CLC Number

UDC Available for reference Yes No



Undergraduate Thesis

**Thesis Title： Metamorphic Testing of Deep**

**Learning-Based NLP Models**

|  |  |
| --- | --- |
| **Student Name：** | **石晗峰** |
| **Student ID：** | **11510700** |
| **Department：** |  |
| **Program：** |  |
| **Thesis Advisor：** | **刘烨庞** |

Date: Month Day， Year

directory

[Metamorphic Testing of Deep Learning-Based NLP Models 1](#_Toc38009485)

[Abstract 1](#_Toc38009486)

[1 Introduction 1](#_Toc38009487)

[2 Testing system 3](#_Toc38009488)

[2.1 Mutation part 3](#_Toc38009489)

[a) Mutating sentences with regularity between sentences. 3](#_Toc38009490)

[b) Mutating sentences by change the active and passive voice 4](#_Toc38009491)

[c) Mutating sentences by mutating keywords of sentences with similar meaning words 4](#_Toc38009492)

[d) Mutating sentences by mutating all words with similar meaning words (if it has) 5](#_Toc38009493)

[2.2 Translation part 5](#_Toc38009494)

[2.3 Text similarity analysis 6](#_Toc38009495)

[a) Levenshtein Distance (Edit Distance)[4] 6](#_Toc38009496)

[b) Corpus-based and Knowledge-based Measures of Text Semantic Similarity[5] 6](#_Toc38009497)

[3 Evaluation 7](#_Toc38009498)

[4 Conclusion 9](#_Toc38009499)

[Reference 9](#_Toc38009500)

Metamorphic Testing of Deep Learning-Based NLP Models

[ABSTRACT]

Machine translate has been wildly used and be an important tool for the whole society. Then ensuring the accuracy of translation are becoming important. Normally the machine learning system are supported to be test with labeled data, which can also be used as training data. This article is supposed to find a way to test the accuracy of translation system with Unlabeled data, saving more labeled data for training.

[**Key words**]: metamorphic testing; machine translation

# 1 Introduction

In recent years, the machine learning has developed quickly, the technology products including translation systems with machine learning technical has been wildly used. As there are many different types of language in the whole world and learning a new kind of language are not an easy job, the machine translation systems played a catalytic role to the communication of people with different culture, promote the developing of Economy and Technology[1].

At that time, the reliability of machine translation software is of great importance. And inferior results will cause great mistake. However, the translation software is not as reliable as it wanted[2].

Normally a machine learning system is built with a dataset with labels. The larger the dataset is, the more accurate the system is. But to ensure the system’s accuracy, the tests are necessary. So, the dataset is normally separated as train data and test data. By input test data into the system and compare the result between labels of test data and output of the system, we get an accuracy. However, the data used for testing can also be used for training. If the methine translation system can be test with unlabeled data, then the labeled data can be saved as training data. According to research[3], the more training data there are, the more accurate the machine learning system are.

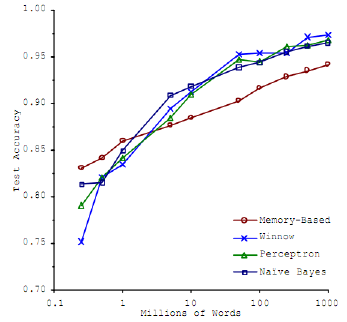


Figure 1 Learning Curves for Confusion Set Disambiguation[3]

This project aims to find a feasible method to test the Chinese-to-English translation system without labeled data. And if the project works, the testing data saved is supposed to be used for training.

The main idea of this project is that sentences with similar meaning have similar meaning translations, so if we can make one sentence become similar meaning sentence, then we can compare the translation of two sentences to ensure the accuracy of translation system. First input one sentence into the project, it mutates the original sentence into a similar meaning sentence, then inputs both the original sentence and similar meaning sentence into a tested translation system, gets two translated sentences. By comparing two translated sentences, we can get the accuracy of the translation system. Normally the two translated sentences have a similar meaning, so the more similar the meanings of the two sentences are, the more accurate the translation system. In this way the test data can be saved as train data, which may help improving accuracy of the translation system.

The project is combined with three parts: 1. Mutation part, which turns Chinese inputted sentences into similar sentences. 2. Translation parts, which stores the translation system needed to be tested. And 3. text similarity analysis, by comparing the translated sentences, calculate the accuracy of translation of each input.

# Testing system

## 2.1 Mutation part

This part mutates the Chinese sentences into similar meaning sentences.

### a) Mutating sentences with regularity between sentences.

Chinese sentences are very flexible. For example, we can omit “的“ after an adjective sometimes while talking, or subject can be omitted and just use verb and particle to finish the sentences. By using the pattern the sentences can be mutated effectively and will not affect the accuracy.

|  |  |  |
| --- | --- | --- |
|  | original sentence | mutated sentence |
| omitted“的” | 今天的天气真好 | 今天天气真好 |
| omitted subject | 你吃饭了吗？ | 吃了吗？ |
| change the order | 你的作业做了吗？ | 做了吗你的作业？ |
| add adverb | 他喜欢这个故事 | 他很喜欢这个故事 |
| others | 猫是一种很可爱的动物 | 猫是很可爱的动物 |

Table 1 Some regularity of Chinese sentences

But this method is not easy to operate. I consider a rule to add “的” between two nouns if the combination of two nouns also is a meaningful noun, like “今天” and “天气” and “今天天气”. But this occurs a problem that some words are not expected to be separated like “人民广场”. Besides, whether two words combine together can be a meaningful word is not a simple problem to solve. As for the omitted subject, it might be easy, but if the sentence does not have a subject, it is not an easy job to add a proper subject.

On the one hand, it is not simple to operate a mutation like this. The regularity between sentences does not always work. Some sentences may have no way to be mutated with the same rule. Or it can not be changed just with sentence structure. Unless the testing system can only use special test cases, or this mutation method is not a recommended method.

### b) Mutating sentences by change the active and passive voice

Modify the active and passive voice of the sentences does least change to meaning of sentences. And normally the sentences with “把” will have similar meaning sentences with “被”, like “我把黑板擦干净了” and “黑板被我擦干净了”. With the regulation of the structure of the sentences, we may do this job easily.

However, to analyzing the structure of the sentences, the sentences should be separated as words with different part of speech. Also, the speech of the combination of words should be able to be analyze, or we may change the original sentence to a semantically unsent sentence. And, not all sentences have a active and passive voice, this can also cause trouble. So, this method is considered as an alternative

### c) Mutating sentences by mutating keywords of sentences with similar meaning words

Levenshtein Distance[4] is a traditional way to calculate the similarity of two sentences. This method compares two sentences and calculates the distance the same words moved to calculate the similarity of them. Based on this method, scientists offer one way to calculate the semantic similarity by calculating the similarity of each word and finally get the similarity. This method inspired me that changing the keyword into a similar meaning word can only slightly change the whole meaning of the sentences. The more similar the two words are, the less change the meaning of the sentence is. Especially for some nouns, the mutation might make no difference because there are many different names for the same item in Chinese.

|  |  |
| --- | --- |
| **original sentence** | **expected mutated sentence** |
| 今天天气真好 | 今日天气真好 |
| 经检验，这些都是具有杀伤力的制式枪支及其配件 | 经检验，这些都是具有杀伤力的制式武器及其配件 |
| 他喜欢这个故事 | 他喜欢这个故事情节 |

Table 2 Expected mutated sentence

So, the method will word 1. segment original sentences into words, 2. choose the keywords, 3. find the synonyms of the keywords and 4. compare the similarity of the original keywords and their synonyms. Replace the one which has the largest similarity. This method asked for a correct segmentation and a dictionary that can offer the synonyms for one word and the similarity of them.

To achieve this method, I use python package ‘pkuseg’ to segment Chinese sentences, which have high accuracy and will separate “今天天气” into “今天” and “天气”，the former has no synonyms but the separated two have. And use the tool ‘synonyms’ package of python to get synonyms and its similarity of one word. By comparing the keywords (noun, verb, adjective) and the closest meaning synonyms of them, the mutation is ready to work.

### d) Mutating sentences by mutating all words with similar meaning words (if it has)

Normally the keyword in Chinese sentence have the synonyms and others do not have, so it is not necessary to find the keyword every time. Just use every word which has synonyms is all right. Then compare every word and its synonyms, if one word has synonyms and the similarity of these two words are larger than any other word pair, it is all right to replace this word even if it is not a keyword.

|  |  |
| --- | --- |
| original sentence | mutated sentence actually |
| 今天天气真好 | 现在天气真好 |
| 这是一个测试用例 | 这是一个试验用例 |
| 这种方法的可行性不高 | 这种方法的可行性不低 |

Table 3 Actually mutated sentence after mutation

Although the mutation part is logical works, there still being problems. 1. Word segmentation is inaccurate. 2. Synonyms are not well chosen. 3. Did not consider the effect of context. To decrease the problems, I built a database to store all the input sentences and offer the position of one mutated sentence. In this way, the problems can be decreased.

## 2.2 Translation part

The translation part is the module my system testing. In this article the translation part contains Google translator and Baidu translator.

As the google translator do not have an api, I use request module to send the sentence to google translate website to simulate google translator. In google translate page, there is one function is to find the mistake of input text and return back. Using this function properly can help fix the illogical sentences after mutated to logical sentences.

## 2.3 Text similarity analysis

This part will calculate the similarity of two sentences. And the larger the similarity is, the more accurate the translation system is.

### a) Levenshtein Distance (Edit Distance)[4]

Levenshtein Distance, proposed by Vladimir Levenshtein in 1965, is the minimum number of editing operations required to change from one string to the other. This method calculates the similarity of two sentences, but do not focus on the meaning of sentences. Sentences with the same meanings sometimes are quite different. But this way can get an approximate result.

### b) Corpus-based and Knowledge-based Measures of Text Semantic Similarity[5]

This method measuring the semantic similarity of texts using corpus-based[5] and knowledge-based[6] measures of similarity. Although this method did not consider the structure of texts, the results of this method is reasonable and do better than some traditional method.

#### i) Corpus-based Measures

This measurement trying to identify the similarity between words with information exclusively derived from large corpora. There are two metrics used:(pointwise mutual information and (2) latent semantic analysis. the basic idea of pointwise mutual information using data collected by information retrieval (PMI-IR)[6] is to count the possibility two words appear in the same text. The more possible the possibility is, the closer the two words are. The PMI calculate with:

Using NEAR query (one of the four different types of queries proposed by Turney) to calculate p:

then:

As the PMI-IRs of all words have same value Web Size, the value of Web Size will not affect the order for all word pair needed to calculate the similarity.

#### ii) Knowledge-based Measures

This measurement using information drawn from semantic networks to calculate the similarity of two words. The semantic networks are the network of relationship of different matters. For example, milk and coffee are all drink, then the relation of them are closer than mild and sofa. To achieve the calculation, this method considered six different metrics, including Jiang & Conrath, Resnik, etc.

To implement the Corpus-based and Knowledge-based Measures of Text Semantic Similarity, first get two text segments. And for each segment word, find the most similar word in the other text segments. Then combine the similarities and their corresponding specificity get the semantic similarity of the two sentences. Then get average value of two asymmetrical score.[7]

Finally, we get the similarity of two sentences. And the less similar the two sentences are, the less accurate the translation system is.

# Evaluation

The score of similarity of two translated sentences are between 0 to 1, so a threshold is needed. By setting the threshold to 0.8, which means while the translated sentences of original sentence and mutated sentence are similarity with score over 0.8, with 332 different Chinese sentences needed as original sentences, the accuracy of google translator is 94.85%. And the total work cost about 8 minutes to finish.

|  |  |  |  |
| --- | --- | --- | --- |
| **Original sentence** | **Original translation** | **Mutated sentence** | **Translated mutated sentences** |
| 象往年一样,数百名乡亲今天上午聚集在钟氏祠堂前的广场举行团拜,互致问候。 | Like, hundreds of folks gathered in the square in front of the Zhong's Ancestral Hall this morning to exchange greetings. | 象往年一样,数百名乡亲今天下午聚集在钟氏祠堂前的广场举行团拜,互致问候。 | Like in previous years, hundreds of folks gathered in the square in front of the Zhongshi Ancestral Hall this afternoon to hold a group worship and exchange greetings. |

|  |  |  |  |
| --- | --- | --- | --- |
| 历史的经验证明,能“打”方能言和,和必须有能“打”为基础。 | Historical experience proves that being able to "beat" can be peaceful, and that peace must be based on being able to "beat". | 历史的经验证明,能“打”方能言和,和必需有能“打”为基础。 | Historical experience has proven that being able to "play" can be peaceful, and that "playing" must be the basis. |
| 来自浙江大学的董石麟委员向工作人员索取了一批统计材料,准备带回驻地仔细研究。 | Dong Shilin, a member of Zhejiang University, asked the staff for a batch of statistical materials and planned to take them back to the site for careful study. | 来自浙江大学的董石麟委员向工作人员索取了一批统计材料,准备送回驻地仔细研究。 | Dong Shilin, a member of Zhejiang University, requested a batch of statistical materials from the staff, and was prepared to return to the station for careful study. |

Table 4 Translation that is not accurate enough

By analyzing the original sentence and mutated sentence and their translated sentences that do not match the threshold, the project works but still have many inaccurate results. For the two sentences have similar meaning but the structure have too much difference, the text similarity analyze might give a low similarity score.

|  |  |  |  |
| --- | --- | --- | --- |
| **Original sentence** | **Original translation** | **Mutated sentence** | **Translated Mutated sentences** |
| 今天天气真好 | The weather is so good today | 今天的天气真好 | Today's weather is great |
| 车灯熄灭后,吉普车就像一个小沙包,静静地卧在离发射架不远的地方。 | After the lights went out, the jeep was like a small sandbag lying quietly not far from the launch pad. | 车灯点燃后,吉普车就像一个小沙包,静静地卧在离发射架不远的地方。 | After the lights were lit, the jeep was like a small sandbag lying quietly not far from the launcher. |
| 西部大开发是我国经济的地区布局在战略上的大转变,其意义巨大而又深远。 | The development of the western region is a strategic change in the regional distribution of China's economy, and its significance is profound and far-reaching. | The great development of the east is a strategic change in the regional layout of China's economy, and its significance is profound and far-reaching. | According to a 1998 statistics, the number of married women of childbearing age in the Tibetan region was 51,400, and 36.29 percent had more than three babies. |

Table 5 Similar meaning sentence may be misjudged

Although getting the similarity of two translated sentences by the similarity of each words is a useful method, the structure can although affect the result. While the two sentences’ structure are quite different, there might be a mistake. Besides, the mutation can be not accurate sometimes. It changes the meaning of sentence while the segment of the proper noun into different words.

# Conclusion

# Reference

1. Pinjia, H. Clara, M and Zhendong, S. 2019. Structure-Invariant Testing for Machine Translation. arXiv preprint arXiv:1907.08710 [cs.SE]
2. Jie M. Zhang, Mark Harman, Lei Ma, and YangLiu.2019. Machine Learning Testing: Survey, Landscapes and Horizons. arXiv preprint arXiv:1906.10742
3. Michele, B. Eric, B. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics July 2001 Pages 26–33
4. E.S. Ristad, P.N. Yianilos, Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence V20N5, May1998, P522-532
5. R Mihalcea, C Corley, C Strapparava, Corpus-based and Knowledge-based Measures of Text Semantic Similarity[J]. Aaai, 2006 - aaai.org
6. Turney, P. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001).
7. Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.