Retrieve Synonymous keywords for Frequent Queries in Sponsored Search in a Data Augmentation Way

Yijiang Lian
lianyijiang@baidu.com
Baidu

Zhenjun You
youzhenjun@baidu.com
Baidu

Fan Wu
wufan_aais@pku.edu.cn
Peking University

Wenqiang Liu
liwenqiang01@baidu.com
Baidu

Jing Jia
jiaying01@baidu.com
Baidu

ABSTRACT
In sponsored search, retrieving synonymous keywords is of great importance for accurately targeted advertising. The semantic gap between queries and keywords and the extremely high precision requirements (>= 95%) are two major challenges to this task. To the best of our knowledge, the problem has not been openly discussed. In an industrial sponsored search system, the retrieved keywords for frequent queries are usually done ahead of time and stored in a lookup table. Considering these results as a seed dataset, we propose a data-augmentation-like framework to improve the synonymous retrieval performance for these frequent queries. This framework comprises two steps: translation-based retrieval and discriminant-based filtering. Firstly, we devise a Trie-based translation model to make a data increment. In this phase, a Bag-of-Core-Words trick is conducted, which increased the data increment’s volume 4.2 times while keeping the original precision. Then we use a BERT-based discriminant model to filter out nonsynonymous pairs, which exceeds the traditional feature-driven GBDT model with 11% absolute AUC improvement. This method has been successfully applied to Baidu’s sponsored search system, which has yielded a significant improvement in revenue. In addition, a commercial Chinese dataset containing 500K synonymous pairs with a precision of 95% is released to the public for paraphrase study.

KEYWORDS
Sponsored search, keyword matching, keyword retrieval, synonymous keywords retrieval, paraphrase generation, paraphrase identification

1 INTRODUCTION
Sponsored search is one of the most major forms of online advertising and also the main source of revenue for most search engine companies. In sponsored search, there are three distinct roles involved: advertiser, user and search engine. Each advertiser submits ads (abbreviated for advertisements) and bids on a list of relevant keywords for each ad. To avoid ambiguity, keywords in this paper is particularly used to denote queries purchased by advertisers. When the search engine receives a query submitted by a user, it firstly retrieves a set of matched keywords. Then an auction is carried out to rank all corresponding ads, taking both the quality and bid price of each ad into account. Finally, the winning ads are presented on the search result page to the user.

Major search engine companies provide a structured bidding language, with which the advertisers can specify how would their purchased keywords be matched to the online queries. In general, 3 match types are supported: exact, phrase, broad. In the early days, exact match requires that query and keyword are exactly the same. Since queries are ever-changing, advertisers usually have to come up with a lot of synonymous forms of their keywords to capture more similar query flows. To ease their burden, modern search engines relax the exact match’s matching requirement to the synonymous level 2, which means under exact match type, the ad would be eligible to appear when a user searches for the specific keyword or its synonymous variants. For example, the keyword how much is iPhone 11 would not only be matched to the identical query but also be matched to other queries like the price of iPhone 11. In phrase match type, the matched queries should include the keyword or the synonymous variants of the keyword. Broad match type further relaxes the matching requirements to the semantic relevance level.

The highly accurate matching makes exact match type greatly welcomed by most customers, and nowadays it still occupies a great portion of the keywords revenue for most search engine companies. In this paper, we focus on the synonymous keywords matching problem under the exact match type. To make it clear, for a given query and a keyword repository (which is a snapshot of all the purchased keywords), we want to retrieve as more synonymous keywords as possible, while keeping a high precision. (On the one hand, retrieving more qualified synonymous keywords can provide a more competitive advertisement queue for the downstream auction system; on the other hand, from the point of advertisers’ fairness, each advertiser’s synonymous keywords should be released.)

The vocabulary mismatch between queries and keywords and the high precision required in the commercial product are two major challenges. Besides, since the volume of keywords and queries might reach billions, how to effectively detect the synonymous relationships between these huge numbers of queries and keywords is not an easy task.

Almost all of the existing published work about keyword matching focused on nonsynonymous matching scenarios happened in phrase match and broad match [1, 17, 20]. As far as we know, few works have tried to tackle the synonymous matching problem in exact match scenario.

It is well known that search queries are highly skewed and exhibit a power-law distribution [27, 33]. Approximately 20% of frequent

1http://ai.baidu.com/broad/subordinate?dataset=paraphrasing

2https://support.google.com/google-ads/answer/2497825?hl=en
queries occupy 80% of the query volume. In an industrial environment, the retrieved synonymous keywords for frequent queries are stored in a Key-Value lookup table, which is computed and updated in an offline mode. When an ad hoc query arrives, corresponding synonymous keywords can be retrieved immediately by looking it up on the table. This pre-retrieval offline architecture provides the industry with a good experimental environment, where lots of web data mining methods and state-of-the-art complex techniques can be tried.

Our main contributions are two folds: Firstly, a practical data-augmentation-like framework is proposed to address the synonymous keywords retrieval problem under exact match type, which includes translation-based retrieval and discriminant-based filtering. The Bag-of-Core-Words transformation trick increases the ∆ substantially while keeping the original precision. And the domain fine-tuned BERT’s performance far exceeds the feature-driven GBDT model’s. To the best of our knowledge, this is the first time to address this important commercial problem. This method has been successfully applied to Baidu’s sponsored search, which yields a significant improvement in revenue.

Secondly, a high-quality Chinese commercial synonym data set containing 500K pairs has been published along with this paper, which might be used in paraphrasing or other similar tasks. As far as we know, this is the first published large scale high-quality Chinese paraphrase data set reaching a precision of 95%.

2 RELATED WORK

The semantic gap between users and advertisers is the most challenging problem in synonymous keywords retrieval. It could lead to
the failure of the standard inverted index-based retrieval and has been widely discussed. Some works introduced query-query transformation (query rewriting) \[5, 10, 20, 43\] and keyword-keyword transformation \[1\]. Direct query-keyword transformation has also been studied \[16, 17\]. \[17\] proposed a Trie-constrained translation method to make sure all generated sentences are valid commercial keywords. However, all of the above work focused on nonsynonymous matching scenarios.

Our framework consists of two main parts: a paraphrase generation model and a paraphrase identification model. Paraphrase generation (PG), a task of rephrasing a given sentence into another with the same semantic meaning, has been used in various Natural Language Processing applications, such as query rewriting \[47\], semantic parsing \[2\], and question answering \[28\]. Traditionally, it has been addressed using rule-based approaches \[21, 45\]. Statistical machine translation has been used in \[40\]. Recent advances in deep learning have led to more powerful data-driven approaches to this problem. \[32\] applied neural machine translation for paraphrase generation to improve Alexa’s ASK user experience. The semantic augmented transformer seq2seq model has also been studied \[38\]. The Variational AutoEncoder based generation model is also a good option \[11, 26\].

Paraphrase identification (PI) aims to determine whether two natural language sentences have identical meanings. With the growing trend of PI, many English paraphrasing datasets have been made for this task, such as Quora Question pairs, and a lot of works have been developed based on them \[4, 29\]. Traditional methods mainly made use of handcrafted features \[6, 19, 30, 37\]. Recently, deep neural network (DNN) architectures have played a part in PI tasks. According to whether the inner interaction between a pair of sentences is modeled, there are mainly two types of methods. The first is encoder-based. RAE \[31\] is a pioneer that introduced recursive AutoEncoders to PI. Both ARC-I \[12\] and Hybrid Siamese CNN \[23\] adopted Siamese architecture introduced in \[3\] but used different loss functions. The other is interaction-based. ARC-II \[12\] and IIN \[9\] utilized the interaction space between two sentences while \[25\] viewed the similarity matrix between words in two sentences as an image and utilized a CNN to capture rich matching patterns. Bi-CNN-MI \[41\], ABCNN \[42\], BiMPM \[39\] and GSMNN \[8\] focused on the effect of introducing multi-granular and multi-direction matching. \[44\] showed that BERT \[7\] pre-trained on a large corpus and then fine-tuned with an additional layer worked quite well on PI tasks.

3 METHOD

Our method can be formalized as 2 main steps: translation-based retrieval and discriminant-based filtering. The seed paraphrasing dataset needed for the augmentation framework can be any dataset of short paraphrasing text pairs. In this paper, we won’t go into much detail about paraphrase extraction techniques. Our augmentation framework aims to retrieve more keywords for each query while ensuring high precision.

3.1 Translation-Based Retrieval

Our translation model follows the common sequence to sequence learning encoder-decoder framework \[35\]. And we implemented it with the Transformer \[36\], considering its state-of-the-art performance in learning long-range dependencies and capturing the semantic structure of the sentences.

To increase the translation model’s retrieval efficiency, a prefix tree for the keyword repository is built ahead of time, and all the beam search decoding is constrained on this prefix tree \[17\]. At each step of the beam search, the prefix tree will directly give the valid suffix tokens following the current hypothesis path, then a greedy top-N selection is performed within the legal tokens. This technique makes sure all the generated sequences are valid keywords.

### Table 1: Some typical trivial synonymous translations for How much does double eyelid surgery cost.

| How much does double eyelid surgery cost generally | How much does double eyelid surgery cost in general? | How much does double eyelid surgery cost probably |
|---------------------------------------------------|----------------------------------------------------|--------------------------------------------------|
| Paraphrases extracted from web data usually include some trivial patterns: reordering of words, insertion of function words and punctuation. A simple implementation of the translation model might generate too many trivial paraphrases, as is shown in Table 1. Common stop words removing method is too coarse to meet our need for synonym keeping, especially in Chinese. We carefully designed a synonym preserving data reduction method called Bag-of-Core-Words (abbreviated as BCW) transformation to reduce each sentence into a compact form without losing semantic information.

For each tokenized sentence, the BCW transformation consists of two sequential steps:

1. **Core-Words Transformation.** Here we consider some part of speech (abbreviated as POS) tags (like interjections, modal particles, etc.) as redundant, which means removing it generally does not change the query’s intention. Table 2 lists the typical redundant POS tags. Tokens with these tags would be removed. It is worth noting that some of the POS removing rules might not be universally applicable to other languages. For example, modal particles and interjections are quite common in Mandarin Chinese, however, these words are unusual in English. And the remaining tokens are considered as core words. For brevity, we refer to this step as CW in the following sections.

| POS tag    | Typical terms        |
|------------|----------------------|
| Interjection | 嗳( humming), 嗯( em), 嘿( hey), 嘘( shh) |
| Auxiliary word | 等等( and so on), 一般 (generally) |
| Punctuation | comma, colon, question mark |
| Modal particle | 啊( ah), 哇( wow), 嗯 (yo), 耶( yeah) |

2. **Bag-of-Words Transformation.** In most cases word order does not affect the meaning of a sentence. In fact, we sample 600 commercial queries from the ad weblog and find that in 94% of cases, the original query and its Bag-of-Words form have the same meaning. Based on this consideration, we sort the remaining core tokens in the sentence literally to
remove the order’s effect in the model training process. To make it more accurate, additional rules have been made to exempt the special cases. For example, *Flights from New York to Beijing and Flights from Beijing to New York* have the same Bag-of-words form, but their meanings are different. The same is true for *Does hypertension cause hyperlipidemia and Does hyperlipidemia cause hypertension*. So when sentences have two location tokens or two disease entity tokens or other token pairs with causality, temporal relation, etc., the order of the paired tokens remains unchanged.

The BCW transformation is simple, fast and very effective. Applying it to our Chinese training data effectively reduces the dataset size by nearly 20% with a synonym precision of 98%. Based on the BCW transformation, we devised a data-augmentation-like method to retrieve synonymous keywords, as is illustrated in Algorithm 1.

```
Algorithm 1: Retrieve synonymous keywords in a data augmentation way.
Input: Synonymous query-keyword pair dataset \( D_{old} \), keyword repository \( K \), frequent queries set \( Q \), beam size \( B \)
Output: Expanded query-keyword pair dataset \( D_{new} \)
1. Apply the BCW transformation to \( D_{old} \) to get \( D_{old} \)
2. Train a neural machine translation model \( M \) on \( D_{old} \)
3. Apply the BCW transformation to \( K \) to get \( K \) and the corresponding relationship for elements in \( K \) and \( K \) is stored in a lookup table \( T \)
4. Set the expanded dataset \( D_{new} \) to be a copy of \( D_{old} \)
5. for each \( q \) in \( Q \) do
6. for each \( k \) in \( R_q \) do
7. Apply the BCW transformation to \( q \) to get \( \tilde{q} \)
8. Using \( M \) to translate \( \tilde{q} \) towards \( K \) with a beam size of \( B \), which results retrieved keywords set \( R_{q} \)
9. Using \( T \) to make inverse BCW transformation on \( R_{\tilde{q}} \)
10. Merge query-keyword pair \( < q, k > \) into \( D_{new} \)
11. end
12. end
```

### 3.2 Discriminant-Based Filtering

In business applications like sponsored search’s matching product, high precision is essential. The BERT-based classifier we use to further filter out bad cases has a similar structure with the classifier used in the sentence pair classification task illustrated in [7], so we skip the exhaustive background description of the architecture and the training progress of the underlying model.

### 4 EXPERIMENTS

#### 4.1 Dataset

**Dataset for translation model.** The seed data \( D_{old} \) is extracted by calculating query-query similarity based on same URL click-through information from the search engine’s weblog [46], keyword-keyword similarity based on same advertiser purchase information from the ad database and synonyms replacement. This data is splitted into 3 parts: \( D_{train}^{PG} \) for training, \( D_{dev}^{PG} \) for developing, and \( D_{test}^{PG} \) for testing. The detailed statistics are shown in Table 3. The keyword repository \( K \) contains 102,025,475 keywords.

| Dataset  | Query Number | Total Pairs   | Average Pairs (/query) |
|----------|--------------|---------------|------------------------|
| \( D_{train}^{PG} \) | 16,578,545 | 6,67,827      | 6.67                   |
| \( D_{dev}^{PG} \)  | 93,467      | 613,143       | 6.56                   |
| \( D_{test}^{PG} \) | 99,588      | 656,602       | 6.59                   |

**Dataset for discriminant model.** To make a balanced domain dataset for human evaluation, three kinds of query-keyword matching weblogs \( D_{exact}, D_{phrase}, D_{broad} \) are used, which correspond to the exact match, phrase match, and broad match respectively. Among them, \( D_{exact} \) is probably synonymous and provides potential positive examples, while \( D_{phrase} \) and \( D_{broad} \) mainly contribute negative examples. These three data sources are merged with the proportion of 2:1:1. Then 170,000 data denoted as \( D_{h}^{PI} \) is sampled from it and sent to professionals for synonymous binary human evaluation. According to the human labels, 42.8% of \( D_{h}^{PI} \) are positive samples and 57.2% are negative. \( D_{h}^{PI} \) is further split into three parts: 90% of it is used as the domain specific data for BERT fine-tuning, which is denoted as \( D_{train}^{PI} \); and 5% of it is used for development, denoted as \( D_{dev}^{PI} \) and the remaining 5% denoted as \( D_{test}^{PI} \) is used for testing.

#### 4.2 Implementation Details

For the translation model, the word embeddings are randomly initialized. The vocabulary contains 100,000 most frequent tokens in the training data \( D_{train}^{PG} \). The word embedding dimension and the number of hidden units are both set to 512. For multi-layer and multi-head architecture of the Transformer, 4 encoder and decoder layers and 8 multi-attention heads are used. And during the training, all layers are regularized with a dropout rate of 0.2. And the model’s cross-entropy loss is minimized with an initial learning rate of \( 5 \times 10^{-3} \) by Adam [15] with a batch size of 128.

The paraphrase discriminant model is implemented with BERT [7], which takes a query-keyword pair separated by a special token as input and predicts a synonymy label. The model contains 12 layers, 12 self-attention heads, and the hidden dimension size is 768. We initialize it with ERNIE [34], which learns Chinese lexical, syntactic and semantic information from a number of pretrained tasks, and fine-tuned it on \( D_{train}^{PG} \). The fine-tuned loss is minimized with an initial learning rate of \( 1 \times 10^{-6} \) by Adam [15] with a batch size of 64. We evaluate our model after each epoch and stop training when the validation loss on \( D_{dev}^{PI} \) does not decrease after 3 epochs.
There are three major concerns: the size of the framework, the CW-M ratio, which proves the feasibility of the data-augmentation-like BASE-M for these three different strategies with a beam size of 30. We can consider for evaluation: and the decoding time for generating a large data increment of 3 different strategies. For all of the strategies, decoding is conducted in a Trie-constrained mode. For the sake of convenience, these strategies are abbreviated as follows:

- **BASE-M**: This is our base strategy, where the translation model is trained on the original training data $D_{PG}^{\text{train}}$. And the prefix tree is built on $\mathcal{K}$.
- **CW-M**: In this strategy, the model is trained on the CW-transformed data. And the prefix tree is built on $\mathcal{CW}$-transformed $\mathcal{K}$. The final result is made by joining the generated hypotheses with $\mathcal{K}$ based on the CW transformation.
- **BCW-M**: In this strategy, the model is trained on the BCW-transformed data. And the prefix tree is built on BCW-transformed $\mathcal{K}$. The final result is made by joining the generated hypotheses with $\mathcal{K}$ based on the BCW transformation.

Each of the trained models is utilized to decode towards those queries in $D_{PG}^{\text{test}}$ to get a result data set $D_{1}^{PG}$. Under the data augmentation framework, we expect the translation model could make a large data increment $\Delta$ while keeping a high precision. There are three major concerns: the size of $\Delta$, the precision of $\Delta$, and the decoding time for generating $\Delta$. The following indicators are considered for evaluation:

- **Diff ratio** is defined as $\frac{|D_{PG}^{\text{train}}-D_{PG}^{\text{test}}|}{|D_{PG}^{\text{test}}|}$, which is an indicator of the generalization ability.
- **BLEU-n** is an indirect indicator of the generation quality. Since most sentences in our scenario are short texts, we only consider BLEU-2.
- **Dist-n** measures the number of distinct N-grams within the set of generated data, which indicates the diversity among the generated paraphrases. Following previous studies [26], we use Dist-1,2.
- **Precision** indicates the proportion of synonymous pairs in generated data. Concretely, for each strategy, 400 query-keyword pairs are sampled from the generated results $D_{PG}^{\text{train}}$ for binary human evaluation. We denote this dataset as $D_{PG}^{\text{1th}}$ for later reference.

Rows 1, 3 and 4 in Table 4 show the retrieval performances for these three different strategies with a beam size of 30. We can see that: BASE-M could already make a certain amount of Diff ratio, which proves the feasibility of the data-augmentation-like framework. CW-M and BCW-M further enlarge the Diff ratio to 2.4 times and 4.2 times, compared with the base method. Meanwhile, the precision of CW-M and BCW-M are almost the same as that of BASE-M. The Dist indicator shows that the results’ diversity has been improved by CW-M and BCW-M.

For further analysis, we evaluate the decoding results of BASE-M with a beam size of 120. As is shown in Table 4, although the diff ratio increases up to 4 times, which is nearly equal to BCW-M with a beam size of 30, the precision and diversity drop significantly. What’s more, BASE-M has to spend more than 10 times of time to make it.

To conclude, BCW-M greatly enlarges the Diff ratio while maintaining a high level precision. It is fast and almost little extra time is consumed.

### 4.3 Results of Translation Model

For the translation model, we compared the retrieval performances of 3 different strategies. For all of the strategies, decoding is conducted in a Trie-constrained mode. For the sake of convenience, these strategies are abbreviated as follows:

- **BASE-M**: This is our base strategy, where the translation model is trained on the original training data $D_{PG}^{\text{train}}$. And the prefix tree is built on $\mathcal{K}$.
- **CW-M**: In this strategy, the model is trained on the CW-transformed data. And the prefix tree is built on $\mathcal{CW}$-transformed $\mathcal{K}$. The final result is made by joining the generated hypotheses with $\mathcal{K}$ based on the CW transformation.
- **BCW-M**: In this strategy, the model is trained on the BCW-transformed data. And the prefix tree is built on BCW-transformed $\mathcal{K}$. The final result is made by joining the generated hypotheses with $\mathcal{K}$ based on the BCW transformation.

Each of the trained models is utilized to decode towards those queries in $D_{PG}^{\text{test}}$ to get a result data set $D_{1}^{PG}$. Under the data augmentation framework, we expect the translation model could make a large data increment $\Delta$ while keeping a high precision. There are three major concerns: the size of $\Delta$, the precision of $\Delta$, and the decoding time for generating $\Delta$. The following indicators are considered for evaluation:

- **Diff ratio** is defined as $\frac{|D_{PG}^{\text{train}}-D_{PG}^{\text{test}}|}{|D_{PG}^{\text{test}}|}$, which is an indicator of the generalization ability.
- **BLEU-n** is an indirect indicator of the generation quality. Since most sentences in our scenario are short texts, we only consider BLEU-2.
- **Dist-n** measures the number of distinct N-grams within the set of generated data, which indicates the diversity among the generated paraphrases. Following previous studies [26], we use Dist-1,2.
- **Precision** indicates the proportion of synonymous pairs in generated data. Concretely, for each strategy, 400 query-keyword pairs are sampled from the generated results $D_{PG}^{\text{train}}$ for binary human evaluation. We denote this dataset as $D_{PG}^{\text{1th}}$ for later reference.

Rows 1, 3 and 4 in Table 4 show the retrieval performances for these three different strategies with a beam size of 30. We can see that: BASE-M could already make a certain amount of Diff ratio, which proves the feasibility of the data-augmentation-like framework. CW-M and BCW-M further enlarge the Diff ratio to 2.4 times and 4.2 times, compared with the base method. Meanwhile, the precision of CW-M and BCW-M are almost the same as that of BASE-M. The Dist indicator shows that the results’ diversity has been improved by CW-M and BCW-M.

For further analysis, we evaluate the decoding results of BASE-M with a beam size of 120. As is shown in Table 4, although the diff ratio increases up to 4 times, which is nearly equal to BCW-M with a beam size of 30, the precision and diversity drop significantly. What’s more, BASE-M has to spend more than 10 times of time to make it.

To conclude, BCW-M greatly enlarges the Diff ratio while maintaining a high level precision. It is fast and almost little extra time is consumed.

### 4.4 Results of Discriminant Model

Our baseline model is a GBDT (Gradient Boosting Decision Tree) model [14] trained on a collection of human designed text similarity features. The features are listed as follows:

1. **Token level matching degree**: max matching length, the proportion of matching and missing tokens, BM25, BLEU1 and BLEU2.
2. **Named entity similarity**: whether the named entities in query and keyword are matched.
3. **Simple Approximate Bigram Kernel** [24] based on dependency parsing tree.
4. **Document class similarity**: whether query and keyword belong to the same document class.
5. **Semantic similarity**: DSSM [13] trained with query-keyword click-noclick pairs shown in the weblog, and Word2vec based cosine similarity [22].
6. **Translation likelihood**: the BASE-M translation score which is calculated by $P(\text{keyword} | \text{query})$.

GBDT is also trained on $D_{PG}^{\text{train}}$, and validated on $D_{PG}^{\text{test}}$. And the model hyperparameters are optimized by grid search. For evaluation, we use two metrics: the area under an ROC curve (AUC) and recall under 95% precision. Table 5 shows the models’ performances on $D_{PG}^{\text{test}}$. To our surprise, BERT greatly exceeds GBDT’s performance. For the AUC indicator, BERT outperforms GBDT by 11.4 percentage points. Under the precision of 95%, BERT’s recall exceeds GBDT by 45 percentage points.

| Strategy | Beam Size | Diff ratio | BLEU-2 | Dist-1/2 | Precision | Decoding Time (ms/query) |
|----------|-----------|------------|--------|----------|-----------|-------------------------|
| BASE-M   | 30        | 17.089%    | 0.446  | 0.0014/0.047 | 83.5%     | 262.5                   |
| BASE-M   | 120       | 69.753%    | 0.356  | 0.0007/0.032 | 62.0%     | 2988.9                  |
| CW-M     | 30        | 40.510%    | 0.404  | 0.0015/0.058 | 83.0%     | 238.8                   |
| BCW-M    | 30        | 72.522%    | 0.387  | 0.0018/0.061 | 82.5%     | 254.1                   |

| Table 4: Results of different strategies. |

| Table 5: Model performances of BERT VS GBDT on $D_{PG}^{\text{test}}$. Recall indicates the recall ratio under the precision of 95%. |

| Strategy | AUC | Recall |
|----------|-----|--------|
| GBDT     | 84.4% | 21.8%   |
| BERT     | 95.8% | 66.8%   |
Table 6: Model performances on $D_{1h}^{PG}$ generated by BCW-M. Recall indicates the recall ratio under the precision of 95%.

| Strategy | AUC  | Recall |
|----------|------|--------|
| GBDT    | 75.0% | 25.2%  |
| BERT    | 94.7% | 91.3%  |

We also considered the models’ performances on $D_{1h}^{PG}$ generated by BCW-M, since our motivation is to remove bad cases in the translations. Here we use the previously human evaluated dataset $D_{1h}^{PG}$ for testing. As is shown in Table 6, BERT achieves a recall of 91.3% under the precision of 95%, which also far exceeds the GBDT’s performance. (The difference of recall ratios in Table 6 and Table 5 comes from the difference of data distribution between $D_{1h}^{PG}$ and $D_{test}^{PL}$. In $D_{1h}^{PG}$, positive examples account for 82.5%, while in $D_{test}^{PL}$, positive examples account for 42.8%)

BERT’s dramatic improvement might come from the following points: a) BERT has a super large parameter space in contrast to the simple semantic similarity features like Word2vec and DSSM. The huge capacity and abundant pretraining make BERT being able to learn a lot of external semantic similarity knowledge, which is especially useful to alleviate the vocabulary mismatch in paraphrase identification. b) The paraphrase relationship between two sentences is usually judged by the alignment of their tokens. To some extent, the multi-head attention mechanism in the Transformer might make soft alignments in multiple semantic spaces.

4.5 Case Study and Discussion

4.5.1 Bag-of-Core-Words Transformation. Table 7 shows some typical cases for the BCW-M strategy. Due to space constraints, we only present the top 5 decoding results. The second column in Table 7 shows the final results of BASE-M, and the last column shows the raw generated results of BCW-M without joining the keyword repository, which is presented as Bag-of-Words form and the symbol ‘|’ is used as the token separator. We can see that with BCW-M, the number of redundant translations decreases a lot. For example, before BCW-M is applied, most of the generated hypotheses for query (the market price, market price of gold) are quite trivial: the market price just removes the auxiliary token ‘of’, and the market price simply further reorders these tokens. When BCW-M is applied, nontrivial bag-of-words paraphrases like (多少黄金钱, Gold) how much) and (查询价格黄金, Gold price) query emerged into the top 5 hypotheses list. Similarly, for query (化妆的培训学校, Training school of makeup), four nearly identical translations (化妆培训学校,化妆的培训学校, 化妆培训学校, 培训化妆的学校) are generated under the base strategy. When BCW-M is applied, the results are not dull anymore.

4.5.2 Good Cases and Bad Cases. Table 8 shows some sampled results generated by our translation model. We can see that in most cases our model greatly captures the synonym relationship between query and keyword. The vocabulary mismatch problem has been alleviated to some extent. For example, the synonym relationships between (减肥, weight loss) and (瘦身, slimming), (油污, oil) and (油渍, greasy dirt), etc. have been captured. Some complex paraphrases like (宝宝身上胎记是怎么回事, 婴儿身上胎记是什么原因) (What causes a birthmark on a baby, what is the reason of a birthmark) have been generated, which could not be simply accomplished by synonymous phrase replacement.

In the meanwhile, bad cases might also be generated. Sometimes, important intentions in the query might be discarded. For example, in the case of (怎么在跑步机上跑步, 怎么跑步减肥) (How to run on a treadmill, How to lose weight through running), the intention of ‘treadmill’ is lost.

4.5.3 Why Translation? People might argue that since we have trained a discriminant model, why not use this model to do direct retrieval? In other words, can we make retrieval by simply scoring all the query-keyword pairs in the Cartesian product set? The reason is the scalability problem in the real industry environment mentioned in the introduction. The keyword volume might reach a magnitude of billions. And the frequent query set size is usually set to reach 10 million to have a great impact on the revenue. Computing so many pairs directly with a BERT model is impossible even in an offline mode, for it costs about 3.5 × 10^7 days on 8 Tesla K40m GPUs to score all pairs. Translation based method gives us a practical way to find more confident query-keyword pairs which are more likely to be synonym candidates, thus greatly narrowing down the possible candidates’ size.

4.6 The published Dataset

Our work has a close relationship with paraphrase generation and paraphrase identification. Most of the existing large scale paraphrasing datasets (MSOCO, SNLI, etc.) are in English and high-quality Chinese dataset is extremely scarce in this domain. The most related dataset is a 24K sized dataset LCQMC [18]. However, it focuses on general intent matching rather than paraphrasing. To promote the Chinese paraphrasing research, we decide to publish a large scale high-quality dataset containing 500K commercial synonymous short text pairs along with this paper. This dataset is produced in the following steps: A translation model trained with the BASE-M strategy is used to make unconstrained decoding for real-world frequent queries, where the prefix tree constraint is discarded. Then our discriminant model is used to filter out bad cases. Finally, some heuristic rules are devised to filter out trivial translations. Manual sampling evaluation shows that this dataset has a precision of 95%.

4.7 Online Experiments

A real online A/B test experiment is deployed on Baidu’s commercial advertising system, where two fractions of search flow are sent to the experimental group and the control group independently. $D_{old}$ is used as the Key-Value table in the control group and $D_{old} + \Delta$ in the experimental group. We use 10,000,000 frequent queries, and the translation model is trained with BCW-M strategy.

For each group, # {searches} is used to denote the total number of queries it received, # {clicks} to denote the corresponding number of clicks, and revenue to denote the search company revenue. The following metrics are calculated independently for each group to evaluate the performance of our method.

- SHOW denotes the total number of ads shown to users.
- $\text{CTR} = \frac{\#\{\text{clicks}\}}{\#\{\text{searches}\}}$, which denotes the average clicks received by the search engine.
which is a significant improvement for our revenue. We give our relevance quality has not been deteriorated. The meanwhile, the human evaluation demonstrates that the ads’ clicks and prices combined to bring about the final CPM growth. In increased by 0.8%, and the clicks increased by 1.02%. The growth in clicks and prices combined to bring about the final CPM growth. In the meanwhile, the human evaluation demonstrates that the ads’ relevance quality has not been deteriorated.

Table 7: Top-5 beam search results of BASE-M VS BCW-M.

| Query                          | BASE-M | BCW-M |
|-------------------------------|--------|-------|
| 市场价格 | 市场价格 | 市场价格 |
| 市场价格 | 市场价格 | 市场价格 |
| 市场价格 | 市场价格 | 市场价格 |
| 市场价格 | 市场价格 | 市场价格 |
| 市场价格 | 市场价格 | 市场价格 |

Table 8: Some typical synonymous keywords generated by the translation model.

| Query                          | Generated Keywords | Label |
|-------------------------------|-------------------|-------|
| 男士减肥机构                 | 男性瘦身机构      | Good  |
| Men’s Weight Loss Agency      | Agency for men slimming | Good  |
| 厨房油污怎样清除             | 厨房油渍如何去除 | Good  |
| How to clean off oil in the kitchen | How to remove kitchen greasy dirt | Good  |
| 宝宝身上胎记是怎么回事       | 婴儿身上胎记是什么原因 | Good  |
| What causes a birthmark on a baby | What is the reason of a birthmark | Good  |
| 怎么在跑步机上跑步             | 怎么跑步减肥      | Bad   |
| How to run on treadmill       | How to lose weight through running | Bad   |

- ACP = \(\frac{\sum_{i} \text{price}}{\#(\text{clicks})} = \frac{\text{revenue}}{\#(\text{clicks})}\), which denotes the average click price paid by the advertisers.
- CPM = \(\frac{\text{revenue}}{\#(\text{searches})} \times 1000 = \text{CTR} \times \text{ACP} \times 1000\), which denotes the average revenue received by the search engine for 1000 searches.
- Quality refers to the query-keyword synonym relationship. For each side of this A/B experiment, 600 query-keyword cases under the exact match type (excluding literally identical ones) are sampled from the system’s ad weblog and are sent for binary human evaluation. And the quality score is calculated as the proportion of the synonymous pairs.

Table 9: Online A/B Test performance of our method.

| SHOW | CTR  | ACP   | CPM  | Quality |
|------|------|-------|------|---------|
| +0.8%| +1.02%| +0.62%| +1.64%| +1.2%   |

Table 9 shows the relative improvements in these indicators. We can see that the incremental dataset Δ increases the CPM by 1.64%, which is a significant improvement for our revenue. We give our intuitive explanation from the demand-supply perspective. The sponsored search system aims to match query flows to the advertisers’ demands. Detecting more synonym relationships between queries and keywords helps to get more ads into the downstream auction phase. The ACP’s 0.62% growth shows the competition in the ad queue has been intensified. Finally, the number of shown ads increased by 0.8%, and the clicks increased by 1.02%. The growth in clicks and prices combined to bring about the final CPM growth. In the meanwhile, the human evaluation demonstrates that the ads’ relevance quality has not been deteriorated.

5 CONCLUSIONS

In this paper, we present a simple but effective data-augmentation-like framework to address the synonymous keyword retrieval problem for frequent queries in sponsored search. Based on a seed paraphrasing dataset, a sequence to sequence translation model is trained and used to decode out more synonymous pairs to expand the data. To ensure high precision for commercial usage, a domain fine-tuned BERT is used to filter out bad cases. During the translation phase, we introduce a novel scheme to make the decoding more effective: Bag-of-Core-words transformation, which enlarges the diff 4.2 times while almost keeping the original precision. During the discrimination phase, BERT outperforms the traditional feature-driven GBDT model by 11 percentage points. To the best of our knowledge, this is the first published work to address the synonymous keywords retrieval problem. Our framework has been successfully applied in Baidu’s sponsored search. As a byproduct, 500K high quality commercial Chinese synonymous pairs have been published along with this paper. To the best of our knowledge, this is also the first published large scale Chinese paraphrasing dataset with a precision of 95%.

REFERENCES

[1] Javad Azimi, Adnan Alam, and Ruofei Zhang. 2015. Ads keyword rewriting using search engine results. In Proceedings of the 24th International Conference on World Wide Web. ACM, 3–4.
[2] Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1415–1425.
[3] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. Machine Learning 94, 2 (2014), 233–259.
[4] Andreas Chandra and Ruben Stefanos. 2020. Experiments on Paraphrase Identification Using Quora Question Pairs Dataset. arXiv preprint arXiv:2006.02648
