Assessing Impact of Method Entities in a Special Task

Xiaole Li
Department of Information Management,
Nanjing University of Science and Technology
Nanjing, China
18260038108@163.com

Yuzhuo Wang
Department of Information Management,
Nanjing University of Science and Technology
Nanjing, China
wangyz@njust.edu.cn

Chengzhi Zhang*
Department of Information Management,
Nanjing University of Science and Technology
Nanjing, China
zhangcz@njust.edu.cn

ABSTRACT

Methods play an important role in the research. Identifying and analyzing entities about research methods can help scholars understand methods used in their field and accelerate the efficiency of scientific research. There are relatively few empirical analysis studies on method entities using quantitative methods. This paper takes named entity recognition (NER) as an example and evaluate the impact of method entities in this domain. This study found that conditional random field (CRF) is the most influential algorithms in NER. Deep learning algorithms have developed rapidly in the past 5 years. F-measure, precision and recall are the most widely used indices and measurements. Scholars do not pay enough attention to use tools and they prefer to use classic datasets.

CCS CONCEPTS

• Information systems →Information extraction

KEYWORDS

Named entity recognition, Impact of method entity, Full-text context analysis

1 INTRODUCTION

Methods play an important role in the science and technology. Different methods need to be used in the process of solving specific tasks. If methods appearing in academic papers can can be marked and evaluated, the current status of the research can be summarized to provide technical reference for beginners and accelerate the efficiency of scientific research.

Research methods comprise data collection techniques and data analysis techniques [1]. Research methods include multiple method entities, e.g. algorithms, tools, data sets and other entities used by scholars in solving problems. In this paper, we take NER as an example, and use full-text context analysis to label the method entities used in NER-related papers and assess their impact.

2 RELATED WORK

For the annotation of method entities, Scholars have used content analysis to label different method entities in academic papers. Zhao extracted the data set [1], and Howison explored the software entities [3]. But their work did not consider the task the method solve. For entity evaluation, Pan assessed software's impact by number of citations and mentions [4] and Wang evaluated algorithms' impact by number of papers, the total number of references and the mentioned location [5]. Therefore, this paper also uses context analysis to identify and assess the impact of method entities.

3 METHOLOGY

The research framework is shown in Figure 1. In order to assess the impact of method entities in a specific task, this paper first obtains academic papers from a website, then annotates method entities in the papers. Finally the impact of method entities is assessed based on different indicators respectively.

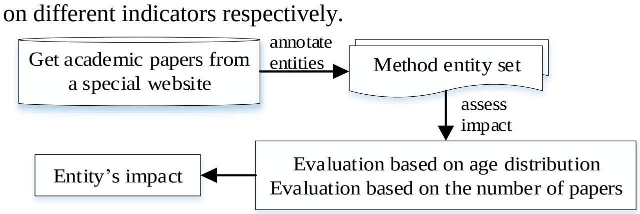


Figure 1: Research Framework

Data collection. We search papers containing 'named entity recognition' or 'extraction or identification' in their title from ACL Anthology (https://www.aclweb.org/anthology/). After deleting non-English papers and literature review, we get full-text content of 426 papers.

Method entity annotation. In the study, we annotate method entities that used by authors in academic papers, including algorithms, tools, data sources, indices and measurements. The entities are labeled by two senior students. Before formal annotation, we compile the annotation specification, and 50 articles are selected randomly for the pre-annotation. We employ Cohen's kappa coefficient to measure the interrater reliability (IRR) between the two students and achieve an IRR of 0.70, which provide sufficient reliability for two coders to code all the papers evenly. Table 1 gives the example of annotated entities.

Table 1: Examples of four types of method entities

Entity	Entity Type	Entity Sentence
Conditional Random Fields	algorithm & model	A Conditional Random Fields model annotates the entities components.
ACE 2005	data source	We used ACE 2005 for our experiments.
F-measure	index & measurement	Performance was measured with the F-measure score.
CRF++	tool	We used the CRF++ to …

Impact assessment of method entities. Two indicators are used to assess impact of method entities. One is *number of papers*: For

* Corresponding author.
JCDL'20, August 1–5, 2020, Virtual Event, China

each entity, we count the number of papers using it, the more papers the greater influence of the entity. Another is *age distribution:* We get the publication time of the paper through the download link and analyze the change in influence of method entities over time.

4 RSULST

After annotating and sorting, we get 345 data sources, 251 algorithms, 235 tools, and 73indices and measurements.

4.1 Evaluation based on the number of papers

Table 2 displays the top 5 highly-used entities in NER papers.

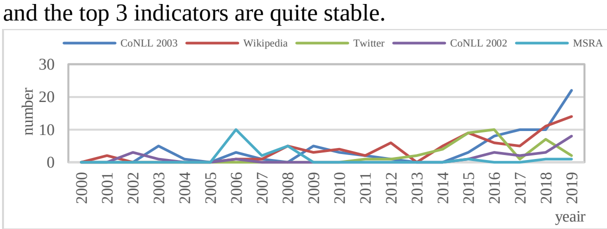
Table 2: Top 5 entities and the number of papers

Data source	Algorithm & model	Tool	Index & measurement
CoNLL 2003(74)	CRF(194)	CRF++(40)	F-measure(371)
Wikipedia(74)	BiLSTM(72)	OpenNLP(11)	Precision(258)
Twitter(37)	SVM(50)	word2vec(11)	Recall(256)
CoNLL 2002(22)	ME(50)	Stanford CoreNLP(10)	cross validation(55)
MSRA(20)	Viterbi(49)	Twitter API(10)	Accuracy(34)

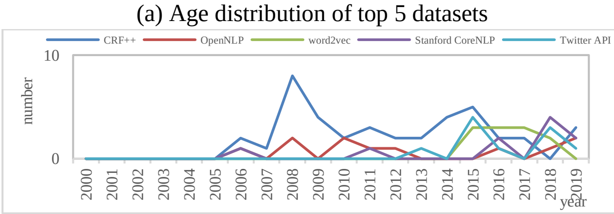
For data source, the data sets generated in classic conferences are used repeatedly by scholars, such as CoNLL 2002 and 2003. MSRA is the most commonly used dataset for Chinese NER. For algorithm, traditional machine learning methods, including the supervised learning algorithm Support Vector Machine (SVM), statistical models Maximum Entropy (ME) and CRF get the highest influence. Recently, the deep learning algorithm Bi-directional Long Short-Term Memory (BiLSTM) has been widely used. For tools, there are tools for specific algorithms, such as CRF++, as well as NLP and machine learning tools (CoreNLP, OpenNLP, etc.). For indicators, the use of F-measure, Precision, and Recall occupies 80% of the total, with the greatest influence.

4.2 Evaluation based on age distribution

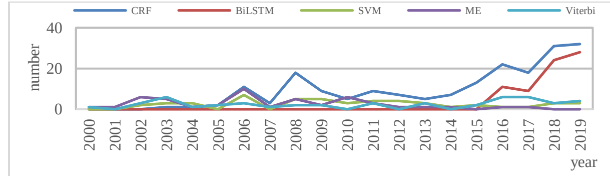
Figure 2 shows the development of impact of various entities. In general, the usage times of tools is low, algorithm is well developed, and the top 3 indicators are quite stable.



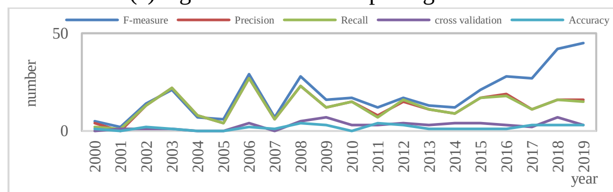
(a) Age distribution of top 5 datasets



(b) Age distribution of top 5 tools



(c) Age distribution of top 5 algorithms



(d) Age distribution of top 5 indices and measurements

Figure 2: The age distribution of top5 in four type entities

Figure 2(a) shows that classic datasets will be used many times in recent years and CoNLL 2003 get the most dramatic growth. In figure 2(b), we find that tools were used commonly before 2015, and they have declined in recent years. On the contrary, the use of algorithms has been greatly improved after 2015(see Figure 2(c)), indicating that scholars began to focus on the algorithm itself to solve complex NER tasks, instead of using tools directly. After 2015, BiLSTM is increasingly used by scholars, and its influence has been greatly improved. As shown in Figure 2(d), F-measure is the most commonly used indicators.

5 CONCLUSION AND FUTURE WORKS

Our results show that CRF is the most influential algorithm in NER related papers; CRF++ is the most commonly used tool; CoNLL2003 is the most commonly used dataset; F-measure, P, and R are widely used indicators. The most used entity is the algorithm and model, and the least is tool.

In terms of entity evaluation, there are still certain deficiencies. In the future, the full text can be used to determine the motivation for using entity, and the entity distribution in different sections can be used to further analyze entity's impact.

ACKNOWLEDGMENTS

Thanks Chen Y. to annotate the corpus.

REFERENCES

[1] Heting Chu and Qing Ke. 2017. Research methods: What's in the name? *Library and Information Science Research*, 39: 284-294. DOI: https://doi.org/10.1016/j.lisr.2017.11.001.

[2] Zhao M, Yan E, and et al. 2018. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 69(1): 32-46. DOI: https://doi.org/10.1002/asi.23919.

[3] Howison J, Bullard J. 2016. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9): 2137-2155. DOI: https://doi.org/10.1002/asi.23538.

[4] Pan X, Yan E and et al. 2015. Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, 9(4), 860-871. DOI: https://doi.org/10.1016/j.joi.2015.07.012.

[5] Wang Y and Zhang C. 2018. Using Full-Text of Research Articles to Analyze Academic Impact of Algorithms. *In International Conference on Information. Springer International Publishing*, Cham, English, 395-401. DOI: https://doi.org/10.1007/978-3-319-78105-1_43.