A Generation Method for Chinese Equipment Name Abbreviations Based On Rules

Jiwei Qin[1]*, Liangli Ma*[1]*, Yanping Wang*[2]
1.Department of computer Engineering, Naval university of Engineering, Wuhan, 43003, China
2. *Unit 92815, Ningbo, 315700, China*

Keywords: NER; Chinese equipment name; abbreviations; finite automate

Abstract

In the field of CALS, Chinese equipment name abbreviations are widely used, and recognizing these abbreviations named entities has an important significance in the CALS knowledge discovery, but there are less research on it. A generation method based on rules is proposed for Chinese equipment name abbreviations. First, we analyses the naming rules and abbreviation characteristics of Chinese equipment name. Next, we establish an automatic generation model by the finite automaton. Then, we set the weight of mapping rules according to a series of factors. Finally, we export the average weight of the mapping rules set isn't less than the threshold to abbreviations dictionary for solving the problem of recognizing the abbreviations named entities. By referring to the experiment result from three types of equipment names data, its F rate reaches 93.88%. The results show that the generation method based on rules is an effective approach and get promising results in equipment named entities recognition.

1 Introduction

Named Entity Recognition (NER) refers to recognize the specific meaning entities in text, such as person name, location name and organization name[1]. The correct recognition of these information, will help to improve the effect of word segmentation and tagging, and lay the foundation for information extraction, information retrieval, machine translation and other fields. The exact meaning of named entities according to the specific application field to decide, and in the field of Continuous Acquisition & Life-cycle Support (CALS)[2-3], the equipment name is an important specie of named entity. But due to the complexity of Chinese equipment name structure, these full names often appears as abbreviations in the documents of CALS field, such as equipment "BGM-109'战斧'巡航导弹", someone may abbreviate it to "'战斧'导弹", and other may abbreviate it to "109 导弹", these situations bring certain difficulties to the NER of Chinese equipment name. So the recognition of these abbreviation is helpful to improve the retrieval performance of information system in CALS field, and enhance the efficiency of maintenance support of equipment.

2 Research Status

In recent years, Chinese named entity recognition research draws more and more attention. For example, Tsai etc. proposed a hybrid method based on maximum entropy[4]; Feng Yuanyong etc. proposed Chinese named entity recognition fast algorithm based single character hints[5]; Chang and Lai proposed the prediction method based on Hidden Markov model HMM abbreviation[6]; Sun et al proposed support vector regression SVR method to measure the difference score[7], and introduced the DPLVM method to improve the original model[8]; Lian Yushun and Zhao Yuming proposed an automatic generation method for Chinese organization abbreviations based on segmentation information for the selection of the optimum abbreviation[9]; Gao Qiang and You Hongliang proposed a cascaded model combining the factors of rules and statistics for NER in the field of defences, and it can recognize parts of equipment full names[10].

The above research on NER mainly oriented the fields of person name, location name and organization name, but there are less research on the NER of Chinese equipment name. First, we divide the Chinese equipment name into each component through analysing the structural characteristics of it. Secondly, according to the composition rules of abbreviation, we construct the finite automata model of the generation of Chinese equipment name abbreviation, and set the weight of each mapping rule of the finite automata. Finally, when the average weight of the mapping rules set isn't less than the threshold, we export the corresponding abbreviation to the abbreviation dictionary for solving the problem of recognition of the Chinese equipment name abbreviation.

3 Abbreviations generation model

3.1 Naming rules of Chinese equipment name

Every type of equipment has a corresponding naming rule. According to the analysis of naming rules of multi equipment types, we get the general naming rules of equipment as follows:

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

$$t / c - ef * m_1 m_2 ... g * m_p ... m_y * h"i"lj_1 .. j_x k_1 ... k_z \qquad (1)$$

The structure of Chinese equipment name is composed of three parts: model name, nickname, the information of multifunctional purpose. In which $t/c\text{-}ef*m_1m_2...g*m_p...m_y*h$, $”i”l$ and $j_1j_2...j_xk_1k_2...k_z$ respectively represent the model name, nickname and the information of multifunctional purpose.

Where service code t represents the application scope of equipment. Function code c is the technical information of equipment basic tasks, equipment type, launching platform, etc. Serial number e illustrates the serial number of equipment. Modification code f illustrates a big improvement of equipment. Smaller modification code $m_1m_2...g*m_p...m_y*$ is used to illustrate a smaller modification or production batch. Other information code h represents the other description information of equipment name, such as the manufacturer information. Nickname i represents the nickname of equipment. Nickname qualifier l is used to qualify the nickname i. Function information $j_1j_2...j_x$ is the information of multifunctional purpose. Type information $k_1k_2...k_z$ is the information of equipment type, in which k_z is suffix word, and must appear in the equipment name.Punctuation explanation: “/” is the sprit, “-” represents the connector, “”” is the quote, and “*” represents the arbitrary punctuation.

e and k_z must appear in the Chinese equipment name, and whether other structure appear in the name determined by the specific equipment category. For example of the aircraft name of US Army. The divided name structure as shown in Fig.1.

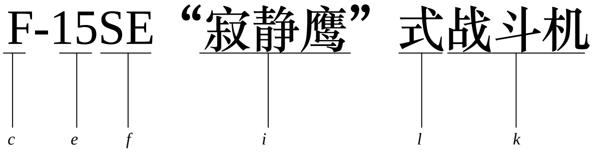

Fig.1: Example of the divided equipment name structure

3.2 Analysis of the composition of abbreviation

The segmentation processes of Chinese equipment name components is the first step of abbreviation generation, the processes are divided into the segmentation processes of model name and Chinese string which are composed by nickname and the information of multifunctional purpose. Model name is segmentation by punctuation, the boundary of digital string, the boundary of letter string and the boundary of other character string. Chinese string is segmentation by using the reversion maximum match algorithm to match the dictionary of universal words. In order to easily treat subsequent processes, we delete the punctuation and nickname qualifier l in the segmentation result.

Abbreviation is composed by parts of the full name components, According to the analysis of equipment name abbreviation, we obtain the composition rules of abbreviations as follow:

i) Service code t attaches on function code c.

ii) Serial number e is the core structure of model name, other codes must attach on e.

iii) g is the core structure of smaller modification code, m_y $(y=1,2…)$ must attach on g.

iv) Individual nickname i can be regarded as a reasonable abbreviation, and also can connect the model name abbreviation to compose abbreviation.

v) Parts or all of j_x must attach on k_z appeared in the abbreviation. The combination of j_x and k_z must attach on the reasonable combination of other parameters to compose abbreviation.

According to the analyses above, we obtain the regular expression of abbreviation generation of Chinese equipment name as follow:

$$\begin{cases} x = ((t \mid \varepsilon)c \mid \varepsilon)e(f \mid \varepsilon)((m_1 \mid \varepsilon)...g...(m_y \mid \varepsilon) \mid \varepsilon)...(h|\varepsilon) \\ y = i \\ z = (x \mid y \mid xy)((j_1 \mid \varepsilon)(j_2 \mid \varepsilon)...(j_x \mid \varepsilon)(k_1 \mid \varepsilon)(k_2 \mid \varepsilon)...k_z \mid \varepsilon) \end{cases} \qquad (2)$$

In which x represents the regular expression of model name, y represents the regular expression of nickname, z is the regular expression of whole equipment name, and $\mathcal{E}$ represents the null character string.

According to the regular expression, we can obtain the application range of used abbreviation. However, this expression generates excess abbreviations through testing multi types of equipment names, and excessive number of abbreviations wastes the scanning time in the matching process. So the generation number need be further reduced.

3.3 The finite automaton model of abbreviations generation

Due to the different importance of different components in equipment name, for example, the importance of nickname is far higher than the importance of Service code, the importance of abbreviation is different at different configurations, and it determines the usage frequency of a certain class of abbreviations in some extent. So according to the importance of elements in regular expression of abbreviation generation, we can set the weight to filter important abbreviations. In order to intuitive obtain results, the regular expression is converted into finite automaton, and each mapping rule of finite automaton is endowed with weight according to the importance analysis of it.

According to the regular expression of abbreviation generation, we obtain the finite automaton of abbreviation generation $DFA=(Q,\Sigma,t,q_0,F)$, in which state set $Q=\{q_0, q_1, q_2, q_3, q_4, q_{50,} q_{51}{}_{...}q_{5y} ,q_6 ,q_7 ,q_{80} ,q_{81} \cdots q_{8x}, q_{91}{}_{...}q_{9z}\}$, $y\geq 0$, $x\geq 0$, $z\geq 1$, The ways of arriving a certain state have two types: recognition of word in input vocabulary and recognition of null character string, specific ways see the mapping rules. The meaning of each state as shown in the Table 1.

State	The meaning of state
q_0	Initial State
q_1	Recognized *t*
q_2	recognized *c*
q_3	Recognized *e*
q_4	Recognized *f*
q_{50}	Recognized *g*
$q_{51}...q_{5y}$	Recognized m_y
q_6	Recognized *h*
q_7	Recognized *i*
q_{80}	Initial State of recognizing j_x
$q_{81}...q_{8x}$	Recognized each j_x
$q_{91...}q_{9z}$	Recognized each k_z

Table 1: The meaning of each state

Input vocabulary $\Sigma=\{t,c,e,f,g,m_{1...}m_y,h,i,j_1...j_x,k_{1...}k_z\}$, $m_{1...}m_y$ can be divided into $m_{1...}m_{p-1}$ and $m_{p...}m_y$. Because the divided number of words of $m_{1...}m_y$, $j_1...j_x$ and $k_{1...}k_z$ is different in every equipment name, the divided number of states of $q_{51}...q_{5y,}$ $q_{81}...q_{8i}$ and $q_{91...}q_{9z}$ is non-fixing. Mapping rules as shown in Fig.2. Initial state is q_0, ending state set $F=\{q_6, q_7, q_{9z}\}$. According to the above rules, we can obtain the finite automaton state diagram of abbreviation generation as shown in Fig.2.

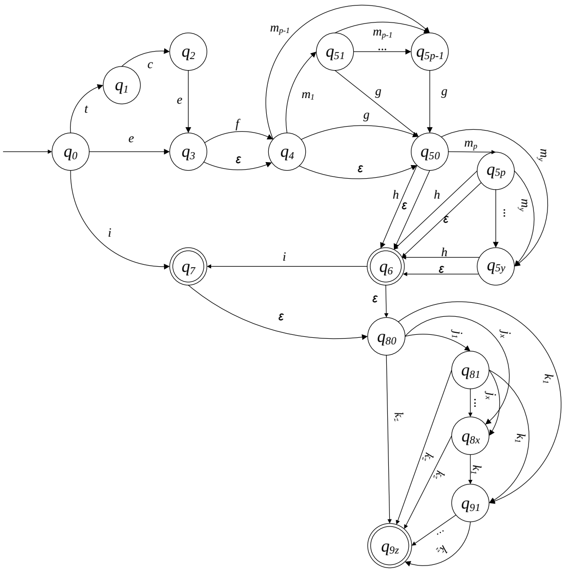

Fig.2: the finite automaton state diagram of abbreviation generation

The factors of weight value distribution of mapping rules in the finite automaton state diagram of abbreviation generation as follow:

i) Whether it can determine the uniqueness of equipment: Whether generated abbreviation by recognizing that mapping rule can uniquely recognize the equipment.

ii) Accurate to the Particle size of equipment: Generated abbreviation by recognizing that mapping rule determines particle size of equipment, particle size of equipment are divided into number level, modification level and smaller modification level, the modification level has the best place, the number level takes second place, and the smaller modification level is the worst.

iii) Users' familiarity: If users are familiar with the recognition elements of those mapping rules, these words usually are core structure of equipment name, such as nickname *i* and serial number *e*.

iv) Abbreviation's length: Generated abbreviation's length by recognizing that mapping rule, and the shorter length of abbreviation is better.

v) Other factors: Such as the recognition elements which can be omitted by the official regulation, those weights of corresponding mapping rules are lower.

Because of the different number of recognized mapping rules in the process of abbreviation generation, we define the weight of abbreviation by calculating the average weight of recognized mapping rules, and output those abbreviations when those weights are not less than specific threshold.

3.4 Example of abbreviations generation

For example of name form *c-ef-g-h"i"lj*$_1$*j*$_2$*...k* of the U.S. military aircraft, we obtain the finite automaton model of abbreviation generation of aircrafts' name through analysing the naming rule, and then set the corresponding weight by the factors of weight value distribution, we can obtain the finite automaton state diagram of aircraft name's abbreviation generation as shown in Fig.3.

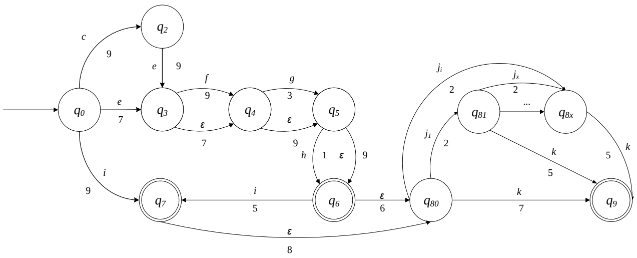

Fig.3: Finite automaton state diagram of example

As for the equipment name which's some optional information is not existing, the corresponding elements of optional information in mapping rules are revised as $\mathcal{E}$. If two mapping rules which's recognition elements are also $\mathcal{E}$ are existing between two same state nodes, the mapping rule of corresponding smaller weight should be delete. According to the aircraft abbreviation generation model in Fig.3, The set of threshold of aircraft abbreviation is 8.0 according to experimentation and results comparison, the abbreviation generation of other types of equipment is similar. The output abbreviations forms as shown in Table 2.

Form	Weight	Form	Weight
i	9.00	*cef*	9.00
cde	8.60	*ef*	8.50
cefi	8.33	*cefk*	8.29
cefik	8.13	*ik*	8.00
e	8.00	*cek*	8.00
cei	8.00		

Table 2: Output abbreviations forms of aircraft name

4 Experiment results and analysis

In order to evaluate the recognition effect of the abbreviation generation model, we obtain 165 names from aircraft, warship and united electronic equipment three types of equipment from Wikipedia and Baidu Encyclopaedia, and export their abbreviations to abbreviation dictionary (abbr-dict) through the abbreviation generation model in this paper. Meanwhile, we use names to obtain 561 relevant pages, 2823 names and abbreviations as the test set on the Internet through fuzzy retrieval. To recognize the named entities of equipment on the conditions of using abbr-dict and not using it respectively, and calculate the averages of recall (R), precision (P) and F-measure (F) of multiple names, the experiment results as shown in the Table 3.

Equipment type	Not use abbr-dict		
	R	P	F
Aircraft	31.23%	100%	47.60%
Warship	29.67%	100%	45.76%
Electronic	26.12%	100%	41.42%
All Types	28.73%	100%	44.64%
Equipment type	Use abbr-dict		
	R	P	F
Aircraft	94.92%	91.45%	93.15%
Warship	90.49%	92.33%	91.40%
Electronic	90.14%	94.80%	92.41%
All Types	91.96%	93.46%	92.70%

Table 3: Comparison of experimental results

From analysing the test set, we find the equipment names usually appear as abbreviation form except the first time due to the complexity of full name， it causes the excessive low recall rate on the condition of not using abbreviation dictionary. But also because of the complexity, the recognition error rate is extremely low when the full name once hits, the F rate is up to 100% in the test set. However, because of the excessive difference between recall and precision, the F rate is on the low side. On the condition of using abbr-dict, because generated abbreviations cover most use forms of abbreviation, this factor makes the rates of R, P and F reach above 90%, the results were comparatively satisfied. Because ambiguous abbreviations like individual serial number abbreviations are considered as reasonable abbreviations, it may other information like digit which is unrelated to equipment name is mistakenly recognized as equipment name abbreviation. Such as aircraft name “F-15SE’寂静鹰’式战斗机”, “15” is a reasonable abbreviation in its abbreviations based by our model, and it causes all of “15” strings are recognized as the abbreviation of this name. Such as this problem makes a certain decrease of precision. On the condition of deleting ambiguous abbreviations in the dictionary, we experimentize anew to obtain results as shown in Table 4.

Equipment type	R	P	F
Aircraft	93.11%	96.25%	94.65%
Warship	89.67%	97.81%	93.71%
Electronic	89.43%	98.07%	93.55%
All Types	90.57%	97.46%	93.88%

Table 4: New experimental results

From the results it can be seen that the recall rates decrease a little while the precision rates and F rate improve in some degree. The method of deleting ambiguous abbreviations in abbreviation dictionary is feasible. In the systems and documents of CALS, due to the austere writing requirements, the ambiguous abbreviations rarely appear in the related system, while the appearance probability of these ambiguous abbreviations greatly improved in the retrieval keywords of user input. Therefore, whether the removal of ambiguous abbreviations is determined according to the different conditions, under the condition of recognizing entities in the related documents of CALS, we eliminate those ambiguous abbreviations from dictionary, while under the condition of retrieval keywords, we introduce them to matching results, and this method has good recognition effect under the multi-conditions. Due to recognition error of part abbreviations’ boundary, the R rate is lower as compared to P rate, the generation method need further improve in studies later.

5 Conclusions

Based on the systematical analysis of the naming rules and abbreviation composition rules of Chinese equipment name, we propose a generation method for equipment name abbreviations, it uses the related rules to construct finite automaton attach weight, and generates the abbreviations which’s weights are not less than specific threshold. According to the experiment results and analysis of different equipment types, generated abbreviations through this method have a good recognition effect. For the boundary recognition, this method has some disadvantages. In the future research, we will focus on combining this method and statistical factors to improve the recall rate of abbreviations recognition.

References

[1] ZHANG Xiaoyan, WANG Ting, CHEN Huowang. Research on Named Entity Recognition. *Computer Science*, 32(4), pp.44-48, 2005.

[2] WANG Qiang, XIE Wenxiu, WEI Chenxi. CALS Theories and Build-up of Information System of Integrated Logistic Support. *Journal of the Academy of Equipment Command & Technology*, 16 (2), pp.33-37, 2005.

[3] PENG Wenjun. Research on Implementation Process o f Weapon Informatization Based on CALS. *Computer and Modemization*, (8), pp.37-39, 2012.

[4] Tsai T, Wu S, Lee C, et al. Mencius a Chinese Named Entity Recognition Using the Maximum Entropy-based Hybrid Model. *International Journal of Computational Linguistics & Chinese Language Processing*, 9(1), pp.65-81, 2004.

[5] FENG Yuanyong, Sun Le, Li Wenbo, et al. A Rapid Algorithm to Chinese Named Entity Recognition Based on Single Character Hints. *Journal of Chinese Information Processing*, 22(1), pp.104-110, 2008.

[6] Chang J, Lai A. A Preliminary Study on Probabilistic Models for Chinese Abbreviation. *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pp.9-16, 2004.

[7] Sun X, Wang H F, Wang B. Predicting Chinese abbreviations from definitions: An empirical learning approach using support vector regression. *Journal of Computer Science and Technology*, 23(4), pp.602-611, 2008[8] Sun X, Okazaki N, Tsujii J. Robust approach to abbreviating terms: A discriminative latent variable model with global information. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, pp.905-913, 2009.

[9] LIAN Yushun, ZHAO Yuming. An Automatic Generation Method for Chinese Organization Abbreviation Based on Segmentation Information. Computer Application and Software, 31(4), pp.153-156, 2014.

[10] Gao Qiang, You Hongliang, Study on Named Entity Recognition Based on Cascaded Model for Field of Defense. New Technology of Library and Information Service, 11: pp.47-52, 2012