Feature Extraction of English Semantic Translation Relying on Graph Regular Knowledge Recognition Algorithm

Lidong Yang
School of Foreign Languages, Jingdezhen University, Jingdezhen, Jiangxi, China, 333000
E-mail: taylorjenny@126.com

Keywords: graph regular knowledge, recognition algorithm, english semantic translation, feature extraction

Received: May 28, 2023

Under the background of big data, people are not only pursuing the quantity but also the accuracy of knowledge in acquiring knowledge, especially for English. Because of the ambiguity, variety, and irregularity of English translation, people's reading has brought a lot of trouble. This paper aims to study the feature extraction of English semantic translation and suggests a recognition algorithm that relies on graph common knowledge. Through the analysis of graph regularization and the construction of the model, the recognition algorithm is improved, and the feature extraction methods are compared and analyzed. At the same time, experiments are intended to investigate the improvement of the English semantic translation of the improved recognition algorithm after feature extraction. The experimental results in this paper show that the improved English semantic translation has increased by 10%-15% in terms of translation accuracy. This degree of improvement has great application significance in actual English semantic translation.

Povzetek: Opisana je raziskava izboljšav angleškega semantičnega prevoda z uporabo izboljšanega algoritma prepoznavanja značilnosti.

1 Introduction

In the study of English semantic translation, we have selected the topic of extracting semantic translation features from the English translation. In recent years, as the global village boom continues, reading foreign language books has grown to be a need in peoples' everyday lives. People urgently need to understand its meaning and the story behind it. People's enthusiasm for English directly contributes to the development of foreign cultures, and English learning is also the most direct medium for English translators to focus on translation. However, the main foreign language activity reports and broadcasts are mainly conducted in English, and most of the domestic translation companies do not have enough ability to perform first-line accurate foreign language translation. The current agency's translation ability is relatively weak, and the requirements of domestic readers are becoming higher. Because of the imbalance between the two, people urgently need to translate the semantics of English, which is inevitable in the context of the development of world integration.

English semantic translation is an important research topic in natural language processing. It is essential to acquire knowledge and analysis in the context of big data. As a critical portion of natural language, the English translation has the characteristics of ambiguity, diversity, and irregularity, which has caused great problems in understanding natural language. The activation signaling technology accurately links the entity reference to the corresponding entity concept based on the narration. It introduces rich background knowledge based on narrating the tasks related to natural language processing. Solving the problems caused by the above characteristics can more appropriately provide services to related academic research and production applications.

As English has become one of the universal languages in the world, Chinese reform and opening up have also brought China into an era in which it is in line with international standards. English has also begun to flood China on a large scale, and foreign companies have also started to take root in China, accompanied by the urgency for English translation. An accurate translation can reduce a lot of trouble, making more and more people begin to invest in studying English semantic translation. Deng A said in his article that given a graph G with a vertex set V(G)=V and an edge set E(G)=E, let G(1) be a line graph and G(c) be the complement of G . Let G(0) be a graph with V(G(0))=V and no edges, G(1) has a complete graph of vertex set V, G(+)=G and G(-)=G(c), Let B(G)(Bc(G)) be the graph of the vertex set VbooleanORE, so that (ve) is an edge in B(G) (correspondingly, in Bc(G)) be vepsilon V, eepsilon E and vertices v, and G Event occurs on edge e in [1]. Bitkina V V raised the issue of studying regular distance graphs. The neighborhood of the vertex is a strongly regular graph. For a given positive integer t, the second

eigenvalue is at most t. This issue is simplified to the description of a regular distance graph, where the neighborhood of the vertex is a strongly regular graph with non-principal eigenvalues t = 1, 2,... [2].

Kaveh A believes that graph theory has a lot of applications in structural mechanics, as well as many topological transformations, to develop similar challenges easier. The skeleton diagram and natural correlation diagram of the finite element model are transformed in this way. These transformations can be effectively used in the ordering of nodes and elements of conventional limited element models. In his article, he proposed an effective method of using graphs and directed graph products to generate the skeleton graph, natural correlation graph, and its grid base of the finite element model [3]. Here, he tried to make the cuckoo search (CS) algorithm parameters free without the Levy step. The algorithm he proposed uses 23 standard benchmark functions for verification [4]. Sahoo SP offered an interest point detection technology based on the local maximum of difference images (LMDI), a selected projection tree with crossing segmentation, and a revised vote score for the acknowledgment of human action. In the interest point detection method based on LMDI, continuous frame difference technology is used to obtain different images. Then 3D peak detection is used to the calculated set of various images. Hough voting technology is applied to test videos to calculate the most significant correlation rating obtained when a single training class [5]. Aiming at the current semantic irregularities in the field of relay protection, Qian H designed an intelligent semantic recognition algorithm for relay protection information based on four modules: dictionary management, semantic matching, retrieval preprocessing, and retrieval. The acquired standard semantic data is examined and confirmed by testing various non-standard semantic data. It is proved that the relay protection information semantic intelligent recognition algorithm has good performance and feasibility [6].

Bracken J believes that translation is usually not directly aligned across languages, and indirect mapping will decrease the accuracy of language learning. He came up with a brand-new ongoing measure to make the examination of this issue easier to quantify the semantic relevance of words with multiple translations. He determined how the correlation between translations affects the learning of translation ambiguities from German to English. Compared with German words with high TSV value, German words with minimum TSV value are noticed as slower and low accurate to translate [7]. Tan Y W pointed out that due to the particularity of legal English, its translation differs from others' translations. Legal translation can be regarded as a dual operation of legal transfer and language transfer. Therefore, legal translation needs to consider many factors. Frame semantics is the perspective of translation, which provides a new view of legal translation. He proposed three legal translation strategies based on frame semantics. These three strategies are frame correspondence, selection, and transfer [8]. The above-mentioned documents mainly involve the introduction of graph common knowledge, recognition algorithms, and English semantic translation. But most of them stay at the research level of the technical level, and not too much research goes deep into the application level. This makes the use of the technology still not clear enough, and the critical points of the relevant technology are still not enough, which leads to the lack of persuasiveness of the article.

The innovation of this article lies in the theoretical support of English semantic translation. At the same time, the feature selection of English semantic translation based on the regular low-rank score of the graph is used as the technical support, and the improved feature extraction recognition algorithm is experimentally explored through design experiments. At the same time, semantic translation and graph regularization are entity-linked, and the accuracy of semantic translation is compared and analyzed. After analysis and comparison, the improved English semantic translation model is 10%-15% higher more accurate than the conventional translation model, which ensures the accuracy and stable operation of translation.

The rest of the portion is structured as follows: part 2 describes the related works, part 3 discusses the methodology of the study, part 4 represents the Graph Regularization and Semantics Entity-Link Experiment, part 5 presents the efficiency analysis, and Part 5 concludes the study with the future work.

2 Related works

Builds a semantic mapping model for interactively optimum English-Chinese translation, creates an English translation model using a feature extraction technique, and works out the best translation strategy utilizing the newly proposed feature extraction algorithm. When put into reality, however, it becomes clear that this approach suffers from slow English translation time. This has resulted in a poor degree of translation efficiency being maintained [26]. To enhance the quality of machine translation, the model combines the language-template-based translation approach with the statistical translation approach of the conditional random field to segment and analyze lengthy phrases along syntactic and statistical dimensions. Unfortunately, the model's implementation method is very complicated, which lengthens the time required to translate from English and reduces translation efficiency [27]. Determining the set of points and their neighbors to form a subgraph. Finally, it calculates the likelihood of local support using the associated relationship acquired by sorting the edges of the two subgraphs based on distance and angle. Various synthetic and actual data were used to verify the suggested method's performance,

demonstrating that it can enhance the resilience and accuracy of conventional methods [28]. A novel probabilistic clustering approach designed to isolate linear groups in datasets. The algorithm is a method for maximizing a mixed probability density function, similar to expectation maximization. A line segment is modeled by each process. The suggested approach is on par with or superior to modern cluster-based methods and conventional line detection techniques in experimental assessments [29]. The unlabeled text vocabulary is vectorized, it is accomplished by combining lexical representation with vector characteristics, and the valuable data about different phrases and their semantics is retrieved using a multilayer neural network model. To finish the construction of the English translation model, a neural network is utilized for Word evaluation and grading inside an online ranking framework, as well as for obtaining the semantic collection of the sample data and predicting the variation of word arrangement. However, English phrases are poorly recognized as parts of speech, leading to an inaccurate translation [30]. A perceptive recognition based on the enhanced GLR algorithm, an English translation model. The results of part-of-speech recognition may be acquired by building the phrase structure via the phrase center, and the English and Chinese structural uncertainty in the part-of-speech recognition outcomes may be improved by the syntactic function of the analytical, linear table. To fulfill the design of the English translation model, the recognized content is finally collected. However, the model struggles to correctly detect the part of speech of English phrases, which harms the quality of the ensuing English translation [31]. Interactive information retrieval issues may be solved using ontology-based techniques for semantic document recognition and representation. The presentation features interactive tools. The ontology is graphically represented by the device through the action of building aspect projections. From a visual and perceptual standpoint, this allows the graph's dimensionality to be reduced to a more manageable level. Keyword or shallow semantic parsing, the two most common efficient and reliable cipher text search techniques, cannot fully meet users' search intents [32]. It outperforms most current RGB-D networks because of its high accuracy and fast inference speed of 22 Hz at full 2048 1024 resolution. The two types of retrieval strategies, text-based and knowledge-based, continue to be at odds with one another. Both fail to adequately handle keyword-based query and ranking retrieval, although the former ignores intricate connections [33]. The issue of long-term dependencies was well-handled by the long short-term memory (LSTM) once gate functions were included in the cell structure. Since its inception, the LSTM has been responsible for almost all of the impressive achievements based on RNNs. Recently, deep learning has shifted its attention to LSTM. To investigate the LSTM cell's potential for learning, the study [34] conduct a systematic study of the LSTM cell and its variations. Sherlock, a multi-input deep neural network for semantic type detection, is presented in the research. By matching $78 in semantic types from DBpedia to column headings, the research also trained Sherlock on $686,765 in data columns that were pulled from the VizNet corpus. Each matched column in the research [35] is given $1,588 attributes that describe its statistical characteristics, character distributions, word embeddings, and paragraph vectors.

Table 1: Literature summary table

References	Methodology	Drawbacks
[28]	Improved Non-Rigid Point Set Registration Algorithm	The computational cost of non-rigid point set registration procedures may be high, especially when working with large-scale point sets or intricate deformations.
[30]	The transformer-based neural machine translation system	To perform well in translation, transformer models often need a lot of training data.
[31]	Improved GLR Algorithm	The enhanced GLR algorithm has trouble addressing frequent syntactic and semantic problems in natural language.
[32]	Ontology graphs	Scalability becomes a problem as the ontology graph's size and complexity rise.
[33]	Real-time fusion semantic segmentation network termed RFNet	It's possible that RFNet won't be able to successfully gather and use contextual data.
[34]	Long short-term memory	Instead of explicitly modeling network topologies, LSTM

	(LSTM)	is optimized for processing sequential data like time series or natural language words. Because of its sequential structure, LSTM may not be able to properly capture the relationships and dependencies between entities represented in a graph, which is a common requirement of graph regular knowledge recognition methods.
[35]	Sherlock, a multi-input deep neural network	Some temporal or dynamic features of the knowledge graph may be lost in Sherlock's graph regular knowledge recognition technique. This might hinder its capacity for learning and recognition of changing information, since it may be unable to accurately represent time-dependent connections or respond to shifts in the graph.

3 English semantic translation method

3.1 English semantic translation theory

This article explores the characteristics and countermeasures of English-Chinese translation under the guidance of text classification and semantic translation theory. The factors mainly include five types of words and sentences and context characteristics, and the countermeasures correspond to one of them.

According to the classification of text types, sports news with universal text characteristics should be classified as informational text. In other words, the translation of sports news should be based on communication translation. However, on the fundamentals of translating (1. Translation principles depends on the target language or the target language. 2. The translation principle oriented towards the author and reader. 3. Aesthetics-oriented translation principles), when the specific language form and content of the original text are equally important, it is also mentioned that semantic translation is needed when it has nothing to do with the type of the original text. When translating such texts, the primary function of news text is to convey information that cannot be ignored [9].

Based on translation theory, this article mainly studies and solves the following two aspects. Analyze the words, sentences, and context characteristics in the text, and explain the corresponding translation strategies. According to the author's translation conventions, this article summarizes the characteristics of almost all 5 kinds of words and sentences in English translation, which are professional terms, idioms, direct quotations, cultural words, and representative words borrowed [10]. Based on these five characteristics, the translation countermeasures of strict observance of norms, obedience to the mainstream, credibility and expressiveness, specific analysis, and classification discussion are proposed, and the context in English translation is explored. The major categories of English translation are shown in Figure 1:

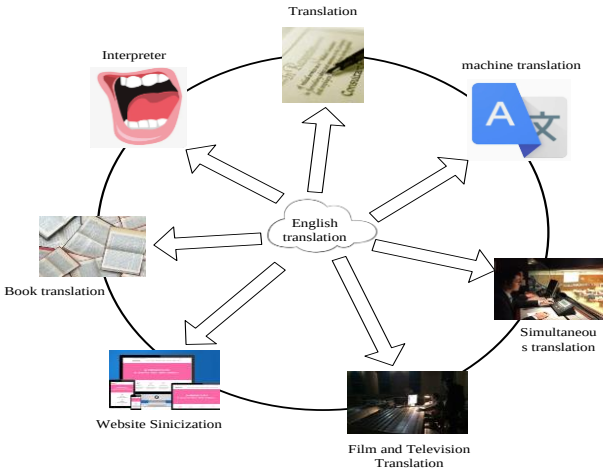


Figure 1: Several types and applications of the English translation.

3.2 Types of texts and division of readers

German scholar Karl Charles [11] advocated the three functions of language, information function, expressive function, and infectious function, and divided text into three types: information type, expression type, and meaning type. At the same time, Newmark proposed three types of information text, formula type text, and call type text [12]. As shown in Figure 2, there are three types of text. Information text explains objective events without personal contact, and information occupies a dominant position in most non-literary works. The pictorial text and text reflect entirely the original author's language style, thoughts, and feelings, as well as the form and content of the language. The author occupies a dominant position in serious literary works, and the circulated texts are aimed at attracting readers because readers occupy a dominant role in advertisements.

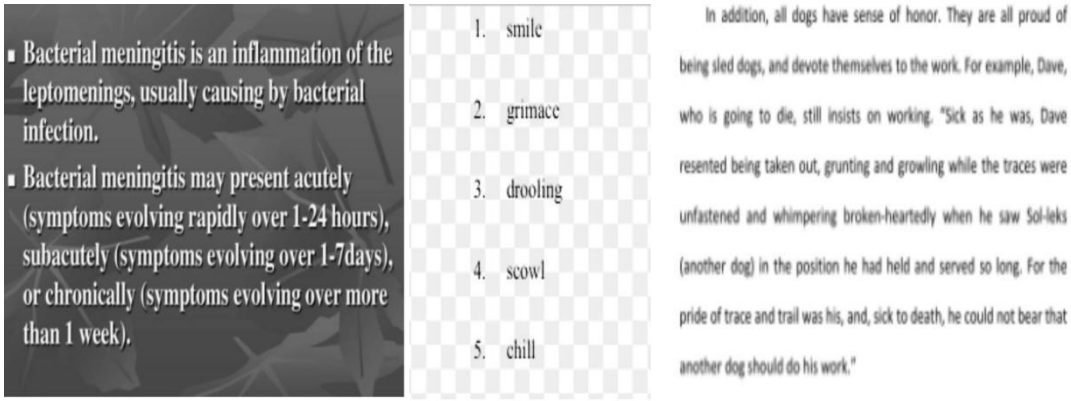


Figure 2: Three different types of text

Newmark divides the readers into three categories based on three factors: knowledge reserve, intelligence level, and learning ability: readers engaged in language work, readers with a certain level of knowledge, and ordinary readers.

The choice of translation method is based on careful consideration of factors such as the type of text, the readership, and the purpose of the translation.

1) Language positioning of the target text

From the point of view of article types, it mainly covers the following two points. First of all, the general function of the text is to spread information and has auxiliary functions to call readers, and secondly, it is the expressive function of English text. Since some of the main work of the text is to spread information, according to the classification of the three text types of tokens, a part of the text first belongs to the information type [13]. If a specific language form and a part of the text are equally important, then no matter what type of text, the text must emphasize the function of expression, and the importance of the meaning unit is very high, so meaning translation must be used [14]. For meaningful translation, we should focus on the following points: 1. Highlight the subject. The topic of a statement is crucial. The heart of a sentence is its topic. If the sentence's topic is incorrect, it will seem quite rambling. 2. Pay attention to the

collocation of words. The English translation does not want to be Chinese; even if the words are not matched correctly, they can still be understood in the wrong order. However, English is different. In English, You need to be aware of how adjectives and nouns, adverbs and verbs, and other combinations are used together. Direct quotations in some texts are related to value judgments. Translators must follow the principle of neutrality, communicate faithfully, and use semantic translation. However, some texts question this point. The contradiction between the randomness of the spoken language and the logic of the written language is that when the spoken language enters the written language, specific logical adjustments are made to the original spoken language according to the solid analytical characteristics of the written language [15]. Unknown semantic language in spoken language often requires translators to make appropriate adjustments based on the original text. This can also make it easy for readers to understand. Of course, all the above adjustments must be premised on not changing the intrinsic meaning and value judgment of the original text.

3.3 Three common feature selection algorithms

1) Variance score

Consider a collection of data X=[x1,x2,⋯,xn]∈Rdxn, d represents the feature dimension of the data, n = a total number of sample points, and xi = data sample points of a column vector. Then the variance scoring model is defined as:

$$H_x(r)=\sum_{i=1}^{n}(f_n-\mu_r)^2 \tag{1}$$

From the variance model 1, it can be seen that the larger the Hs® value of a feature, the more sufficient the amount of information contained in the quality. Different types of samples are distinguished by the difference in the information contained in this feature point of other data sample points. That is, the more significant the difference between different data sample points in this feature value, the easier it is for this feature to distinguish other sample points. Therefore, the variance scoring is to evaluate each feature point by formula 1, select those feature points with high scores and rich information, and discard those with low scores and less information.

2) Laplace score

Laplace scoring is to increase the similarity constraint between data on the variance scoring model; that is, the feature points with the same category have similar spatial distributions, and the scoring model is expressed as:

$$L_s(r)=\frac{\sum_{x,y}(f_{rx}-f_{ry})^2 W_{xy}}{V_s(r)} \tag{2}$$

From the formula, we can see that for a good feature, the Laplace score Ls ® should be smaller. By scoring each feature value, the features with lower scores are selected to form a new feature subset, thereby reducing dimensionality [16].

3) Sparse scoring

In sparse representation, a data sample point is linearly reconstructed by a few other data points under an over-complete dictionary to obtain a more concise data representation. The reconstructed coefficient matrix thus obtained replaces the similarity matrix in the Laplace score to get a sparse scoring model, which can fully express the local topological structure information between the data.

If these sample points come from the same subspace, there is a high similarity and correlation between them, which can play a more significant part in the reconstruction method. On the contrary, if they come from different subspaces, the similarity correlation between the sample points is weak, so they play a small role in the reconstruction process [17]. Therefore, the linear reconstruction coefficients of the sample points exhibit sparsity. Thus, in the process of reconstructing the coefficients of the sample, the sparsity constraint is added to the coefficients, and the resulting model is as follows:

$$\min\|\alpha_x\|_0 \quad s.t. x_x=\sum_{y=1,y\neq i}^{n}\alpha_{xy}x_y \tag{3}$$

Where n represents the total number of sample points.

For the formula, we consider the influence of noise, so the above equation constraint model is relaxed into the following form: $\min\|\alpha_x\|_1 \quad s.t.\left\|x_x-\sum_{y=1,y\neq i}\alpha_{xy}x_y\right\|\leq\varepsilon$ (4)

Therefore, the similarity between the two sample points is:

$$W_{xy}=W_{yx}=\frac{1}{2}(\alpha_{xy}+\alpha_{yx}) \tag{5}$$

After understanding the similarity between the two collected samples, we can better carry out a comprehensive screening of their characteristics.

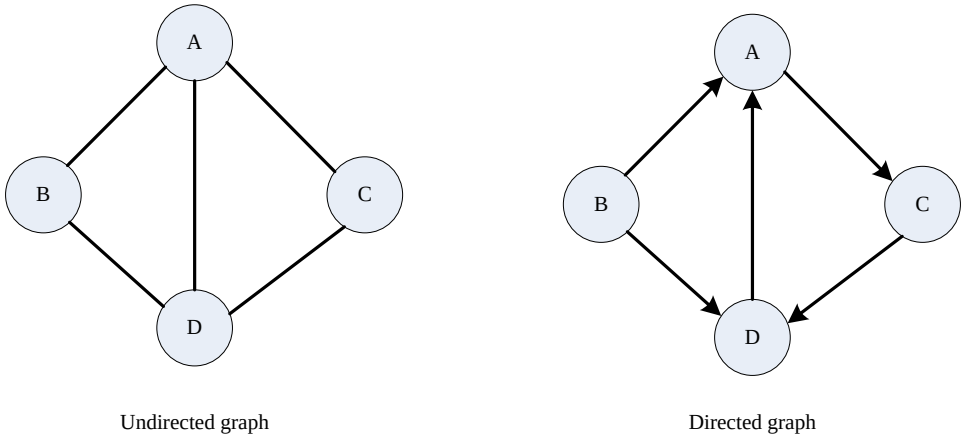
By comparing and analyzing the formulas of three standard feature selection algorithms, we conclude that the coefficient scoring algorithm among the three selection algorithms has better feature selection. This can be well applied to our research topic. So we finally adopted the sparse score selection method for feature

screening.

3.4 Feature selection of english semantic translation based on graph regular low-rank score

1) Basic concepts of graphs

A graph is a data structure composed of a collection of vertices and a collection of relations between vertices, which can be represented by the symbol Graph=(V, E). V is the set of vertices, and E is the edges between vertices. If the edges of any two vertex times shown in Figure 3 are undirected, then the graph is called an undirected graph [18].



Undirected graph Directed graph

Figure 3: Undirected graph and directed graph.

A directed graph D refers to an ordered triple (V(D), A(D), ψD), where ψD) is the correlation function, and each element in A(D) is termed a directed element (called produced an edge or arc) is equivalent to an ordered element (called a vertex or point) in V(D).

In image processing, the so-called graph usually refers to an undirected graph. In processing, the corresponding graph is generally calculated under manifold assumption. The manifold hypothesis means that samples in a small neighborhood have similar properties.

The low-rank representation model uses its data as a dictionary to learn the lowest-rank coefficient matrix, which has a robust global description and anti-interference abilities. However, many studies in the field of manifold learning have shown that the local structural information of the data also plays a significant role in accurately expressing the essential attributes of the data. Manifold wisdom refers to Manifold learning is the process of discovering the low-dimensional manifold within the high-dimensional space and then finding the corresponding embedding mapping to accomplish dimensionality reduction or data visualization under the assumption that the data was uniformly sampled in the low-dimensional manifold. The goal is to get to the heart of things by analyzing events and discovering the underlying rules that produce information. The graph-based algorithm can reflect the local structure data of the high-dimensional sample space very well, therefore, selecting discriminative translation methods conducive to clustering and classification from the massive English semantic data. In this part, a new data representation model is constructed by combining the original LRR model and the graph regularization item reflecting the local similarity structure of the data that is called the graph regular low-rank representation [19]. After that, the coefficient matrix obtained by the solution model is used to construct the graph weight matrix, and a brand-new scoring method is accepted for feature selection in English semantic translation. This is called a graph regular low-rank scoring algorithm, and the specific process is shown in Figure 4.

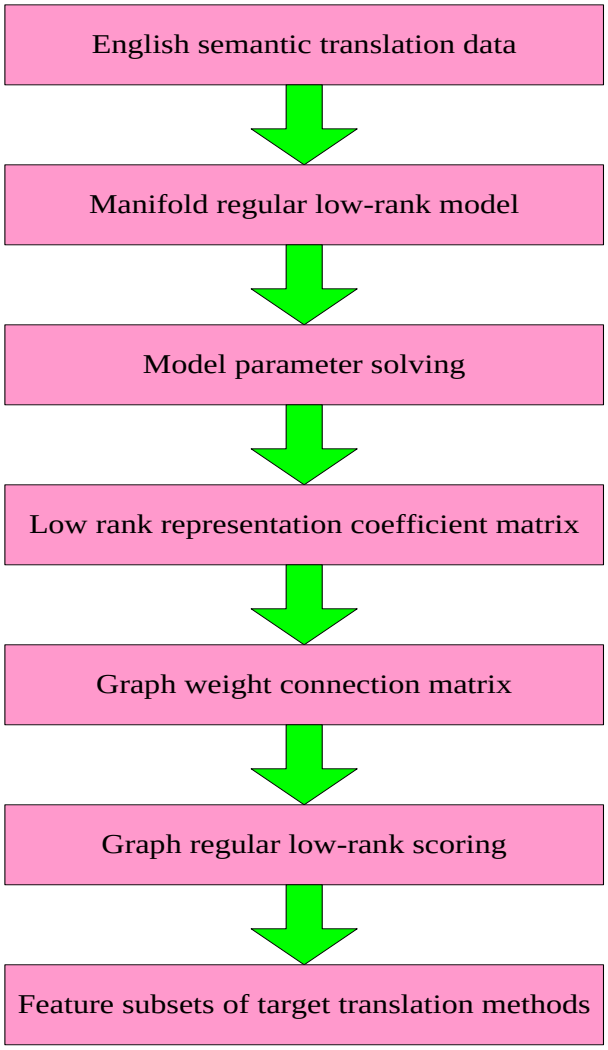


Figure 4: Translation feature selection method based on graph regular low-rank scoring.

2) Constructing the regular term of manifold

The geometric structure information of the data plays a vital role in the discrimination of information. To maintain the local geometric structure between samples in the neighboring space, according to the principle of various hypotheses, the two sample points, xp, and xq, in the high-dimensional data space are adjacent points of each other. Then the coefficients of their corresponding low-dimensional space indicate that zp and zq also have a neighbor relationship. Therefore, manifold learning still maintains the geometric topological structure of the high-dimensional space after dimensionality reduction, thus simplifying the operation [20]. Here we use the more popular Gaussian kernel function to express, namely:

$$W_{pq} = e^{\frac{\left\|x_p - x_q\right\|_2^2}{\sigma}} \tag{6}$$

Based on the manifold hypothesis, high-dimensional data is embedded in a low-dimensional manifold. When two samples are distributed in a small local neighborhood in the low-dimensional manifold, they are assigned the same category. To achieve this, a reasonable way is to minimize the following functions:

$$R(H) = \frac{1}{2}\sum_{p,q=1}^{N}\left\|z_p - z_q\right\|_2^2 W_{pq} = \sum_{p=1}^{N} z_p^T z_p D_{pp} - \sum_{p,q=1}^{n} z_p^T z_q W_{pq} \tag{7}$$

After reducing, you can get the following:

$$R(H)=Tr\left(ZLZ^{T}\right) \tag{8}$$

Where Tr represents the trace of the matrix, and D is a diagonal matrix.

The significance of Equation 8 is to use the weight between sample points to reflect their distance in the accordingly low-dimensional space. That is, when the weight between them is more significant, the length in the accordingly low-dimensional space is closed. On the contrary, when the weight between them is small, the distance in the accordingly low-dimensional space is far [21]. According to the low-rank representation model and Equation 8, the objective function obtained is as follows:

$$\min_{Z,E}\|Z\|+\lambda\|E\|_{2,1}+\frac{B}{2}Tr\left(ZLZ^{T}\right) \tag{9}$$

$$s.t.X=XZ+E \tag{10}$$

3) Model solution

To make each variable in the objective function easy to separate during the alternate update process, a new auxiliary variable, J, is first introduced into the model, and the model becomes:

$$\min_{Z,j,E}\|J\|+\lambda\|E\|_{2,1}+\frac{\beta}{2}Tr\left(ZLZ^{T}\right) \tag{11}$$

$$s.t.X=XZ+E,\ \ Z=J \tag{12}$$

The structure of the model is shown in Figure 5:

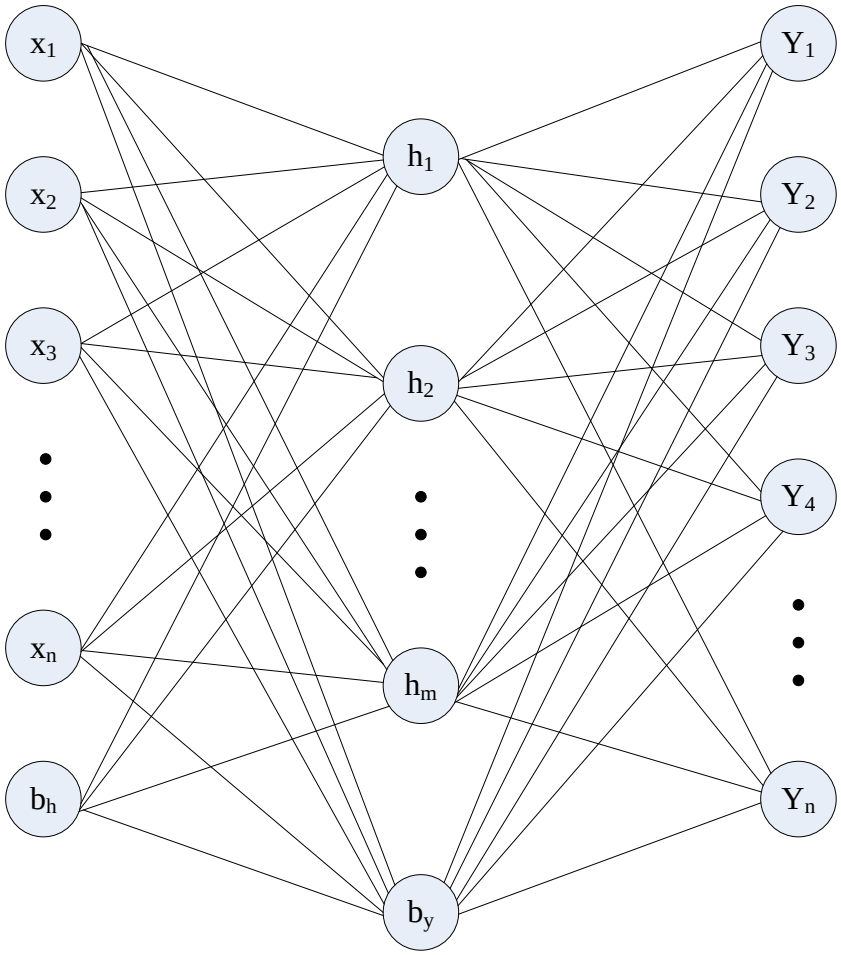


Figure 5: Model structure diagram.

And it is solved by minimizing the following augmented Lagrangian function:

$$L(E,J,Z,Y,U,\mu)=\|J\|+\lambda\|E\|_{2,1}+\frac{\beta}{2}Tr\left(ZLZ^{T}\right)+\langle Y,X-XZ-E\rangle \tag{13}$$

Derivatives can be obtained:

$$L(E,J,Z,Y,U,\mu)=\|J\|+\lambda\|E\|_{2,1}+\frac{\beta}{2}Tr\left(ZLZ^{T}\right) \tag{14}$$

By fixing two variables, the parameters J, Z, and E can be updated alternately, and then the parameters Y and U can be edited. The above problems can be divided into the following sub-problems.

1) Update J

$$J_{k+1}=\arg\min_{J}\|J\|+\frac{\mu}{2}\|Z-J+U/\mu\|_{F}^{2} \tag{15}$$

$$J_{k+1}=\Theta_{\mu^{-1}}(Z+U/\mu) \tag{16}$$

2) Update Z

$$Z_{k+1}=\arg\min_{Z}\frac{\beta}{2}Tr\left(ZLZ^{T}\right)+\frac{\mu}{2}\left(\|X-XZ-E+Y/\mu\|_{\mu}^{2}+\|Z-J+Y/\mu\|_{F}^{2}\right) \tag{17}$$

$$Z_{k+1}=\left(I+X^{T}X+\frac{\beta}{\mu}L\right)^{-1}\left(X^{T}(X-E)+J+\frac{X^{T}Y-U}{\mu}\right) \tag{18}$$

3) Update E

$$E_{k+1}=\arg\min_{E}\lambda\|E\|_{2,1}+\frac{\mu}{2}\|X-XZ-E+Y/\mu\|_{F}^{2} \tag{19}$$

$$E_{k+1}=\Omega_{\mu\text{-}}(X-XZ+Y/\mu) \tag{20}$$

4) Update Y

$$Y=Y+\mu(X-XZ-E) \tag{21}$$

5) Update U

$$U=U+\mu(Z-J) \tag{22}$$

6) Update parameter μ

$$\mu=\min(\rho\mu,\mu_{\max}) \tag{23}$$

7) Verify the convergence criteria

$$\|X-XZ-E\|_{\infty}<\varepsilon \tag{24}$$

$$\|Z-J\|<\varepsilon \tag{25}$$

The updated graph regularization model is shown in Figure 6:

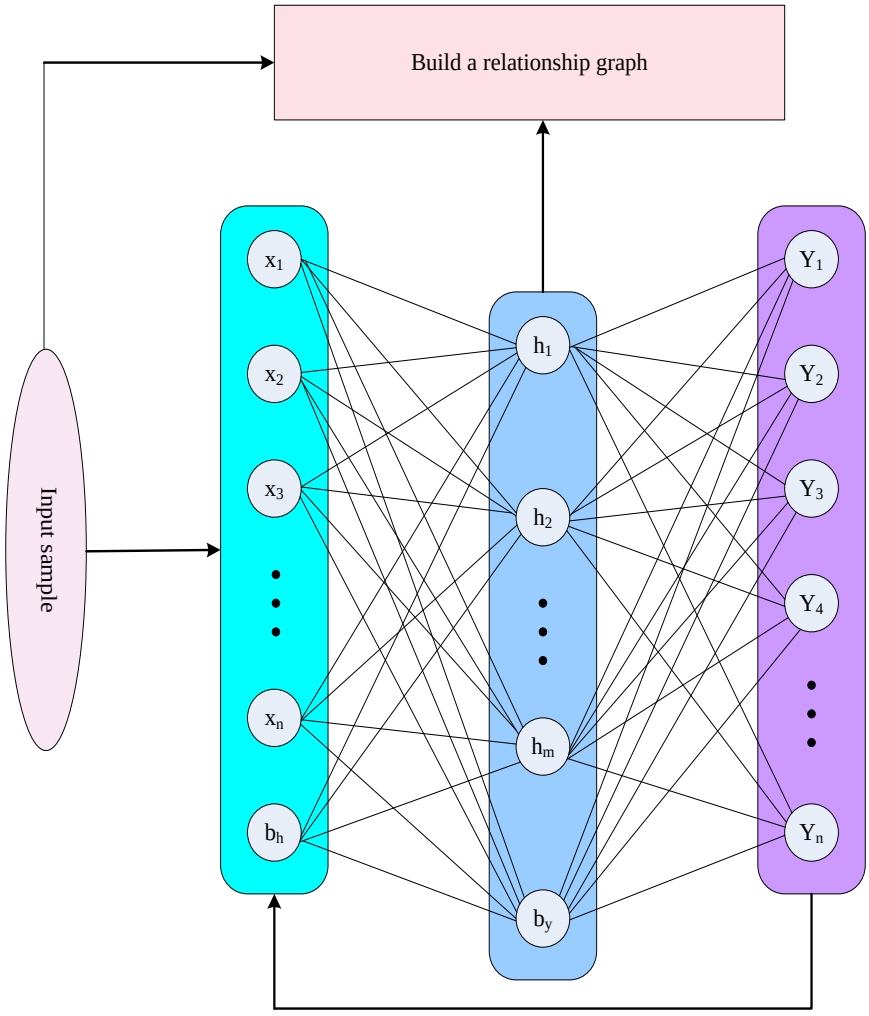


Figure 6: Schematic diagram of the updated graph regularization model.

4. Semantic representation and graph regularization entity link experiment

4.1 Entity link system structure

According to the characteristics of the task, the entity link system is primarily split into two modules, namely candidate entity generation and candidate entity disambiguation. As shown in Figure 7, it is a schematic diagram of a microblog entity linking system [22].

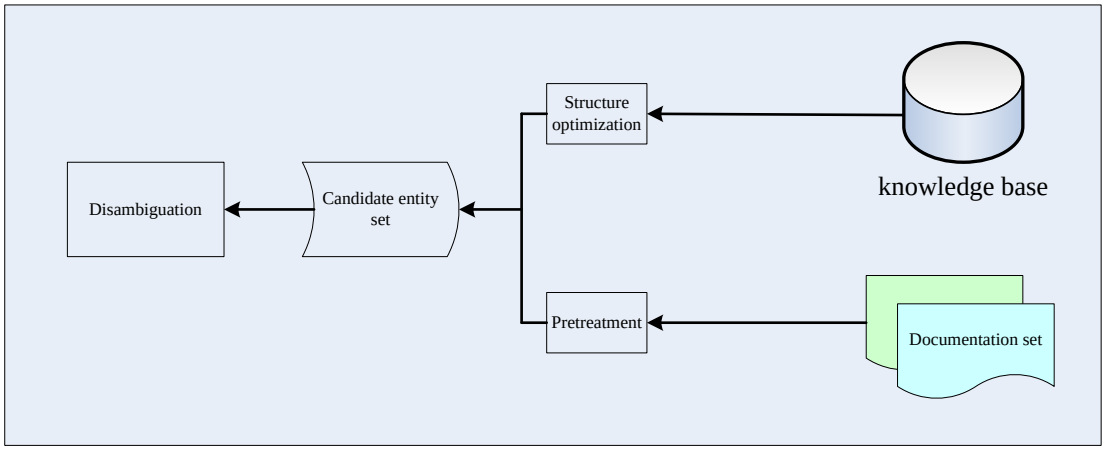


Figure 7: Schematic diagram of the entity link system.

1) Difficulties of entity linking

Due to the diversity and complexity of named entities, entity linking faces various problems. For example, the ambiguity of the entity reference, the referential variety of the entity, etc.

1)The ambiguity of entity reference

The ambiguity of entity reference generally means that an entity reference has multiple meanings, and it is impossible to determine which entity the reference refers to only from the surface form of the entity reference. The phenomenon of duplicate names is one of the most representative ambiguity problems of entity referents. As shown in Figure 8, the entity refers to "Zhang San," with 3 different persons corresponding to it. If there is no further effective information, it is difficult for us to judge what it specifically refers to. In addition, place names and organization names also have the problem of entity ambiguity [23].

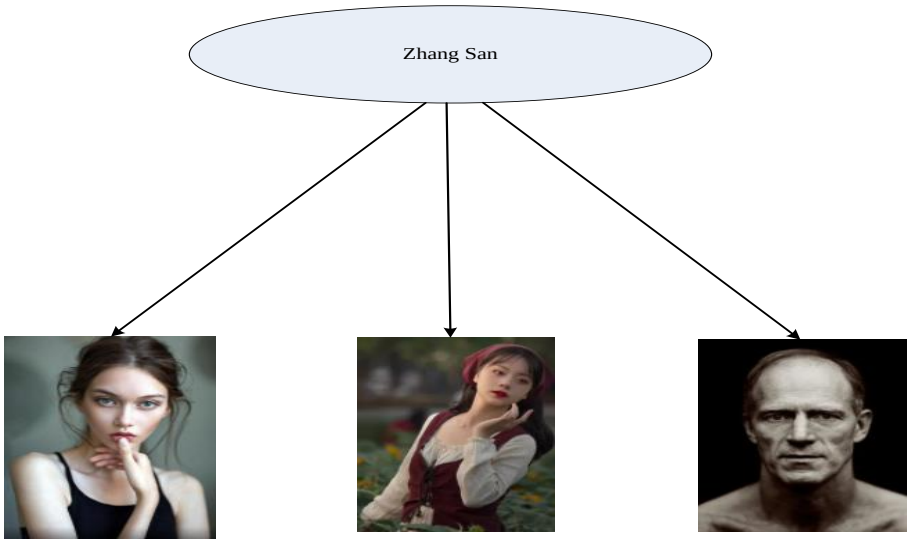


Figure 8: Ambiguity of names

2) Referential diversity of entities

Entity referent diversity generally means that a named entity in the knowledge base often has many entity referents. If an entity reference not covered in the knowledge base is used in the background document, it will be challenging to link the connection to the corresponding named entity. For example, American basketball star Stephen Curry (Figure 9) has as many as 8 physical references (including nicknames, nicknames, etc.), and there will be new references [24].

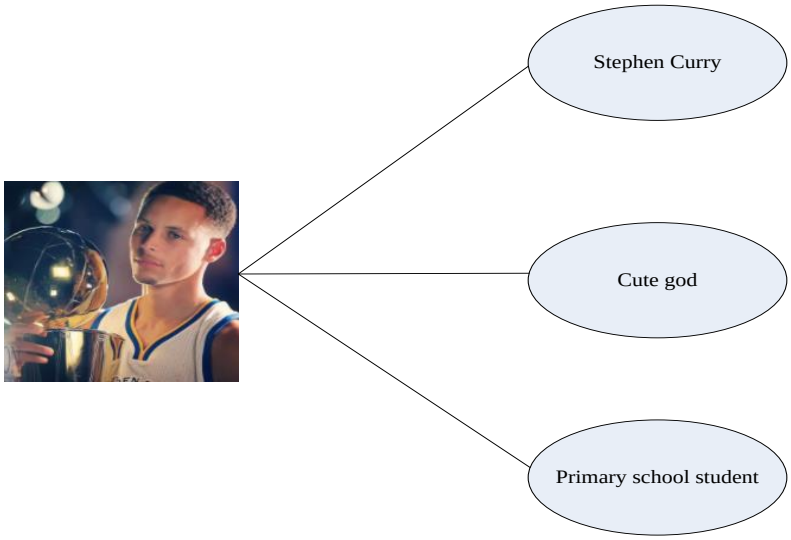


Figure 9: Examples of referential entity diversity

3) The deep semantic relationship model

To calculate the relevance of entities in terms of local consistency, this paper advocates learning latent semantic entity representations, which can reflect the latent semantics of entities.

The difference is that when we construct the feature vector layer of the model, we comprehensively use the four types of information in the knowledge base to represent each entity. They are related entities, entity relationships, entity types, entity descriptions, etc. Above the word hashing layer, we set up multiple hidden layers to perform the non-linear mapping. Concerning the objective function designed for entity relations, the deep neural network can learn useful semantic features using the back-propagation algorithm [25].

4) Deep semantic relationship model training

To train a deep semantic relationship model that can obtain entity semantics sensitive to entity relationships, we first automatically extract training data based on the knowledge base and Wikipedia annotations. In addition to using the linked entity pairs in the knowledge base as positive training samples, we will also draw more training samples from Wikipedia, significantly negative training samples. In training the model, we use the highest likelihood estimation strategy to evaluate the model parameters to maximize the probability of the occurrence of positive training samples and minimize the loss function.

4.2 Description of experimental data set

The data set selected in this experiment are all Chinese Weibo data sets, which were provided by the entity link tasks of the Natural Language Processing and Chinese Computing Conference in 2013 and 2014, respectively. All data are given in XML format. There is no correlation between two different data, and the entities between the data sets are also not correlated.

In addition, we conducted statistics on the entity link data sets of the Natural Language Processing and Chinese Computing Conference in 2013 and 2014. The 2013 data set consists of an overall of 964 Weibo data, including a total of 1,498 entities. The detailed statistical outputs are depicted in Table 1.

Among them, the 2014 data set consists of 1257 Weibo data, including 1402 entities. The detailed statistical outputs are depicted in Table 2.

Table 1 Detailed statistics of the 2013 entity link evaluation data set

	Training data set	Test data set
Total number of Weibo	187	760
Total number of entities	239	1276
Linked entities	92	-
No linked entity	157	-

Table 2 Detailed statistics of the 2014 entity link evaluation data set

	Training data set	Test data set
Total number of Weibo	237	1124
Total number of entities	157	1118
Linked entities	201	-
No linked entity	72	-

Use the mentioned entities to link datasets, knowledge bases, and evaluation methods. This article uses the Lucene-based entity linking method, vector space model-based entity linking approach, and entity linking process based on semantic representation and graph regularization to conduct experiments. Aiming at the experimental outcome, this study analyzes the overall accuracy rate, the accuracy rate of logged-in entities, and the accuracy rate of unlogged-in entities.

1) Comparative analysis of overall data accuracy

This article first conducted entity link experiments on the 2013 and 2014 data sets to compare and analyze the accuracy of different methods on the overall data set. Because the named entities that need to be linked are already given in the data set, follow the usual practice. The use of accuracy to measure the practical effect of entity-linking strategies on the overall data is of

reference. The recall rate and F value are not considered here.

The experimental results of each method on the 2013 comprehensive data set are shown in Table 3. Among them, the Best_2013 system is the best score on this data set in the evaluation.

The experimental results of each experiment method on the 2014 comprehensive data set are shown in Table 4. Among them, the Best_2014 system is the best score on this data set in the evaluation.

Table 3: Accuracy statistics of the overall data set in 2013

	Lucene-EL	VSM-EL	DNN-EL	Best-2013
Total number of correct results	588	762	742	687
Total number of entities to be linked	900	879	831	847
Accuracy	0.653	0.867	0.893	0.811

Table 4: Accuracy statistics of the overall data set in 2014

	Lucene-EL	VSM-EL	DNN-EL	Best-2014
Total number of correct results	381	520	537	522
Total number of entities to be linked	587	600	608	627
Accuracy	0.649	0.867	0.883	0.833

From the Table, it is not difficult to see that in the overall data set accuracy statistics in 2013, the DNN-EL method has the highest accuracy, reaching 89.3%. Followed by the VSM-EL method, the accuracy rate reached 86.7%, and Best-2013 has the third accuracy rate, reaching 81.1%. The worst is the Lucene-EL method, with an accuracy rate of 65.3%. In the overall data set accuracy statistics in 2014, the accuracy of the DNN-El method is also the highest, reaching 88.3%. Next is VSM-EL, which has an accuracy rate of 86.7%. Best-2013 has an accuracy rate of 83.3%, and the lowest is Lucene-EL, which has an accuracy rate of 64.9%. Through the analysis of the above data set, it is not challenging to see that the accuracy of DNN-EL is relatively high and can be used. The accuracy statistics of the overall data set of the trained relational model are depicted in Table 5:

Table 5: Accuracy statistics of the overall data set after training

2013				
	Lucene-EL	VSM-EL	DNN-EL	Best-2013
Total number of correct results	542	713	779	748
Total number of entities to be linked	828	831	832	816
Accuracy	0.655	0.858	0.936	0.917
2014				
	Lucene-EL	VSM-EL	DNN-EL	Best-2014
Total number of correct results	368	508	545	531
Total number of entities to be linked	613	617	620	599
Accuracy	0.6	0.823	0.879	0.886

From the Table, we can see that the accuracy after training and improvement has improved to varying degrees. In the 2013 data set, the accuracy rate of DNN-EL reached 93.6%, the accuracy rate of Best-2013 reached 91.7%, the accuracy rate of VSM-EL reached 85.8%, and the accuracy rate of Lucene-EL reached 65.5%. In the 2014 data set, DNN-EL has an accuracy rate of 87.9%, Best-2014 has the highest accuracy rate of 88.6%, VSM-El has an accuracy rate of 82.3%, and Lucene-EL has an accuracy rate of 60%. It can be seen that there is still a specific improvement in accuracy, and the progress in the accuracy of semantic translation can significantly improve the problem of semantic inaccuracy in the actual translation.

4.3 Translation efficiency analysis

1) English Semantic Translation Analysis

The experimental outcome show that the DNN_EL method has achieved the highest entity link accuracy rate, achieving good results of 89.3% and 88.3% on the 2013 and 2014 data sets, respectively, which is better than the best results on each data set; the accuracy rate of the VSM_EL method is second, with a success rate of 86.7% in the two overall data sets; the worst performer is the Lucene_EL method, with accuracy rates of 65.3% and 64.9%, respectively. Through data analysis, we found that, except for unregistered entities that do not need to perform entity disambiguation and return to NIL directly, the results will reflect the effect of entity disambiguation by various methods. The entity link method based on Lucene only uses query keywords for entity disambiguation, and the experimental results are not particularly ideal; the process depends on the vector space model, further uses the context information of the named entity, and the experimental output is greatly improved; the best is the method based on semantic representation and graph regularization, which incorporates more features and achieves an accuracy of more than 90%. The visual display of the experimental results of each method is shown in Figure 10.

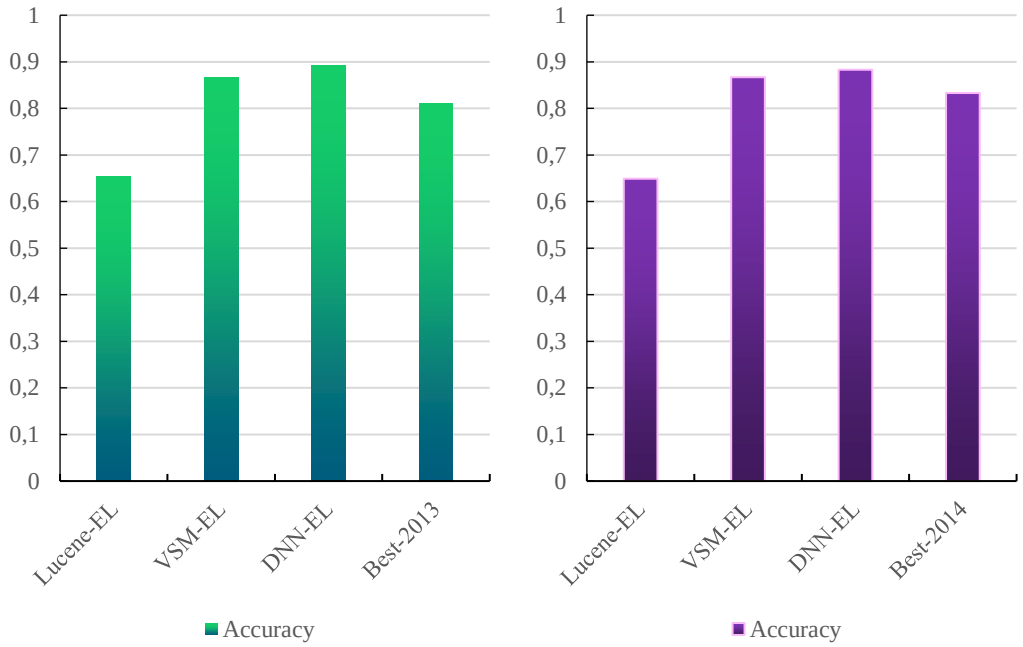


Figure 10: Display of the accuracy rate of all data in 3013 and 2014.

The method based on semantic representation and graph regularization not only performs knowledge-base matching but also optimizes regularization on graphs constructed based on context and entity semantic similarity. This avoids false matching of unregistered entities to the greatest extent and improves the recognition accuracy and F value of unregistered entities. The visual display of the experimental results of each method is shown in Figure 11.

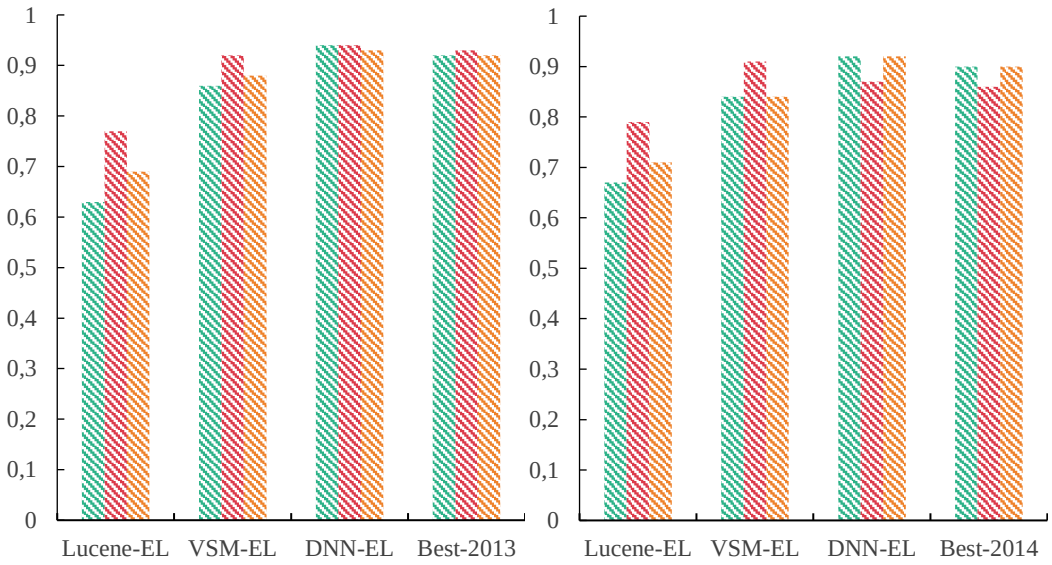


Figure 11: Display diagram of various indicators of the link results of unregistered entities in the knowledge base.

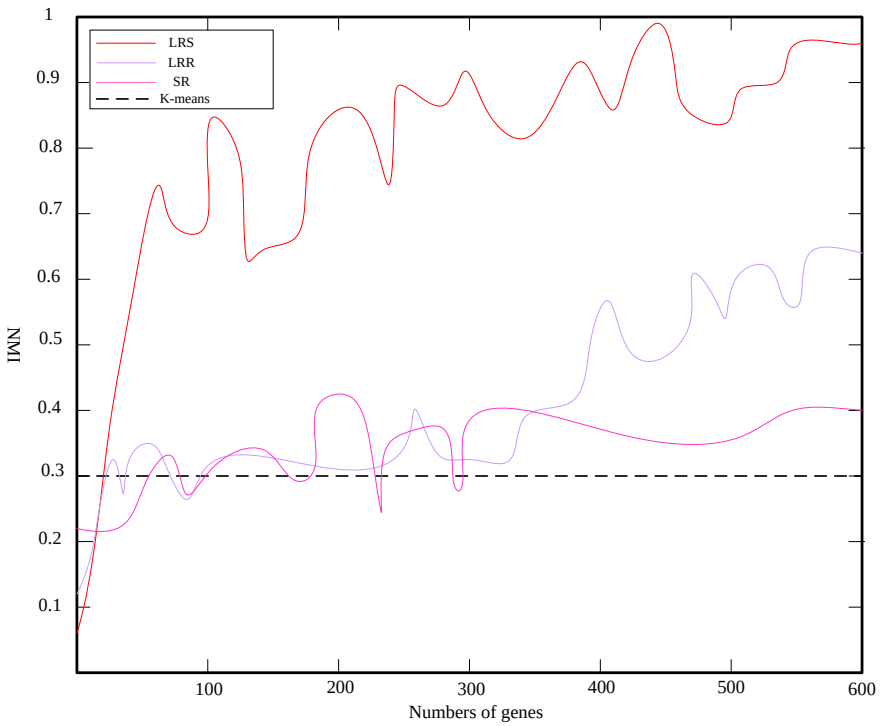
By comparing the two graphs, we can find that among the four methods, DNN-EL has the highest accurate data rate, followed by VSM-EL, Best-2013, and the lowest Lucene-EL. In the 2014 data set, DNN-EL has the highest translation semantic accuracy, and Lucene-EL has the lowest. It can be seen that in the actual use process, it is best to use DNN-EL to perform English semantic translation, which can better ensure the accuracy of our English translation.

In the display diagram of the various indicators of the entity link results, we can see that in the 2013 data set, the multiple indicators of DNN-EL, whether in accuracy, recall, or F-value, are among the top three indicators. The accuracy rate of the same Best-EL is also among the best, followed by the accuracy rate of the semantic translation of VSM-EL. The indicators of both are still good, and the worst is Lucene-EL. Among the four methods, all his indicators are of relatively low data, so they are not applicable.

In the 2014 data set, the indicators of DNN-EL and Best-2014 are relatively high. The overall English semantic translation accuracy rate is still relatively high, and it is instead used, followed by VSM-EL. The indicators are generally average and are at a reasonably high level. The lowest is Lucene-EL. All three indicators belong to the lowest category, so this article does not use this method.

2) Feature Extraction Analysis

To obtain reliable experimental results, firstly, different recognition algorithms are used to score the feature points of each segment of English semantic translation. Correspondingly the feature point scores and their importance is sorted from low to high, and then the first 600 features with low scores are selected to form a subset of the translation target. Finally, use K-means to perform 20 clustering experiments on the obtained target feature subset and choose the best clustering result; on the other hand, K-means is used directly for clustering experiments on all the original data sets and contrast with the previous method; Finally, NMI and ACC are used as evaluation indicators to determine the performance of each algorithm in the clustering experiment. Figure 12 is the NMI and ACC trend charts obtained by the three scoring methods and the K-means clustering method without feature selection on four standard English semantic translation data sets under their respective parameters.



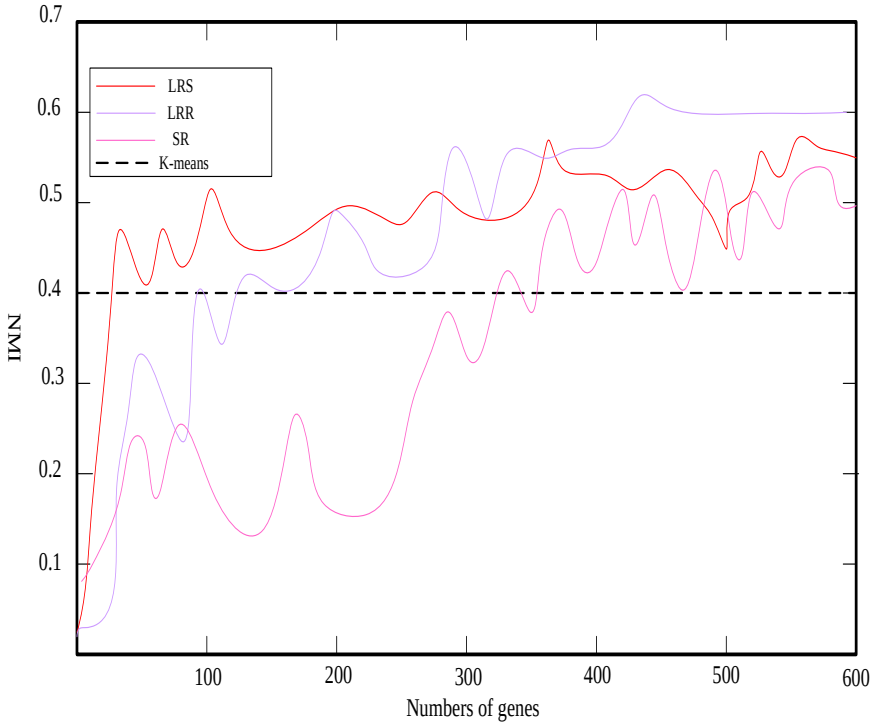


Figure 12 NMI and ACC charts without feature selection and feature selection.

Through comparison, we found that:
(1) In the expression data set, for the clustering accuracy rate, the overall LRS score is on the rise.
When the feature points are less than 400, his fluctuations are relatively large. When the number of feature points is between 400-600, the overall trend is stable, and the overall clustering accuracy is higher than other algorithms. It is lower than the low-rank scoring algorithm on individual feature points. For normalized mutual information, the LRS score shows more substantial superiority than the other three algorithms;
(2) When the number of selected features is less than 300, the two indicators of the LRS score on the data set are significantly better than other scoring methods; when the number of elements is minimal, the clustering algorithm without feature screening is considerably better than different algorithms.

4.4 The correct rate of english semantic translation

In English semantic translation, we not only pursue the speed of translation but also ensure the accuracy of the translation. For English translation, the translation of English semantics is the most important thing. Here we compare the traditional English semantic translation and the improved English semantic translation. Compare the translation efficiency, translation speed, and accuracy of the two translation modes. To this end, we design an experiment for comparison by comparing a large number of data sets and testing the stability of their translation; the test results are shown in Figure 13:

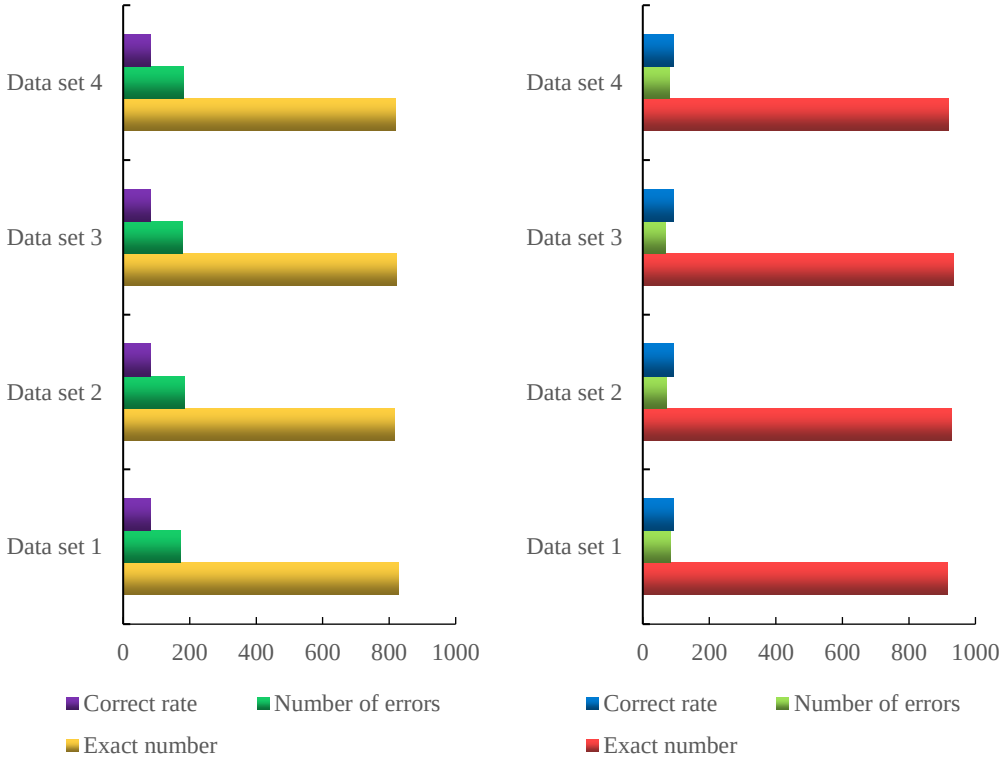


Figure 13: Traditional English semantic translation and improved English semantic translation

From Figure 13, it is not difficult to see that the improved English semantic translation is significantly better than the traditional English semantic translation in terms of translation rate and accuracy. The translation accuracy rate of the conventional translation mode is maintained between 80%-85%, the improved semantic translation accuracy rate is maintained at 90-95%, and the accuracy rate is increased by 10%-15%. This translation mode can be well applied in actual translation, with highly high translation accuracy.

4.5 Discussion

According to Figure 10, the DNN_EL technique has the greatest entity link accuracy rate, reaching excellent results of 89.3% and 88.3% on the 2013 and 2014 data sets, respectively. The VSM_EL method and Lucene_EL method came in second and third on each data set, respectively. Figure 11 shows a visual depiction of several indications of the knowledge base's connection findings for unregistered entities. DNN-EL has the most correct data rate among the four models, followed by VSM-EL, Best-2013, and Lucene-EL, which have the lowest accuracy rate. DNN-EL and Lucene-EL both have poor translation semantic accuracy in the 2014 data set. Concerning four typical English semantic translation data sets, we examined the NMI and ACC trend charts produced by the three scoring techniques and the K-means clustering approach without feature selection in Figure 12. The total LRS score is improving in the expression data set for the clustering accuracy rate. The variations are quite substantial when the feature points are under 400. The general trend is constant and the overall clustering accuracy is greater than that of other methods when the number of feature points is between 400 and 600. The LRS score exhibits more significant superiority than the other three techniques for normalized mutual information. The two indicators of the LRS score on the data set perform significantly better than other scoring methods when the number of selected features is under 300. When the number of elements is low, the clustering algorithm without feature screening performs significantly better than other algorithms. Figure 13 demonstrates that the improved semantic translation accuracy rate is maintained at 90-95% and the accuracy rate is raised by 10%-15% while the traditional translation accuracy rate is maintained between 80% and 85%.

DNN-EL has the highest accurate information rate when compared to the Improved GLR Algorithm, RFNet strategy, It may be difficult for the improved GLR algorithm to resolve syntactic and semantic difficulties clearly, which makes it difficult to extract precise and contextually relevant information. Additionally, the improved GLR algorithm often just evaluates statements in their immediate context, without taking conversation or larger context into account. It is often necessary to capture contextual dependencies during feature extraction for semantic translation, such as anaphora resolution, co-reference, or knowledge of discourse interactions. It may be difficult for the algorithm to

extract characteristics that effectively reflect these contextual subtleties due to its low context sensitivity. In contrast, RFNet was designed mainly for visual activities and may not have any innate language comprehension ability. Additionally, RFNet must analyze visual data in real time, which places demands on the efficiency and availability of processing resources.

5 Conclusions

This article mainly studies the feature extraction of English semantic translation. Through the construction of graph regularization knowledge, model construction, and the comparison of the three feature extraction methods, comparatively excellent feature extraction methods are compared, and popular regularization terms are constructed to analyze the graph regularization. At the same time, it explores the recognition pattern of the recognition algorithm, makes the most efficient English semantic translation method, and investigates the accuracy of the enhanced English semantic translation method. In the end, it is concluded that the accuracy of the improved English semantic translation is 10%-15% greater than the previous translation. This is excellent data in actual English translation, which can effectively enhance the semantic understanding of English translation. Taking up the difficulties of ambiguous language and words with numerous meanings. In the future, we may be able to use sophisticated natural language processing algorithms and semantic analysis to categorize words and sentences depending on their context.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest

Funding Statement

This study did not receive any funding in any form.

References

[1] Deng A, Kelmans A, Meng J . “Laplacian spectra of regular graph transformations”, Journal of Donghua University(English Edition), vol. 161, no.3,pp. 118-133,2017.

[2] Bitkina V V , Makhnev A A . “On Automorphisms of a Distance-Regular Graph with Intersection Array {125, 96, 1; 1, 48, 125”, Lobachevskii Journal of Mathematics, vol. 39,no. 3, pp. 458-463,2018.

[3] Kaveh A , Koohestani K . “Formation of graph models for regular finite element meshes”,Computer Assisted Mechanics & Engineering Sciences, vol. 16,no.2, pp. 101-115,2017.

[4] Naik M K, Panda R . “A novel adaptive cuckoo search algorithm for intrinsic discriminant analysis based face recognition”, Applied Soft Computing, vol. 38, no. C, pp. 661-675,2016.

[5] Sahoo S P, Ari S. "On an algorithm for human action recognition", Expert Systems with Application, vol. 115,no. JAN, pp. 524-534,2019.

[6] Qian H , Qiu J , D . Zhang, et al. “Research and application of semantic intelligent recognition algorithm for relay protection information”, Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control, vol. 46,no. 3, pp. 83-88,2018.

[7] Bracken J , Degani T , Eddington C , et al. “Translation semantic variability: How semantic relatedness affects learning of translation-ambiguous words”, Bilingualism Language & Cognition, vol. 20, no. 4, pp. págs. 783-794,2016.

[8] Tan Y W . “A Study of Legal Translation from the Perspective of Frame Semantics”, Overseas English, vol. 000,no. 002, pp. 199-200,2017.

[9] Nagar A K , Sriram S . “On Eccentric Connectivity Index of Eccentric Graph of Regular Dendrimer”, Mathematics in Computer Science, vol. 10,no. 2, pp. 229-237,2016.

[10] Chua C C, Lim T Y, Soon L K, et al. "Meaning preservation in Example-based Machine Translation with structural semantics", Expert Systems with Applications, vol. 78, no. JUL, pp. 242-258,2017.

[11] Jurko P . Pragmatic meaning in contrast: semantic prosodies of Slovene and English[J]. Perspectives: studies in translatology, 25(1): pp.157-176,2017.

[12] Zinszer B D , Anderson A J , Kang O , et al. “Semantic Structural Alignment of Neural Representational Spaces Enables Translation between English and Chinese Words, Journal of Cognitive Neuroscience, vol. 28,no. 11, pp. 1749-1759,2016.

[13] Novikova A V , Mylnikov L A . “Problems of machine translation of business texts from Russian into English”, Automatic Documentation & Mathematical Linguistics, vol. 51,no. 3, pp. 159-169,2017.

[14] Abudalbuh M . “The Acquisition of English Articles by Arabic L2-English learners: A Semantic Approach”, Arab World English Journal, vol. 7,no. 2, pp. 104-117,2016.

[15] Yang M , Liu S , Chen K , et al. A” Hierarchical Clustering Approach to Fuzzy Semantic Representation of Rare Words in Neural Machine Translation” IEEE Transactions on Fuzzy Systems, vol. 28,no. 5, pp. 992-1002,2020.

[16] Anible B . Iconicity in American Sign Language–English translation recognition[J]. Language and Cognition, 12(1): pp.138-163,2020.

[17] Dahan N A, Ba-Alwi F M . Extending a model for ontology-based Arabic-English machine translation[J]. International Journal of Artificial Intelligence & Applications, 2019, 10(01):

pp.55-67.

[18] Shu H . “A Model for English Translation of Chinese Classics”, Language and Semiotic Studies, vol. 4, no. 0, pp. 109-133,2018.

[19] Junhui L I, Zhu M, Wei L U , et al. "Improving Semantic Parsing with Enriched Synchronous Context-Free Grammars in Statistical Machine Translation", ACM transactions on Asian language information processing, vol. 16,no. 1, pp. 6.1-6.24,2017.

[20] Hartmann E C . “Discovering the Coulisses of Artistic Collaboration: A Genetic Reading of the English Translation of Saint-John Perse's Poem Amers”, Ilha Do Desterro A Journal of English Language Literatures in English & Cultural Studies, vol. 71,no. 2, pp. 153-164, 2018.

[21] G Stankeviiūt, Kasperaviien R , Horbaauskien J . “Issues in Machine TranslationA case of mobile apps in the Lithuanian and English language pair”, Nephron Clinical Practice, vol. 4,no. 1, pp. 75-88, 2017.

[22] Yoon-cheol, Park. “The Korean-English Translation of Proper Nouns in the Information Board of Cultural Properties”, The Journal of Modern British & American Language & Literature, vol. 35,no. 2, pp. 135-155, 2017.

[23] Nagyeong. “A Study on English Translation of the Muk'am Collection - Focusing on the Elements of literary style", The Journal of Translation Studies, vol. 18no. 1, pp. 65-93, 2017.

[24] Mondrzak R, Reinert C, Sandri A, et al. "Translation and cross-cultural adaptation of the Rating Scale for Countertransference (RSCT) to American English" Trends Psychiatry Psychother, vol. 38,no. 4, pp. 221-226, 2016.

[25] Copeland, Jack B . Prior, "Translational semantics, and the Barcan formula”, Synthese, vol. 193,no. 1, pp. 3507-3519, 2016.

[26] Yu, Q., 2018. Design of interactive English Chinese translation system based on feature extraction algorithm. Modern electronic technology, 41(4), pp.169-171.

[27] Jing, L., 2021. Research on translation methods based on the fusion of syntactic features. Electronic design engineering, 29(16), pp.153-157.

[28] Sang, Q., Huang, T., Tang, H. and Jiang, P., 2021. An improved non-rigid point set registration algorithm by preserving local topology. Pattern Recognition and Image Analysis, 31(4), pp.646-655.

[29] Stylianopoulos, K. and Koutroumbas, K., 2021. A probabilistic clustering approach for detecting linear structures in two-dimensional spaces. Pattern Recognition and Image Analysis, 31, pp.671-687.

[30] Currey, A. and Heafield, K., 2019, August. Incorporating source syntax into transformer-based neural machine translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers) (pp. 24-33).

[31] Dang, S.S. and Gong, X.T., 2020. Design of intelligent recognition English translation model based on improved GLR algorithm. Computer measurement and control, 28(4), pp.161-164.

[32] Maksimov, N.V., Golitsina, O.L., Monankov, K.V., Lebedev, A.A., Bal, N.A. and Kyurcheva, S.G., 2019. Semantic search tools based on ontological representations of documentary information. Automatic Documentation and Mathematical Linguistics, 53, pp.167-178.

[33] Sun, L., Yang, K., Hu, X., Hu, W. and Wang, K., 2020. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. IEEE Robotics and automation letters, 5(4), pp.5558-5565.

[34] Yu, Y., Si, X., Hu, C. and Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), pp.1235-1270.

[35] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan. 2019. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. *SIGKDD explorations : newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*.