Improvement of English-Chinese Translation Method by Feature Reduction and Rule Optimization Based on Rough Set Theory

Hongxia Wei
Foreign Languages School of Anhui Polytechnic University
Anhui Wuhu 241000
China
whxahgf@163.com

Journal of Digital Information Management

ABSTRACT: *The traditional machine translation algorithm faces low efficiency and low accuracy due to complex grammar and multiple rules of English noun phrases. A kind of noun phrase identification method based on a rough set was proposed to improve the accuracy of English noun phrases. The rough set method regarded the identification of English noun phrases as a decision problem. We used rough set theory to reduce features, optimize rules for English noun phrases, and finally identify them. Then, a simulation experiment was carried out on an English noun phrase sample on the Wall Street Journal (WSJ) using rough set theory. The stimulation demonstrated that the accuracy of the noun phrase improved by the rough set was higher than another translation method; therefore, it is an effective machine identification method for English noun phrases, providing a basis for practical design.*

Subject Categories and Descriptors: I.1.2[Algorithms] Translation algorithm; I.6.1[Simulation Theory] Simulation data set

General Terms: Rough set theory, Pattern Recognition

Keywords: Noun Phrase, Machine Translation, Rough Set, WSJ

Received: 18 January 2023, Revised 2 May 2023, Accepted 21 May 2023

Review Metrics: Review Scale- 0/6, Review Score- 4.60, Inter-reviewer consistency- 88.2%

DOI: 10.6025/jdim/2023/21/3/83-87

1. Introduction

Reform, opening, and foreign cooperation constantly deepen, and international tourism, cultural exchange, and business contacts become increasingly frequent. However, language differences brought great inconvenience. Machine translation is an automatic translation that converts one natural language to another natural one using a computer under the condition of meaning equivalence. Hence, machine translation provides a new approach to solving the problem [1-3]. Noun phrase identification, a key technology in machine translation, is the basis of syntactic analysis, and its identification effect directly affects the accuracy.

Rough sets are a kind of machine studying method developed in recent years. It functions as an inducing decision. It can not only reduce properties and data of the knowledge system and acquire decision rules from the decision table but also classify the derived decision rule [4-5]. It is widely applied in various identification and classification fields since it is well-suited for English noun phrase identification. Given this, this paper proposed a kind of English noun phrase identification method based on a rough set to solve the low identification accuracy of the current English noun phrase identification method. This method is to learn the decision rule of English noun phrases by rough set and obtain identification results. The simulation experiment results demonstrated that this method improved the identification accuracy of English noun phrases.

2. Principle of English Noun Phrase Identification

English noun phrase is the essential constituent unit of English sentences and is the unit of information transmission during speech communication. English noun phrase identification is one of the leading research con-

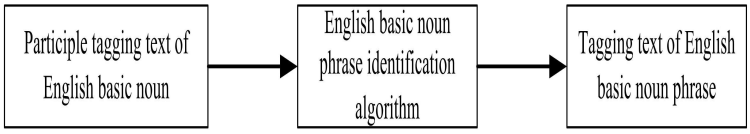

Figure 1. Principle of English noun phrase identification

tent in machine translation. Its purpose is to identify non-recursive noun phrases without a post-modifier. Its input is the English text labelled by a participle, and output is the English text that has been identified and whose phrases have been labelled. The principle of English noun phrase identification is shown in Figure 1.

Suppose B is expressed as an English noun phrase, I is an internal part of a noun phrase, and O is another situation. By doing that, identification of English noun phrases can be expressed in the form of $\{B,I,O\}$, we have:

$$y = f\{B, I, O\} \qquad (1)$$

From formula (1), it is known that identifying English noun phrases is a problem of decision rule. The current identification methods for English noun phrases cannot obtain satisfactory results since these methods have no function of decision rule analysis. Rough set is a kind of new-type method with the function of decision rule learning and is well suited for noun phrase identification. Therefore, this paper tried to identify English noun phrases by rough set.

3. Design of English Noun Phrase Identification Algorithm

3.1. Reduction of noun phrase property

Definition 1: An English noun phrase decision system S is defined as a tetrad.

$$f(x): U \times R \longrightarrow V \qquad (2)$$
$$R = p \cup \{d\} \qquad (3)$$
$$S = \{U, R, V, f\} \qquad (4)$$

Where $f(x)$ is expressed as information function acquired from the training corpus, U is expressed as the set composed of every punctuation and vocabulary from the English noun corpus to be identified, R is expressed as attribute set, p is expressed as noun phrase tagging and part of speech tagging of context within certain range of the current vocabulary; d is expressed as noun phrase tagging of the current vocabulary, V is expressed as union set of noun tagging set and part of speech tagging of context.

Definition 2: for an English noun phrase decision-making system, attribute subset $B \subseteq R$ is an indiscernible relation.

$$ND(B) = \{(x,y) \mid (x,y) \in U^2, \forall b \in B(b(x) = b(y))\} \qquad (5)$$

Definition 3: $a_i(x_j)$ is expressed as the value of x_j of the sample to be identified on the attribute a_i. Then, the discernibility matrix of the English noun phrase decision system S can be expressed as:

$$C_D : U \times U \to \rho(R) \qquad (6)$$

Definition 4: suppose U is expressed a domain of discourse, P *and* Q are expressed as two equivalence relation clusters on U.

If $POS_p(Q) = POS_{(p\{t\})}(Q)$, then r is the attribute that Q can omit in P; otherwise, r is the attribute that Q cannot obligate in P, that is, it can not be omitted. If every r in P is the attribute that Q can not omit in P, then P is the independent attribute of Q. For the independent subset S of Q in P, if $POS_S(Q) = POS_p(Q)$, then S is the attribute reduction of Q in P. All attribute reduction of Q in P is expressed as $RED_Q(P)$. Therefore, all Q in P that cannot omit the original relationship cluster is termed as Q core of P, and denoted by $CORE_Q(P)$.

Detailed procedures for the rough set to carry out attribute reduction on English noun phrases are as follows: calculate discernibility matrix C_D of decision table of English noun phrase decision-making system; solve $CORE_Q(P)$ and core logical expression is: $L_{CORE} = a_i \varepsilon CORE_Q(p) a_i$ (7); calculate discernibility matrix C'_D that omits core:

$$C'_D = \begin{cases} C_D(i,j) & if CD(i,f) \cap COREQ(P) = \phi \\ 0 & otherwise \end{cases} \qquad (8)$$

set up disjunction logical expression L_{ij} for all nonempty set element in discernibility matrix C'_D, that is: $L_{ij} = \bigvee_{a_i \in C_{ij}} (C_{ij} \neq 0, C_{ij} \neq \phi)$ (9); carry out conjunction calculation on all disjunction logical expression L_{ij} and L_{CORE} that is set up, and then a conjunction normal form is acquired, that is $L = (\bigwedge_{Ci_j \neq 0, Ci_j \neq 0} L_{ij}) \wedge L_{CORE}$ (10); convert the acquired conjunction normal form into disjunctive normal form L', that is: $L' = \bigvee_i L_i$ (11); output attribute reduction result of the English noun phrase.

3.2. Set up decision-making Rule of Noun Phrase

Rough set generates decision-making rule of English noun phrase as follows: obtain T through selecting an attribute reduction result from attribute reduction table according to the attribute reduction algorithm of the rough set; obtain rule set T' through reducing attribute value, the details are: calculate value core of the rules and merge the rules with the same value core; test all rules, and if the

decision-making rule composed by value core of the rules are identical, then the rule remains unchanged; otherwise, a non-value-core attribute is generated based on value core of the rule. Generate a minimum rule set of English noun phrases and obtain a minimum rule set that covers the whole information system; output the optimal rule set of English noun phrases.

3.3. Tagging of English Noun Phrases

The description method of English noun phrases is to use square brackets of "O+C", that is, to insert "[" to express the staring of an English noun phrase and "]" to express the end of a English noun phrase. Tagging of English noun phrases goes on based on the decision-making rules that have been generated and the context characteristics of the current words.

Therefore, English noun phrases, in nature, select the best matching rule from the decision-making rule set established above and tagging based on decision-making attribute value.

The detailed procedures for tagging a noun are as follows: first is to confirm the condition attribute value of the noun phrase, obtain part-of-speech tagging information of context from the English noun phrase training corpus and select noun phrase tagging information from the results that have been tagged; screen the candidate rule set, match the condition attribute of all rules in decision-making rule set with the noun phrase to be identified; select out rules with the least amount of inconsistent attributes to make up candidate rule set; second is to confirm the optimal matching rule: classify all the rules in candidate rule set based on decision making attribute value, select the first rule among the category sets with the most rules and take it as the optimal matching rule.

3.4. Process of English Noun Phrase Identification

A tagged noun is acquired by identifying English noun phrase samples by the tagging rule generated by the rough set algorithm. This paper adopts IOB tagging symbol sequence to obtain a single word; however, an English noun phrase may contain one or more words. Therefore, the IOB tagging symbol sequence should be converted into an English noun phrase sequence. A state transition recognizer realizes that conversion phrase, and then the identification is carried out.

The flow of English noun phrase identification algorithm is shown in Figure 2.

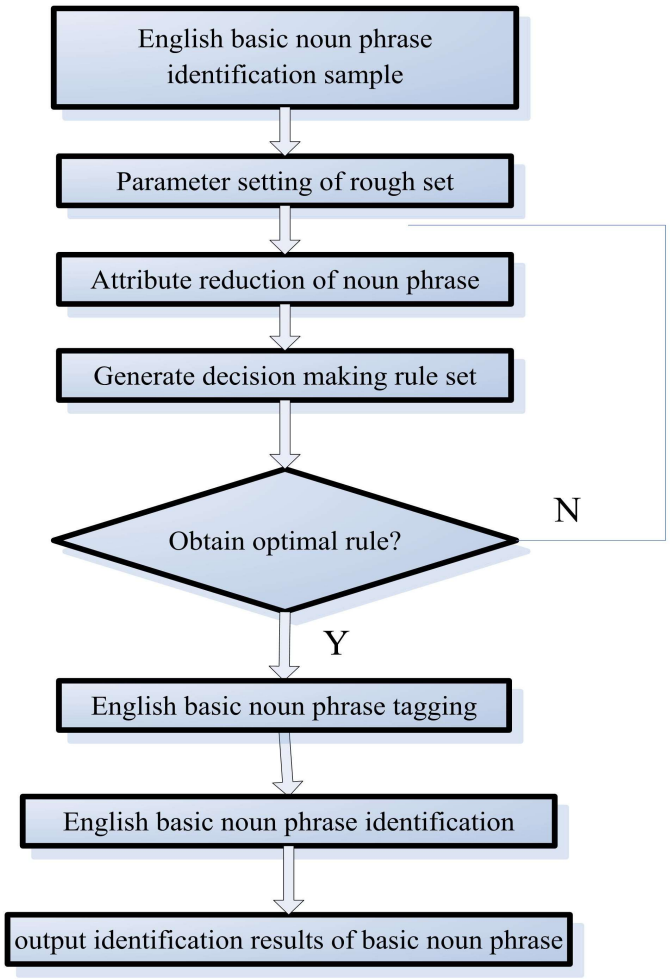

Figure 2. Flow of English noun phrase identification

3.5. Evaluation standard of English noun phrase identification results

In order to better evaluate various identification methods for English noun phrase, this paper adopted precision rate and recall rate as the evaluation standards of English noun phrase identification algorithm. The detailed definition is as follows:

Precision rate = (right noun phrases that have been identified/ noun phrase that have been identified)*100% (12)

Recall rate= (right noun phrases that have been identified/ all noun phrases in the text)*100%)

4. Stimulation Experiment

4.1. Stimulation Data Set

Data of stimulation experiment in this paper come from wall street journal (WSJ) corpus. WSJ corpus is the most authoritative English corpus for English phrase identification. It not only contains a large amount of text with tagged part of speech, but also contains many tagged phrases and syntactic structures [6-9]. Training corpus samples come from sections 15-18 of the WSJ corpus. The test samples come from sections 20-22 of the WSJ corpus. The hardware environment of the stimulation experiment is a dual-core CPU 2.5MHz, internal storage is 2Gbyte, and the operating system is Windows XP. The procedure is compiled under the VB environment.

4.2. Identification Results and Analysis

First, obtain the optimal English noun phrase identification model through training English noun phrase training set by rough set algorithm. Then, the model is adopted to identify the English noun phrase test sample. The identification results are shown in Table 1. Table 1 shows that the identification results obtained by the English noun phrase identification algorithm are satisfactory, and precision and recall rate are more than 90%. It illustrated that the method proposed by this paper can identify English noun phrases well.

4.3. Comparison with the Performance of Other Identification Methods

Some typical identification methods are selected to compare and verify the advantages and disadvantages of rough set algorithm and other English noun phrase identification algorithm. Those algorithms include Endong algorithm, Brill algorithm, TKS algorithm, TBL algorithm, Bayesian network algorithm and HMM algorithm. Precision

Test corpus	Precision rate /%	Recall rate /%
Section 20	93.84	93.53
Section 21	92.81	93.88
Section 22	92.77	92.97

Table 1. Identification results of rough set algorithm

Test corpus	Brill algorithm	Endong algorithm	TKS algorithm	Bayesian network algorithm	HMM algorithm	Rough set algorithm
Section 20/%	73.91	74.93	82.25	85.76	92.67	93.84
Section 21/%	73.22	74.14	81.38	84.85	91.69	92.81
Section 22/%	73.13	74.15	81.30	84.77	91.60	92.77

Table 2. Comparison of precision rate of identification of various algorithms

Test corpus	Brill algorithm	Endong algorithm	TKS algorithm	Bayesian network algorithm	HMM algorithm	Rough set algorithm
Section 20/%	73.57	74.61	82.98	85.48	92.37	93.53
Section 21/%	73.93	74.94	81.26	84.77	93.68	93.88
Section 22 /%	73.35	74.24	81.74	84.92	91.79	92.97

Table 3. Comparison of recall rate of various algorithms

rate of identification of all algorithms is shown in table 2 while recall rate of all algorithms is shown in table 3. From table 2 and 3, it is known that the experimental results of the rough set algorithm proposed by this paper are the best. That indicates rough set is a kind of effective and accurate noun phrase identification method, and is very suitable for solving some problems in natural language processing field.

5. Conclusion

English noun phrase identification problem existing in machine translation research is the basis and key point of machine translation. Based on the discussion on the problems of the current English noun phrase identification algorithm, this paper designed a English noun phrase identification method based on rough set theory according to the characteristics of English noun phrase. This method first finds decision making rule from training samples, then optimizes learning rule in the perspective of the whole system, and finally identifies English noun phrases. Stimulation and comparison experiments prove that the English noun phrase identification method proposed by this paper is an effective method with high efficiency and recall rate, therefore, it is very suitable for solving various natural language processing problems such as noun phrase identification, part of speech tagging and shallow analysis.

References

[1] Li, Yingjun. (2014). Status and Prospect of Machine Translation and Translation Technology Research. *Chinese Science & Technology Translators Journal*, 24-27.

[2] Chen, Yun., Zhang, Penghua., Ren, Lihua. (2013). A Review on Machine Translation. *Value Engineering*, 174-176.

[3] Feng, Ke. (2014). *Computer Aided Translation. Brightness*, 27, 289.

[4] Li, Tianrui., Chen, Hongmei., Yang, Y. (2013). Rough Set Theory and Application. *International Academic Development*, 13-15.

[5] Li, Xiang., Cheng, Yusheng., Ding, Meiwen. (2014). The Method of Bayesian Network Based on the Rough Set. *Theory Journal of Anqing Teachers College (Natural Science Edition)*, 36-40.

[6] Wang, Shuai. (2014). Review of Corpus Translation Development in China. *Chinese Editorials*, 29-32.

[7] Yin, Le . (2014). Acquisition of Translation Knowledge in Machine Translation based on Corpus. Beijing Jiaotong University.

[8] Xue, Huiwen. (2014). *An Analysis on the Significance of Corpus to Translating Research*. *Reading Digest*, 45.

[9] Lu, Zhenghai. (2014). Webpage Machine Translation Based on Corpus. *Anhui Literature* (Second Half Month), 12, 42-43.