.30	90.58	89.32	90.19	88.35
飞狐外传	95.96	94.95	94.95	92.93	95.86	95.96	96.77	95.96
连城诀	97.30	97.30	97.30	97.84	97.30	97.30	97.30	100.00
雪山飞狐	92.31	100.00	92.31	92.31	92.31	92.31	96.15	92.31
鸳鸯刀	91.67	91.67	91.67	87.50	92.50	91.67	100.00	75.00
Average	94.75	95.49	95.43	95.12	96.35	96.32	96.39	94.79

Our LDA results are shown in Fig. 6.2. From Fig. 1(a), we can observe that there is a weak correlation between test performance (using F1-score as our metric) the number of topics in our LDA models. In our view, we argue that Jin Yong's novels are all about Chinese Wuxia and Jianghu (which refers to the environment where the martial artists live) with similar writing style, thus it's difficult to find the optimal topic numbers. We find our LDA model with 60 topics can achieve satisfactory results among all three tokenizers. From Fig. 6.2, we can conclude that test sample size has a strong impact on the performance of LDA, which makes sense because texts with more words always contain more semantic information.



(a) Performance of LDA with regard to the number of topics

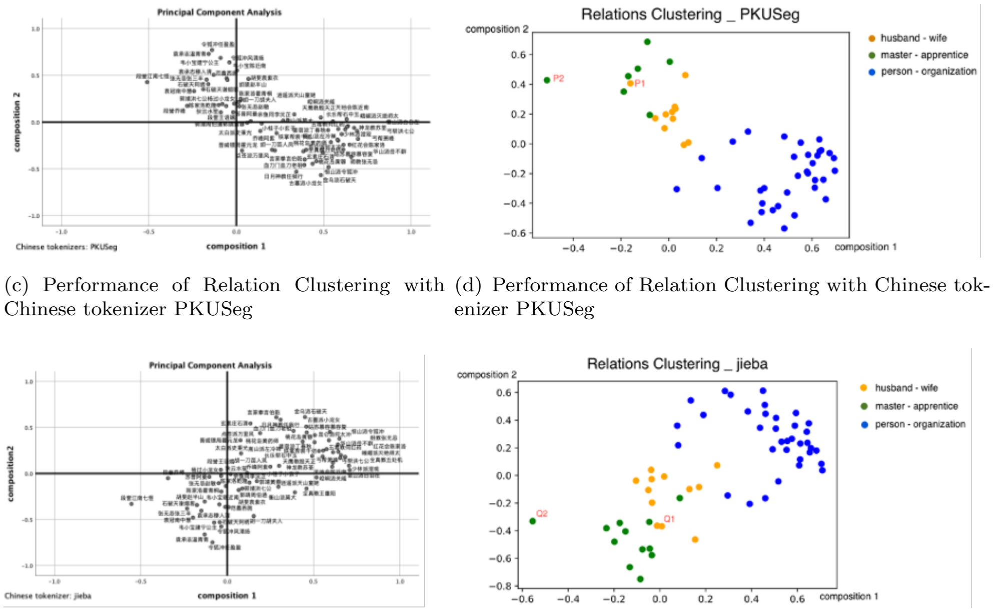
(b) Performance of LDA with regard to sample size

Besides, results also highlight the superiority of Jiagu tokenizer and we should be aware that tokenization is a key issue for Chineses NLP tasks.

6.3 Text Clustering

6.3.1 Relations Clustering

After novel text data preprocessing such as word segmentation and stop words removal, we count the number of occurrences of each word. We choose the most frequent words in the novel. Every two words form an entity pair. Based on word embedding technique, we extracted contextual information from entity pairs and cluster these entity pairs simply.



(c) Performance of Relation Clustering with Chinese tokenizer PKUSeg

(d) Performance of Relation Clustering with Chinese tokenizer PKUSeg

(e) Performance of Relation Clustering with Chinese tokenizer jieba

(f) Performance of Relation Clustering with Chinese tokenizer jieba

As a consequence of Fig. 1(c), we deal with Jin Yong' s Wuxia' s novel as a text data set with PKUSeg, a Chinese tokenizer, and relations clustering visualizes the results with Principal Component Analysis (PCA). As is depicted in Fig. 1(d),it is obvious that entity pairs are divided into three clusters. Furthermore, we find that these clusters represent three relations: husband and wife, master and apprentice, person and organization. Considering the impact of the tokenizer on relations clustering, we use another tokenizer jieba instead of PKUSeg. As is shown in Fig. 1(e),we get the similar result in the same way. There are also three clusters in Fig. 1(f).

There are some points (eg.P1 and Q1) at the junction of clusters caused by multiple relationships between entity pairs. For example, Yang Guo and Xiao Longnv (the hero and heroine of one of Jin Yong's Novels) are both masters and apprentices as well as husband and wife. There are some points (eg.P2 and Q2) out of clusters. For instance, Jiangnanqiguai, Guo Jing' s master, is a group of seven people. The relationships between them is difficult to be described.

In general, we find that, for one thing, the more different between two entities such as entity pairs: person-person and person-organization, the better the relation clustering effect presents. For another, idef two entities have an ordering such as master-apprentice and parents-children, the relation clustering will be better.

7 Discussion

7.1 Issues on Chinese Text Mining

Chinese is different from English which are studied most thoroughly by the NLP community. There are no spaces between words in Chinese written texts, and Chinese grammatical relations are indicated by word order. These factors have multiplied the difficulty of word segmentation, Part-of-Speech tagging, and semantic disambiguation at lexical, syntactic and semantic levels, since modern linguistic concepts and principles are more suitable for English than for Chinese.

7.2 Issues on Chinese Literature Text Mining

Chinese literature text mining is slightly different from other Chinese text mining tasks. For example, as shown in 7.1, Chinese segmentation is a key issue which is highly accurate on news data, but the accuracies drop significantly on other domains, such as science and literature. For scientific domains, a significant portion of out-of-vocabulary words are domain-specific terms, and therefore lexicons can be used to improve segmentation significantly. For the literature domain, however, there is not a fixed set of domain terms. (Qiu & Zhang, 2015) For example, each Jin Yong's novel may contain a specific set of person, Kungfu move and clan (which refers to a group of people who share the same interest and learn Kungfu moves together) names.

8 Conclusion

Text mining technology is now broadly applied to a wide variety of government, research, and business needs. In our studies, We perform dozens of text mining techniques, including classification, clustering, and frequency-based relation extraction on Jin Yong's 15 Wuxia novels. Moreover, we compare various methods in each task and give our explanation and discussion. Behind those results, we should be aware that there remains opportunities and challenges on extracting interesting features and learning an effective representations from Chinese literature.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Igor V. Cadez, Scott Gaffney, and Padhraic Smyth. A general probabilistic framework for clustering individuals and objects. In Raghu Ramakrishnan, Salvatore J. Stolfo, Roberto J. Bayardo, and Ismail Parsa (eds.), *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, August 20-23, 2000*, pp. 140–149. ACM, 2000. doi: 10.1145/347090.347119. URL https://doi.org/10.1145/347090.347119.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In Usama M. Fayyad, Surajit Chaudhuri, and David Madigan (eds.), *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15-18, 1999*, pp. 63–72. ACM, 1999. doi: 10.1145/312129.312198. URL https://doi.org/10.1145/312129.312198.

Thomas Hofmann. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013.

Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In Chee Yong Chan, Beng Chin Ooi, and Aoying Zhou (eds.), *Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007*, pp. 593–604. ACM, 2007. doi: 10.1145/1247480.1247546. URL https://doi.org/10.1145/1247480.1247546.

Zhenhui Li, Jae-Gil Lee, Xiaolei Li, and Jiawei Han. Incremental clustering for trajectories. In Hiroyuki Kitagawa, Yoshiharu Ishikawa, Qing Li, and Chiemi Watanabe (eds.), *Database Systems for Advanced Applications, 15th International Conference, DASFAA 2010, Tsukuba, Japan, April 1-4, 2010, Proceedings, Part II*, volume 5982 of *Lecture Notes in Computer Science*, pp. 32–46. Springer, 2010. doi: 10.1007/978-3-642-12098-5_3. URL https://doi.org/10.1007/978-3-642-12098-5_3.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. Pkuseg: A toolkit for multi-domain chinese word segmentation. *arXiv preprint arXiv:1906.11455*, 2019.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. In *Advances in Neural Information Processing Systems*, pp. 2746–2757, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Ownthink. Jiaguiagu: A toolkit chinese word segmentation. , 2019.

Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. Relation extraction : A survey. *CoRR*, abs/1712.05191, 2017. URL http://arxiv.org/abs/1712.05191.

Likun Qiu and Yue Zhang. Word segmentation for chinese novels. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Junyi Sun. Jieba. *Chinese word segmentation tool*, 2012.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. Revisiting Unsupervised Relation Extraction. *arXiv e-prints*, art. arXiv:2005.00087, April 2020.

Nianwen Xue. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pp. 29–48, 2003.