Jie Zhou, Bi-cheng LI, Gang Chen, 2015. Automatically building large-scale named entity recognition corpora from Chinese Wikipedia. *Frontiers of Information Technology & Electronic Engineering*, **16**(11):940-956. [doi:10.1631/FITEE.1500067]

Automatically building large-scale named entity recognition corpora from Chinese Wikipedia

Key words: NER corpora, Chinese Wikipedia, Entity classification, Domain adaptation, Corpus selection

Corresponding author: Jie Zhou
E-mail: zhoujie.nlp@gmail.com

ORCID: http://orcid.org/0000-0001-5615-9334

Introduction

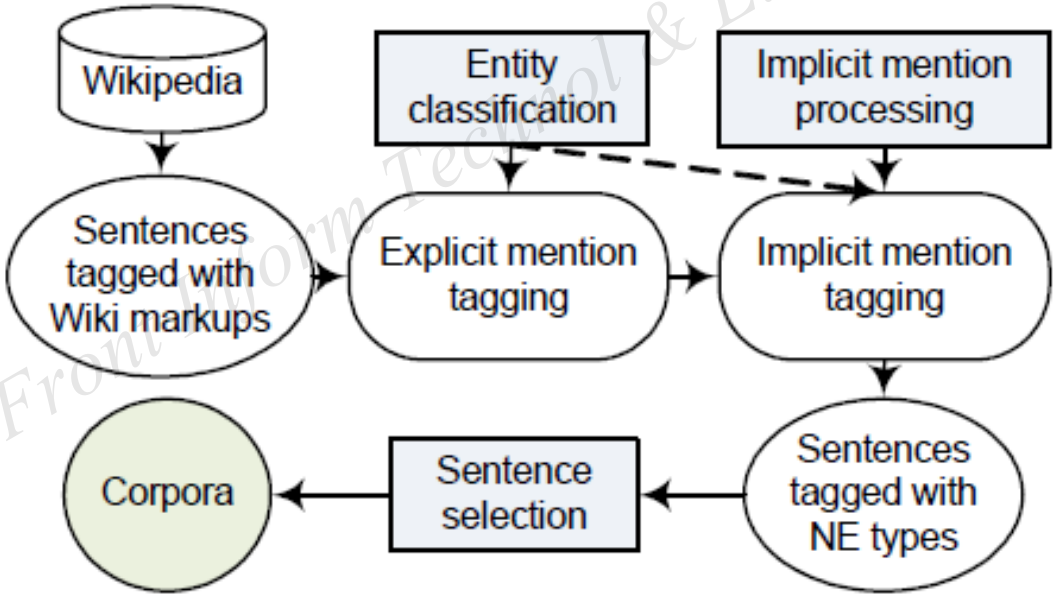
- Named entity recognition (NER) is a core component in many nature language processing applications.

- Most NER systems rely on supervised machine learning methods, which depend on time-consuming and expensive annotations in different languages and domains.

- This paper presents a method for building silver-standard NER corpora from Chinese Wikipedia automatically.

Framework of our method

- An overview of the Chinese NER corpora generation process



Design method (I)

- Entity classification of explicit mention
  - Currently, the work on entity classification has mainly focused on *heuristic rule based* and *supervised classifier based*.
  - To achieve high performance and large coverage, a method that combines heuristic rules of multi-faceted information with supervised NE classifier is designed to determine NE types of all articles in Chinese Wikipedia.

$$\boldsymbol{C}_{\text{multi}} = \alpha_1 \boldsymbol{C}_{\text{category}} + \alpha_2 \boldsymbol{C}_{\text{infobox}} + \alpha_3 \boldsymbol{C}_{\text{language}} + \alpha_4 \boldsymbol{C}_{\text{title}}$$

Design method (II)

- Type identification of implicit mention
  - To avoid missed annotations caused by unlabeled links, we present a method for finding implicit mentions in article content by using boundary information of outgoing links.
  - This method identified NE type of ambiguous mentions based on EL method.

Design method (III)

- Tagged corpus selection approach
  - The NER model is generally trained by gold-standard corpora, most of which come from newswire about national contemporary politics. However, this model may be applied to process documents from other domains in practice, thus resulting in poor performance.
  - We design an approach based on core article extending to select data that are related to current domains. The process contains core article extraction, core article extending, article ranking, and corpus building.

Major results

(1) Evaluation of entity classification:

Table 1 Entity classification performance of heuristic rules, the supervised NE classifier, and the combined method

Method	Article number	Precision (%)	Recall (%)	*F*-score (%)
Heuristic rules	2391	95.36	95.46	95.40
Supervised NE classifier		87.67	87.14	87.28
Classifier (content)		81.09	80.21	80.37
Classifier (+structured)	3678	83.91	83.47	83.59
Classifier (+category)		85.25	84.75	84.79
Classifier (+article title)		85.69	85.51	85.52
Combined method	3678	91.31	90.56	**90.73**

The weighted average of each measure (precision, recall, and *F*-score) is used to evaluate the overall performance

Major results

(2) Evaluation of NER corpora:

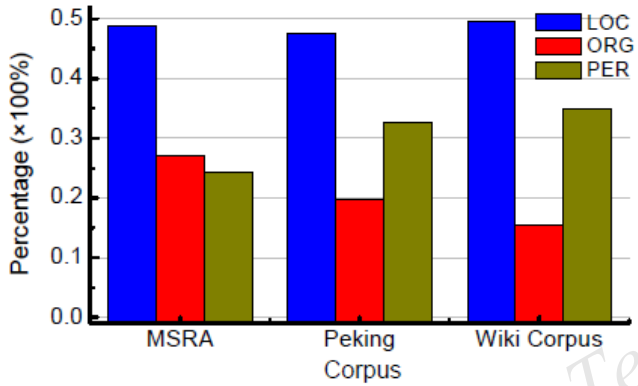


Fig. 6 NE distribution of different NE types across the gold-standard corpora and the Wiki corpus

Table 5 Evaluation results of automatic annotation for MSRA and Peking test data

NE type	Precision (%)		Recall (%)		*F*-score (%)	
	MSRA	Peking	MSRA	Peking	MSRA	Peking
PER	83.93	76.44	65.92	42.84	73.84	54.91
ORG	66.91	49.07	54.24	51.25	59.92	50.14
LOC	79.41	77.79	67.87	86.51	73.19	81.92
Total	78.14	71.64	64.31	65.01	70.55	68.16

Major results

(3) NER performance:

Table 6 Evaluation results (*F*-score) of NER when training on different corpora

ID	Training corpus	*F*-score (%)											
		MSRA				Peking				Domains			
		PER	ORG	LOC	ALL	PER	ORG	LOC	ALL	PER	ORG	LOC	ALL
G1	MSRA	92.94	83.62	91.17	90.15	89.17	68.73	80.71	81.08	86.48	67.17	78.67	77.70
	Peking	93.12	59.44	76.61	79.22	94.49	84.01	91.38	91.18	85.66	50.00	76.51	72.53
G2	CW-R	54.57	62.65	74.95	66.36	55.91	47.71	80.25	67.20	53.16	63.95	68.41	62.38
	CW-P	66.91	50.39	76.12	68.46	69.37	49.25	81.40	71.75	72.88	50.39	69.57	64.55
	CW-E	72.57	61.38	80.89	74.45	70.96	55.12	82.84	74.65	78.15	66.39	75.56	73.68
G3	MSRA+CW-E	–	–	–	–	88.51	68.05	82.25	**81.49**	86.36	73.41	80.71	**80.02**
	Peking+CW-E	92.73	65.02	77.71	**80.38**	–	–	–	–	87.16	64.54	77.66	76.64

Major results

(4) Evaluation of cross-domain:

Table 7 Evaluation results (*F*-score) of NER for special cross-domain corpora

Test corpus	Number of tokens	*F*-score (%)				
		MSRA	Peking	CW-E	MSRA+CW-E	Peking+CW-E
Economics	24 346	76.16	70.11	70.14	**81.60**	75.45
Technology	18 488	76.34	70.68	72.03	**82.99**	79.19
Politics	20 883	81.26	77.58	75.47	**81.98**	78.54