doi:10.3772/j.issn.1006-6748.2021.03.007

Towards consistent machine translation of abbreviated terms in scientific literature[①]

He Yanqing (何彦青), Sun Yueying, Wu Zhenfeng, Pan You, Zhang Junsheng[②]
(Research Center for Information Science Theory and Methodology, Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract

Scientific literature often contains abbreviated terms in English for brief. Machine translation (MT) systems can help to share knowledge in different languages among researchers. Current MT systems may translate the same abbreviated term in different sentences into different target terms. MT systems translate the abbreviated term in two ways: one is to use translation of the full name, the other is to use the abbreviated term directly. Abbreviated terms may be ambiguous and polysemous, and MT systems do not have an explicit strategy to decide which way to use without context information. To get the consistent translation for abbreviated terms in scientific literature, this paper proposes a translation model for abbreviated terms that integrates context information to get consistent translation of abbreviated terms. The context information includes the positions of abbreviated term and domain attributes of scientific literature. The first abbreviated term is translated in full name while the latter ones of the same abbreviated term will show the abbreviated form in the translation text. Experiments of translation from Chinese to English show the effectiveness of the proposed translation model.

Key words: abbreviated term, context information, domain information, machine translation (MT)

0 Introduction

Scientific literature is the carrier for recording production practice and scientific experiment. Foreign scientific literature is an important resource to analyze the technical trends and latest developments in foreign countries. Scientific literature exists in different forms, such as scientific papers, patents, scientific reports, etc. It is inefficient to rely solely on manual translation, however, high-quality machine translation (MT) system can help rapidly acquire foreign scientific information.

Scientific literature usually contains many abbreviated terms that are composed of multiple words and often shown as abbreviations. It is concise and convenient for abbreviated terms to economically express ideas, diffuse knowledge and transmit information. When a scientific document is input into a MT system, it is segmented into sentences, which are then translated one by one and combined to get the final translation. Thus, the translation system has two methods for the abbreviated terms in each sentence in the text: the first method is direct translation, which means that the abbreviated terms are directly shown in the translation; the second method is the restitution translation, which chooses the full name of the abbreviated term as its translation. However, abbreviated terms are generally ambiguous and polysemous. For instance, the abbreviated term ‘AST’ has multiple full names, such as ‘amorphous-silicon TFT’, ‘Average Seek Time’, ‘Asynchronous Shared Terminal’, ‘Atomized Suspension Technique’ and ‘Abstract Syntax Tree’. Translation system does not have an explicit strategy to decide which translation method to use, direct translation or reduction translation. It also has no enough knowledge to choose which full name is the correct restitution translation. Without the context information and domain attributes of the abbreviated term, MT system may not produce translation correctly or even lead to domain inconsistency.

This paper proposes a translation model for abbreviated terms that integrate context information to get consistent translation of abbreviated terms. Context in-

[①] Supported by the National Key Research and Development Program of China (No. 2019YFA0707201) and ISTIC Research Foundation Project (No. ZD2020-10).

[②] To whom correspondence should be addressed. E-mail: zhangjs@ istic. ac. cn

Received on Aug. 31, 2020

formation includes the positions of abbreviated term and domain attributes of scientific literature. The first abbreviated term in the text is translated in full name, while the latter ones of the same abbreviated term will show the abbreviated term in the translation text. Domain attributes use Chinese Library Classification (CLC) number to ensure that its restitution translation is consistent with the domain attributes of the sentence or the chapter, thereby improving the translation of the chapter. The model consists of four parts. Firstly, a domain polysemous dictionary of abbreviated terms is constructed in which an abbreviated term has its full names and CLC numbers. Secondly, the dictionary is used to find each abbreviated term in the chapter to be translated, and replace them with specific serial number labels to get labeled chapter. At the same time, context information is generated to record the position information of each abbreviation label in the chapter. A pre-trained MT system translates the labeled chapter sentence by sentence to get the initial translation. All the full names of abbreviated terms labeled in the chapter are obtained by searching the dictionary, and those full names whose CLC number is consistent with the literature are used as candidate full name translations. Each label is replaced with the candidate full name translations respectively to obtain candidate chapter translations. Finally, long-short term memory (LSTM) language model is used to help select the best translation of the abbreviated term.

The rest of this paper is organized as follows: Section 1 presents the related work; Section 2 gives the construction method of domain polysemous dictionary of abbreviated terms; Section 3 details each steps of the model; Section 4 is experimental verification, and finally conclusions are obtained in Section 5.

1 Related work

In 'The New Oxford Dictionary of English', an acronym is defined as a word formed from the initial letters of the several words in the name, and it is a simplified phenomenon in language[1]. Longer words or more complex phrases are often abbreviated into several capital letters to keep the original meaning and easier expression.

Acronyms have the characteristics of conciseness, ambiguity, irregularity and instability[2]. They are concise for communication, writing and memory due to the reduction of syllables. The remaining letters in acronyms are not closely related in its sound, form and meaning. The creating habits vary from person to person, which leads to its ambiguity and informality. Instability means that acronyms are often affected by its use frequency and change rapidly over time.

According to the application scenario, acronyms can be divided into two categories: daily acronyms and technical acronyms. Daily acronyms are used for daily conversation, reading and writing, such as 'Ms' (Miss, Ms.), 'Mon' (Monday), 'No' (number), etc. Such acronyms have functions of ellipsis, euphemism, humor and emotion[3]. Technical acronyms are used to describe technology-related information, such as 'ACU' (antenna control unit), 'ALS' (amyotrophic lateral sclerosis), 'RTILs' (room temperature ionic liquids), etc. Technical acronyms cannot only convey scientific information, but is also simple to write, and save writing space. Therefore, researchers who tend to use technical acronyms are increasing dramatically, as well as some journalists and magazine's editors declare the principles of using acronyms in papers[4]. This paper pays the best attention to technical acronyms, which specifically refer to abbreviated terms in the scientific literature whose full names are technical terms.

According to the word building method, abbreviated terms can be divided into 10 types: (1) by the first letter of some words, such as 'ABS' (antilock braking system), 'HSF' (heat shock transcription factor), etc. (2) By the beginning letters of the words, such as 'MUSIC' (multiple signal classification), 'ID' (identification), etc. (3) By the letters inside the words, such as 'GNAT' (Gcn5-related N-acetyltransferase), 'DNA' (deoxyribonucleic acid), etc. (4) By the letters at the beginning and end of the words, such as 'Dr.' (doctor), 'Rd' (road), etc. (5) Generated after abbreviating some words, such as 'NF-kappaB' (nuclear factor-kappa B), e-mail (electronic mail), etc. (6) Generated after cutting the head or tail, sometimes with some morphological changes, such as 'plane' (aeroplane), 'chute' (parachute), 'Feb.' (February) 'Dorm' (dormitory), 'flu' (influenza), 'frig' (refrigerator), 'biz' (business), etc. (7) Generated from keywords in truncated phrases, such as 'Med' (Mediterranean Sea), 'Pop' (popular music) , etc. (8) concatenation of partial letters from the words, such as 'brunch' (breakfast + lunch), 'telecast' (television + broadcast), etc. (9) Digital abbreviations, such as 'C4ISR' (command, control, communication, computer, intelligence, survival, reconnaissance), etc. (10) else. There is no specific rules for the type, such as 'ADRB2' (beta 2 adrenergic receptor).

Abbreviated terms are also terms[5]. There are two cases when MT system translates terms. One is to translate only terms. At this time, it is not necessary to consider the context information of the source language in which the terms are located. The other one is to translate the terms in the context of the sentence or chapter. In order to improve sentence translation, terms translation needs to consider their sentence context. This paper focuses on the second case.

Nowadays, the most popular MT methods are statistical MT (SMT)[6-8] and neural MT (NMT)[9-11]. Both need to be trained on parallel corpora and then batch translation can be performed. When SMT prevailed, term translation was a very difficult problem. Document-topic distribution information of bilingual parallel corpora was learned as the domain information and a consistent calculation method was embedded as a feature in SMT[12]. The topic was tacit and had no explicit label, so the domain consistency of term translation cannot be well guaranteed. Wikipedia is a vocabulary resource to automatically identify bilingual terms and was integrated into SMT[13]. But Wikipedia method did not utilize the domain information of the terms. Ref. [14] proposed a combining method of a term database and a word inflection sdatabase. Since the term database is provided by the user, the method has weak generalization ability. Domain-specific term database was also put into the training corpus to train translation model[15]. Whenever the term is updated the model needs to be restrained. So the cost is relatively high. Compared with SMT, NMT improves the fluency of translation results. However, its vocabulary size is limited and it does not improve the terms translation to the same extent. Ref. [16] proposed a method using SMT phrase table as a supplement, replacing terms, translating and restoring them. The detailed steps of this method are: (1) Construct a bilingual term database to identify the terms in the source language of the parallel corpus and the test set, and replace the terms with serial number tags; (2) Train the translation model I with tagged parallel corpus; (3) Use translation model I to translate the tagged test set to get translation 1; (4) Query the bilingual term database to find the translation 2 of the tagged terms in the test set 2; (5) Combine translation 1 and translation 2 to get the final translation. This method expands vocabulary of NMT, but does not use domain information. Manual terminology constraints[17,18] were provided to NMT with term translations that were more in line with users' needs. This method requires manual interaction, and it is more difficult for non-specialists to obtain accurate and consistent terms.

When MT system translates chapter-level scientific literature, they usually segment each chapter into sentences firstly, then translate sentences and abbreviated terms inside, finally combine the sentence translations to obtain chapter translations. Therefore, the translation system has two main translation methods for abbreviated terms, direct translation and reduction translation. Abbreviated terms are often ambiguous and have multiple full names. Regardless of whether it is translated directly or reductively, the system does not consider the chapter context information and domain attributes, which leads to inconsistent problem. When terms or abbreviations are translated in the context of a sentence, it is necessary to consider context information in order to facilitate the translation of terms and abbreviations, thereby improving the translation of sentences and chapters.

This study is similar to Ref. [16]'s method in replacing terms, translating and restoring. The differences are: 1) The proposed method focuses on abbreviated terms. This paper introduces context information and domain attributes of abbreviated terms to help MT system to choose more correct and consistent translation. 2) The LSTM language model is employed to make the translation more fluent.

2 Construction of domain polysemous abbreviated terms dictionary

A general dictionary of abbreviated terms usually contains abbreviated terms and their full names, as shown in Table 1.

Table 1 An example of a general dictionary of abbreviated term

Abbreviated term	Full name
MPPT	maximum power point tracking
MRF	Markov random field
MSC	mesenchymal stem cells
NAR	nitrite accumulation rate

Researchers developed automated methods to collect such abbreviated terms dictionaries. Word formation of abbreviated terms were classified and rules were designed to identify the abbreviated terms[19]. The abbreviated terms combined with some auxiliary words such as 'abbreviation' and 'stands for' formed a list of query words to retrieve web pages to extract candidate phrases, match their full names and finally obtain the full name with an accuracy rate of 90%. Ref. [20] proposed a scanning method of reverse order scanning abbreviated terms and their full name. For standard abbreviation forms, they comprehensively consider the length, frequency, and inverse word frequency of the candidate full name. Extraction formulas are designed and the overall accuracy is as high as 96%, and the recall rate reaches 97%.

This paper mainly focuses on Chinese-English translation and how to correct English translation of English abbreviated terms in Chinese chapter. Therefore, based on Ref. [20]'s extraction method, a domain dictionary of abbreviated terms was built from network data and ascientific paper knowledge base. The construction of the dictionary includes three parts: (1) Recognition of abbreviated terms; (2) Collection of their full names: reverse order fast scanning collection method, the co-occurrence analysis collection method and the network resource method are used here; (3) Collection of their corresponding CLC numbers.

2.1 Recognition of abbreviated terms

Chinese scientific papers usually include Chinese titles, English titles, Chinese abstracts, English abstracts, Chinese keywords, English keywords, and CLC number. In scientific papers, abbreviated terms and their full names exist mainly in the form of 'full names (abbreviated terms)' or 'abbreviated terms (full names)'. The contents in the parentheses are generally abbreviated terms, but some scholars will add comments in the parentheses such as ',', '.', ';', such as 'Adaptive Pulse Code Modulation (APCM, This is adaptive pulse)'; At the same time, some abbreviated terms also contain "," such as "3,4-DCBN" (3,4-dichlorobenzonitrile) and "2,3-BD" (2,3-butanediol). Therefore, the recognition process of abbreviated terms is to first segment text into sentences, then recognize the strings in parentheses. There are three types of strings in the parentheses: (1) if there is no ',', '.' or ';' in the string, the string is directly extracted into the abbreviated terms set. (2) if there is '.' or ';' in the string, the string is split into two parts by '.' or ';' and the first part is extracted into the abbreviated terms set. (3) If there is ',', first determine whether the two sides of ',' are numbers. If so, go to (1), otherwise to (2).

2.2 Collection of full names

For each abbreviated term in the abbreviated terms set, three methods are used to obtain their full names: reverse order scanning method, co-occurrence analysis method and network resource method.

Each sentences including abbreviated terms in scientific papers are reversely scanned from right to left, and its full name must satisfy three conditions. Firstly, the full name must be longer than the abbreviated term; secondly, the first letter of the full name must match first letter of the abbreviated term; and finally, the matching letters of the full name must be in the same order with abbreviated term as Fig. 1. In Fig. 1 after the abbreviated term 'RNN' and '(' are recognized, the text before '(' is scanned from right to left, first matching the letter 'N' of position s1 (the last letter in the word 'RNN') with letter 'N' of position 1 (the first letter in the word 'network'), then the letter 'N' of position s2 (the last but one letter in the word 'RNN') with letter 'N' of position 2 (the first letter in the word 'Neural'). For the letter 'R' of position s3 (the first letter in the word 'RNN'), the letter 'r' at position 3 or 4 is not the first letter of the word 'Recurrent', thus, scanning is continued until position 5, which matches the first letter 'R' of 'RNN'. All the matched words form the full name of the abbreviated term 'RNN' and are included in the abbreviated term dictionary.

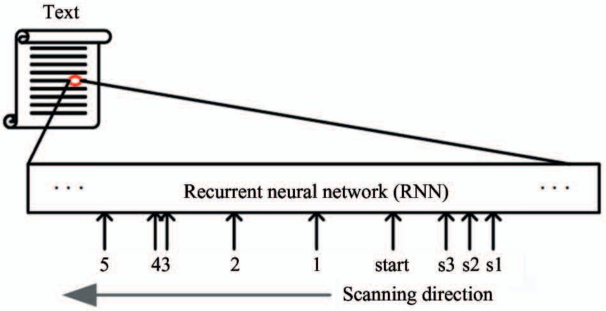


Fig. 1 Reverse order scanning matching

Co-occurrence analysis method is based on the co-occurrence frequency of abbreviated term and their full names in the text. The higher the frequency is, the bigger their matching degree becomes. Similar to the TF × IDF method, the frequency, length, and inverse word frequency of the candidate full name are used as features in the score to measure the importance of candidate full name.

$$Score(s) = [freq(s) - \sum_{t \in Ts} len(t) \times \tau] \times ISF \qquad (1)$$

$$\tau = freq(t)/freq(Ts) \qquad (2)$$

$$ISF = \log[freq(Ts)/freq(s) + 0.01] \qquad (3)$$

where, s is a candidate full name of abbreviated term t, $len(t)$ represents the length of t, $freq(s)$ represents the co-occurrence frequency of candidate full name s and abbreviated term t, Ts is a nested set of candidate full names which is related to s and include s and all its predecessors, $freq(Ts)$ represents the total occurrence frequency of all the full names in Ts, τ represents the probability of the occurrence of the nested term t, ISF is the inverse word frequency, 0.01 is the adjustment coefficient. Those with the highest score in Eq. (1) is chosen as full name of the abbreviated term of *t*.

In the network resource method, this work looks up the abbreviation module on the website (https://abbr. dict. cn/), and extracts full names for each abbreviated term.

All the abbreviated terms and their full names obtained from three methods above are combined and removed duplication.

2.3 Collection of domain attributes

After retrieving all the full names to each abbreviated term in the scientific paper database automatically, CLC number of the paper containing the full name is collected as the domain attributes to construct the domain polysemous abbreviated terms dictionary. It can be seen from Table 2 that each abbreviated term has multiple full names and each full name has several CLC numbers to represent its domain attributes.

Table 2 The domain polysemous abbreviated terms dictionary

Abbreviated term	Full name	Domain information
ABS	American Broadcasting System	D,G,O,P,Q,R,S,TB,TD,TE,TF,TG,TH,TJ,TK,TL,TM,TN,TP,TQ,TS,TU,U,V,X
ABS	Antilock Braking System	D,E,F,G,H,N,O,R,S,TD,TE,TF,TH,TJ,TM,TN,TP,TQ,TS,U,V,X
ABS	Acrylonitrile Butadiene Styrene	F,G,H,O,TB,TP,TQ,U,X
ABS	Alkyl Benzene Sulfonate	D,F,G,I,O,Q,R,S,TB,TD,TE,TG,TH,TK,TQ,TS,TU,U,X,Z
BMA	Bayesian Model Averaging	C,D,E,F,G,H,N,O,P,Q,R,S,TB,TD,TE,TG,TH,TJ,TK,TM,TN,TP,TQ,TU,TV,U,V,X,Z
BMA	British Medical Association	H,O,Q,TE,TN,TP,TQ
BMA	Block Matching Algorithm	E,G,O,P,R,S,TD,TE,TG,TH,TJ,TN,TP,U,V

3 The proposed model

The proposed abbreviated term translation model integrates chapter context information and domain attributes into a MT system. The flowchart is shown in Fig. 2. When a translation system translates chapter-level scientific literature, it usually segments the chapter into sentences, then translates each sentence one by one, and then combines the translation of each sentence to get the chapter translation. This process is shown in the left part of Fig. 2. This model adds the recognition of abbreviated term, records context information, labels abbreviated term before sentence translation, then uses domain attributes and LSTM language model to help choose the best full name as abbreviated term translation, finally combines the improved sentence translation results to get chapter translation.

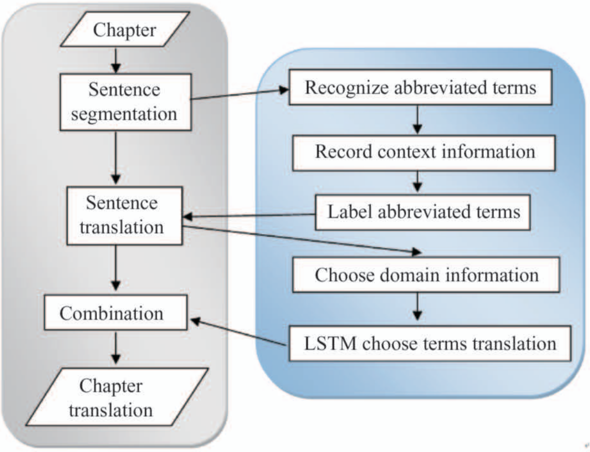
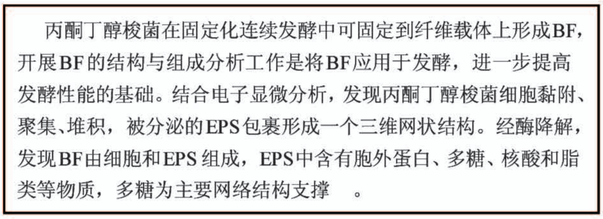

Chapter
Sentence segmentation
Sentence translation
Combination
Chapter translation
Recognize abbreviated terms
Record context information
Label abbreviated terms
Choose domain information
LSTM choose terms translation

Fig. 2 The flowchart of the proposed model

After a Chinese chapter to be translated (Fig. 3) is first segmented into sentences according to ‘。’, the sentences are numbered as shown in Table 3. Each sentence is recorded with label replacement number and the initial value of the label replacement number is set as 0.



丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成BF，开展BF的结构与组成分析工作是将BF应用于发酵，进一步提高发酵性能的基础。结合电子显微分析，发现丙酮丁醇梭菌细胞黏附、聚集、堆积，被分泌的EPS包裹形成一个三维网状结构。经酶降解，发现BF由细胞和EPS组成，EPS中含有胞外蛋白、多糖、核酸和脂类等物质，多糖为主要网络结构支撑 。

Fig. 3 The chapter to be translated

3.1 Recognize abbreviated terms

Here some rules are designed to recognize the abbreviated terms in each sentence and get the set of abbreviated terms for the chapter. Recognition rule for abbreviated terms uses Python regular expression: “pattern = re. compile(‘[AZ]+[az]*[–]*[·]*[0-9]*[AZ]*[0-9]*’)”. Each abbreviated term in the set is searched in the domain polysemous abbreviated terms dictionary to obtain its full names and their domain attributes, as shown in Table 4. The CLC numbers and their names are shown in Table 5.

Table 3 Sentence segmentation

Sent. id	Sentence	Label replacement number
1	丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成BF,开展BF的结构与组成分析工作是将BF应用于发酵,进一步提高发酵性能的基础。	0
2	结合电子显微分析,发现丙酮丁醇梭菌细胞黏附、聚集、堆积,被分泌的EPS包裹形成一个三维网状结构。	0
3	经酶降解,发现BF由细胞和EPS组成,EPS中含有胞外蛋白、多糖、核酸和脂类等物质,多糖为主要网络结构支撑。	0

Table 4 Full names and domain information

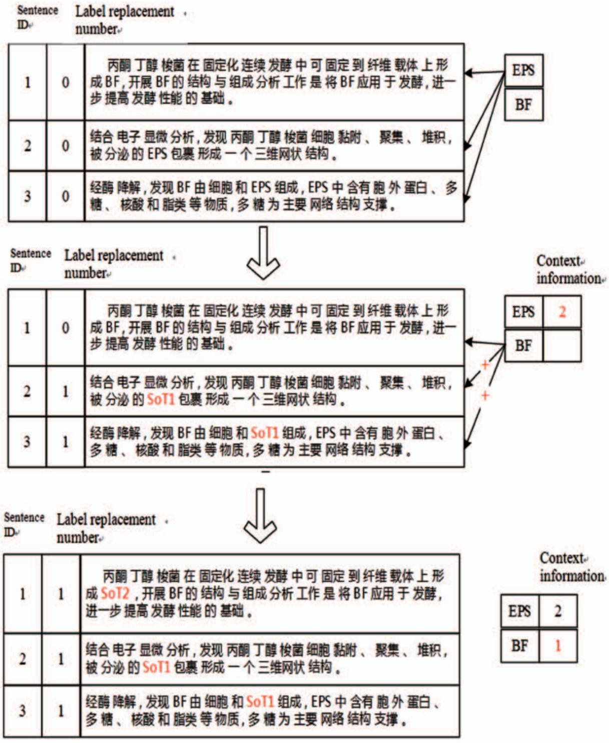
Abbreviated terms	Full names	Domain attributes
EPS	extracellular polymeric substances	Q;R;TB;TQ;TU;X
BF	Biofilm	F; G; O; P; Q; R; S; TB; TG; TM;TP;TQ;TS;TU;TV;V;X
	Bridging Fault	E; F; G; O; P; S; TB; TG; TH; TJ;TM;TN;TP;U;V;X
	Boundary Function	C;D;F;G;H;I;J;K;O;P;Q; R;S;TB;TD;TE;TG;TH;TJ; TK;TM;TN;TP;TQ;TS;TU; TV;U;V;X

Table 5 CLC numbers and their names

CLC No.	Name	CLC No.	Name	CLC No.	Name
C	Social science	Q	Biological sciences	TN	Radio electronics, telecommunication technology
D	Politics, law	R	Medicine and health	TP	Computer technology, automation technology
E	Military	S	Agricultural Science	TQ	Chemical industry
F	Economics	TB	General industrial technology	TS	Light industry, handicraft industry
G	Culture, science, education, sports	TD	Mining engineering	TU	Architecture
H	Language and Literature	TE	Oil and gas industry	TV	Water conservancy project
I	Language	TG	Metallics and metalworking	U	Transportation
J	Art	TH	Mechanical instrument industry	V	Aviation and aerospace
K	History, geography	TJ	Arms industry	X	Environmental Science and Safety Science
O	Mathematical science and chemistry	TK	Energy and power engineering		
P	Astronomy, earth science	TM	Electrical engineering technology		

3.2 Label abbreviated terms

After the abbreviated terms of each sentence are identified, they are replaced with serial number label SoTi ($i = 1, 2\cdots$). Here two principles need to be satisfied: (1) at most one label in one sentence; (2) label the abbreviated term that appear for the first time under the conditions of (1). An example is shown in Fig. 4. Abbreviated term ‘EPS’ is searched in each sentence and replaced with serial number label SoT1 in sentence 2 and sentence 3. The label replacement number for sentence 2 and sentence 3 where the replacement occurs is increased by 1. At the same time, the context information of ‘EPS’ is recorded, that is sentence 2, the sentence id where ‘EPS’ first appears in the chapter. For the second abbreviated term ‘BF’, it is replaced with serial number label SoT2 in sentence 1. Since label replacement number in sentence 2 and sentence 3 has reached the maximum value 1, only the first ‘BF’ in sentence 1 is labeled and its context information is sentence 1. Some abbreviated terms that appear fairly late in the chapter are not labeled after searching all the sentences because the previous abbreviated terms have made label replacement reach the maximum value 1 for each sentence. At this time, it must label the abbreviated term where it appears and also record its context information.



Sentence ID	Label replacement number	
1	0	丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成BF,开展BF的结构与组成分析工作是将BF应用于发酵,进一步提高发酵性能的基础。
2	0	结合电子显微分析,发现丙酮丁醇梭菌细胞黏附、聚集、堆积,被分泌的EPS包裹形成一个三维网状结构。
3	0	经酶降解,发现BF由细胞和EPS组成,EPS中含有胞外蛋白、多糖、核酸和脂类等物质,多糖为主要网络结构支撑。

Sentence ID	Label replacement number	
1	0	丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成BF,开展BF的结构与组成分析工作是将BF应用于发酵,进一步提高发酵性能的基础。
2	1	结合电子显微分析,发现丙酮丁醇梭菌细胞黏附、聚集、堆积,被分泌的SoT1包裹形成一个三维网状结构。
3	1	经酶降解,发现BF由细胞和SoT1组成,EPS中含有胞外蛋白、多糖、核酸和脂类等物质,多糖为主要网络结构支撑。

	Context information
EPS	2
BF	

Sentence ID	Label replacement number	
1	1	丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成SoT2,开展BF的结构与组成分析工作是将BF应用于发酵,进一步提高发酵性能的基础。
2	1	结合电子显微分析,发现丙酮丁醇梭菌细胞黏附、聚集、堆积,被分泌的SoT1包裹形成一个三维网状结构。
3	1	经酶降解,发现BF由细胞和SoT1组成,EPS中含有胞外蛋白、多糖、核酸和脂类等物质,多糖为主要网络结构支撑。

	Context information
EPS	2
BF	1

Fig. 4 Label abbreviated term and record context information

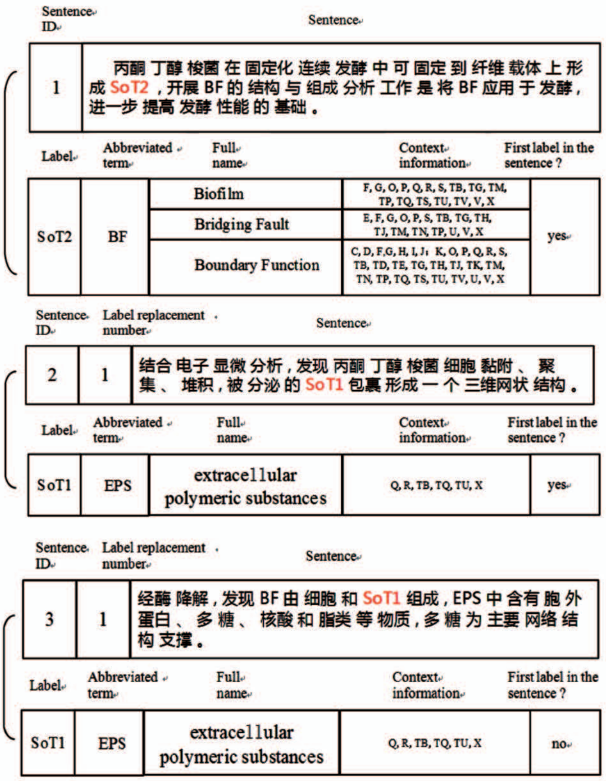
In the end, each sentence is labeled with a serial number label, and each label corresponds to a small dictionary of abbreviated terms including domain information and context information as Fig. 5 shows. The labeled sentences are translated with a pre-trained translation system to get the initial translations as Table 6 shows.

3.3 Choose abbreviated term translation

In order to provide clear meaning to the abbreviated terms in final translation, the position where each abbreviated term is first replaced with label in the chapter is given in the form of ‘full name (abbreviated term)’, and other position is only given in the abbreviated form.

3.4 Translate labeled sentences

Each labeled abbreviated term has obtained all full names and domain attributes . Those full names which are not consistent with the chapter are removed, and the rest are used as candidate full-name translations. Here a classification model can be used to label chapter or sentence[21,22] to assign CLC number as their do-



Sentence ID	Sentence
1	丙酮 丁醇 梭菌 在 固定化 连续 发酵 中 可 固定 到 纤维 载体 上 形成 SoT2 ,开展 BF 的 结构 与 组成 分析 工作 是 将 BF 应用 于 发酵,进一步 提高 发酵 性能 的 基础。

Label	Abbreviated term	Full name	Context information	First label in the sentence ?
SoT2	BF	Biofilm	F, G, O, P, Q, R, S, TB, TG, TM, TP, TQ, TS, TU, TV, V, X	yes
		Bridging Fault	E, F, G, O, P, S, TB, TG, TH, TJ, TM, TN, TP, U, V, X	
		Boundary Function	C, D, F,G, H, I, J; K, O, P, Q, R, S, TB, TD, TE, TG, TH, TJ, TK, TM, TN, TP, TQ, TS, TU, TV, U, V, X	

Sentence ID	Label replacement number	Sentence
2	1	结合 电子 显微 分析,发现 丙酮 丁醇 梭菌 细胞 黏附 、 聚集 、 堆积,被 分泌 的 SoT1 包裹 形成 一 个 三维网状 结构。

Label	Abbreviated term	Full name	Context information	First label in the sentence ?
SoT1	EPS	extracellular polymeric substances	Q, R, TB, TQ, TU, X	yes

Sentence ID	Label replacement number	Sentence
3	1	经酶 降解,发现 BF 由 细胞 和 SoT1 组成,EPS 中 含有 胞 外 蛋白 、 多 糖 、 核酸 和 脂类 等 物质,多 糖 为 主要 网络 结 构 支撑。

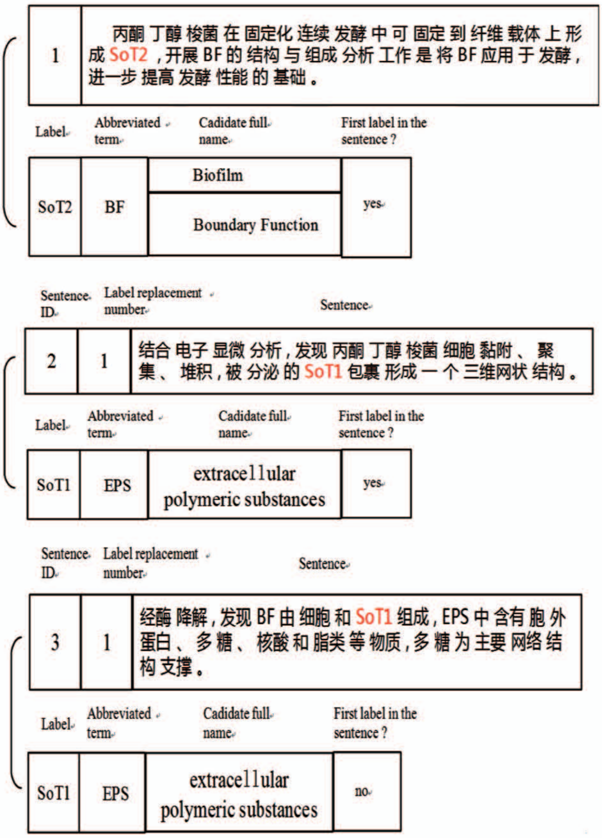
Label	Abbreviated term	Full name	Context information	First label in the sentence ?
SoT1	EPS	extracellular polymeric substances	Q, R, TB, TQ, TU, X	no

Fig. 5 Each sentence with a small abbreviated term dictionary

Table 6 Initial translations

Labeled sentence	Translation of labeled sentence
丙酮丁醇梭菌在固定化连续发酵中可固定到纤维载体上形成SoT2,开展 BF 的结构与组成分析工作是将 BF 应用于发酵,进一步提高发酵性能的基础。	SoT2 could be formed by immobilized Clostridium acetone and butanolin continuous fermentation. The analysis of structure and composition of BF was the basis of applying BF to fermentation and further improving fermentation performance.
结合电子显微分析,发现丙酮丁醇梭菌细胞黏附、聚集、堆积,被分泌的 SoT1 包裹形成一个三维网状结构。	Combined with electron microscopic analysis, it was found that Clostridium acetone-butanol cells adhered, aggregated and accumulated, and were wrapped by secreted SoT1 to form a three-dimensional network structure.
经酶降解,发现 BF 由细胞和 SoT1 组成,EPS 中含有胞外蛋白、多糖、核酸和脂类等物质,多糖为主要网络结构支撑。	It was found that BF was composed of cells and SoT1, EPS contained extracellular proteins, polysaccharides, nucleic acids and lipids, and polysaccharide was the main network structure support.

main information. The chapter domain in the example is TQ (chemical industry). The translation of the candidate full-name is shown in Fig. 6.



1	丙酮 丁醇 梭菌 在 固定化 连续 发酵 中 可 固定 到 纤维 载体 上 形成 SoT2 ，开展 BF 的 结构 与 组成 分析 工作 是 将 BF 应用 于 发酵，进一步 提高 发酵 性能 的 基础。

Label	Abbreviated term	Cadidate full name	First label in the sentence ?
SoT2	BF	Biofilm Boundary Function	yes

Sentence ID	Label replacement number	Sentence
2	1	结合 电子 显微 分析，发现 丙酮 丁醇 梭菌 细胞 黏附 、 聚集 、 堆积，被 分泌 的 SoT1 包裹 形成 一 个 三维网状 结构 。

Label	Abbreviated term	Cadidate full name	First label in the sentence ?
SoT1	EPS	extracellular polymeric substances	yes

Sentence ID	Label replacement number	Sentence
3	1	经酶 降解，发现 BF 由 细胞 和 SoT1 组成，EPS 中 含有 胞 外 蛋白 、 多 糖 、 核酸 和 脂类 等 物质，多 糖 为 主要 网络 结构 支撑。

Label	Abbreviated term	Cadidate full name	First label in the sentence ?
SoT1	EPS	extracellular polymeric substances	no

Fig. 6 The translation of the candidate full-name

If the abbreviated term's context information indicates the current sentence is the sentence of its first appearance and there is only one candidate full name translation, its serial number label will be replaced directly with the form of 'full name (abbreviated term)'. Other serial number labels for the abbreviated term are still replaced with the abbreviated form.

If there are multiple full names, the corresponding serial number labels are replaced with each candidate full name to obtain multiple sentence translations. LSTM language model[23-24] is used to select the best sentence translation. The calculation equation of LSTM language model score is as follows:

$$Score\ (canditate\ sentence\ translation) = lm(canditate\ full\ name) \quad (4)$$

$$lm(candidate\ full\ name) = \log(P(s_0 s_1 s_2 \cdots s_{t+1})) \quad (5)$$

$$P(s_0 s_1 s_2 \cdots s_{t+1}) = P(s_0) \times P(s_1 \mid s_0) \times \cdots \times P(s_{t+1} \mid s_0 s_1 \cdots s_t) \quad (6)$$

$$P(s_{t+1} \mid s_0 s_1 \cdots s_t) = predict(s_0 s_1 \cdots s_t,\ state_t, s_{t+1}) \quad (7)$$

The score of the candidate sentence translation $(s_0 s_1 s_2 \cdots s_{t+1})$ depends on language model score of candidate full name which refers to the fluency score of the sentence $(s_0 s_1 s_2 \cdots s_{t+1})$ after the candidate full name replaces term label, $P(s_0 s_1 s_2 \cdots s_{t+1})$ means the probability of sentence $(s_0 s_1 s_2 \cdots s_{t+1})$, $P(s_{t+1} \mid s_0 s_1 \cdots s_t)$ is the function of $s_0 s_1 s_2 \cdots s_t$, s_{t+1} and $state_t$, which is calculated by a recurrent neural network (RNN). When $t = 0$, there is no previous state, so set $P(s_0) = 1$. That means the fluency score of the sentence is calculated from the second word. The final translation is shown in Table 7.

Table 7 Final translations after LSTM choosing

Sent. ID	Initial translation (Domain: TQ)	Final translation
1	SoT2 could be formed by immobilized Clostridium acetone and butanol in continuous fermentation. The analysis of structure and composition of BF was the basis of applying BF to fermentation and further improving fermentation performance.	Biofilm(BF) could be formed by immobilized Clostridium acetone and butanol in continuous fermentation. The analysis of structure and composition of BF was the basis of applying BF to fermentation and further improving fermentation performance.
2	Combined with electron microscopic analysis, it was found that Clostridium acetone-butanol cells adhered, aggregated and accumulated, and were wrapped by secreted SoT1 to form a three-dimensional network structure.	Combined with electron microscopic analysis, it was found that Clostridium acetone-butanol cells adhered, aggregated and accumulated, and were wrapped by secreted extracellular polymeric substances(EPS) to form a three-dimensional network structure.
3	It was found that BF was composed of cells and SoT1, EPS contained extracellular proteins, polysaccharides, nucleic acids and lipids, and polysaccharide was the main network structure support.	It was found that BF was composed of cells and EPS, EPS contained extracellular proteins, polysaccharides, nucleic acids and lipids, and polysaccharide was the main network structure support.

4 Experiments

4.1 Experimental setup

The baseline MT system is a T2T Chinese-English translation model built on Transformer. The hyper parameters are set as follows: batch size is 4096; 2 hidden layers, and the number of neurons is 512; In order to avoid over fitting, the dropout rate is set to 0.2.

The statistics of the MT system are shown in Table 8. The training corpus involves multiple domains whose CLC numbers cross from A to Z; The domain of the test set are: E, F, O, P, Q, S, TB, TD, TE, TF, TG, TH, TJ , TK, TL, TM, TN, TP, TQ, TS, TU, TV, U, V, X.

Table 8 Data statistics

Data	Language	Chapter number	Sentence number	Shortest sentence length	Longest sentence length	Average sentence length	Vocabulary
Training data	En		69445873	1	1706	26	6098603
	Ch		69445873	1	2759	39	4421049
Development data	En		2200	1	117	27	7793
	Ch		2200	4	141	41	8123
Test data	En	145	599	102	588	199	288 24
	Ch	145	599	96	336	181	26230

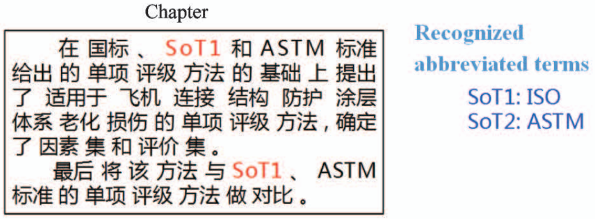
4.2 Model construction

According to the method described in Section 2, a domain polysemous abbreviated terms dictionary is constructed by using a scientific paper knowledge base and network resources. Here only the abbreviated terms dictionary used in the test set are filtered and the statistical information is shown in Table 9.

Table 9 Statistical information of dictionary filtered by test data

Number of abbreviated terms	Average length of abbreviated terms	Average number of full names	Average number of domain attributes
340	3.43	2.58	2.65

In order to extract abbreviated terms without destroying the semantic information of the sentence to a better extent, each sentence is labeled according to two replacement rules: (1) at most one label for each sentence; (2) guarantee (1) and label abbreviated term which first occurs in the sentence. Therefore, when the label replacement number of the sentence is not full, replace the abbreviated term in it with the serial number label and record the sentence id where the label is located, namely the chapter context information of the abbreviated term. According to these two rules, some later terms cannot be labeled. There are two sentences in the chapter, and two abbreviated terms are recognized as Fig. 7 expressed: ‘ISO’ and ‘ASTM’. When the abbreviated term ‘ISO’ is replaced firstly with serial number label ‘SoT1’ after traversing two sentences, the chapter already satisfies rule (1) which means there is at most one label for each sentence. In this case, the abbreviated term ‘ASTM’ cannot be labeled. In order to provide as many clear full names as possible, the position where abbreviated term ‘ASTM’ first appears in the chapter is forcibly labeled here with serial number label ‘SoT2’, and its chapter information is also recorded.



在 国标 、 SoT1 和 ASTM 标准 给出 的 单项 评级 方法 的 基础 上 提出 了 适用于 飞机 连接 结构 防护 涂层 体系 老化 损伤 的 单项 评级 方法，确定 了 因素 集 和 评价 集 。

最后 将 该 方法 与 SoT1 、 ASTM 标准 的 单项 评级 方法 做 对比 。

Fig. 7 A labeling example

In the experiment 365 abbreviated terms were eventually replaced by labels in the test set. Then the pre-trained baseline Chinese-English translation model is used for translating the labeled sentences to get the labeled initial translation. At this time, the position of each label, the corresponding abbreviated terms, different full names, their domain information, and the chapter information are known. Table 10 shows an example.

Domain information is used for selecting candidate full names for each abbreviated term. It can be seen from Table 10 that ‘ISO’ has two full names and the domain attributes of its second full name ‘International Student Organization’ does not contain ‘TQ’ where the chapter’s domain lies, so the full name does not meet domain consistency and is removed as a candidate full name translation. Similarly, the first full name 'International Organization for Standardization' is reserved as candidate full name. The full translation of 'ASTM' is also consistent with the domain. Thus, it also retains as a candidate full translation.

Table 10 An example of abbreviated term with domain information and context information

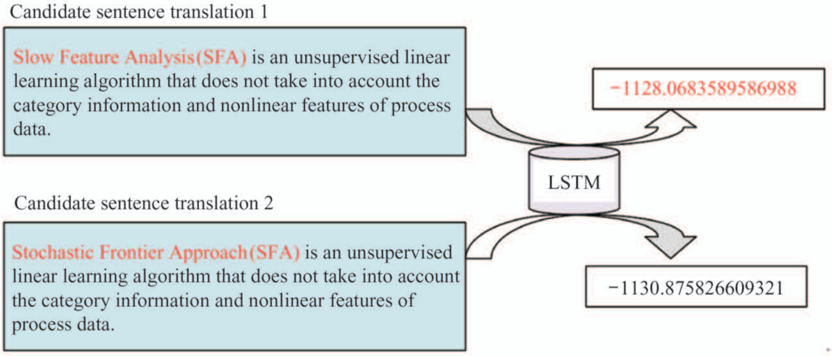
Sentence No.	Initial Sentence translation	Chapter domain	Abbreviated term	Full names	Domain information	Context information
1	On the basis of the single rating method given by the national standard, SoT1 and SoT2 standards, a single rating method for aging damage of protective coating system for aircraft connection structure is proposed, and the factor set and evaluation set are determined.	TQ	ISO	International Organization for Standardization	C, D, E, F, G, H, J, O, P, Q, R, S, TB, TE, TG, TH, TK, TL, TM, TN, TP, TQ, TS	Sent. 1
				International Student Organization	C, D, G, H, O	
2	Finally, the method is compared with SoT1 and ASTM standard single rating methods.		ASTM	American society for Testing and Material	D, F, G, H, J, O, P, Q, R, S, TB, TD, TE, TF, TG, TH, TK, TL, TM, TN, TP, TQ, TS, TU, TV, U, VX	Sent. 1

If there is only one candidate full name translation for the abbreviated term, the serial number label is replaced with full name based on their chapter context information which point out that the label is the first appearing in the chapter. Some abbreviated terms obtain multiple candidate full names after the domain information filtering, as shown in the example in Table 11. The context attributes of 'SFA' is sentence 1, and the current sentence ID is also 1, thus it is necessary to translate 'SFA' with the full name. All the candidate full names are used for replacing the serial number label to generate sentence candidate translations.

Table 11 An example of multiple candidate full names

Sentence No.	Initial sentence translation	Abbreviated term	Full name	Context attributes
1	SoT2 is an unsupervised linear learning algorithm that does not take into account the category information and nonlinear features of process data.	SFA	(1) Slow Feature Analysis (2) Stochastic Frontier Approach	Sentence 1

For all the sentence candidate translations, LSTM language model is used to score the fluency of the candidate translations, and the higher score becomes the final sentence translation, as shown in Fig. 8. In the example of Table 11 the two candidate full-names of 'SFA' are used for replacing the serial number label


Candidate sentence translation 1
-1128.0683589586988
LSTM
Candidate sentence translation 2
-1130.875826609321

Slow Feature Analysis(SFA) is an unsupervised linear learning algorithm that does not take into account the category information and nonlinear features of process data.

Stochastic Frontier Approach(SFA) is an unsupervised linear learning algorithm that does not take into account the category information and nonlinear features of process data.

Fig. 8 LSTM language model selection

'SoT1' to generate sentence candidate translations. Thus, the candidate translation 1 has the higher score of LSTM language model and output as the final sentence translation.

At last all the sentence translations are combined to obtain the final chapter translation.

4.3 Model evaluation

Selection accuracy here is used to evaluate the performance of domain information and LSTM language model, and BLEU[25] is used to measure translation effects. Domain selection accuracy and language model selection accuracy is given in Eqs(8) and (9).

$$\textit{Domain selection accuracy} = \frac{\textit{number of abbreviated term with correct full name by domain selection}}{\textit{number of abbreviated term}} \tag{8}$$

$$\textit{LSTM selection accuracy} = \frac{\textit{number of abbreviated term with correct full name by LSTM selection}}{\textit{number of abbreviated term}} \tag{9}$$

After analyzing the final translations, it is found that the constructed dictionary is still limited in candidate translations of full names, and some full names have not been collected. Therefore, the dictionary is supplemented manually later. The experimental results are shown in Table 12. From BLEU values in Table 12, it can be seen that the abbreviated term translation model can indeed provide the correct full name of the abbreviated term and the integrity of the dictionary can directly affect translations effects. The dictionary provides multiple full name translations for abbreviated term and adds some full name translations which the original MT system cannot give. Secondly, most irrelevant term translations are filtered out by the domain attributes, which reduces the problem of domain inconsistency. Finally the LSTM language model helps to select candidate translations whose domain information is similar. During the process, context information is served as the guide of substitution. In case of a complete dictionary, the abbreviated term translation method can obtain 93.8% in the accuracy of domain consistency, 68.9% in accuracy of language model selection, and 0.82 BLEU score in translation improvement.

Table 12 Accuracy and BLEU

Dictionary	BLEU baseline =0.2383	Domain selection accuracy	LSTM selection accuracy
Initial dictionary	0.2443 (+0.60%)	86.9%	51.1%
Supplemented dictionary	0.2465 (+0.72%)	93.8% (+6.9%)	68.9% (+17.8%)

4.4 Error analysis

After error analysis the following problems are found: (1) The missing label. After some abbreviated terms are replaced with labels and the labeled sentence is translated, some labels are lost directly, resulting in the missing translation of the abbreviated term (about 13.4%). If the abbreviated terms are not replaced with labels, the result may also be missed by NMT, but loses slightly (roughly 4.6%). (2) The accuracy of language model selection in the current method is not ideal. In the experiment, the language model is trained based on data in the chemical industry, but the full names of abbreviated terms may belong to other domains and have different language habits. Therefore, multiple language models in different domains may help to filter the full names that is consistent in the domain[26].

5 Conclusions

In the process of translating scientific documents, MT systems usually use direct translation or reduction translation for abbreviated terms, which leads to domain inconsistencies or front and behind differences. To address this problem, an abbreviated term translation model that integrates context information and domain attributes is proposed. The constructed extra knowledge, such as the abbreviated term dictionary, is used for supervising term translation and alleviate the problem of inconsistent translations within the chapter to a certain extent. Context information ensures that the abbreviated terms that appear in the chapter for the first time are translated in full name and in the form of abbreviations that appear later, which can solve the problem of unclear translation of abbreviated terms and further guarantee the translation is clean and economical. Finally, the effectiveness of the translation model on the test set is evaluated and the results show that the proposed method can greatly improve the problem of domain inconsistency, and can also provide a relatively correct full name of the abbreviated terms.

This paper mainly corrects English translation of English abbreviated terms in Chinese chapter. This method is also applicable to other translation directions, as long as the dictionary of abbreviated terms is changed.

References

[1] Hu H. English acronym[J]. *Journal of Harbin University*, 2011, 32(11): 79-82

[2] Xue H Q. Research on English acronym[J]. *Kaoshi Zhoukan*, 2015, (31):86-87

[3] Zhang W H. English abbreviatations and their pragmatic functions[J]. *Yangtze River Series*, 2017(20):88

[4] The editorial office. Principles for the use of abbreviations in medical scientific papers[J]. *Chinese Journal of Laparoscopic Surgery* (*Electronic Edition*), 2014 (5):431-431

[5] Teresa M, Castellvi C, Zhu B, et al. Terminology and translation[J]. *China Terminology*, 2020, 22(4): 31-35

[6] Och F J, Ney H. Discriminative training and maximum entropy models for statistical MT[C]//Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002: 295-302

[7] Brown P F, Pietra S D A, Pietra V D J, et al. The mathematics of statistical MT: parameter estimation[J]. *Computational Linguistics*, 1993:19(2):263-311

[8] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003:48-54

[9] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems, Montreal, Canada, 2014: 3104-3112

[10] Bahdanau D, Cho K, Bengio Y. Neural MT by jointly learning to align and translate[J]. *arXiv*:1409.0473v7, 2015

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//The 31st Annual Conference on Neural Information Processing Systems, Los Angeles, USA, 2017: 6000-6010

[12] Meng F D, Xiong D Y, Jiang W B, et al. Methods for evaluating domain consistency of terminology translation and statistical MT[P]. China patent, CN201410520322, 2015

[13] Arcan M, Giuliano C, Turchi M, et al. Identification of bilingual terms from monolingual documents for statistical MT[C]//Proceedings of the 4th International Workshop on Computational Terminology, Dublin, Ireland, 2014: 22-31

[14] Pinnis M. Dynamic terminology integration methods in statistical MT[C]//The 18th Conference of the European Association for MT, Antalya, Turkey, 2015:89-96

[15] Scansani R, Bernardini S, Ferraresi A, et al. Enhancing MT of academic course catalogues with terminological resources[C]//International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria, 2017:1-10

[16] Long Z, Utsuro T, Mitsuhashi T, et al. Translation of patent sentences with a large vocabulary of technical terms using neural MT[J]. *arXiv*:1704.04521, 2017

[17] Hasler E, Gispert A D, Iglesia G, et al. Neural MT decoding with terminology constraints[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Melbourne, Australia, 2018: 506-512

[18] Hokamp C, Liu Q. Lexically constrained decoding for sequence generation using grid beam search[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017: 1535-1546

[19] Zhu J T, Cai D F, Zhang G P. A method of abbreviation bilingual information automated extraction based on the web [C]//Proceedings of the 8th China National Conference on Computational Linguistics, Nanjing, China, 2005:687-689

[20] Wang J D, Zhang Z X. Rapid extraction algorithm of abbreviation based on reverse scanning and co-occurrence analysis[J]. *Application Research of Computers*, 2018, 35(3):700-704

[21] Ding L, Li Y, He Y Q. Comparison study of domain adaption methods in statistical machine translation[J]. *Technology Intelligence Engineering*, 2016, 2(4):80-88

[22] Ding L, Yao C Q, He Y Q, et al. Application of deep learning in statistical machine translation domain adaptation[J]. *Technology Intelligence Engineering*, 2017, 3(3):64-76

[23] Olah C. Understanding LSTM networks [EB/OL]. http://colah.github.io/posts/2015-08-Understanding-LSTMs/:Git Hub, 2017

[24] Sundermeyer M, Schluter R, Ney H. LSTM neural networks for language modeling[C]//Annual Conference of the International Speech Communication Association, Portland, USA, 2012:194-197

[25] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of MT[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002:311-318

[26] Lin Q, Liu Q, Su J S, et al. Focuses and frontiers tendency in neural machine translation research[J]. *Journal of Chinese Information Processing*, 2019,33(11):1-14

He Yanqing, born in 1974. She received the B. S. degree in mathematics education in Hebei Normal University and M. S. degree in probability and mathematical statistics from Renmin University in 2005 and the Ph. D. degree in pattern recognition and intelligence system from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2009. From 2009 to 2012, she was a research assistant with Information Technology Support Center, Institute of Scientific and Technical Information of China (ISTIC). Since 2012, she has been a researcher in Research Center for Information Science Theory and Mechodology of ISTIC. Her research interests include MT, semantic role labeling, machine learning, natural language processing.