Research on Entity Relation Extraction for Military Field

Chen Liang[12], Hongying Zan[1], Yajun Liu[1], Yunfang Wu[3]
[1]Natural Language Processing Laboratory, School of Information and Engineering,
Zhengzhou University, Zhengzhou, China
[2]Key Laboratory of Universal Wireless Communications, Ministry of Education,
Beijing University of Posts and Telecommunications, Beijing, China
[3]Key Laboratory of Computational Linguistics, Peking University, Beijing, China
liangchenee@163.com, iehyzan@zzu.edu.cn,
liuyajun_gz@163.com, wuyf@pku.edu.cn

Abstract

The purpose of entity relation extraction is to obtain the relation between named entities from the text. However，most of the researches focus on the tourism, music, medical and other fields, seldom relate with entity relation extraction in the military field. This paper preprocesses military data sets, defines entity relations and completes classification. Based on dependency syntax relations, we extract the core verb features and dependency triples features in the sentences, using SVM as the classifier. The experiment results show that the features extracted in this paper can effectively improve the performance of relation detection and relation extraction. The F1 value for relation extraction can achieve 89%.

1 Introduction

As an important technology in natural language processing, entity relation extraction refers to the relation between named entities obtained from the corpus expressed by natural language. For example, the relation between people and articles (ART), the relation between people and organizations (ORG-AFF) are entity relations. Entity relation extraction is a more in-depth study based on named entity recognition(NER). It is widely used in question answering and machine translation. Nowadays, with the rapid development of the Internet, the relation extraction technology not only has profound theoretical significance, but also has broad application prospects.

The study of relation extraction was originally conducted on the main line of message understanding conference(MUC). Later, automatic content extraction(ACE) replaced it. A large number of excellent extraction methods have been proposed in the evaluation, which has effectively promoted the development of relation extraction research. Currently, the methods for entity relation extraction mainly include machine learning methods, hybrid extraction methods, and deep learning methods.

2 Related Works

Entity relation extraction is a key issue in information extraction. Researchers have made many breakthroughs in this field currently. In the machine learning based relation extraction methods, some studies use unsupervised methods to extract relations through clustering. YanY et al., (2009) presented a clustering methods based on pattern combination. Experiments on two different types of texts in Wikipedia had achieved good results. Shi et al., (2016) constructed a corpus for relational discriminant research based on ACE2005 task Chinese news text corpus, combining with the word vector trained by Sohu news data and using the convolutional neural network model for relation discrimination. The F1 value can reach 81.78%. The unsupervised methods have shown its good cross-domain versatility in the above research, but its ability to describe the types of relations and the extraction result in special fields are poor.

The methods based on supervised relation extraction have developed rapidly in recent years. Kambhatla (2004) proposed a method based on logistic regression. Rink et al., (2010) proposed extracting features of vocabulary and semantic, using support vector machine (SVM) for semantic relation extraction. Tratz et al., (2013) used the maximum entropy model to extract discrete features and classified them through WordNet external dictionary features. This method can seamlessly use various existing classifiers, but requires a large number of feature templates. Culotta et al., (2004) used a dependency tree to generate a kernel function of shallow parsing, first detecting whether there is a relation between two entities, and then finding out the specific relation between them.

The selection of features has a great influence on supervised learning methods. The types of entities, part-of-speech, word relative position in the sentence can be used as features. However, all of the above mentioned are shallow features and cannot fully express the relation between entity pairs. Therefore, we use grammar and syntactic features in this paper.

Some researchers also used the supervised methods based on feature vector. Zhou et al., (2009) used SVM classifier to extract the entity relation in the music field, they combined some features such as part-of-speech features, location features and context features, The F1 value reached 80.65%. On the basis of the two methods proposed by Dong et al., (2007) and Guo et al., (2014), Gan et al., (2016) added the dependency characteristics and the most recent syntax-dependent verb characteristics，and conducted experiments on the three data sets of Lushan, Taishan and Jinggangshan for the texts in the field of tourism, the F1 value of relation detection and extraction can reach 80.13%.

There are different performances on relation extraction in different fields. The researchers have done a lot of work in the general field and some specific fields of the entity relation extraction task, and finally get a richer corpus resources. However, there are relatively few studies in military field, and related research is mostly directed to NER. Feng et al., (2015) used a semi-supervised method based on conditional random field (CRF) model to extract special entities such as military titles and material names in military texts. Song et al., (2016) used military documents as data set, established a matching dictionary and designed 18 kinds of category labels, and used the Tri-Training algorithm for NER. The F1 value was up to 92%. Shan et al., (2016) used the word rules to integrate the military entities appearing in the combat documents and constructed the relation, and then used SVM to extract relation. Experiments showed that the word rules can effectively improve the performance.

This paper uses the military news text as a corpus for experimentation. First, the ANSJ system is used to preprocess the corpus, such as word segmentation and part-of-speech, and then use language cloud platform (LTP-Cloud) of Harbin Institute of Technology for dependency syntax analysis.[1] Core verbs and dependency triples are added as features. The experiment uses SVM to classify. The results show that core verb features and dependency triple features can effectively improve the performance of relation detection and relation extraction. The F1 value can reach 89%.

3 Methods

The selection of features directly determines the performance of the relation extraction. This section describes the algorithm that extracts verb features and dependency triple features, then introduces the classification method.

3.1 Feature Extraction Based On Syntax Analysis

In recent years, the method that has achieved good results is based on the feature vector extraction method. More common features include part-of-speech features, semantic features, and so on. This section preprocesses the corpus, and then extracts the core verb features and the dependency triples features.

3.1.1 Core Verb Feature Extraction

We uses Chinese word segmentation tool ANSJ for word segmentation and part-of-speech tagging.[2] The standard effects are as follows:

[1] http://www.ltp-cloud.com/intro/

[2] http://maven.nlpcn.org/org/ansj/

中国/ns 空军/n 轰 6K/nx 等/u 战机/n 执行/v 东海/ns 防空/vn 识别/vn 区/n 警巡/v (/w 图 /n)/w 。/w
（*The Chinese Air Force H-6 Strategic Bomber and other fighters perform the police patrol in the East China Sea Air Defense Identification Zone.*）

Where **n** is a noun, **ns** is a place name, **v** is a verb, **w** is a punctuation mark.

Through word segmentation and part-of-speech tagging, the whole text is traversed. According to the part-of-speech v, the core verbs in the sentence are extracted, and then added to the feature set. For example, "执行" and "警巡" are extracted as core verbs.

3.1.2 Dependency Triple Feature Extraction

In the dependency parsing analysis, predicates in sentences are the core of other parts, and are not subject to any other parts. All major components depend on the verb. Dependency analysis can reflect the semantic modification between the components of a sentence, and it can obtain long-distance collocation information, regardless of the relative physical position of the sentence components.

The extraction of relation triples has always been a very important part of the work of entity relation extraction. In this paper, the method of extracting the triples is to use the LTP-Cloud to perform the dependency parsing analysis after the word segmentation, and then complete the extraction through the dependency relation. The result of the relation labeling is shown in Figure 1:

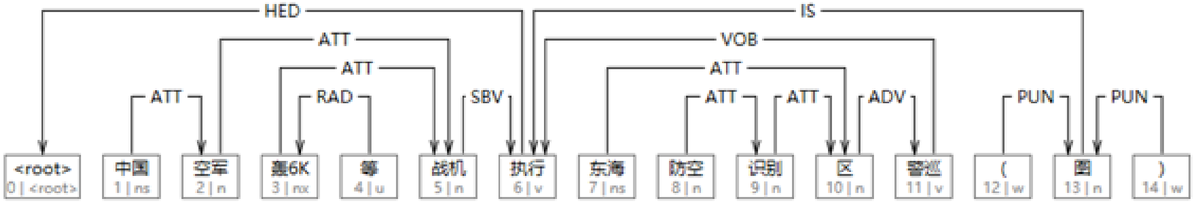

Figure 1. Dependency parsing analysis and part-of-speech tagging example

Where SBV, ATT, etc. are the dependencies between the words. The specific meaning is shown in Table 1:

Description	Tag
subject-verb	SBV
verb-object	VOB
Attribute	ATT
Adverbial	ADV
Head	HED
right adjunct	RAD
Coordinate	COO
Punctuation	WP
Independent structure	IS

Table 1. Tag set of dependency relations

After the word segmentation and dependency of corpus are determined, the extraction algorithm of dependency relation triples is as follows:

Step 1: generate a dictionary that holds the dependency child nodes for each word in the sentence. That is, find the child node corresponding to each word by "head" and record the dependency in the dictionary.
Step 2: extract the verbs in a sentence according to the existing part-of-speech tagging. The entities e1 and e2 can be found by searching the dependency child nodes dictionary of the verb. If the dictionary of an entity contains ATT, COO and other dependencies, and then the entity should be improved to get the longest.
Step 3: after completing entity e1 and e2, according to the dependency relation dictionary of this verb, extract the dependency triples relations, and finally obtain the dependency and determine the syntactic relation (e.g., SBV-VOB corresponds to the subject-predicate relation, SBV-CMP corresponds to the prepositional object verb complement relationship, etc.).
Step 4: according to the NER annotation completed by LTP-cloud, the entities of the name, place and institution in the sentence are extracted.

Through this algorithm, we have selected some samples of dependency triples extraction, as shown in Table 2.

Example	Dependency relation triple	Sentence type
中国空军轰 6K 等战机执行东海防空识别区警巡。 （*The Chinese Air Force H-6 Strategic Bomber and other fighters perform the police patrol in the East China Sea Air Defense Identification Zone.*）	（中国空军轰 6K 战机，执行，东海防空识别区警巡） *(The Chinese Air Force H-6 Strategic Bomber, perform, the police patrol in the East China Sea Air Defense Identification Zone)*	*Subject predicate object relation*
对进入防空识别区的外国军机及时识别、判明类别，并进行了全程监视。 *(The foreign military aircraft entering the air defense identification zone were identified in a timely manner, and the whole process was monitored.)*	（外国军机，进入，防空识别区） *(The foreign military aircraft, enter, the air defense identification zone)*	*Attributive postposition Verb object relation*
中国海军 851 号情报收集舰采用长首楼船型，首楼一直延伸到直升机甲板处，这样既增加了干舷高度，又加大了船内空间。 *(The Chinese Navy's 851 intelligence collection ship adopts the long-first ship type, and the first floor extends to the helicopter deck, which increases the height of the freeboard the space inside the ship.)*	（中国海军 851 号情报收集舰，采用，长首楼船型） （首楼，延伸到，直升机甲板处） *(The Chinese Navy's 851 intelligence collection ship, adopts, the long-first ship type)* *(the first floor, extends, the helicopter deck)*	*Subject predicate object* *prepositional object* *verb complement*

Table 2. Dependency Relation triples

3.2 Entity Relation Detection and Relation Extraction SVM Model with Syntactic Features

3.2.1 SVM Model

SVM is a supervised binary classifier，which was officially released in 1995 and soon became the mainstream technology for machine learning. SVM is mainly used to analyze linear separability problems, and linear indivisible problems can be solved by using nonlinear mapping algorithms. The strategy used in the learning process is to maximize the interval between the two types of sample sets. The result of the learning is the set SV of support vectors, the associated weights α_i and constants b.

The basic SVM is shown in Figure 2, which is a linear classifier. But by introducing a kernel function, data from the Euclidean space can be nonlinearly converted into a high-dimensional space for performing nonlinear classification. The kernel function determines the final performance of the SVM. This paper compares them by using several different kernel functions:

poly kernel function:

$$K(x, z) = (x \cdot z + 1)^p \qquad (3\text{-}1)$$

The corresponding SVM is a multi-classifier of the power of p. Based on this situation, the classification decision function is:

$$f(x) = sign(\textstyle\sum_{i=1}^{N_s} a_i^* y_i (x_i \cdot x + 1)^p + b^*) \qquad (3\text{-}2)$$

Gaussian kernel function:

$$K(x, z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2}) \qquad (3\text{-}3)$$

The corresponding support vector machine is a Gaussian radial basis classifier. Based on this situation, the classification decision function is:

$$f(x) = sign(\textstyle\sum_{i=1}^{N_s} a_i^* y_i \exp(-\frac{\|x-z\|^2}{2\sigma^2}) + b^*) \qquad （3\text{-}4）$$

Linear kernel function：

$$K(x, z) = x \cdot z \qquad （3\text{-}5）$$

Linear classification decision function:

$$f(x) = sign(w^* \cdot x + b^*) \qquad （3\text{-}6）$$

In this paper, SVM is selected as the classifier. And linear kernel is proven to have the best results.

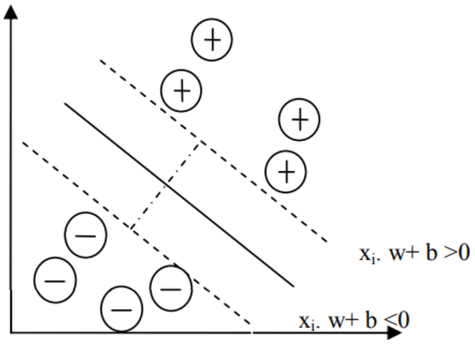

Figure 2. SVM model

3.2.2 Entity Relation Detection and Relation Extraction SVM Model with Syntactic Features

This article uses the TF-IDF method to process the text and use the result of putting it into the SVM as a baseline. Then we add the extracted verb features and the dependency ternary features to each pre-processed sentence, perform TF-IDF processing on the corpus of the merging feature, put it into the SVM for classification. Finally, the experimental results were compared. The schematic diagram of the algorithm flow in this paper is shown in Figure 3.

The experimental algorithm analyzes from two perspectives: relation detection and relation extraction. Relation detection divides experimental data into two categories: the " NON " category is treated as a separate category, and the remaining types of entity relations are combined into a correlation, defined as a " HAS " type. The task of relation extraction is to classify the data sets of the "HAS" type after removing the "NON" type in the data sets. In the next section, this paper will prove the improvement of relation detection and extraction performance by verb feature and dependency triple feature through detailed data analysis.

4 Experiment

4.1 Data Set

The experimental data used military news obtained from People's Daily Online, China News, Netease News, Sohu News and other websites. Its contents cover diplomatic speeches, military exercises, war conflicts, visits, etc., including a variety of relations of people, places, weapons. The data set used in this experiment has a total of 10,697 sentences. For this dataset, the ANSJ system is used for word segmentation and part-of-speech tagging. The LTP-Cloud is used for dependency parsing analysis. And then through the prediction and analysis of verbs and entities in the military field, five types of entity relations are formulated. Other classes (None) contains entity relations such as "是" and "有" that have no obvious semantic association.

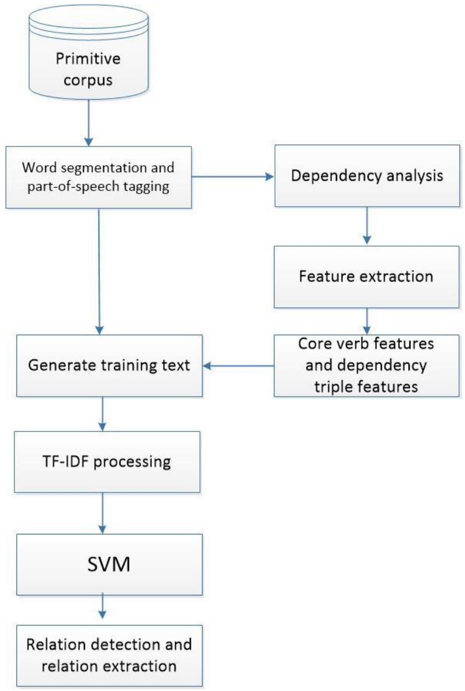
Primitive corpus
Word segmentation and part-of-speech tagging
Dependency analysis
Feature extraction
Core verb features and dependency triple features
Generate training text
TF-IDF processing
SVM
Relation detection and relation extraction

Figure 3. Workflow Diagram

The name and quantity information of the entity relation type in the experimental data set are shown in Table 3.

Entity Relation	Symbol	Numbers
diplomatic speak	SPE	2,620
military action	ACT	415
displacement action	DIS	310
negotiation	NEG	55
visit	VIS	36
other (none)	NON	6,842

Table 3. Statistics Information of Entity Relation of Military Dataset

4.2 Evaluation metrics

The experimental uses commonly used evaluation metrics: P (precision rate), R (recall rate) and F1 value. P is the proportion of the correct result in the returned result, R is the ratio of the return results to all the results. The specific evaluation formula is:

$$P = \frac{TP}{TP+FP} \quad (4\text{-}1)$$

$$R = \frac{TP}{TP+FN} \quad (4\text{-}2)$$

$$F1 = \frac{2\times P\times R}{P+R} \quad (4\text{-}3)$$

Where TN is true positive, TN is true negative, FP is false positive, FN is false negative.

4.3 Experiment and result analysis

In this paper, the entire data set was randomly selected, 80% of which was used as the training set, and 20% was the test set. The experiments use the Taiwan University LIBSVM toolkit.

4.3.1 Relation Detection

Since there are many types of non-relation types in the corpus used in this data set, the relation detection is first performed on the text, that is, identifying whether there is a semantic relation between an entity pair. Table 4 shows information about whether entities in the dataset have relationships.

Entity Relation	Numbers
HAS	3,855
NON	6,842

Table 4. Entity Relation Information for Relation Detection

Firstly, use the linear kernel, polynomial kernel (POLY), and Gaussian kernel (RBF) of SVM to compare the performance of relation detection on the basic features. See Table 5 for specific comparisons.

Kernel	P	R	F1
POL	73%	71%	66%
RBF	74%	73%	69%
Linear	82%	65%	72%

Table 5. Comparison of relation detection performance of different SVM kernels (%)

The results show that when the linear kernels is selected, the precision, recall rate and F1 value of the relation detection are higher than other kernels, and there is no over-fitting. Therefore, the subsequent comparison experiments all use linear kernels.

This paper chooses the core verb feature and the dependency relation triple feature, comparing it with the featureless relation detection result, the comparison results are shown in Table 6.

Entity Relation	P	R	F1
Baseline	82%	65%	72%
+Verb	90%	73%	81%
+Relation triples	90%	73%	81%

Table 6. Comparison of relation detection results with core verbs and dependency triple features (%)

For this table, "+Verb" means adding a core verb feature, and "+Relation triples" means adding a dependency triple feature.

After comparison, it can be found that in the relation detection task, the precision, recall rate and F1 value are obviously improved after the core verb features and the dependency triple features are added, but the two features have the same improving effect on relation detection. We do further comparisons in the relation extraction task.

4.3.2 Relation extraction

For the performance of relational extraction, this paper uses the same method as relation detection. Since there are many unrelated data set type, the relation between a small number of samples cannot be identified when the whole classification is performed, and the overall precision, recall rate and F1 value are low. Therefore, this paper only classifies those data that have relations, which is divided into five categories. This paper extracts core verb features and dependency triples features, and compares them with the featureless extraction. The comparison results are shown in Table 7.

Entity Relation	Baseline			+Verb			+Relation triples		
	P	R	F1	P	R	F1	P	R	F1
action	76%	53%	62%	77%	72%	74%	82%	75%	79%
move	77%	47%	58%	81%	76%	79%	81%	74%	77%
negotiate	87%	61%	72%	85%	79%	81%	91%	74%	82%
speak	84%	95%	89%	90%	94%	92%	91%	95%	93%
visit	75%	54%	63%	87%	67%	75%	84%	67%	74%
avg / total	82%	83%	82%	88%	88%	88%	89%	89%	89%

Table 7. Comparison of relation extraction results with core verbs and dependency triple features (%)

It can be seen from the comparison that in the task of relation extraction, the precision, recall rate, and F1 value are significantly improved by adding the core verb features and the dependency triples features. And the result is extracted after the ternary feature is added, The F1 value increased from 82% to 88%, and the lifting effect is very obvious.

Through experiments, we came to the following conclusions:

1）In entity relation detection and extraction, the linear kernel effect is much better than the POLY core and RBF core.

2）After adding the core verb feature, the relation detection result F1 value increases by 5%, and the relation extraction F1 value increases by 6%.

3）Adding the dependency relation to extract the ternary feature, the relation detection result F1 value increases by 5%, and the relation extraction F1 value increases by 7%.

Further analysis of the above data reveals that linear kernel SVM is still the preferred technology for text classification. The reason is that each word is an attribute of the data, whose attributes are high space dimensionality and large redundancy. It is enough to “shatter” and classify it. In terms of integration features, this paper adds the word frequency of named entities and relational verbs after the integration of core verbs feature and dependencies extracting ternary feature, and the ability to discriminate keywords is improved for TF-IDF methods. Therefore, both of the relation detection and extraction achieved good results.

5 Conclusion

Chinese relation extraction has the characteristics of complex corpus and sparse data, which greatly increase the difficulty of relation detection and extraction. In this paper, we use the method of supervised entity relation extraction based on the feature vector, and propose a new relation extraction model for military prediction. The work of this paper mainly includes:

(1) Pretreatment of military field corpus

In terms of experimental datasets, this paper uses military news texts as a corpus for experimentation. First, the ANSJ system is used to preprocess the corpus such as word segmentation and part-of-speech tagging, and then the LTP-Cloud is used for dependency syntax analysis.

(2) The extraction algorithm of core verb features and dependency triples features

In terms of feature selection, this paper extracts the core verb features of each sentence by doing word segmentation and dependency analysis. The dependency triples feature is generated according to the algorithms of the longest entity and dependency relation combination. In this paper, the feature weights are generated by using the TF-IDF method, and the core verb features and the dependency triple features are added for experimental comparison

The military corpus used in this paper is unstructured information, so the processing of the data set in this paper is very cumbersome. Of course, the new data set also has the problem of less training data and uneven distribution of semantic relations, which is why the final experimental F1 value has not been able to reach more than 90%. In future research, we can try to use deep learning methods, using models such as CNN and RNN to extract relations. It is also possible to focus on the exploration of weakly supervised machine learning methods to generate large-scale, high-quality markup corpora, reducing the reliance on large-scale hand-marked corpora.

References

Kambhatla, & Nanda. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. ACL 2004 on Interactive Poster and Demonstration Sessions (pp.22). Association for Computational Linguistics.

Rink, B., & Harabagiu, S. (2010). UTD: Classifying semantic relations by combining lexical and semantic resources. International Workshop on Semantic Evaluation (pp.256-259). Association for Computational Linguistics.

Ran, C., Shen, W., Wang, J., & Zhu, X. (2015). Domain-Specific Knowledge Base Enrichment Using Wikipedia Tables. IEEE International Conference on Data Mining (pp.349-358). IEEE Computer Society.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network.

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. *Meeting on Association for Computational Linguistics* (pp.415). Association for Computational Linguistics.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. *ACL 2009, Proceedings of the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore* (Vol.2, pp.1021-1029). DBLP.

Shi Feng.(2016).Research On Relation Extraction In Chinese News Text. (Master's thesis, Harbin Institute of Technology).

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural Relation Extraction with Selective Attention over Instances. *Meeting of the Association for Computational Linguistics* (pp.2124-2133).

Tratz, S., & Hovy, E. (2013). ISI: Automatic classification of relations between nominals using a maximum entropy classifier. *International Workshop on Semantic Evaluation* (pp.222-225). Association for Computational Linguistics.

Culotta, A., & Sorensen, J. (2004). Dependency tree kernels for relation extraction. *Meeting on Association for Computational Linguistics* (pp.423-429).

Zhou Lanjun. (2009). Research Of Chinese Relation Extraction In The Field Of Music. (Doctoral dissertation, Harbin Institute of Technology).

Gan, L., Wan, C., Liu, D., Zhong, Q., &Jiang, T.(2016). Chinese Named Entity Relation Extraction Based On Syntactic and Semantic Features. *Journal of Computer Research and Development, 53*(2), 284-302.

Dong, J., Sun, L., Feng, Y., & Huang, R. (2007). Chinese Automatic Entity Relation Extract ion. *Journal of Chinese Information Processing, 21*(4), 80-85.

Guo, X., He, T., Hu, X., & Chen, Q. (2014). Chinese Named Entity Relation Extraction Based On Syntactic And Semantic Features. *Journal of Chinese Information Processing, 28*(6), 183-189.

Feng, Y., Zhang, H., Hao, W. (2015). Named Entity Recognition for Military Text. *Computer Science, 42*(7), 15-18.

Song,R. (2016). Research On The Key Technology Of Named Entity Recognition And Relation Extraction In Military Field. (Doctoral dissertation, Harbin Institute of Technology).

Shan, H., Wu, Zhaolin., Zhang, H, & Liu Peilei. (2016). A Military Named Entities Relation Extraction Extraction Method Based on SVM Integrated with Word Rules. *Command control and simulation, 38*(4), 58-63.

Zhou, Z.(2016). *Machine learning*. Tsinghua University Press.

Li H.(2012). *Statistical learning method*. Tsinghua University Press.

Chang, C. C., & Lin, C. J. (2011). *LIBSVM: A library for support vector machines.* ACM.