Character Profiling in Low-Resource Language Documents

Tak-sum Wong
The Hong Kong Polytechnic University
Hong Kong SAR, China
egwts@polyu.edu.hk

John Lee
City University of Hong Kong
Hong Kong SAR, China
jsylee@cityu.edu.hk

ABSTRACT

This paper focuses on automatic character profiling — connecting "who", "what" and "when" — in literary documents. This task is especially challenging for low-resource languages, since off-the- shelf tools for named entity recognition, syntactic parsing and other natural language processing tasks are rarely available. We investigate the impact of human annotation on automatic profiling, based on a Medieval Chinese corpus. Experimental results show that even a relatively small amount of word segmentation, part-of- speech and dependency annotation can improve accuracy in named entity recognition and in identifying character-verb associations, but not character-toponym associations.

KEYWORDS

information extraction, low-resource language, named entity recog- nition, dependency parsing

ACM Reference Format:
Tak-sum Wong and John Lee. 2019. Character Profiling in Low-Resource Language Documents. In *Australasian Document Computing Symposium (ADCS 2019)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As more literary and historical texts become digitized, scholars increasingly complement traditional, manual research methodol- ogy with natural language processing (NLP) on a variety of topics, including literary style [14], literary genres [22], intertextuality [6], authorship [24], and document structure [9]. Recent research has also attempted to analyze characters in terms of their persona [3], social networks [1], and the location, time and nature of the events in which they participate [31]. This paper focuses on character profiling, centering on three "Wh-questions" — *who* the characters were; *what* they did; and *where* they were.

Specifically, we investigate the impact of human annotation for character profiling in literary text written in low-resource lan- guages. Automatic profiling involves NLP tasks such as named entity recognition (NER) [10] and syntactic analysis [5]. Although syntactic features have been shown to be beneficial for information extraction [5, 21, 34], off-the-shelf parsers are rarely available for

Character	Association	Frequent words
Ananda Pratyeka-buddha Mahāyānadeva	verb	□ 'to address', □ 'to speak' □ 'not have', □ 'to attain' □ 'to serve', □ 'to translate'
Buddha	toponym	Jetavana, Śrāvastī, Rājagṛha

Table 1: Frequent character-verb and character-toponym as- sociations extracted from the Chinese Buddhist Canon

low-resource languages, and manual annotation may not yield suf- ficient data to train a high-accuracy automatic parser. Given these constraints, syntactic features may not always be helpful in ana- lyzing documents in low-resource languages. This paper addresses the following questions:

- Q1: Can small-scale annotation on word segmentation im- prove NER?
- Q2: Can small-scale part-of-speech (POS) and dependency annotation improve character profiling?

Focusing on retrieving relations between characters, verbs and to- ponyms, we answer these questions in the context of a Medieval Chinese corpus, using a small training set with word segmenta- tion, POS and dependency annotation on about 50,000 Chinese characters.[1]

After a review of previous work (Section 2), we present our dataset (Section 3). To address Q1, we investigate the effect of limited training data for word segmentation on NER performance (Section 4). To address Q2, we report experimental results in extract- ing character-verb and character-toponym associations (Sections 5 and 6), some examples of which are shown in Table 1.

2 PREVIOUS WORK

Information extraction (IE) models have mostly been evaluated on the news domain. Typically, a named entity recognizer first identi- fies the entities, such as people, organizations and locations [12]. Then, a statistical classifier, trained on annotated corpora, deter- mines if two entities in a text participate in a target relation. A wide range of properties including lexical features, regular expressions and POS tags have been explored [2, 15, 29].

A number of studies have shown that syntactic features can help improve NER and IE. Zhou et al. [34] reported that syntac- tic features can improve IE accuracy. Mintz et al. [21] found that the combination of syntactic and lexical features provides better

performance than either feature set on its own. However, these improvements may not hold in low-resource domains due to lower parsing accuracy in general.

[1]We use the term "characters" to refer to people in a literary document, and use the same term to refer to Chinese writing only with the expression "*Chinese* characters".

This research question is especially under-investigated for the lit- erary domain. In the Namescape project, named entity recognizers were trained on Dutch literary texts to recognize names in Dutch fiction [10]. Grammatical structure was exploited in recognizing proper names in French novels [5]. Although automatic dependency parsing was performed on the Chinese Buddhist Canon [28], the resulting treebank has not been applied to character profiling. A previous study analyzed "nexus points" [4], i.e., people present at the same locations, but it addressed manual annotation rather than automatic profiling.

3 DATA

The Chinese Buddhist Canon (henceforth, "the Canon") is the largest digitized corpus of Medieval Chinese text, whose size ex- ceeds 40 million Chinese characters. The Canon provides a useful context for studying character profiling in low-resource languages for two reasons. First, existing NLP tools for Chinese, trained mostly on Modern Standard Chinese [13, 25], tend not to perform well on Medieval Chinese. Second, there is not yet any attempt to automat- ically mark up the myriad of characters and places in the Canon.

Lee and Kong built a dependency treebank on a small subset of the Canon, consisting of about 50,000 Chinese characters drawn from four sutras [19]. This treebank (henceforth, the "L&K Tree- bank") uses a digital version [17] of the *Tripiṭaka Koreana*, the Ko- rean Edition of the Canon, produced from the most complete set of currently available blocks, stored at Haein Monastery in Korea [18]. The L&K Treebank largely follows the guidelines for the Penn Chi- nese Treebank for word segmentation and POS tagging [32], and adopts the Stanford Dependencies for Modern Chinese [7]. We created three evaluation datasets from this treebank.

3.1 Named entity recognition dataset

Each word tagged as a proper noun (NR) in the L&K Treebank was automatically classified as a person or toponym, based on its membership in the Person or Place Authority Database [11]. In cases where the word is found in neither database, an annotator with a background in Medieval Chinese performed the classification. This NER dataset, with a total of 1,914 characters and 114 toponyms, facilitates the experiments in Section 4.

3.2 Character-verb association dataset

We retrieved the child and parent words of the "nominal subject" (nsubj) dependency relations in the L&K Treebank where the child word is annotated as a personal name in Section 3.1. This dataset, consisting of 896 character-verb pairs, will be used in Section 5.

3.3 Character-toponym association dataset

Since the L&K Treebank contains too few toponyms for a reliable evaluation, we also utilized *Pi nai yeh* (Vinaya) (K936), an important text on vinaya. An annotator with a background in Medieval Chi- nese examined each toponym, and determined whether it indicates the location of a person. Out of the 663 toponyms present in the text, the annotator identified 201 character-toponym pairs. This dataset will be used in Section 6.

Method	Characters		Toponyms	
	Precision	Recall	Precision	Recall
Baseline	0.77	0.51	0.58	0.19
CRF	**0.87**	**0.69**	**0.82**	**0.48**

Table 2: Named entity recognition performance on character names and toponyms

4 NAMED ENTITY TAGGING

We implemented a lexicon-based named entity tagger with the following lexica: the Person/Place Authority Database [11], which contains 39,277 personal names and 18,017 geographical names; the *Dictionary of Chinese Buddhist Terms* [27], with 16,687 entries; and 720 Sanskrit-transliterated terms [8].

Naive use of these lexica would lead to many false alarms be- cause they cover a wide range of people and locations related to Buddhism, many of which can also serve as common nouns and verbs. Since the Canon is predominantly a translation from texts in Indic languages, most named entities are of non-Chinese origins. We therefore filtered out terms of Chinese origins from the lexica to increase NER precision.

4.1 Approach

Baseline. Our baseline uses Stanford CoreNLP for Chinese word segmentation [7]. Since most segmentation errors involved the split- up of a word, we performed Forward Maximal Matching (FMM) on the segmentation output with the Person/Place Authority Data- base [11], in order to improve recall.

CRF. Using the CRF++ implementation [16], we trained con- ditional random fields on the L&K Treebank to perform Chinese word segmentation. We adopted the features in [33], with the four resources listed above as external lexica. Compared to Modern Chinese, fewer words in Medieval Chinese contain more than two syllables. Therefore, we adopted a 2-tag set for word segmenta- tion [23, 30]. We

also performed FMM on the CRF output.

4.2 Results

Table 2 shows the NER precision and recall on the dataset in Sec- tion 3.1. For recognizing personal names, the baseline achieved 77% precision and 51% recall. Since the baseline word segmentation tool is trained on Modern Chinese, its relatively poor performance on Medieval Chinese is not unexpected. The CRF model, trained on a smaller but in-domain dataset, improved the precision to 87% and the recall to 69%.

Toponym recognition turned out to be more challenging. The baseline achieved 58% precision and 19% recall. The CRF model performed substantially better, at 82% precision and 48% recall. For both personal names and toponyms, recall suffered from limited coverage of the lexica.

5 CHARACTERS AND VERBS

To connect "Who" and "What", we identify the verbs that are most frequently associated with each person.

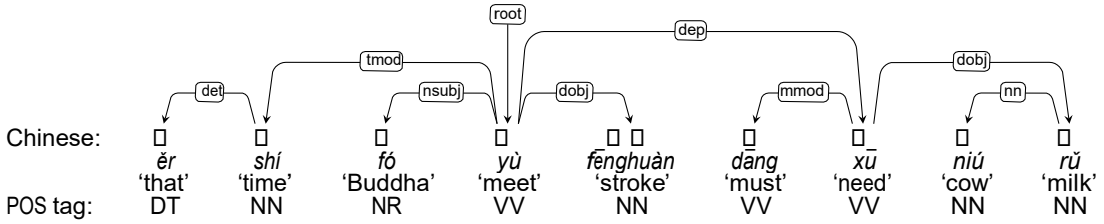

Figure 1: Two character-verb pairs, 'Buddha'-'meet' and 'Buddha'-'need', are extracted from this dependency tree using the dependency-based approach (Section 5.1)

5.1 Approach

POS-based approach. We trained a part-of-speech (POS) tagger with CRF++ on the L&K Treebank, using the settings in [28]. While other Medieval Chinese corpora such as *Huainanzi* [26] could have provided additional training data, we chose not to include them since the vocabulary was significantly different. In addition to the standard unigram and bigram features, we also included a feature for proper nouns, based on the lexica described in Section 4.

We identified character-verb pairs by retrieving each verb (VV) that either immediately follows a personal name, or separated only with an adverb (AD). For example, the words 'Buddha' and 'meet' in the sentence in Figure 1 form such a pair, since *yù* 'meet' imme- diately follows the name *fó* 'Buddha'.

Dependency-based approach. We trained a Minimum-Spanning Tree parser [20] on the L&K Treebank, using the settings in [28]. To collect character-verb pairs from the trees, we retrieved all "nomi- nal subject" (nsubj) relations where the child word is tagged as a character (Section 3.1), and the parent word is a verb. The pair 'Bud- dha' *fo*-'meet' *yu*) is extracted in Figure 1 since the nsubj relation indicates that *fó* is the subject of the verb *yù*.

Dependency structures facilitate extraction of character names separated by longer distances from their verbs. They are helpful for Medieval Chinese since serial verb constructions are common. In such a construction, the "dependent" (dep) relation links the verbs, such as *yù* 'meet' and *xū* 'need' in Figure 1. Although *fó* 'Buddha' is the subject for both verbs, the second verb (*xū*) is not directly linked to it. We attributed the subject to all verbs in a serial verb construction, hence in this case also recognizing 'Buddha' *fó*- 'need' *xū*) as a character-verb pair.

Baseline. Our baseline method considers each character name and the following word to be a character-verb pair.

5.2 Results

Table 3(a) shows the subject-verb pair extraction performance on ten-fold cross-validation on the dataset in Section 3.2. The precision and recall of the baseline approach were 0.46 and 0.64. Despite the small amount of training data, the POS-based approach improved the precision and recall to 0.77 and 0.71, respectively.

Dependency information further improved the precision to 0.91 and the recall to 0.93. The gain in recall was mostly due to recognition of character-verb pairs in serial verb constructions. The remaining recall errors resulted from the parser's mislabeling of

Task	Method	Precision	Recall
(a) Character-verb pair extraction	Baseline	0.46	0.64
	POS-based	0.77	0.71
	Dependency-based	**0.91**	**0.93**
(b) Character-toponym pair extraction	Baseline	0.99	**0.73**
	POS-based	**1.00**	0.66
	Dependency-based	**1.00**	0.66

Table 3: Extraction performance for (a) character-verb pairs and (b) character-toponym pairs

nominal subject relations as noun modifier (nn) or adverbial modi- fier (advmod). Vocatives mistaken as nominal subjects led to most precision errors.

Table 1 shows some sample results. The verbs associated with Ananda reflect his many conversations with Buddha. Pratyeka- buddha was the "lone buddha" who sought to 'attain' enlightenment only for himself. The Canon often describes Mahāyānadeva, an eminent translator, as having 'served' the Chinese emperor and
'translated' the scriptures.

6 CHARACTERS AND TOPONYMS

To connect "Who" and "Where", we extract toponyms that most frequently indicate the location for each character.

6.1 Approach

POS-based approach. We identified location markers by examin- ing verbs and prepositions that immediately precede the toponyms. The two most common such verbs, *zài* 'dwell in' and *zhù* 'stay in', account for almost a third of the instances; the most frequent preposition, *yú* 'at', constitutes more than half of the instances. All verbs and prepositions whose frequency exceeded 0.1% were included as location markers. When a location marker is present between a personal name and a toponym, they are considered to be a character-toponym pair.

Dependency-based approach. Toponyms typically serve as the direct object of a verb (e.g., *zài* 'dwell in') or the object of a prepo- sition (e.g., *yú* 'at'). We retrieved all verbs and prepositions that participate in a dobj or prep relation with a toponym. Similar to the POS-based approach, all verbs and prepositions whose frequency exceeded 0.1% were included as location markers.

From the trees underlying the character-verb pairs (Section 5.1), we identified character-toponym pairs by extracting cases where (i)

the verb is a location marker and takes a toponym as direct object; or (ii) the verb is modified by a prepositional phrase containing a location marker and a toponym.

Baseline. The baseline method takes all personal names and toponyms within the same sentence as character-toponym pairs.

6.2 Results

Table 3(b) shows the character-toponym extraction performance on ten-fold cross-validation on the dataset in Section 3.3. The strong performance of the baseline, at 0.99 precision and 0.73 recall, can be attributed to the regularity in toponym usage: when a toponym appears in the same sentence as a character, it almost always indicates the location of the character. Recall was lower since the extraction algorithm did not attempt to identify character-toponym pairs where the personal name and toponym occur in different sentences, which constitute 15% of the gold data. This would have necessitated semantic analysis and likely lowered the precision.

The POS-based and dependency-based approaches did not improve the overall accuracy in extracting character-toponym pairs, unlike character-verb pairs. Their more stringent requirements on lexical choice and dependency structure boosted their precision to 100%, but degraded recall to 0.66. A number of verbs were not recognized as location markers due to infrequency.

The toponyms most associated with Buddha are listed in Table 1. They include Śrāvastī, the city where Buddha spent much of his monastic life, and Jetavana, where the monastery was located.

7 CONCLUSION

This paper investigated character profiling in literary text written in a low-resource language. Experiments on a Medieval Chinese corpus showed that even a small amount of in-domain annotation — word segmentation, POS and dependency annotation on 50,000 Chinese characters — improved accuracy in named entity recognition and in the extraction of character-verb pairs. These results suggest that annotation on a similar scale is well worth considering in future literary analyses of low-resource language documents.

We have applied these techniques to produce the largest Medieval Chinese corpus to date that is automatically annotated with named entities, character-verb and character-toponym associations.[2] In future work, we intend to apply character profiling on other seminal works of literature.

ACKNOWLEDGMENTS

This work was partially supported by a grant from the Research Grants Council of the Hong Kong SAR (Project No. CityU 155412) and by CityU Internal Funds for ITF Projects (no. 9678104).

REFERENCES

[1] A. Agarwal, A. Corvalan, J. Jensen, and O. Rambow. 2012. Social Network Analysis of Alice in Wonderland. In *Proc. Workshop on Computational Linguistics for Literature*.
[2] A. Agarwal and O. Rambow. 2010. Automatic detection and classification of social events. In *Proc. EMNLP*.
[3] David Bamman, B. O'Connor, and N. A. Smith. 2013. Learning latent personas of film characters. In *Proc. ACL*.

[2] This corpus may be obtained from the first author for research purposes.

[4] Marcus Bingenheimer, Jen-Jou Hung, and Simon Wiles. 2011. Social Network Visualization from TEI Data. *Literary and Linguistic Computing* 26, 3 (2011), 271–278.
[5] C. Bornet and F. Kaplan. 2017. A simple set of rules for characters and place recognition in French novels. *Frontiers in Digital Humanities* 4, 6 (2017), 1–21.
[6] M. Büchler, A. Gessner, T. Eckart, and G. Heyer. 2010. Unsupervised detection and visualisation of textual reuse on Ancient Greek texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1, 2 (2010), 117.
[7] P. C. Chang, H. Tseng, D. Jurafsky, and C. D. Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proc. 3rd Workshop on Syntax and Structure in Statistical Translation*.
[8] C. N. Chu. 1996. Vocabulary of Buddhist sutras of the West Chin dynasty. In *NSC Project Report*. National Chung Cheng University, Taipei.
[9] T. E. Clement. 2008. 'A thing not beginning and not ending': Using digital tools to distant-read Gertrude Stein's The Making of Americans. *Literary and Linguistic Computing* 23, 3 (2008), 361381.
[10] K. Van Dalen-Oskam, J. de Does, M. Marx, I. Sijaranamual, K. Depuydt, B. Verheij, and V. Geirnaert. 2014. Named Entity Recognition and Resolution for Literary Studies. *Computational Linguistics in the Netherlands* 4 (2014), 121–136.
[11] DDBC. 2008. *Buddhist Studies Authority Database Project*. Dharma Drum Buddhist College, http://authority.ddbc.edu.tw.
[12] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proc. LREC*.
[13] J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local informa- tion into information extraction systems by Gibbs sampling. In *Proc. ACL*.
[14] David I. Holmes. 1994. Authorship Attribution. *Computers and the Humanities* 28 (1994), 87–106.
[15] H. Ji and R. Grishman. 2008. Refining event extraction through unsupervised cross-document inference. In *Proc. HLT-NAACL*.
[16] Taku Kudo. 2005. *CRF++: Yet Another CRF Toolkit*. http://taku910.github.io/crfpp/. [17] Lewis Lancaster. 2010. From Text to Image to Analysis: Visualization of Chinese Buddhist Canon. In *Proc. Digital Humanities*.
[18] Lewis Lancaster and S. Park. 1979. *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley University Press, Berkeley, CA.
[19] John Lee and Yin Hei Kong. 2016. A dependency treebank of Chinese Buddhist texts. *Digital Scholarship in the Humanities* 31, 1 (2016), 140–151.
[20] Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual Depen- dency Analysis with a Two-stage Discriminative Parser. In *Proc. 10th Conference on Computational Natural Language Learning (CoNLL-X)*.
[21] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. ACL*.
[22] F. Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, London, UK.
[23] F. Peng, F. Feng, and A. McCallum. 2004. Chinese segmentation and new word detection using Conditional Random Fields. In *Proc. COLING*.
[24] H. Sayoud. 2012. Author discrimination between the Holy Quran and Prophet's statements. *Literary and Linguistic Computing* 27, 4 (2012), 427–444.
[25] C. W. Shih, T. H. Tsai, S. H. Wu, C. C. Hsieh, and W. L. Hsu. 2004. The construction of a Chinese named entity tagged corpus: CNEC1.0. In *Proc. 16th Conference on Computational Linguistics and Speech Processing*.
[26] Y. Song and F. Xia. 2014. Modern Chinese helps Archaic Chinese processing: Finding and exploiting the shared properties. In *Proc. LREC*.
[27] W. E. Soothill and L. Hodous. 1937. *A dictionary of Chinese Buddhist terms: With Sanskrit and English equivalents and a Sanskrit-Pali index*. Kegan Paul, Trench, Trubner Company, Limited, Carter Lane, EC.
[28] Tak sum Wong and John Lee. 2016. A Dependency Treebank of the Chinese Buddhist Canon. In *Proc. Linguistic Resources and Evaluation Conference (LREC)*. 1679–1683.
[29] M. Surdeanu and M. Ciaramita. 2007. Robust information extraction with per- ceptrons. In *Proc. NIST 2007 Automatic Content Extraction Workshop (ACE07)*.
[30] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan Bakeoff 2005. In *Proc. Fourth SIGHAN Workshop on Chinese Language Processing*.
[31] P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, SK. Hsieh, and CR. Huang. 2008. KYOTO: A system for mining, structuring and distributing knowledge across languages and cultures. In *Proc. Language Resources and Evaluation Conference (LREC)*.
[32] N. Xue, F. Xia, F. D. Chiou, and M. Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11 (2005), 207–238.
[33] H. Zhao, C. N. Huang, and M. Li. 2007. An improved Chinese word segmentation system with Conditional Random Field. In *Proc. 5th SIGHAN Workshop on Chinese Language Processing*.
[34] G. Zhou, M. Zhang, D. Ji, and Q. Zhu. 2007. Tree Kernel-based Relation Extrac- tion with Content-sensitive Structured Parse Tree Information. In *Proc. EMNLP- CoNLL*.