# I Can Guess What You Mean: A Monolingual Query Enhancement for Machine Translation

Chenxi Pang[1,2], Hai Zhao[1,2,⋆] and Zhongyi Li[1,2]

[1] Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
{pangchenxi,rival2710}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

**Abstract.** We introduce a monolingual query method with additional webpage data to improve the translation quality for more and more official use requirement of statistical machine translation outputs. The motivation behind this method is that we can improve the readability of sentence once for all if we replace translation sentences with the most related sentences generated by human. Based on vector space representations for translated sentences, we perform a query on search engine for additional reference text data. Then we rank all translation sentences to make necessary replacement from the query results. Various vector representations for sentence, TFIDF, latent semantic indexing, and neural network word embedding, are conducted and the experimental results show an alternative solution to enhance the current machine translation with a performance improvement about 0.5 BLEU in French-to-English task and 0.7 BLEU in English-to-Chinese task.

## 1 Introduction

Research on statistical machine translation (SMT) has achieved remarkable progress on various ways [1–15]. Huge effort has been paid to improve quality and confidence of translation sentences to make them more similar to human translation. These works include automatic translation quality measure [16–18], preordering [19–21], neural network based SMT training and decoding [22–25]. However, most SMT system outputs can only serve as an assistant role for nearly all applications even though many translation sentences may be amazingly and natively accurate. For most cases, people can guess what the translation outputs mean, but feel hard to officially use them in any way.

⋆ Corresponding author. This paper was partially supported by Cai Yuanpei Program (CSC No. 201304490199 and No. 201304490171), National Natural Science Foundation of China (No. 61170114 and No. 61272248), National Basic Research Program of China (No. 2013CB329401), Major Basic Research Program of Shanghai Science and Technology Committee (No. 15JC1400103), Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04), and Key Project of National Society Science Foundation of China (No. 15-ZDA041).

Here is a translation example in Table 1. Without knowing any knowledge about the source language or source sentence, readers can easily guess the true meaning of the translation sentence despite it is not a perfect sentence and such a guess indeed matches the true meaning of the respective source sentence.

| | |
|---|---|
| source sentence | 我会在明天之前完成要求的任务 |
| machine translation | *I'll be tomorrow before the completion of the tasks required* |
| human translation | *I will finish the required task by tomorrow.* |

**Table 1.** A translation example

The example in Table 1 shows that there is still a gap between machine translation and human translation that results in machine translation sentence can not be officially used. At least, to our best knowledge, no reports are obtained to show that there is an application case for official use of SMT outputs, such as this paper itself being fully translated from its Chinese source only using an SMT system but without any human polishing Work (We cannot run such a risk!). It is even worse that the gap will still exist in the near future. As the current research on SMT is not likely to reach such an ideal aim to eliminate the gap, we propose to directly use human-generated sentences to replace those poorly translated ones. The motivation can be explained still from the above example in Table 1, if one can speculate the meaning of a sentence with improper word order or word usage, then it is quite possible to retrieve more accurate or authentic expression from a human-generated text dataset, only if it is large enough. Such a database will be referred to the relevant dataset hereafter in this paper.

In this work, we use the retrieval results from web search engine as search engine database is supposed to be the largest corpus that computational linguistician can ever find. However, our preliminary experiments show that even the most powerful search engine would fail to provide sufficient text data for every sentences in this task. Therefore we will limit our process only to a small number of machine translation sentences that are much more possible to find matching sentences in the retrieval results.

A lot of works are about post editing and even manual labor is used to produce better quality machine translations. Both rule-based and statistical automatic post editing methods have been proposed over the years [26–32], but most of them focus on evaluating a specific method and have a common trait that the reported methods are only suitable for very limited cases. There are two major differences between this work and all previous post-editing like methods. The first is that there is no any 'editing' operation inside our work, our approach just makes full replacement for a translated sentence if necessary. The second is that our method does not rely on the source side of machine translation and any specific language characteristics, which has been an obvious advantage other than all previous methods. In fact, all relevant data are automatically retrieved from web search engine. All languages are treated equally without discrimination so it has the potential to be generally used.

# 2 Our Model

Searching from the Internet for each translation sentence output by an SMT system, we look forward to finding a well-formed sentence that expresses the same or similar meaning. Usually, these translation sentences may be quite long and few retrieval results will return as searching all words inside them. Therefore instead of full long sentence replacing, we separate sentences that have few retrieval results into short ones and then carry out replacing based on these shorter sentences. All our later process will be based on translation sentences and these short sentences, which will be referred to *segments* hereafter.

After stop words are filtered out, all words in each sentence or segment will be separately put into search engine to collect returned webpages and a relevant sentence dataset $S$ will be obtained by putting all sentences inside webpages together. We also build a relevant segment dataset $SG$ by spliting all sentence in dataset $S$ into segments and let $W$ denote the set of all word types in $S$. Every sentence will be compared to the sentence in $S$ to find the most similar sentence for possible whole sentence replacement at first, and every segment in the sentence will be compared to the segment in $SG$ for possible segment replacement of the sentence if the previous process fails to achieve a reasonable result.

Vector space model (VSM) is a classical tool for representing text as a vector. For a sentence $s_i = w_1 w_2 ... w_m$ in $S$, its vector representation $v_i$ is:

$$v_i = (v_{i1}, v_{i2}, ..., v_{in}),$$

and each dimension $v_{ij}$ is related to a separate word. The value of $v_{ij}$ in the vector is non-zero if $w_i$ occurs in the segment. We consider a typical similarity or distance measures between vectors, Euclidean, as the following:

$$sim_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

There are multiple strategies to build sentence vector from word vectors. The following will give a list of vector representation formalizations for sentences and segments, including TFIDF, Latent Semantic Indexing and Neural network word embedding.

## 2.1 TFIDF

TFIDF, short for term frequency-inverse document frequency, is a numerical statistic to indicate how important a word is to a document in a collection or corpus. The TFIDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. The obvious shortcoming of TFIDF is the ignorance of word order information is ignored.

We use the following detailed formalizations to calculate TFIDF value of word $w_i$ in sentence $s_j$, where $count_{s_j}(w_i)$ is how many times word $w_i$ occurs in sentence $s_j$,

$|s_j|$ is the sentence length or the number of words inside sentence $s_j$, $|s|$ is the sentence size and $count_s(w_i)$ is the number of sentences that contains word $w_i$.

$$TF = \frac{count_{s_j}(w_i)}{|s_j|}, \tag{2}$$

$$IDF = \log(\frac{|s|}{count_s(w_i) + 1}), \tag{3}$$

$$TFIDF = TF \cdot IDF \tag{4}$$

For each sentence, we construct a one-hot vector with its TFIDF values. To make the representation more smoothing, we set an empirical threshold, TFIDF values below this threshold will be forced to set to zero and dismissed. Then we select part of remain words in the sentence to construct TFIDF vector of the sentence.

### 2.2 Latent Semantic Indexing

Latent semantic indexing (LSI) or latent semantic analysis (LSA) is an indexing and retrieval method that uses singular value decomposition (SVD) to identify patterns in the relationships between terms and concepts contained in an unstructured collection of text [33]. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of text by establishing associations between those words that occur in similar contexts.

The LSI model for this work will be based on our previous TFIDF vectors. The initial matrix for LSI is from all sentences of $S$ in which the $i$-th column is TFIDF vector of $i$-th sentence. Then SVD is used to obtain vectors with certain number of topics, which will be regarded as sentence vectors.

### 2.3 NN Word Embedding

In recent works, learning vector representations of words using neural network has been proved effective for various natural language processing tasks. Mikolov proposed two models, Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skip-gram), for computing continuous vector representations of words from very large data sets by using efficient NNs without hidden layer [34, 35]. After the model is trained, the word vectors are supposed to be mapped into a vector space, semantically similar words have similar vector representations and their word vector have a similar position in the vector space.

**Sentence Representation without Word Order Information** After we have vector representations for each word in sentence our task is to compare different sentence pairs, we need to consider how to combine word vectors for sentence representation and model sentence at last. The most intuitive method is to sum or calculate average of all the word vectors in the sentence [3].

$$v_s = \sum_{k=1}^{n} tfidf_{w_k} \cdot v_{w_k} \tag{5}$$

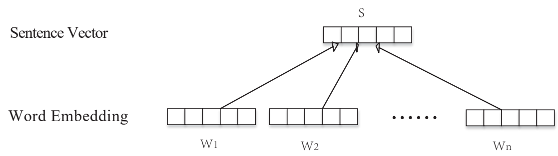


**Fig. 1.** A framework for learning sentence vector without word order information.

Figure 1 just illustrates such a simple strategy that utilizes the TFIDF weighted average of all word vectors as sentence vector, where $n$ is number of words in sentence, and $tfidf_{w_k}$ is the corresponding TFIDF values.

Note that all the above process can be either applied to sentences or segments so that we can obtain vector representations for both. Once sentence or segment vector has been computed, their similarity or distance can be directly computed through predefined measures.

**Segment Vector with Sentence Context** Besides the above word order free integration from word vectors, we still consider an effective method that may introduce useful word order information for sentence vector building. The motivation is simple, as machine translation system outputs sentences according to target language model constraints, the original word order in the translated sentences should still make sense to some extent. Therefore our exact query purpose is to find a human-generated sentence that has the most word overlapping and least word order change compared to the original machine translation sentence.

Le and Mikolov [37] proposed an unsupervised learning algorithm that learns vector representations for variable length pieces of texts by adding paragraph vector as an additional feature for the word embedding learning framework, which has shown effective in text classification and sentiment analysis tasks. The paragraph vector and word vectors are averaged or concatenated to predict the next word given many contexts sampled from the paragraph.

Figure 2 illustrates Le and Mikolov's method, which is adopted for our task. The sentence vector will be integrated into the word embedding learning. Every sentence $sen_j$ is mapped to unique vector, represented by a column in a matrix $M$. The column

[3] We are aware that there are many other effective method such as [36] who used a parse tree and matrix-vector operations to retain word order information. However, this work is about machine translation sentence processing, we need robust and simple strategy to handle various possible defective sentences.

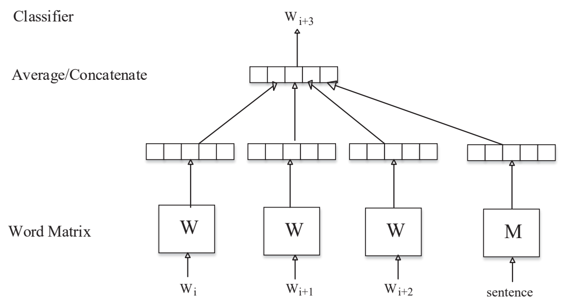


**Fig. 2.** Word embedding incorporated with sentence vector.

is indexed by the position of the sentence in the relevant sentence dataset $S$. Every word is also mapped to a unique vector, represented by a column in a matrix $W$. The column is indexed by position of the word in the vocabulary. The sentence vector and word vectors are averaged or concatenated to predict the next word given many contexts sampled from the sentence.

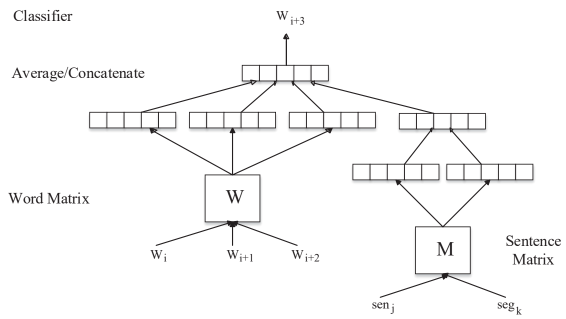


**Fig. 3.** Integrating sentence vector for learning segment vector. The input words $(w_i, w_{i+1}, w_{i+2})$ are mapped to columns of the matrix $W$ and input sentence $(sen_j)$ and segment $(seg_k)$ are mapped to vectors via matrix $M$. The segment vector acts as a memory of what missing from the current context and the sentence vector acts as a memory that remembers what missing from other segments.

However, while applying this method to segments we still lose the contextual information of other segments in the same sentence. Therefore we propose a combinational approach to help ease such a drawback by adding a sentence vector for segment vector learning as shown in Figure 3.

Segment and sentence vectors are weighted averaged to a context vector to contribute to the prediction task of the next word:

$$v = \alpha v_{seg_k} + (1 - \alpha) v_{sen_j} \tag{6}$$

where $v_{seg_k}$ is segment vector of the segment $seg_k$ that input words are located and $v_{sen_j}$ is sentence vector of the sentence $sen_j$ that the input segment $seg_k$ are located. The segment-sentence matrix $M$ is shared across all segments and sentence in $SG$ and $S$, and the word vector matrix $W$ is shared across all words in $W$. Sentence, segment and word vectors are trained using stochastic gradient descent and the gradient is obtained through backpropagation [38].

# 3 Experiments

## 3.1 Experimental Settings

We conduct experiments on the French-to-English translation task. The baselines of the IWSLT2014 evaluation campaign are followed. The dataset with $186k$ sentence pairs are used to train a phrase-based MT system, and dev2010 and tst2010 are selected as development data and evaluation data. We also conduct experiments on the English-to-Chinese machine translation tasks. MultiUN parallel corpus [39, 40] with $200k$ sentence pairs is used to train a phrase-based MT system. We run GIZA++ [41] on the training corpus in both directions [42] to obtain the word alignment for each sentence pair and MERT [43] for tuning on the development data. Using the SRILM Toolkits [44] with interpolated Kneser-Ney smoothing, we train a 5-gram language model for the tasks on the target side of its training data. In our experiments, the translation performances are measured by case-sensitive BLEU4 metric with a single reference.

There are many methods to separate sentence to segments (short sentences) [4], We adopt a simply and straightforward strategy by splitting sentence into segments according to comma appearance.

We retrieve 500 records at most for each sentence and 1000 records at most for each segment from two search engines, Google and Baidu. Then we remove all sentences that do not contain any word in the sentence. Our preliminary experiments indicate that if the Euclidean distance of two 100-dimension sentence vectors is larger than 2, then these two sentences are nearly irrelevant. So after filtering out all these sentences, we put all the rest sentences together as our relevant sentence dataset and split all sentences into segments as our relevant segment dataset for later process. There are many ways of word segmentation [47] and two different ways are used in English-to-Chinese task for these Chinese sentence corpus and datasets, one is done by Stanford Word Segmenter [48, 49] and another is done by trivially splitting each character into single-character words.

We construct a 20 dimension vector for both TFIDF and LSI model. Word embedding is done on both English and Chinese Wikipedia corpus via CBOW model. 5 previous words and 5 next words are used as context and corresponding vectors are concatenated to predict the next word. Each word is projected to a 100-dimension vector and all sentences and segments are also mapped to a 100-dimension vector.

[4] A sophisticated approach is cutting sentence into several relative independent parts according to parse tree of sentence [45, 46], which can be regarded as a further improvement over the current simple segmentation strategy.

In the replacement process, we use the following formalization:

$$P_{sen} = \log(\prod_{i=1}^{t} p(w_i | w_{i-4} w_{i-3} w_{i-2} w_{i-1})) \tag{7}$$

to calculate the 5-gram value of all the translation sentences, where $p(w_i|w_{i-4}...w_{i-1})$ is the conditional probability of $w_i$ given previous four words. Intuitively, the higher the 5-gram language model score is, the better the translation sentence is. Therefore we only focus on those sentences with language model scores lower than -1.3 (it is empirically determined according to BLEU scores over development set) and replacement operations will be only done over them.

### 3.2 Results

The experiments are done as we measure the similarity between sentences and then proceed replacement. For Euclidean distance, we set a threshold value of sentence similarity from 0 to 2 with interval 0.5 and sentences will be adopted only if the similarity is lower than the threshold. Too high Euclidean distance threshold will let too many sentences replaced, even some already very good sentences, while too low Euclidean distance threshold will result in no replacement of sentence. For Euclidean distance, the smaller the distance is, the higher similarity between sentences is. The results are given in Table 2.

| Word Vector | Sentence Vector | French-to-English | | | | | English-to-Chinese | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 0.5 | 1.0 | 1.5 | 2.0 | 0 | 0.5 | 1.0 | 1.5 | 2.0 |
| TFIDF | | 31.82 | **31.89** | 30.68 | 27.58 | 26.51 | 26.46 | **26.51** | 25.38 | 21.77 | 18.46 |
| LSI | | 31.82 | **31.93** | 30.77 | 27.72 | 26.59 | 26.46 | **26.59** | 25.41 | 22.63 | 18.65 |
| WE | TFIDF-Weight | 31.82 | **32.22** | 31.91 | 29.07 | 28.26 | 26.46 | **26.85** | 26.52 | 23.81 | 21.46 |
| WE | SEG-SEN | 31.82 | **32.34** | 32.08 | 30.25 | 28.63 | 26.46 | **27.03** | 26.85 | 23.89 | 22.05 |

**Table 2.** BLEU scores in terms of Euclidean distance, where WE and SEG-SEN correspond to word-embedding and segment-sentence respectively.

We compare the results from different vector representations with the baseline in terms of BLEU scores. It is well known that the translation quality of MT system highly depends on the language model and MT systems could generate very good translations if there is a strong LM. Original translation sentence have already preserve enough semantic and word order information. However in our TFIDF model, we only focus on word occurrence and ignore all of these information. As shown in Table 3, when we try to remain some useful semantic information in our LSI model, we can improve translation quality slightly. In our word embedding model, we suppose to preserve both semantic and word order information that contained in translation sentence by adding sentence into our word embedding framework.

| Word Vector | Sentence Vector | French-to-English | English-to-Chinese |
|---|---|---|---|
| TFIDF | | **31.89** | **26.51** |
| LSI | | **31.93** | **26.59** |
| Word Embedding | TFIDF-Weight | **32.22** | **26.85** |
| Word Embedding | Sengment-Sentence | **32.34** | **27.03** |
| Baseline | | **31.82** | **26.46** |

**Table 3.** Comparisions among different similarity calculation methods. BLEU score is the best result of corresponding method.

Our experiments indicate that our word embedding models outperform traditional model and achieve a better translation quality. Word embedding with both segment and sentence is better than sentence representation without word order information as it supposed to remain word and segment order information. An improvement in BLEU score also proves that our model successes in preserving semantic and order information of translation sentence and the refined sentences preserve the same meaning of the original one from another aspect.

| Word Vector | Sentence Vector | Standford-Word-Segmenter | Single-Character-Segmenter |
|---|---|---|---|
| TFIDF | | **26.51** | **26.33** |
| LSI | | **26.59** | **26.47** |
| Word Embedding | TFIDF-Weight | **26.85** | **26.89** |
| Word Embedding | Sengment-Sentence | **27.03** | **27.13** |
| Baseline | | **26.46** | |

**Table 4.** Comparisions among different word segmentation methods. BLEU score is the best result of corresponding method with Euclidean distance.

We also investigate the impact of different Chinese word segmentation over the SMT performance. Table 4 indicates that single-character segmentation slightly outperforms Stanford Word Segmenter and achieve about 0.1 BLEU score improvement in word embedding with segment-sentence method. Here, single character segmentation means just trivially segment each Chinese character as a word without truly considering they are words.

The improvement over BLEU scores indicates that the corresponding sentence replacement is helpful, and the returned sentences have successfully approached the meaning of those corresponding translation sentences as BLEU score improvement has demonstrated that the returned sentences are more closed to the reference sentence than the respective translation sentence.

### 3.3 Discussion

Our experiment has processed about 13% and 15% translation sentences in French-to-English and English-to-Chinese experiment, about 9% and 10% among them are originally the same as human translation sentences, the reference sentence is exactly the same as the returned results by search engines. However, we have set a good start, for all these processed sentences, they can be officially used in any way only if we carefully limit the processed sentences onto a small range to guarantee the accuracy. With the process of the proposed method, we have automatically put all machine translation sentences into two parts, one part is still from SMT system outputs, but the other part is right from human translation. If we can access larger and larger relevant dataset such as much more webpages crawled by search engines, then we can enlarge the portion of human translation part more and more.

The above has already shown that our process brings about BLEU improvement, but our work is more than the higher score, as we introduce human-generated sentences to replace those poorly machine translations. As well known, BLEU score is more reliable a metric only for lower quality translation matching. Effective and accurate translation for one source sentence can be multiple. We observe all non-reference-matching replacements by our model, they are exactly accurate translation for the corresponding source sentences. However, BLEU scores for these part are not 100% as it simply gives matching rate on $n$-grams but not semantic equivalence.

| | French-to-English | English-to-Chinese |
|---|---|---|
| source sentence | *la plupart sont complètement ignorés par notre ' moi du souvenir '.* | *Namibia will deal with perpetrators of terrorist acts according to the ordinary criminal law.* |
| translation sentence | most are completely ignored by our ' me the memory . ' | 纳米比亚 将 <u>涉及 普通 刑法 的</u> 恐怖 行为 的 实施者 。 |
| reference sentence | most of them are completely ignored by the remembering self . | 纳米比亚 将 根据 一般 刑法 处理 从事 恐怖 行为者 。 |
| retrieval sentence | most of them are completely ignored by the remembering self . | 纳米比亚 将 根据 一般 刑法 处理 从事 恐怖 行为者 。 |

**Table 5.** A-hole-in-one example. Underlines indicates words that are improperly translated. For words in translation and retrieval sentences, the same color and font indicate that the translation is nearly exact as the alignment part in reference sentence, while the same color but different font indicate closed but inaccurate meaning that are translated.

The following shows a series examples how our model replaces and improves those machine translation sentences. Sometimes, our methods are lucky enough as the example in Table 5. The search engine just returns exactly the same sentence as the reference sentence, which has been shown 2/3 replacements are so lucky. Of course, this also shows that both the selected search engine and our query methods work effectively so that they can jointly return the right reference sentences.

| | French-to-English | English-to-Chinese |
|---|---|---|
| source sentence | *c'est très difficile d'évaluer correctement son bien-être . j'espère vous avoir montré combien cela est difficile .* | *the production of sanitary napkins, another basic reproductive health commodity, is also hampered by restrictions on imports of raw materials.* |
| translation sentence | it's very difficult to evaluate properly its well-being . I hope I've shown you how much this is hard . | 另 一 个 基 本 生 殖 保 健 商 品 ， 也 受到 限制 ， 卫生 napkins 的 生产 原料 进口 。 |
| reference sentence | it is very difficult to think straight about well-being , and I hope I have given you a sense of how difficult it is . | 另外 一 种 基本 的 生殖 健康 商品 ， 即 卫生巾 的 生产 也 由于 原料 的 进口 受到 限制 而 遭到 损害 。 |
| retrieval sentence | it's very difficult to *evaluate its well-being properly* . I hope I've *shown you how hard it is.* | 另 一 种 基本 生殖 保健 用品 卫生巾 的 生产 也 由于 对 原材料 进口 的 限制 而 受到 影响 。 |

**Table 6.** A standard replace, similar but different, where OOV words are well handled.

However, it is impossible to always retrieve reference sentence for every machine translation sentence. Many times, our methods just come up with a sentence quite similar with translation sentence. Table 6 demonstrates such a case. It is actually very hard for readers to understand the meaning of machine translation sentence which suffers from unrecognized word order and out-of-vocabulary word (OOV) problem.

| | | |
|---|---|---|
| source sentence | *presque toutes les techniques pour produire aujourd'hui de l'électricité , en dehors desénergies renouvelables et du nucléaire , rejettent du CO2 .* | *Cuba has been engaged in a process of institutional and economic reforms for almost 10 years now.* |
| translation sentence | *almost all the techniques to* produce today of electricity , *outside of renewable energies and nuclear , dismiss CO2 .* | 近 10 年 来 ， 古巴 一直 在 体制 的 进程 和 经济 改革 。 |
| reference sentence | almost every way we make electricity today , except for the emerging renewables and nuclear , puts out CO2 . | 古巴 进行 体制 改革 和 经济 改革 迄今 已 差不多 十 年 。 |
| retrieval sentence | *almost all the techniques to produce electricity today* , except renewable energies and nuclear , *dismiss CO2 .* | 近 10 年 来 ， 古巴 一直 在 实行 体制 改革 和 经济 的 改革 。 |

**Table 7.** An example about incomplete translation.

The proposed method is an approximate search, so it is quite robust for such malfunctions inside machine translation sentences. The result shown in Table 6 just demonstrates that OOV word has been perfectly handled according to the query-replacement process in an English-to-Chinese translation example.

Readers can guess the general meaning of the translation sentence from words contained in the sentence, but can not understand its full true meaning or convince that this is a complete sentence. However, this difficulty can be partially solved by our methods and Table 7 shows such a translation example. Our methods find a segment from our relevant segment dataset that is semantically similar with the second segment of translation sentence and also a segment that is the same as the first translation sentence segment. Even through our retrieval sentence is not completely the same with reference sentence, but it has been easy enough to let readers full understand the sentence. In fact, the returned sentence is semantically consistent with the source sentence.

| | |
|---|---|
| source sentence | *Samuel Zbogar, Secretary of State, Ministry of Foreign Affairs of Slovenia* |
| translation sentence | Samuel Zbogar 外交部 国务 秘书 斯洛文尼亚 |
| reference sentence | 斯洛文尼亚 外交部 国务 秘书 塞缪尔・日博加尔 |
| retrieval sentence | 斯洛文尼亚 外交部 国务 秘书 Samuel Zbogar |

**Table 8.** An example for word reording.

We also find that our methods perform well in some English-to-Chinese translation sentences that need reordering. As shown in Table 8, the only difference between machine translation sentence and reference sentence is word order. Though failing to make the foreign name translation, our retrieval sentence successfully recovers the right word order for target language.

## 4 Conclusion

We have proposed a simple and effective method to enhance machine translation by replacing them with human translation sentences and expect to make them available for later official use. Our relevant sentences are queried from search engine and TFIDF, LSI, NN word embedding and specially designed segment vectors are used to calculate the similarity between sentences. The results show that our approach indeed gives better translation performance. In addition, this work shows a convenient start to improve the quality of machine translation by the roots.

## References


[1] Huang, Shujian and Chen, Huadong and Dai, Xin-Yu and Chen, Jiajun: Non-linear Learning for Statistical Machine Translation. ACL, 825–835, (2015)

[2] Yu, Heng and Zhu, Xuan: Recurrent Neural Network based Rule Sequence Model for Statistical Machine Translation. ACL, 132–138, (2015)

[3] Lu, Shixiang and Chen, Zhenbiao and Xu, Bo: Learning New Semi-Supervised Deep Auto-encoder Features for Statistical Machine Translation. ACL, 122–132, (2014)

[4] Xiong, Deyi and Zhang, Min: A Sense-Based Translation Model for Statistical Machine Translation. ACL, 1459–1469, (2014)

[5] Neubig, Graham and Duh, Kevin: On the Elements of an Accurate Tree-to-String Machine Translation System. ACL, 143–149, (2014)

[6] Riezler, Stefan and Simianer, Patrick and Haas, Carolin: Response-based Learning for Grounded Machine Translation. ACL, 881–891, (2014)

[7] Wang, Rui and Zhao, Hai and Lu, Bao-Liang: Bilingual continuous-space language model growing for statistical machine translation. IEEE, 1209–1220, (2015)

[8] Zhang, Jingyi and Utiyama, Masao and Sumita, Eiichro and Zhao, Hai: Learning local word reorderings for hierarchical phrase-based statistical machine translation. MT, 1–18, 2016

[9] Wang, Rui and Utiyama, Masao and Goto, Isao and Sumita, Eiichiro and Zhao, Hai and Lu, Bao-Liang: Converting Continuous-Space Language Models into N-gram Language Models with Efficient Bilingual Pruning for Statistical Machine Translation. ACM,2016.

[10] Wang, Rui and Zhao, Hai and Ploux, Sabine and Lu, Bao-Liang and Utiyama, Masao: A Bilingual Graph-Based Semantic Model for Statistical Machine Translation.

[11] Zang, Shuo and Zhao, Hai and Wu, Chunyang and Wang, Rui: A novel word reordering method for statistical machine translation. FSKD, 843–848, 2015

[12] Zhang, Jingyi and Utiyama, Masao and Sumita, Eiichro and Zhao, Hai: Learning word reorderings for hierarchical phrase-based statistical machine translation. ACL-IJCNLP, 542–548, 2015

[13] Wang, Rui and Zhao, Hai and Lu, Bao-Liang and Utiyama, Masao and Sumita, Eiichiro: Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation. EMNLP, 189–195, 2014

[14] Zhang, Jingyi and Utiyama, Masao and Sumita, Eiichiro and Zhao, Hai: Learning Hierarchical Translation Spans. EMNLP, 183–188, 2014

[15] Wang, Rui and Utiyama, Masao and Goto, Isao and Sumita, Eiichiro and Zhao, Hai and Lu, Bao-Liang: Converting Continuous-Space Language Models into N-Gram Language Models for Statistical Machine Translation. EMNLP, 845–850, 2013

[16] Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation. ACL,311–318, (2002)

[17] Guzmán, Francisco and Joty, Shafiq and Màrquez, Lluís and Nakov, Preslav: Pairwise Neural Machine Translation Evaluation. ACL,805–814, (2015)

[18] Graham, Yvette: Improving Evaluation of Machine Translation Quality Estimation. ACL,1804–1813, (2015)

[19] Miceli-Barone, Antonio Valerio and Attardi, Giuseppe: Non-projective Dependency-based Pre-Reordering with Recurrent Neural Network for Machine Translation. ACL,846–856,(2015)

[20] Zhang, Jingyi and Utiyama, Masao and Sumita, Eiichro and Zhao, Hai: Learning Word Reorderings for Hierarchical Phrase-based Statistical Machine Translation. ACL,542–548, (2015)

[21] Nakagawa, Tetsuji: Efficient Top-Down BTG Parsing for Machine Translation Preordering. ACL,208–218, (2015)

[22] Bengio,Yoshua, Réjean Ducharme, Pascal Vincent and Christian Jauvin: A neural probabilistic language model. JMLR,1137-1155, (2003)

[23] Vaswani, Ashish and Zhao, Yinggong and Fossum, Victoria and Chiang, David: Decoding with Large-Scale Neural Language Models Improves Translation. EMNLP,1387–1392, (2013)

[24] Devlin, Jacob and Zbib, Rabih and Huang, Zhongqiang and Lamar, Thomas and Schwartz, Richard and Makhoul, John: Fast and Robust Neural Network Joint Models for Statistical Machine Translation. ACL,1370–1380, (2014)

[25] Liu, Shujie and Yang, Nan and Li, Mu and Zhou, Ming: A Recursive Recurrent Neural Network for Statistical Machine Translation. ACL,1491–1500, (2014)

[26] Simard, Michel and Ueffing, Nicola and Isabelle, Pierre and Kuhn, Roland: Rule-based Translation with Statistical Phrase-based Post-editing. SMT,203–206, (2007)
[27] Dugast, Loïc and Senellart, Jean and Koehn, Philipp: Statistical Post-editing on SYSTRAN's Rule-based Translation System. SMT,220–223, (2007)
[28] Simard, Michel and Goutte, Cyril and Isabelle, Pierre: Statistical Phrase-based Post-editing. NAACL,508–515, (2007)
[29] Isabelle,Pierre and Goutte, Cyril and Simard, Michel: Domain Adaptation of MT Systems through Automatic Post-editing. MT Summit,255-261, (2007)
[30] Lagarda, A-L and Alabau, Vicente and Casacuberta, Francisco and Silva, Roberto and Diaz-de-Liano, Enrique: Statistical post-editing of a rule-based machine translation system. ACL, 217–220, (2009)
[31] Béchara, Hanna and Ma, Yanjun and van Genabith, Josef: Statistical post-editing for a statistical MT system. MT Summit,308–315, (2011)
[32] Huang, Zhongqiang and Devlin, Jacob and Matsoukas, Spyros: BBN's Systems for the Chinese-English Sub-task of the NTCIR-10 PatentMT Evaluation.
[33] Deerwester,Scott Susan T.Dumals, George W.Furnas, Thomas K.Landauer and Richard Harshman: Indexing by Laten Semantic Analysis. JASIS,391–407, (1990)
[34] Mikolov Tomas , Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781, (2013)
[35] Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff: Distributed Representations of Words and Phrases and their Compositionality. NIPS,3111–3119, (2013)
[36] Huang, Eric and Socher, Richard and Manning, Christopher and Ng, Andrew: Improving Word Representations via Global Context and Multiple Word Prototypes. ACL,873–882, (2012)
[37] Le, Quoc V. and Mikolov Tomas: Distributed Representations of Sentences and Documents. Eprint Arxiv,1188-1196,(2014)
[38] Werbos, Paul J: Backpropagation through time: what it does and how to do it. IEEE,1550–1560, (1990)
[39] Andreas Eisele and Yu Chen: MultiUN: A Multilingual Corpus from United Nation Documents. LREC,2868-2872, (2010)
[40] Jörg Tiedemann: Parallel Data, Tools and Interfaces in OPUS. LREC, (2012)
[41] Franz Josef Och and Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. CL, 19–51, (2003)
[42] Koehn,Philipp Franz Josef Och, and Daniel Marcu: Statistical phrase-based translation. NAACL,127–133, (2003)
[43] Och, Franz Josef: Minimum error rate training in statistical machine translation. ACL,701-711, (2003)
[44] Stolcke, Andreas: SRILM – An extensible language modeling toolkit. ICSLP,901–904, (2002)
[45] Soricut, Radu and Marcu, Daniel: Sentence level discourse parsing using syntactic and lexical information. ACL, 149–156, (2003)
[46] Zhang, Zhisong and Zhao, Hai and Qin, Lianhui: Probabilistic Graph-based Dependency Parsing with Convolutional Neural Network ACL, 1382–1392, 2016
[47] Cai, Deng and Zhao, Hai: Neural Word Segmentation Learning for Chinese. ACL, 409–420, 2016
[48] Tseng, Huihsin Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning: A Conditional Random Field Word Segmenter. SIGHAN, (2005)
[49] Chang, Pi-Chuan and Galley, Michel and Manning, Christopher: Optimizing Chinese Word Segmentation for Machine Translation Performance. WMT,224–232, (2008)