Chinese Named Entity Recognition using Conditional Random Fields

Xilun Chen[1,2] Yuanfei Zhu[1] Zeyu Wang[1]

[1]Shanghai Jiao Tong University

[2]On his visiting at Natural Language Computing Group, Microsoft Research Asia

December 16, 2012

Table of contents

1. Introduction
   - Named Entity Recognition
   - Sequence Labeling
2. Proposed Approach
   - Preprocessing
   - Features
3. Experiments
   - Experiment Settings
   - (Preliminary) Results

Section 1

Introduction

Subsection 1

Named Entity Recognition

Named Entity Recognition

- NER is a task aimed at extracting named entities from raw text and classifying into their corresponding categories, such as persons, locations, organizations, etc.

Named Entity Recognition

- NER is a task aimed at extracting named entities from raw text and classifying into their corresponding categories, such as persons, locations, organizations, etc.
- In this task, we are required to identify three entity categories: PER, LOC, ORG given a Chinese text.

NER as a Sequence Labeling Task

Sequence Labeling has been the state-of-the-art approach to a number of NLP tasks:

- Word Segmentation(Tseng et al., 2005).
- POS Tagging(Toutanova and Manning, 2000).
- Chunking(Shallow Parsing)(Sha and Pereira, 2003).
- Named Entity Recognition(McCallum and Li, 2003).

Subsection 2

Sequence Labeling

Sequence Labeling

Hidden Markov Model (HMM)

- Generative Model
- Cannot incorporate long distance features

Maximum Entropy Markov Model (MEMM)

- Discriminative Model
- Can incorporate long distance features
- Optimize conditional probability

Conditional Random Field (CRF) — *Our Choice*

- Discriminative Model
- Can incorporate long distance features
- Optimize joint probability (Has looser assumption of independency)
- High cost, long training time

Section 2

Proposed Approach

Subsection 1

Preprocessing

Tagging Scheme

- We need to convert the labels to fit the sequence labeling task.

Tagging Scheme

- We need to convert the labels to fit the sequence labeling task.
- We adopt the classic IOB tagging method with totally 7 tags: B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG and Others

Tagging Scheme

- We need to convert the labels to fit the sequence labeling task.
- We adopt the classic IOB tagging method with totally 7 tags: B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG and Others
- Gain produced by more complex tagging methodologies like IOBES remains elusive.(Collobert et al., 2011)

Word Segmentation & POS Tagging

Word Segmentation

We adopt a word-based approach rather than character-based one under the rationale:

- In Chinese, words can better encode semantics atomics than characters.
- By word segmentation, additional useful features such as POS tags can be introduced.

POS Tagging

- Named Entities are more likely to have particular POS tags (e.g. Proper Nouns).
- POS Tagging can ameliorate the problem of data sparsity.

Subsection 2

Features

Lexical Features

- Lexical features are the most basic features that everyone would exploit.

Lexical Features

- Lexical features are the most basic features that everyone would exploit.
- But extravagant utilization of lexical features would result in severe sparsity problem.

Lexical Features

- Lexical features are the most basic features that everyone would exploit.
- But extravagant utilization of lexical features would result in severe sparsity problem.
- Hence we incorporate some but limited lexical features of unigram, bigram and trigram features.

Lexical Features

- Lexical features are the most basic features that everyone would exploit.
- But extravagant utilization of lexical features would result in severe sparsity problem.
- Hence we incorporate some but limited lexical features of unigram, bigram and trigram features.
- We also set a relatively large cutoff value to avoid the bias induced by lexical features of extremely low frequency.

POS Tags

As we said before, POS tags can benefit the NER task.

- Named Entities are very likely to possess a specific subset of POS tags (NR, for example).

POS Tags

As we said before, POS tags can benefit the NER task.

- Named Entities are very likely to possess a specific subset of POS tags (NR, for example).
- POS tags does not suffer from the data sparsity as the lexical features.

Prefixes & Suffixes

In Chinese, much helpful knowledge can be revealed by examining the prefix and suffix of a word.

Prefixes & Suffixes

In Chinese, much helpful knowledge can be revealed by examining the prefix and suffix of a word.

- e.g. Patterns like Zhang XX, Li XX suggests a high probability of persons.

Prefixes & Suffixes

In Chinese, much helpful knowledge can be revealed by examining the prefix and suffix of a word.

- e.g. Patterns like Zhang XX, Li XX suggests a high probability of persons.
- e.g. Patterns like XX Shi, XX Xian suggests a high probability of locations.

Prefixes & Suffixes

In Chinese, much helpful knowledge can be revealed by examining the prefix and suffix of a word.

- e.g. Patterns like Zhang XX, Li XX suggests a high probability of persons.
- e.g. Patterns like XX Shi, XX Xian suggests a high probability of locations.
- e.g. Patterns like XX Ju, XX Chu suggests a high probability of organizations.

Gazetteers

(Ratinov and Roth, 2009) argued that numerous works have reported that gazetteers would substantially help the performance of a NER system.(Cohen and Sarawagi, 2004; Kazama and Torisawa, 2007; Florian et al., 2003)

Gazetteers

(Ratinov and Roth, 2009) argued that numerous works have reported that gazetteers would substantially help the performance of a NER system.(Cohen and Sarawagi, 2004; Kazama and Torisawa, 2007; Florian et al., 2003)

We collected a dozen of gazetteers from various sources (Experiments are still on-going, not final choice)

- A list of 300 most common Chinese Surnames
- Classified gazetteers from *Sogou Cell Lexicon*.

Word Clustering

- Word clustering is another approach that deals with the sparsity.

Word Clustering

- Word clustering is another approach that deals with the sparsity.
- Socher and Chris Manning from Stanford pointed out in their tutorial at ACL 2012(Socher et al., 2012) that Word Clustering can improve the accuracy of NER, as shown in Figure 6.

	POS WSJ (acc.)	NER CoNLL (F1)
Supervised NN	96.37	81.47
NN with Brown clusters	96.92	87.15

Word Clustering (Cont.)

Weiwei Sun from PKU reported similar gain by using word clustering on POS tagging (Sun and Uszkoreit, 2012) in her talk at MSRA last week.

Features	Brown	MKCLS
Supervised	94.48%	
+ #100	94.82%	94.93%
+ #500	94.92%	94.99%
+ #1000	94.90%	95.00%

Word Clustering (Cont.)

Weiwei Sun from PKU reported similar gain by using word clustering on POS tagging (Sun and Uszkoreit, 2012) in her talk at MSRA last week.

Features	Brown	MKCLS
Supervised	94.48%	
+ #100	94.82%	94.93%
+ #500	94.92%	94.99%
+ #1000	94.90%	95.00%

Therefore, we are trying to incorporate the **mkcls** clustering by Franz Josef Och (Och, 1999).

Word Clustering (Cont.)

Weiwei Sun from PKU reported similar gain by using word clustering on POS tagging (Sun and Uszkoreit, 2012) in her talk at MSRA last week.

Features	Brown	MKCLS
Supervised	94.48%	
+ #100	94.82%	94.93%
+ #500	94.92%	94.99%
+ #1000	94.90%	95.00%

Therefore, we are trying to incorporate the **mkcls** clustering by Franz Josef Och (Och, 1999).
Experiments on Word Clustering are still on-going.

Section 3

Experiments

Experiment Settings

- We use Stanford Segmenter(Tseng et al., 2005) for word segmentation

Experiment Settings

- We use Stanford Segmenter(Tseng et al., 2005) for word segmentation
- We use Stanford POS Tagger(Toutanova and Manning, 2000) for POS tagging.

Experiment Settings

- We use Stanford Segmenter(Tseng et al., 2005) for word segmentation
- We use Stanford POS Tagger(Toutanova and Manning, 2000) for POS tagging.
- We use CRF++ as implementation of CRF.

Experiment Settings

- We use Stanford Segmenter(Tseng et al., 2005) for word segmentation
- We use Stanford POS Tagger(Toutanova and Manning, 2000) for POS tagging.
- We use CRF++ as implementation of CRF.
- Data Selection:

  Training Set First 950/1000 documents in the training data.

  Dev Set Last 50/1000 documents in the training data.

Preliminary Results

We are still keep polishing our models and conducting further experiments, hence the results presented here are not our final version.

Table below shows the F1 score on Dev Set.

Setting	F1(Term-level)	F1(Character-level)
Lexical	66.91	67.51
+POS	79.02	81.95
+Pre/Suffix	85.74	88.74
+Gazetteer	Not Ready	Not Ready
+mkcls	Not Ready	Not Ready

The End

Thank you!

Cohen, W. and Sarawagi, S. (2004). Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In *Conference on Knowledge Discovery in Data: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, volume 22, pages 89–98.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Kazama, J. and Torisawa, K. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Och, F. (1999). An efficient method for determining bilingual word classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 71–76. Association for Computational Linguistics.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Socher, R., Bengio, Y., and Manning, C. (2012). Deep learning for nlp (without magic). In *Tutorial Abstracts of ACL 2012*, pages 5–5. Association for Computational Linguistics.

Sun, W. and Uszkoreit, H. (2012). Capturing paradigmatic and syntagmatic lexical relations: Towards accurate chinese part-of-speech tagging.

Toutanova, K. and Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.