Correction of English Translation Accuracy Based on Poisson Log-linear Model

ChengXiao Xiao¹,²,*
¹Public English Teaching and Research Department, Harbin Normal University
²School of Liberal Arts, Heilongjiang University
*Corresponding author e-mail: ChengxiaoXiao@163.com

Abstract. In the conventional machine translation methods, pipelined sequential operation is used to perform part-of-speech identification and syntactic analysis on the raw corpus to obtain the syntactic structure of English language, thereby reducing the iterative transfer error between translation tasks and the accuracy of structured instances, resulting in reduced accuracy in English language and literature translation. In this paper, a Poisson log-linear model that saves the corresponding bilingual corpus by means of Chinese-English dependency-tree-to-string is designed to implement dependent structured processing on the source language end and ensure that the accurate translation of English language is further proofread through the data-oriented translation model in the Chinese-English bilingual correspondence. The experimental results show that translation with high accuracy can be obtained based on the proposed method, which is highly accurate and stable.

Keywords: Translation Accuracy, Semantic Similarity, Correction Algorithm, Iterative Transfer

1. Introduction
Machine translation is the key to the natural language operation scope with a high application value [1-2]. The case-based machine translation is an empirical English language and literature translation strategy, which does not require complex deep-level grammar and semantic analysis. It can improve the efficiency of English language translation [3-4]. However, the instance-based machine-translation method has higher requirements on the quality of the instance library. Conventional machine translation methods use pipelined sequential operations to implement part-of-speech identification and syntactic analysis on the raw corpus to obtain the syntactic structure of the English language, which makes iterative transmission of errors between translation tasks and reduces the accuracy of structured
instances [5-6]. To solve this problem, we studied the method of machine translation accuracy in English language and literature, shaped and implemented a machine translation system based on the instances of Chinese-English dependency-tree-to-string in this paper, which improved the accuracy of English machine translation.

2. Poisson log-linear model

2.1. Dependency-tree-to-string model
The model of dependency-tree-to-string is <D, S, A>, where <D, S> is a translation pair, D represents the dependency tree of the source language, S represents the target word string of the source language, and A is used to describe the relationship between D and S. An instance of the word alignment relationship based on the dependency-tree-to-string bilingual alignment model is shown in Figure 1.

![Figure 1. Instance of the dependency-tree-to-string model](image)

The source tree dependency tree D at the top of the instance in Figure 1 includes words and part-of-speech features in each street. English under each word indicates the part-of-speech corresponding to the vocabulary, such as NN for nouns, VV for verbs, and JJ for adjectives. Lines in words are used to describe the dependencies between words. At the bottom of the instance, the English string sequence S corresponding to the Chinese sentence. The upper and lower dashed lines are used to describe the alignment relationship between Chinese word nodes and English words.

2.2. Lexical semantic similarity of poisson log-linear model
The range of similarity is [0,1], and the semantic similarity between different words $W_1$, $W_2$ is shown in formula (1):

$$ Sim_{semantic}(W_1, W_2) = \max_{j=1,2,…,n, j=1,2,…,m} Sim(S_{1j}, S_{2j}) $$ (1)
Where $S_{li}(i = 1,2,\cdots,n)$ and $S_{lj}(i = 1,2,\cdots,m)$ are used to describe the $n$ and $m$ concepts existing in the words $W_1$ and $W_2$. The semantic similarity of two words is the highest value for the similarity in their concepts.

The conceptual similarity of words can be described by the similarity of the meaning of the concept, and the similarity of the meanings of $p_1$ and $p_2$ is calculated using formula (2):

$$ Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} $$

(2)

Where $\alpha$ represents a controllable parameter; $d$ represents the path distance of two Yiyuan in the Yiyuan tree, and its value is non-negative.

The Poisson log-linear model uses a multi-feature thinking judgment model. For a set sentence $f_i^d = f_i^\prime, f_j^\prime, \cdots, f_j^\prime$, a translation $e_i^d = e_i^\prime, e_j^\prime, \cdots, e_j^\prime$ is formed. The maximum entropy translation model is as shown in equation (3):

$$ e_i^d = \sum_{m=1}^{M} \lambda_m h_m(e_i^d, f_i^d) $$

(3)

The Poisson log-linear model is highly scalable, can set corresponding features for different target requirements, and can apply a variety of linguistic methods to machine translation. Feature functions such as forward and backward translation probabilities and target language models are the main forms of machine translation systems. Based on the actual requirements of the translation system, a feature function and corresponding privilege weights are automatically set, and the optimal translation with the highest score on the generated translation is obtained according to formula (3).

2.3. Implementation of the machine translation system

The Sato & Nagao method is used to describe the dependency mechanism. The source language dependency tree of the dependency-tree-to-string-aligned instances is formalized. The matching description method is used to detect instance fragments in the instance database, and to obtain input sentences to achieve similar instance detection. Matching expressions can be replaced, filtered, or added in three ways. In the target word string without the dependency tree structure relationship layer, the corresponding translation expression changes accordingly. The following shows the structure of the source language dependency-tree-to-string instance D3 and instance D5 in the instance library:

- D3: [c2], [purchase], [c22, [she, PN]], [c23, [XX]]
- %% I bought XX
- S3: [e21, she]
- [e22, buy]
- [e23, XX]
- %% i buy XX
- A3 ([c21, e22], [c22, e21], [c23, e23])
%% c21↔e22, c22↔e21, c23↔e23
D5: [c51, [Book NN]], [e52, [一，CD]], [c53, [English NN]], [c54, this DT]
%% a political book
S5: [e51, a]
[e52, English]
[e53, book]
%% an English book
A5 ([c51, e53], [c52, e51], [c53, e52], [c54, e51])
%% c51↔e53, c52↔e51, c53↔e52, c54↔e51

In the instance database, identifiers such as e21 and e51 are labeled in word order, and the target word string S is labeled with “e” in the prescript. For the instance sentence “She bought an English book”, integrating source language instances D3 and D5 to detect and obtain [c21, [r, c23 [c51]]] is one of the corresponding expressions. The target translation of the input sentence obtained through the target matching expression is: I buy a politics book.

According to the Poisson log-linear model in this paper, the eigenfunctions used are:
1) Probability of THE forward and reverse translation. When the number of words is the same, there are more words in the same sentence between the translated sentence and the translated sentence instance. The applied feature function will produce a more accurate translation.
2) Language model. The quality of the generated translation is measured by this function, which improves the fluency of the translation. In this paper, the language model of the target language can be used to find the probability of translation fragments in the target language.

3. Experimental analysis

3.1. Experimental settings
The experimental corpus is the Chinese-English news corpus used in the official evaluation of CWMT 2018. About 420,000 pairs of English-Chinese parallel corpora were collected from it and used as the initial corpus of the bilingual instance database. The test set for the official evaluation of CWMT 2018 is used as the test set, and the experimental corpus is shown in Table 1.

3.2. Experimental results and analysis
In order to test the effectiveness of the system in this paper, the experiment is based on the corpus of Table 1. A comparative analysis of the translation results of the system, semantic language-based machine translation system, and open-source statistical machine translation system is shown in Table 2.

| Table 1. Experimental corpus situation |
|---------------------------------------|
| Language | Development set | Test set |
| Chinese  | 422 700         | 955 (including reference translation) |
| English  | 422 700         | —       |
The BLEU in Table 2 is a comparative analysis of n-unit fragments of the translation to be evaluated and the reference translation. The higher the number of matching fragments calculated, the better the quality of the translation to be evaluated. NIST is a measurement standard for translation quality assessment. It is used to evaluate the quality of translations per unit quantity. The higher the value, the better the translation quality. Analysis of Table 2 shows that the BLEU and NIST values of the system in this paper are higher than those of the other two systems, which indicates that the proposed machine translation system in this paper has better performance. Hence, it is an effective method for English language and literature translation.

The experiment collects partial translations from the translation results of the three translation systems and analyzes them. Table 3 shows the translations obtained from the test sentence “The information industry shows a rapid development trend”.

The translation differences between the three translation systems in terms of “fast development” are analyzed, as shown in Table 3. The open-source statistical machine translation system translates the translation as “fast change”, and the semantic language-based machine translation system translates the translation as “keeping the “Moment going” has a high deviation from the raw word and does not conform to the grammar and semantics of English. Although the translation results of the word in this system are inconsistent with the reference translation, the semantics meets the requirements and have high accuracy.

Table 2. Comparison results of translation systems

| Translation system                                      | BLEU values | NIST values |
|---------------------------------------------------------|-------------|-------------|
| Machine translation system based on semantic language   | 25.42       | 5.7672      |
| Open-source statistical machine translation system      | 25.18       | 5.7518      |
| Text system                                            | 25.53       | 5.7824      |

Table 3. Translations obtained by different translation systems

| Raw                                                                 | Information industry shows the rapid development |
|---------------------------------------------------------------------|---------------------------------------------------|
| Reference Translation                                               | The information industry is developing rapidly    |
| Machine translation system based on semantic language               | The information industry is keeping the momentum going |
| Open-source statistical machine translation system                  | The information industry is developing rapidly    |
| Text system                                                         | The information industry is a high-speed development situation |

Table 4 and Table 5 are the translation results of the system in this paper and the semantic language-based machine translation system from English to Chinese and from Chinese to English. The first column in the two tables is the average of the translation results of the two systems for each sentence. The average number of the systems in this paper is smaller than that of machine translation systems based on semantic languages, indicating that the system in this paper has fewer inaccurate results. The recall rates of accurate translations in the translation results of the second column in the two tables are analyzed, i.e., the proportion of accurate translations, it can be seen that the recall rate of the system in this paper is higher. Analysis of the 3rd and 4th columns in the two tables shows that the correct translation rate of the first and first two translation results of the system in this paper is 8-9 percentage points higher than the semantic language-based machine translation system and 11-13 percentage point. Comprehensive analysis of these results suggest that this system systematically improves the accuracy of translation results and has high English language and literature translation.
performance and stability.

**Table 4. English-Chinese translation results**

| Translation system                | Average translation result per sentence | Recall rate for accurate translation in results % | The first accurate translation rate in the results % | Accurate translation rate of the first two digits in the result % |
|-----------------------------------|----------------------------------------|---------------------------------------------|---------------------------------------------------|----------------------------------------------------------|
| Text system                       | 1.4                                    | 92                                          | 81                                                | 94                                                       |
| Machine translation system based on semantic language | 4                                      | 86                                          | 69                                                | 83                                                       |

**Table 5. Chinese-English translation results**

| Translation system                | Average translation result per sentence | Recall rate for accurate translation in results % | The first accurate translation rate in the results % | Accurate translation rate of the first two digits in the result % |
|-----------------------------------|----------------------------------------|---------------------------------------------|---------------------------------------------------|----------------------------------------------------------|
| Text system                       | 3.8                                    | 86                                          | 58                                                | 80                                                       |
| Machine translation system based on semantic language | 8                                      | 79                                          | 47                                                | 67                                                       |

4. Conclusions
In this paper, the machine translation accuracy methods in English language and literature are studied, and a machine translation system based on the Poisson log-linear model is created and implemented to accomplish the accurate translation of English language and literature.

Acknowledgement
1. Research Program of Philosophy and Social Science of Heilongjiang Province (19XWC153)
2. Harbin Normal University's First "Course Ideology and Politics" Teaching Reform Demonstration Project (KCSZ2019005)
3. Heilongjiang Province Economic and Social Development Key Research Project Foreign Language Special Key Project (WY2018010-A)

References
[1] Tracy Babb, Adrianna Gillman, Sijia Hao. An accelerated Poisson solver based on multidomain spectral discretization[J]. Bit Numerical Mathematics, 2016,3(1):1-29.
[2] Septimus Edward, Hickok Jason, Moody Julia. Closing the Translation Gap: Toolkit-based Implementation of Universal Decolonization in Adult Intensive Care Units Reduces Central Line – associated Bloodstream Infections in 95 Community Hospitals[J]. Clinical Infectious Diseases An Official Publication of the Infectious Diseases Society of America, 2016,2(2):112-120.
[3] Haiyang Yu, Rakel Kongsmo, Nilesh Patil. On determining the Poisson's ratio of viscoelastic polymer microparticles using a flat punch test[J]. International Journal of Mechanical Sciences, 2017, 128(2):150-158.
[4] Jörg Stiller. Nonuniformly Weighted Schwarz Smoothers for Spectral Element Multigrid[J]. Journal of Scientific Computing, 2017, 72(1):81-96.
[5] Ali Akbar Jafari, Rasool Roozegar. On Bivariate Generalized Exponential-Power Series Class of Distributions[J]. Journal of the Iranian Statistical Society, 2015, 17(1):1-11.

[6] Gaultier Lambert. Incomplete Determinantal Processes: From Random Matrix to Poisson Statistics[J]. Journal of Statistical Physics, 2016, 5(4):110-121.