information

MDPI

Article

Study on Unknown Term Translation Mining from Google Snippets

Bin Li [1,*] and Jianmin Yao [2,*]

[1] School of Information Engineering, Anhui Open University, Hefei 230041, China
[2] Provincial Key Laboratory of Computer Information Processing Technology, Soochow University, Suzhou 215006, China
[*] Correspondence: szbinlee@126.com (B.L.); jyao@suda.edu.cn (J.Y.)

Received: 16 May 2019; Accepted: 12 August 2019; Published: 28 August 2019

Abstract: Bilingual web pages are widely used to mine translations of unknown terms. This study focused on an effective solution for obtaining relevant web pages, extracting translations with correct lexical boundaries, and ranking the translation candidates. This research adopted co-occurrence information to obtain the subject terms and then expanded the source query with the translation of the subject terms to collect effective bilingual search engine snippets. Afterwards, valid candidates were extracted from small-sized, noisy bilingual corpora using an improved frequency change measurement that combines adjacent information. This research developed a method that considers surface patterns, frequency–distance, and phonetic features to elect an appropriate translation. The experimental results revealed that the proposed method performed remarkably well for mining translations of unknown terms.

Keywords: unknown term; translation mining; web mining; google snippets

1. Introduction

The rapid development of Web 2.0 and constant expansion of the network size has led to a great increase in the amount of informational resources in multiple languages on the Internet. Because most populations tend to express their demands in their local language, linguistic differences have become a major obstacle. Cross-language information retrieval (CLIR) enables people to retrieve documents written in multiple languages through a single query. Despite the rapid advancements in this field, CLIR still has a major choking point when a query involves the translation of unknown terms, which is known as an out-of-vocabulary (OOV) problem. Searching for and finding the correct translations of those terms is bound to enhance the performance of CLIR, machine translation systems, and question-answering systems. Otherwise, if unknown terms are translated incorrectly, the performance of CLIR or other systems is greatly reduced.

This research suggests a new approach for mining translations of web-based terms. It adopts co-occurrence information to expand the cross-linguistic terms so as to more effectively extract bilingual web snippets that are more relevant. In order to extract valid candidates, the research employed various term extraction methods with frequency change measurement. All kinds of features were studied during the stage of ranking the translations to assign weights to the candidates.

This paper is organized as follows: Section 2 introduces the relevant research. Section 3 proposes a system for thoroughly mining translations of web-based terms. The experimental evaluation is discussed in Section 4, and Section 5 draws conclusions.

2. Related Works

In traditional CLIR, queries are translated based on a bilingual dictionary. Exploring the recognition of phrases in queries, Ballesteros et al. [1] investigated the role of phrases in query expansion by using local feedback and local context analysis with dictionary-based methods. Pirkola et al. [2] evaluated the problems of dictionary-based approaches, which include untranslatable search keys, phrase translation, inflected keys, and ambiguous translation. They described several effective approaches to deal with the above issues, such as using special dictionaries for extending the base of translatable terms and part-of-speech (POS) tagging for phrase translation. Pirkola et al. [3] adopted recognition of transformation rules and translation equivalents based on frequency to enhance the performance of the dictionary-based approach. Sharma and Mittal [4] built a query translation system based on dictionaries in which the queries were tokenized, and N-gram techniques were adopted to create terms for multiword queries. In the dictionary-based approach, even an optimal dictionary cannot solve the OOV problems related to short, real, diverse, and dynamic queries.

Term translation extraction from a bilingual corpus (including parallel and comparable corpora) is a topic of considerable interest. Many researchers [5–8] have extracted translation equivalents at the sentence level. Some researchers [9,10] have described methods for learning word translations from the article-level-aligned corpus. Otero and Campos [11] proposed an approach to extract word translations. Their approach depends on bilingual pairs for lexico-syntactic templates that are previously extracted from parallel corpora. By using an intermediary language, scholars have improved the extraction of bilingual lexicon from parallel corpora [12,13]. Vuli'c and Moens [14,15] applied a skip-gram model that learns bilingual word embeddings from data that are not aligned with nonparallel documents. Based on general-domain data, Hazem and Morin [16] improved the word co-occurrence of comparable corpora, and thereby, context representation. Hazem and Daille [17] extended the work of [18] and explored the synonym extraction of single-word terms and multiword terms of variable lengths. For a complete comparable corpus, Rigouts et al. [19] proposed a gold standard involving manual annotations with the support of a new method developed for accommodating inherent difficulties in the task. Although these methods are capable of realizing the highly accurate extraction of translations, one does not always have access to large bilingual corpora for specific domains. The translation pairs extracted from such corpora are usually for specific domains, and do not have the ability to translate short dynamic queries. So, the problem related to these methods is how to obtain a larger corpus that contains an abundant domain lexicon.

Due to the emergence of more bilingual resources on the Internet (e.g., Chinese to English, Korean to English, and Japanese to English), web texts generally contain the co-occurrence of some terms with corresponding translations. This makes it feasible and more reliable to explore web data and then translate queries for mining OOV translations.

Nagata et al. [20] were the first to use the Internet to translate Japanese words without clues in dictionaries. After querying a search engine with Japanese terms waiting for translation, the top 100 web pages were downloaded. Appropriate translations of Japanese words were extracted by taking the empirical function of the byte distance of Japanese and English words as the criterion. Lu et al. [21] proposed a method to extract translations of words from a web query by mining link structures and web anchor texts. Due to their dependence on abundant web data, these methods consume a great deal of time and space, and they require huge storage capacity and considerable network bandwidth, as well as a long computation and access time.

Cheng et al. [22] proposed the use of chi-squared and context vector approaches for selecting appropriate query translations. The returned top 100 snippets (containing titles and abstracts) were used as bilingual corpora rather than the returned top 100 pages. Huang et al. [23] discovered a search engine for extracting translations of pivotal phrases. Using punctuations to divide snippets and taking continual English strings as candidates, they evaluated potential translations by a combined model. Developing an approach for partitioning source terms, Fang et al. [24] expanded these terms in order to extract web pages. Building each English term as a beginning index, candidate strings were constructed as the string of a unit of English words increases in a 100-byte window centered on the keyword. Candidate translations were evaluated by multiple features. Sun et al. [25] divided source terms and looked up translations of divided units by proposing a forward–backward maximum matching method to extend source terms for the collection of bilingual snippets. Scholars have also used frequency and word-overlap features to mine translations. For query translation, Ge et al. [26] proposed a combined method to enhance the mining of OOV translations. Pal et al. [27] presented a CAT tool based on the web (i.e., CATaLog Online) that provides an online CAT environment for posteditors/translators. The goal of the system was to provide support to distributed translations, in which translator teams work together on diverse sections in the same text, thereby reducing effort and time for postediting, improving postediting experience, and capturing data used for research on the translation process and incremental MT/APE (machine translation/automatic post-editing). Berger and Lafferty [28] proposed a probabilistic method for retrieving information in accordance with the methods and ideas of statistical machine translation. All such approaches search more valid bilingual resources by exploiting query expansion. In many situations, however, a source term does not simply mean the combined meanings of component words; this brings additional noise in query expansion, and results in errors in translation mining.

Despite extensive studies on the mining of term translations based on the web, as mentioned above, there is still a huge gap in our ability to reach satisfactory performance. First, existing studies do not expand the query to collect sets of web pages, or the expansion approaches are defective and not applicable to diverse OOV terms. Second, continuous strings were regarded as translation candidates in these studies, while these strings are generally ineffective lexical units. Finally, apart from frequency, other features also exist for selecting potential translation candidates, such as surface and phonetic patterns as well as distance.

3. Web-Based Term Translation Mining System

3.1. Architecture of the System

The mining system for translations of unknown terms based on the web has the following architecture (Figure 1).

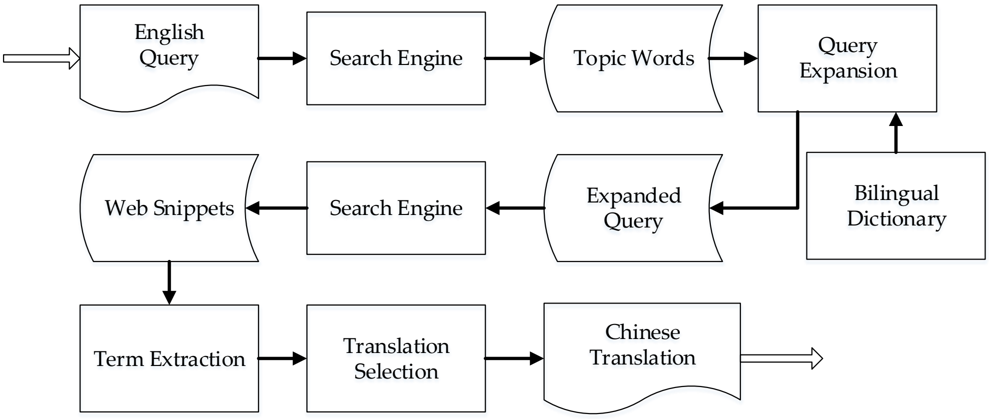

Figure 1. The architecture of the mining system for translations of unknown terms based on the web.

Taking the mining of English–Chinese translations as an example, the system works in the following steps: Firstly, the system sends source queries in English to a search engine for the retrieval of English documents. Secondly, related topic terms—serving as hint words for topic or subject of the queries—are mined from English snippets returned in the first step. Thirdly, the source queries, along with translations of subject terms obtained from the English–Chinese bilingual dictionary, are put into the search engine again to attain related snippets from the bilingual web. The "expanded query" usually consists of "the source query plus translated topic words". In the next step, valid terms are extracted from bilingual snippets returned, followed by the last step for sorting the ranks of candidate terms for obtaining final translations. In brief, three main parts are included in the system:

- Collection of bilingual snippets. Retrieve bilingual snippets containing source terms in English together with translations in Chinese via a search engine, and download the snippets as bilingual resources. Effective techniques for obtaining highly related snippets lay the foundation for the extraction of translations.
- Extraction of candidate terms. Extract valid lexical and multi-lexical units (MLUs) from the returned set of snippets in step 1. However, this process is indirect. To begin with, no spaces exist between characters in Chinese texts, so a snippet of two or three sentences is smaller relative to authoritative corpora. Second, OOV terms are usually contained in snippets. Thus, specific research remains to be conducted for the extraction of terms from returned snippets.
- Selection of proper translations. Rank and sort the translation candidates yielded in step 2. As a candidate set is probably extremely large, the most suitable translations are expected to be chosen therefrom.

3.2. Collection of Bilingual Snippets

For the purpose of mining Chinese translations of source terms in English, one first needs to collect enough related snippets containing not only the source term but also corresponding translations via a search engine. However, it is unfortunate that not all snippets collected via a search engine have both pieces of information mentioned above. Take “stealth fighter” as an example. When sending the phrase to a search engine, only a few of the snippets contain effective information for extracting the corresponding translation because most snippets are in English, which are useless for mining translations in Chinese. It was proposed by Ballesteros [1] that query expansion establishes a substantial foundation for the extraction of translations. Source queries are inevitably divided in expansion approaches in existing works, while unfortunately, the connotation of source terms is generally not simply the combination of the meanings of individual words therein. For example, in “风凉话”, which means sarcastic comments, while the component characters therein separately refer to “wind”, “cool”, and “talk”. Segmentation will introduce additional noises and even errors into the mining of translations. Therefore, this study employed a co-occurrence information-based expansion approach. For gathering bilingual snippets with both source terms and corresponding translations of high rankings, one first needs to send source terms to a search engine, followed by extracting subject terms from snippets returned in the language of the source terms. Afterwards, both the source term and translations of the subject terms in the target language are sent to a search engine to gather bilingual resources of higher relevance and validity.

Irrelevant information (e.g., HTML tags, punctuation marks, and so on) were removed from the returned snippets of source terms through processing. Non-noun words were also screened via POS tagging and English stop words. In this way, we could obtain a list of English noun words. The simple TF*IDF metric in the proposed system was used to mine topic terms from the above list. Afterwards, the top five were chosen from the list as topic terms. The weight is defined as:

$$\begin{aligned} &Weight_{tfidf}(w) = tf(w) \times idf(w) \\ &= f(w) \times (\log \tfrac{f_{\max}}{f(w)} + 1) \end{aligned} \tag{1}$$

where w represents a word, $f(w)$ denotes the total frequency of w appearing in the corpora, and $f_{\max}$ refers to the maximum frequency of words in the corpora.

It was found that topic terms mined from the first 20 snippets showed the highest quality for mining translations. In the above example “stealth fighter”, topic words mined from the first 20 snippets included “fighter”, “military”, “security”, “politics”, and “state”. The corresponding translations (note: to simplify the process, the first meaning was chosen from the dictionary if multiple meanings existed) were then adopted for cross-linguistic expansion. Therein, words for the cross-language expansion corresponded to “战机”, “军事”, “安全”, “政治”, and “国家”. Then, we sent the source term with translations of these topic words to retrieve bilingual snippets. Returned snippets of the expanded query of “stealth fighter” were discussed, and it revealed that most of the snippets contained useful information in Chinese. We greatly improved the quality of the bilingual resources compared with the case in which merely the source term was used. In the next step, these snippets served as a basis for the extraction of translation candidates.

3.3. Extraction of Candidate Terms

Candidate terms should be extracted from bilingual snippets to judge the Chinese terms with appropriate translations of the source terms. A candidate set comprises both words and MLUs (e.g., compound nouns, phrases, idioms, etc.). A source term can probably be translated into a single word or an MLU. Therefore, we need to label the boundaries of MLUs in Chinese texts. Traditional partitioning methods based on a dictionary are not applicable to identifying OOV terms in Chinese sentences; so, these approaches generally fail to obtain translations of the source terms. We also cannot obtain satisfactory results in a search engine using methods to mine translations of OOV terms from large corpora. This is because the size of corpora for search-engine-based methods have a rather small size in most cases. Generally, only two or three sentences are contained in one snippet. Besides, snippets basically appear as incomplete sentences, to which the traditional method to segment Chinese words is not applicable because of the absence of contextual information.

A method named frequency change measurement and adjacent information (FCMAI) was proposed in the section for extracting MLUs from relatively small bilingual collections containing noise. The method integrates frequency change measurement (FCM) [29] with adjacent information. FCM is based on the following two observations. Firstly, component characters constituting a term present similar frequencies in collections returned from search engines. In the previous example of “stealth fighter”, the frequencies of “隐形战机” were similar to the frequencies of “隐”, “形”, “战”, and “机” in the snippets. Secondly, as effective MLUs expand with an extra character, the frequency noticeably decreases for extended terms. For example, the MLU of “隐形战机” was extended with the character “了”, and the frequency of “隐形战机了” dropped significantly. FCM employs the equation below to judge a string S as an MLU:

$$R(S) = \frac{f(S)}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \tag{2}$$

where S denotes a Chinese string; $f(S)$, x_i, and $\overline{x}$ represent frequencies of S and each character in S, as well as the average frequency of all characters in S, respectively.

Candidate terms mined using FCM still comprise several partial terms that are subsequences in effective MLU terms, and several high-frequency Chinese characters may break the second observation. For example, when we added the high-frequency character “王” after some MLUs, the frequencies of extended MLUs did not drop. Thus, using FCM alone failed to extract correct candidates. To solve this problem, adjacent information was introduced to increase the quality of terms, as it was discovered that effective terms commonly involve various adjacent characters, but their sub-terms possess relatively fewer and unchanged adjacent characters. Using the FCMAI method, Equation (2) is rewritten as:

$$R'(S) = \frac{LN(S) \times f(S) \times RN(S)}{1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \tag{3}$$

where $LN(S)$ and $RN(S)$ represent the total quantities of unique characters neighboring S in the left and right, respectively. The extraction algorithm for candidate terms is described as Algorithm 1.

Algorithm 1 FCMAI

```
Input: s = a_1, a_2, . . . , a_n, which is an English string containing $n$ characters
Output: M, which is an assembly of MLUs. /* M is the assembly of candidate translation terms. */
BEGIN Procedure FCMAI
Use symbol "+" in s to replace English stop words like "the", "is", "are", "a", etc. and split s to substrings
according to "+" and punctuation marks. In this way, s contains several substrings: s_1, s_2, . . . , s_m.
M = None, $Threshold = 0.5$, $\omega = 7$ /* The value of Threshold is chosen empirically. */
For each substring s_i in s
    Let $b = 1$, $e = 1$, first-term = true
LOOP: Let t1 = a_b, a_2, . . . a_e, t2 = a_b, a_2, . . . a_(e+1)
   If $R'(t1) > R'(t2)$ and $R'(t1) > Threshold$ then M = M∪{t1} /* Add substring t1 to M */
If first-term = true, then
first-position = $e$ and first-term = false
 If $e - b + 1 > =\omega$, then
      $e$ = first-position, $b = e + 1$, first-term = true
      $e = e + 1$
   If $e + 1 <=$ length of s_i then goto LOOP.
END For.
Return M.
END Procedure FCMAI
```

English stop words are not simply deleted from a string directly in Algorithm 1, which makes it possible to extract characters in the left and right neighbors of stop words along with MLUs, thereby introducing additional noise. Take the sentence "F117 is the first stealth fighter of the world that can be formally combated" as an example. If stop words "is", "the", "of", and "that" are directly eliminated, an MLU may be extracted. Therefore, we split the sentences into substrings based on punctuation marks and stop words. For the purpose of improving quality and decreasing the amount of candidate terms, a threshold was employed to screen useless MLUs. We used another parameter, ω, referring to the largest quantity of characters waiting check, for mining longer terms. The set of candidate terms extracted using Algorithm 1 also comprised several words that can commonly be discovered in dictionaries. Because source queries were OOV terms, their translations were also OOV terms. If the candidate terms could be found in our bilingual dictionary, they were deleted from the candidate list to improve the quality of the list. For example, "world" could be found in the bilingual dictionary, so we deleted it from the list of candidate terms.

3.4. *Selection of Translations*

In the last step, appropriate translation(s) of the source term were chosen in the candidate set mined in foregoing steps. To remove noise, we utilized some features in the system. An approach which integrates the surface pattern, frequency–distance, and phonetic features of the Chinese candidates was adopted to choose the ultimate translation(s).

3.4.1. Frequency–Distance Model

In this model, the most effective features were taken into account, that is, the frequencies of candidate terms and the distance from candidates to the source term. It is intuitive that authentic translations of the source term frequently appear at the same time as the source term. The shorter the distance from a candidate to the source term, the higher the probability that the candidate is the proper translation. The calculation for a measurement of a Chinese candidate and the source term was

$$F\_D(s,t) = \frac{\sum\limits_{J}\sum\limits_{k}\frac{1}{d_k(s,t)}}{\max_{fre-dis}}, \tag{4}$$

where s, t, and $d_k(s,t)$ denote the source term, a candidate, and, the kth distance between s and t in one snippet (because s and t may occur at the same time several times in a snippet), respectively. J and K separately denote the quantities of the snippets and co-occurrences between s and t; and the parameter $\max_{fre-dis}$ is the maximum reciprocal for the distances of all candidates. The distance was represented by the word amount between t and s, instead of the byte distance, because snippets probably comprise various differently encoded symbols, including Chinese and English characters, punctuation marks, and so forth. Therefore, taking the word amount to measure the distance mirrors more linguistic information. In the case that there were no characters between t and s, the distance measured was 1. This model was used as the baseline.

3.4.2. Match Modeling of Surface Patterns

Some users of Asian languages (Chinese, Japanese, etc.) often add annotations of terms by placing English translations in parentheses. Lin et al. [30] put forward the heuristic information for extracting translations from web pages. Such punctuation marks between source terms and translations are carriers of key information. Therefore, we can utilize the useful information to improve the quality of eventual translation(s). In the proposed solution, we submitted pairs of several English–Chinese terms into a search engine to automatically obtain surface patterns. Table 1 displays some of the surface patterns obtained using this process.

Table 1. Surface patterns obtained from the results of the search engine.

No.	Surface Patterns
1	E (C, C(E, E (C, C (E
2	E [C, C[E, E [C, C [E
3	E.C, C.E
4	E,C C,E
5	E>>(C
6	E』(C
7	E|C
8	E-C

In the case that candidate terms are found fitting with surface patterns in a set of bilingual snippets, there will be a much higher possibility of obtaining a correct translation. The formula for the contribution of matching with surface patterns is

$$SP(s,t) = \frac{N_{matching}}{\max_{num}} \tag{5}$$

where s, t, and $\max_{num}$ respectively denote a source term, a candidate term (the numerator represents the amount of times for s and t matching surface patterns), and the largest quantity of matching surface patterns of all candidates.

3.4.3. The Transliteration Model

Proper nouns (i.e., names of persons, places, and so forth) occupy a great proportion of OOV terms. The translation of many of these words is realized according to pronunciation—that is, transliteration. Some relevant studies have been conducted to mine the translations of terms in accordance with transliteration technologies [31]. According to these techniques, an English name is converted to phonetic representations, which are converted to symbols of Chinese pin-yin (phonetic sequences). Finally, pin-yin sequences are translated to sequences of Chinese characters. Our transliteration model is different from others in two perspectives. Firstly, it is a kind of matching problem. As Chinese candidates were obtained, it was not necessary to yield transliterations in Chinese. As for the second difference, to avoid the occurrence of dual errors in conversions from English phonetic representations to pin-yin and from Chinese pin-yin to characters, an idea was applied to partition English names into a sequence of syllables, and the probability of the co-occurrence of English syllables and Chinese characters was calculated to estimate the probability. This aimed to calculate phonetic similarity and finally to choose correct translations. Firstly, English terms were partitioned into a sequence of syllables according to heuristic rules, followed by using the following equation to calculate the transliteration cost:

$$Trl(s,t) = \frac{P(s,t)}{D(s,t)}, \tag{6}$$

where $P(s,t)$ denotes the possibility of the co-occurrence of s and t, and its definition is

$$P(s,t) \approx \prod_{i=1}^{\min(m,n)} (1 - \gamma_1) prob(e_i, c_i), \tag{7}$$

where γ_1, $prob(e_i, c_i)$, and $D(s,t)$ represent the smoothing weight, the possibility of the co-occurrence of an English syllable e_i and a Chinese character c_i (calculated in accordance with dynamic programming in a training corpus consisting of 37,665 pairs of proper nouns), and the number of syllable differences between the English term s and the Chinese candidate t, respectively. The definition of $D(s,t)$ is

$$D(s,t) = \varepsilon + \left| m - n \right|, \tag{8}$$

where ε, m, and n represent a decaying parameter, the total quantity of syllables of English terms, and the total quantity of Chinese characters, respectively.

To avoid false transliteration mapping of English syllables with Chinese characters, forward and backward mappings were combined. The final transliteration cost is

$$Trl(s,t) = \frac{Trl_F(s,t) + Trl_B(s,t)}{2}, \tag{9}$$

where $Trl_F(s,t)$ and $Trl_B(s,t)$ are forward and backward transliteration values, respectively.

3.4.4. Combination of Features

According to Fang et al. [24], the overall feature cost is the linear combination of all features used. This research proposes a different strategy for integrating features. Taking the frequency–distance model as the baseline, the results were reranked using the matching model of surface patterns. Because source terms were not all transliteration terms, the overall performance may have degraded in the experiments of the current research if the value of the transliteration model was directly integrated. If the source terms were transliterated, the value of the transliteration model would become more useful. Experiments were performed to calculate the threshold of the transliteration model. Afterwards, once transliterated values of candidates exceeded the threshold, candidates were reranked by transliterated values, or values of the transliteration model were omitted.

4. Experimental Evaluation

Evaluation experiments were carried out for the proposed method of term translation. The test set contained query terms taken from NTCIR5 and NTCIR4 CLIR tasks. In total, 110 queries were collected, in which each English query corresponded to a Chinese query. English query terms were mined from the field of English Title, including 129 terms not available in the English–Chinese dictionary (286,932 English–Chinese pairs). In other words, there were 129 OOV terms in the test set. We utilized Chinese translation as the gold standard. The evaluation metric was the top n inclusion rates, having the definition of the percentage of terms. Translations of the metric were included in the top n translations returned. For the purpose of evaluating the effects of the quantity of snippets, experiments were performed using test sets consisting of 50, 100, and 150 snippets. Results are reported in Table 2.

Table 2. Top *n* inclusion rates in the mining results of term translations obtained by changing the number of snippets.

No.	Method	Top 1	Top 3	Top 5	Top 10
50	Baseline	45.8%	68.2%	75.2%	89.1%
	Our approach	58.1%	77.5%	80.6%	89.1%
100	Baseline	66.7%	80.6%	82.2%	91.5%
	Our approach	73.6%	87.6%	89.1%	94.6%
150	Baseline	66.7%	77.5%	82.9%	90.7%
	Our approach	73.6%	82.2%	87.6%	91.5%

Here, the baseline system was based on the frequency–distance model, and our approach was proposed by integrating the three models described in Section 3. It can be observed from Table 2 that, despite the good performance of the baseline alone, the combined method significantly enhanced the performance of results under the condition of different numbers of snippets. That is, the matching model of surface patterns and the transliteration model were integrated effectively with the baseline. Our approach improved the top 1 inclusion rate to 73.6% from 58.1% (an improvement of 15.5%). When there were 100 snippets, our approach showed a higher accuracy of approximately 94.6% or so for the top 10 candidates, increasing by 5.5% in comparison with the condition using 50 snippets. Meanwhile, the top 3, 5, and 10 inclusion rates under the condition of 150 snippets decreased relative to those when there were 100 snippets. This was because the last 50 had less relevance, and introduced greater noise and wrong Chinese candidates. For this reason, the application of a larger number of snippets will not always lead to performance improvement. If there were 100 snippets, the integrated approach achieved the highest performance: 87.6% for the top 3 and 94.6% for the top 10 inclusion rates.

The proposed method was also compared to BabelFish (https://www.babelfish.com/, a machine translation engine on the web), the chi-squared and context vector method in [22], and the combined method in [26]. Chi-squared and context vector is a popular web-based translation mining method. Among the remaining three methods, they expanded each term using different models and then collected 100 snippets to extract candidates. Table 3 shows the results.

Table 3. Top *n* inclusion rates of term translations mined using machine translation and translation based on web pages.

Method	Top 1	Top 3	Top 5	Top 10
BabelFish	41.9%	N/A	N/A	N/A
Chi-squared and context vector	30.2%	44.2%	55.8%	79.1%
Combined method in [26]	61.2%	74.4%	82.2%	90.7%
Our approach	73.6%	87.6%	89.1%	94.6%

Despite its high performance in translating paragraphs and sentences, BabelFish is not applicable to short terms with no contextual information. The cross-linguistic expansion process is not used in the chi-squared and context vector method, and sometimes the context vector misleads the selection if incorrect translations are distributed in a manner conforming to that of source terms. The combined method in [26] performed better, for which the top 1 increased by 31% and the top 3 inclusion rate increased by 30.2%. Based on the combined method, we updated the POS tool, the algorithm of FCMAI, and the strategy of the combined method to be more effective. Our combined method performed the best, improving the top 1 inclusion rate by 12.4%, while the top 3 inclusion rate had a 13.2% improvement. This indicates that the three models in our combined method complement each other and the improvement of the algorithm was effective.

5. Conclusions

This study proposed an approach for mining the translations of unknown terms based on the web and it is an extension of our research team's work. The first key step was a method based on co-occurrence information for expanding source queries to gather valid bilingual snippets. The study then employed the FCMAI to mine valid candidates from small bilingual corpora containing noises. To select the correct translations, we adopted an approach based on surface patterns, frequency–distance, and transliteration modeling. Meanwhile, we made improvements in our experiment based on the work [26], especially in the areas of the FCMAI algorithm, the strategies of the combined method, the preprocessing of corpus data (i.e., POS tagging), and so on. It can be observed from the experimental results that the proposed approach performed well in mining the translations of unknown terms.

To improve the quality of translation mining and to carry out experiments on larger test sets, more features will be considered in further research, including semantic analysis results. For better query translation, performance in retrieving cross-linguistic information must be enhanced. We also plan on validating this concept further by conducting experiments.

Author Contributions: Conceptualization, B.L. and J.Y.; methodology, B.L. and J.Y.; software, B.L. and J.Y.; validation, B.L. and J.Y.; formal analysis, B.L.; investigation, B.L.; resources, J.Y.; writing—original draft preparation, B.L.; writing—review and editing, J.Y.; project administration, B.L. and J.Y.; funding acquisition, B.L. and J.Y.

Funding: This research was funded by the National Natural Science Foundation of China under Grant No. 61672367, the Key Program of the Foundation for Young Talents in the Colleges of Anhui Province under Grant No. 2013SQL097ZD, the Program of Quality Engineering in the Colleges of Anhui Province under Grant Nos. 2017SJJD101 and 2018XFSYXM010, the Key Program of the Natural Science Foundation of Anhui Education Committee under Grant KJ2019A0969, and the Foundation of Anhui Open University under Grant No. JY19-01.

Acknowledgments: The authors express their gratitude to the anonymous reviewers for their carefulness and patience.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ballesteros, L.; Croft, W.B. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, 27–31 July 1997; pp. 84–91.
2. Pirkola, A.; Hedlund, T.; Keskustalo, H.; Järvelin, K. Dictionary-based Cross-language Information Retrieval: Problems, Methods, and Research Findings. *Inf. Retr.* **2001**, *4*, 209–230. [CrossRef]
3. Pirkola, A.; Toivonen, J.; Keskustalo, H.; Järvelin, K. FITE-TRT: A High Quality Translation Technique for OOV Words. In Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, 23–27 April 2006; pp. 1043–1049.
4. Sharma, V.K.; Mittal, N. Cross-Lingual Information Retrieval: A Dictionary-Based Query Translation Approach. In *Advances in Computer and Computational Sciences*; Springer: Singapore, 2018; Volume 2, pp. 611–618.
5. Tufiş, D.; Barbu, A.M.; Ion, R. Extracting Multilingual Lexicons from Parallel Corpora. *Comput. Hum.* **2004**, *38*, 163–189. [CrossRef]
6. Piperidis, S.; Harlas, I. Mining Bilingual Lexical Equivalences out of Parallel Corpora. In Proceedings of the 4th Hellenic Conference on Artificial Intelligence, Heraklion, Greece, 18–20 May 2006; pp. 311–322.
7. Liu, L.; Ge, Y.D.; Yan, Z.X.; Yao, J.M. A CLIR-oriented OOV Translation Mining Method from Bilingual Webpages. In Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, China, 10–13 July 2011; pp. 1872–1877.
8. Macken, L.; Lefever, E.; Hoste, V. TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology* **2013**, *19*, 1–30. [CrossRef]
9. Widdows, D.; Dorow, B.; Chan, C.K. Using Parallel Corpora to enrich Multilingual Lexical Resource. In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, 29–31 May 2002; pp. 240–245.

10. Morin, E.; Hazem, A. Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 23–25 June 2014; pp. 1284–1293.
11. Otero, P.G.; José, R.P.C. An Approach to Acquire Word Translations from Non-parallel Texts. In Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence, Covilha, Portugal, 5–8 December 2005; pp. 600–610.
12. Kwon, H.; Seo, H.; Kim, J. Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, 8 August 2013; pp. 11–15.
13. Linard, A.; Daille, B.; Morin, E. Attempting to Bypass Alignment from Comparable Corpora via Pivot Language. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, Beijing, China, 30 July 2015; pp. 32–37.
14. Vulić, I.; Moens, M.F. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 719–725.
15. Vulić, I.; Moens, M.F. Bilingual Distributed Word Representations from Document-Aligned Comparable Data. *J. Artif. Intell. Res.* **2016**, *55*, 953–994. [CrossRef]
16. Hazem, A.; Morin, E. Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 3401–3411.
17. Hazem, A.; Daille, B. Word Embedding Approach for Synonym Extraction of Multi-Word Terms. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018; pp. 297–303.
18. Hazem, A.; Daille, B. Semi-Compositional Method for Synonym Extraction of Multi-Word Terms. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 1201–1207.
19. Rigouts, T.A.; Hoste, V.; Lefever, E. A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. In Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018; pp. 1803–1808.
20. Nagata, M.; Saito, T.; Suzuki, K. Using the Web as a Bilingual Dictionary. In Proceedings of the Workshop on Data-Driven Methods in Machine Translation, Toulouse, France, 7 July 2001; pp. 1–8.
21. Lu, W.H.; Chien, L.F.; Lee, H.J. Translation of Web Queries Using Anchor Text Mining. *ACM TALIP* **2002**, *1*, 159–172. [CrossRef]
22. Cheng, P.J.; Teng, J.W.; Chen, R.C.; Wang, J.H.; Lu, W.H.; Chien, L.F. Translating Unknown Queries with Web Corpora for Cross-language Information Retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004; pp. 146–153.
23. Huang, F.; Zhang, Y.; Vogel, S. Mining Key Phrase Translations from Web Corpora. In Proceedings of the Conference on Human Language Technology & Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 483–490.
24. Fang, G.; Yu, H.; Nishino, F. Chinese-English Term Translation Mining Based on Semantic Prediction. In Proceedings of the COLING/ACL on Main Conference Poster Session, Sydney, Australia, 17–18 July 2006; pp. 199–206.
25. Sun, J.; Yao, J.; Zhang, J.; Zhu, Q. Web Mining of OOV Translations. *J. Inf. Comput. Sci.* **2008**, *6*, 97–103.
26. Ge, Y.D.; Hong, Y.; Yao, J.M.; Zhu, Q.M. Improving Web-Based OOV Translation Mining for Query Translation. In Proceedings of the Asia Information Retrieval Symposium, Taipei, China, 1–3 December 2010; pp. 576–587.
27. Pal, S.; Naskar, S.K.; Zampieri, M.; Nayak, T.; Van, G.J. Catalog Online: A Web-based CAT Tool for Distributed Translation with Data Capture for Ape and Translation Process Research. In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 98–102.
28. Berger, A.; Lafferty, J. Information Retrieval as Statistical Translation. In *ACM SIGIR Forum*; ACM: New York, NY, USA, 2017; Volume 51, pp. 219–226.

29. Lu, C.; Xu, Y.; Geva, S. Web-based Query Translation for English-Chinese CLIR. *Comput. Linguist. Chin. Languist. Process.* **2008**, *13*, 61–90.
30. Lin, D.; Zhao, S.; Van Durme, B.; Paşca, M. Mining Parenthetical Translations from the Web by Word Alignment. In Proceedings of the ACL-08: HLT, Columbus, OH, USA, 16–18 June 2008; pp. 994–1002.
31. Lam, W.; Huang, R.; Cheung, P.S. Learning Phonetic Similarity for Matching Named Entity Translations and Mining New Translations. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25–29 July 2004; pp. 289–296.