Bilingual Product Name Dictionary Construction Using a Two Stage Method

Yatian Shen, Xuanjing Huang
School of Computer Science, Fudan University, Shanghai, China
shenyatian@gmail.com, xjhuang@fudan.edu.cn

Abstract

This paper proposes a novel two-stage method for bilingual product name dictionary construction from comparable corpora. In previous work, some researchers study the problem of expanding a set of given seed entities into a more complete set by discovering other entities that also belong to the same concept, it just solves the problem about expansion of entity set in a monolingual language, but the expansion of bilingual entity is really blank problem from comparable corpora. A typical example is to use “Honda-本田”as seed entity, and derive other entities(e.g., “Ford-福特”) in the same concept set of product name. We address this problem by utilizing a two-stage approach based on entity set expansion and bilingual entity alignment from comparable corpora. Evaluations using English and Chinese reviewer corpus verify that our method outperforms conventional methods.

1 Introduction

Bilingual lexicons are important resources for bilingual tasks such as machine translation (MT) and cross-language information retrieval (CLIR). Therefore, the automatic building of bilingual product name lexicons from corpus is one of the important issues, however it has not attracted many researchers. As a solution, a number of previous works have been proposed for extracting bilingual product name lexicons from comparable corpora, in which documents are not direct translations but share the same topic or domain. The use of comparable corpora is motivated by the fact that large parallel corpora are only available for a few language pairs and limited domains.

Bilingual product name lexicon is similar to traditional bilingual lexicon extraction, what they are all common on is extract bilingual entity translation pair from comparable corpora, but there is some difference between them. Our problem is: first given an seed set for semantic classes, finding the conceptually entities by extending semantic classes. Then, the bilingual entity translation pairs are extracted from comparable corpora. Traditional bilingual lexicon extraction approaches can only find entity translation pairs from comparable corpora, but not expand semantic set.

Set expansion systems provide us a useful solution to the above problem because they create a more perfect set of name entities by expanding the small number of seed words given for the target domain. Google Sets is a well-known example of a web-based set expansion system. Another prominent work is the SEAL system (Wang and Cohen, 2007; Wang and Cohen, 2008; Wang and Cohen, 2009), which adoptes a two-phase strategy, where they first build customized text wrappers based on the input seeds in order to exact candidate entities from web pages. Then a graph-based random walk approach is used to rank candidate entities based on their closeness to the seeds on the graph. The third method is set expansion by iterative similarity aggregation (He and Xin, 2011), in which a set of given seed entities is expanded into a more complete set. All these methods are entity expansion from monolingual data sources.

Another meaningful work is the bilingual lexicon extraction (Fung and McKeown, 1997; Rapp, 1999; Andrade et al., 2010; Fišer et al., 2011; Daille and Morin, 2005; Vulic et al., 2011; Andrade et al., 2011; Bo et al., 2011). Most of the

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 61–69,
Wuhan, China, 20-21 October 2014

previous methods are based on the assumption that a word and its translation tend to appear in similar contexts across languages. Based on this assumption, many methods calculate word similarity using context and then extract word translation pairs with a high context similarity. while their researches aim to generate a general bilingual lexicons, our work is bilingual entity extraction of the same semantic category, these entities are refer to product name.

Considerable progresses have been made in developing high-quality set expansion systems in the monolingual setting. while bilingual product name dictionary construction and extraction still do not attract much research attention. For bilingual product name dictionary construction, there are two major fundamental problems. The first is generating an extensive list of the same semantic entity, while some seed entities of the same concept are given as input. The second problem is to find bilingual entity translation from comparable corpora.

Facing the above problems, we present a novel approach to construct bilingual product name dictionary in this paper. In order to express the simplification, we will replace word ”product name” with ”entity” each other. Following the common practice, our system proceeds in two stages, which first expands the entity set for the semantic category by giving some bilingual set pairs and then finds bilingual product name translation pair from comparable corpora. Semantic category set expansion is carried out through the bootstrapping algorithm. In this stage ,our goal is to discover relevant entities by giving some entity seed set. In the second stage, we use this assumption that a word and its translation tend to appear in similar context across languages (Rapp, 1999). Our method calculates entity similarity using context and then extract entity translation pairs with a high context similarity. We call this method as context-similarity-based methods. The context similarity is usually computed using machine translation model by mapping contexts expressed in two different languages into the same language space. In the mapping process, information not represented by the seed lexicon is discarded.

The main contributions of this paper are as follows: 1) we propose a bilingual product name extraction method that can get the set of semantic category by bootstrapping. At the same time, we can find bilingual product name translation pairs based on context similarity from comparable corpora. 2) we propose an the algorithm that can not only build set of semantic category by giving some bilingual seed set but also find entity translation pairs from comparable corpora. 3) we construct a dictionary of the bilingual product name from comparable corpora, which do not need fully parallel data that is seldom.

2 Related Work

There is a significant body of related work in the broad space of information extraction and named entity extraction. We will only summarize work most relevant to set expansion and bilingual entity extraction due to the limit of space.

Google sets does set expansion using propriety algorithms which are not publicly available. (He and Xin, 2011) expand seeds by iterative similarity aggregation. (Talukdar et al., 2006) studied the problem of set expansion of open text, which proposes to automatically identify trigger-words which indicate patterns in a bootstrapping manner. (Ghahramani and Heller, 2005) used the method of Bayesian inference to solve the problem of set expansion. In comparison, our approach expands bilingual entity seeds set by using bootstrapping algorithms ,which learn entity candidates and their corresponding patterns iteratively. Our goal is to find the same semantic concept set .

(Fung and McKeown, 1997) present a statistical word feature that is said to the word relation matrix, which can be used to find translated pairs of words and terms from non-parallel corpora across language groups. (Daille and Morin, 2005) proposes a method of extracting bilingual lexicon composed of single-word terms (SWTs) and multi-word terms (MWTs) from comparable corpora of a technical domain. First, this method extracts MWTs in each language, and then uses statistical methods to align single words and multiword terms by exploiting the term contexts. The alignment of words in translated texts are well established, this algorithm is used to identify word translations (Rapp, 1999). (Andrade et al., 2010) suggest a new method which selects a subset of words (pivot words) associated with a query and then matches these words across languages, a new Bayesian method for estimating Point-wise Mutual Information is used to detect word associations. (Fišer et al., 2011) presents a series of exper-

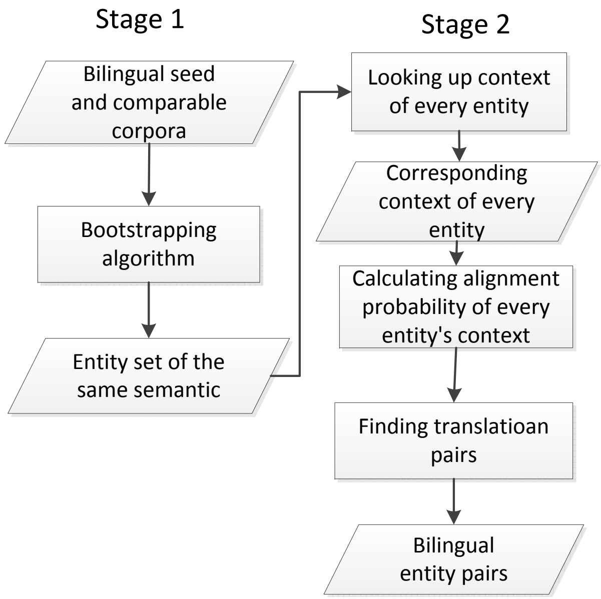
Stage 1
Stage 2
Bilingual seed and comparable corpora
Bootstrapping algorithm
Entity set of the same semantic
Looking up context of every entity
Corresponding context of every entity
Calculating alignment probability of every entity's context
Finding translatioan pairs
Bilingual entity pairs

Figure 1: Flow chart of our two-stage system.

iments aimed at inducing and evaluating domain-specific bilingual lexicon from comparable corpora. (Vulic et al., 2011) investigate the algorithm of bilingual topic models, which finds translations of terms in comparable corpora by using knowledge from word-topic distributions. (Andrade et al., 2011) propose to perform a linear transformation of the context vectors, the new word translations are found by context similarity. (Bo et al., 2011) introduce a clustering-based approach for enhancing corpus comparability which exploits the homogeneity feature of the corpus, and preserves most of the vocabulary of the original corpus. (Tamura et al., 2012) proposes a novel method for lexicon extraction that extracts translation pairs from comparable corpora by using graph- based label propagation.

All the methods mentioned above may potentially extract entities translation pairs when context of entities are similarity. We are also based on this assumption, but we are different from the previous models where we use machine translation model to map the context of entity to the same language space, which can improve performance and illustrate robustness.

3 Proposed Method

Figure 1 illustrates the framework of our proposed methods. The proposed method has the following components: Bootstrapping algorithm is used to get entity sets and the patterns for Chinese and English respectively, then we can find entity translation pairs by calculating context similarity and construct bilingual product name lexicon.

Step 1. Using Bootstrapping algorithm gets entity sets and the patterns for Chinese and English respectively.

Step 2. Based on the assumption that the word and its translation tend to appear in similar contexts across languages, we can find translation pair.

Step 3. Construct bilingual product name lexicon.

4 Bootstrapping for Entity Set Expansion

In this paper, we expand seed entity set into a more complete set by discovering other entities that also belong to the same concept set. A typical application is to use seed entities to derive other entities in the same concept set of brands. In order to discover such relevant entities, we expand seed entities to assign semantic similar entities to the same semantic set using plenty of user reviews.

4.1 Growing Seed Dictionary

We focus on the problem of how to grow the seed dictionary and discovering new product names from user reviews. In this section, we use the seed entity to automatically generate semantic lexicons. For the specific case of brand discovery, this initial list used to generate semantic lexicons must contain only names that are unambiguously. We hence remove ambiguous names or phrases that belong to multiple entity types from the dataset, and only choose those entities as entity seed that it owns definite semantic. We used a weakly supervised bootstrapping algorithm that automatically generates semantic lexicons (Thelen and Riloff, 2002).

Bootstrapping algorithms hypothesizes the semantic class of entity by gathering collective evidence about semantic associations from extraction pattern context. For our representation of extraction patterns, we used the AutoSlog system (Riloff, 1996), AutoSlog's extraction patterns represent linguistic expressions that extract a noun phrase in one of three syntactic roles: subject, direct object, or prepositional phrase object. Before bootstrapping begins, we run AutoSlog exhaustively over the corpus to generate an extraction pattern for every noun phrase that appears. In these, noun or noun phrase are entities, we will possibly extract them as production name. The patterns are then applied to the corpora and all of theirs extracted noun phrases are recorded. For every iteration, the top 20 extraction patterns are put into a pattern pool. Every pattern used the R-logF metric that has been used for extraction pattern learning (Riloff, 1996).

All entities in the candidate entity pool are scored and the top five words are added to the semantic lexicon. Bootstrapping algorithm learns pattern that associate entity to their correct expansions, the intuition of our work is that the algorithm learns context that can associate some entities that have the same semantic.

5 Finding Translation Pairs

Translating domain-specific entities from one language to another is challenging because they are often not listed in a general dictionary. In this section, we are based on this assumption that context similarity is helpful since two words with identical meaning are often used in similar contexts across languages(Rapp,1999). Let us briefly recall the main idea for using context similarity to find translation pairs. First, the context pattern of every entity is found because the context of a entity is usually defined by the word which occur around it(bag-of-words model), we use ten forward and backward window of word as context. Second, we use machine translation model to translate context, the context of two entities can be aligned about their probability. At last, if the context of two entities is similar, so they corresponds to entity pairs as bilingual product name pairs. The detail algorithms is as follows:

5.1 Looking Up Context of Every Entity

With the bootstrapping algorithm, we get the set of semantic category entity in English and Chinese comparable corpora. For every entity, we look up their context, and use the method of string matching in the corpora. We use 3 forward and backward window of word as context, That is what we call the context. The context and their corresponding entities have great relevance. As an example, it is easier for us to find some words around ”Camera” name,such as the pixel,the screen and cmos, these words are the context of entity, which often appear near the name of camera. By context, we are able to find their corresponding entities.

Algorithm 1 Finding translation pairs in bilingual

```
Input: $I = (x_i), i = 1, 2, \ldots, l$ in which $x_i$ is
    the ith entity of the same semantic entity set,
    BilingualData is bilingual comparable corpora
Output: Entity tanslation pairs
repeat
  for $i = 1\ to\ n$ do
    Looking up the context of every entity
      $context_i$ in BilingualData ;
    Calculating the alignment probability of
      every entity's context in different languages
    Computing similarity of the context between $context_i$ and $context_j$
    if Similarity($context_i$,$context_j$)is maximal then
      For the highest similarity value of context to corresponding entity pair,
        extracting them as an entity translation pair
    end if
  end for
until no $x_i$ is in the $I$ during iteration
```

5.2 Aligning the Context of two Entities

To bilingual context, how they are aligned with each other is a major problem. This component is to identify equivalence relation in every entity corresponding to bilingual context. We assume that the same context appears around the same entity. Thus, our aim is to find translation pairs between Chinese and English corpora. Machine translation is commonly used to complete the task. By the tool of machine translation, two different language context of entity is mapped to the same language space.

Many studies on machine translation use GIZA++ as their underlying word-by-word alignment system. Machine translation systems have also benefited from such alignment, performing it at the character level (AbdulJaleel and Larkey, 2003), (Virga and Khudanpur, 2003), (Gao et al., 2005). GIZA++ is a statistical machine translation toolkit freely available for research purposes. The extended version of this toolkit is called GIZA++ and was developed by (Och and Ney, 2003). We employ the word-based translation model to perform context alignment, we get the alignment probability between the context pattern of two different entities. GIZA++ alignment system is trained on parallel corpora English and Chinese reviews, we manually annotate the context of bilingual entity pair on 3000 parallel sentence pairs about car domain reviews. A probability table about the context of bilingual entity pair is generated by training GIZA++ model.

5.3 Entity Translation Extraction

In order to find entity translation pairs in different languages, we use statistical machine translation toolkit GIZA++ to calculate the alignment probability of every entity's context in different languages. A pair of entity is treated as a bilingual product name pair when the alignment probability of their context is high. In this, if the alignment probability of four words which is said to context is greater than threshold, we will think that entity pairs which have this context are bilingual entity pair, We found that the word alignment probability threshold of the context is set to 0.53 is a good choice by experiment.

6 Experiments

6.1 Dataset and Evaluation Metrics

In order to evaluate our approach, we conduct experiments on two real data sets, which are from collection of brand reviews including digital cameras and car domains. For the target language of English, the product dataset contains 9542 reviews which are collected from www.buzzilions.com and www.carreview.com. For the source language of Chinese, the product dataset contains 8432 reviews which are collected from www.Amazon.cn and www.xche.com.cn. For our experiment,we use a Oxford English-Chinese bilingual dictionary to match similarity semantic reviewer sentence, any two of them are used as comparable corpus,the copora are non-parallel, but loosely compara in term of its content. Though the scale of Chinese corpora is large, most of the reviews are short texts and there are a lot noise in the content. For Chinese, we use the ICTLAS 3.0 (Zhang et al., 2003) toolkit to conduct word segmentation over sentences.

To evaluate the effectiveness of our algorithms, we select two semantic entity sets in camera domain and car domain as seeds, where set expansion experiments are conducted . We select these two categories because (1) they are from different domains; and (2) they have different degree of difficulty for finding entity translation pairs.

Language	Domain	#Sentence	#Reviews
Chinese	Camera	2480862	1566
	Car	3526109	2103
English	Camera	1090862	4506
	Car	2563120	5036

Table 1: Statistics on English corpus about Camera and Car domain. # denotes the size of the reviews/sentences

In experiments, each English review is segmented into sentences according to punctuation. Then sentences are tokenized and the part-of-speech of each word is assigned. Stanford NLP tool is used to perform POS-tagging. Next, function words were removed since function words with little semantic information spuriously co-occurred with many words. Table 1 shows the size of each corpora.

We measure the performance on product name translation pair extraction as Top N accuracy (Acc_N), which is the number of test words whose top N translation candidates contain a correct translation equivalent over the total number of test words. We randomly select 50 Chinese words as our test data. We manually evaluate whether translation candidates contained a correct translation equivalent. We do not use recall because we do not know whether the translation equivalents of a test word appear or not in the corpus.

6.2 Example Output

Table 2 lists the top 20 ranked results produced by two stage algorithm for the two domains that we experiment with. In each domain, those terms in boldface are the input seeds. The underlined terms are the results that do not belong to the ground truth set and thus counted as incorrect results. While the remaining terms are correct results expanded from the input seeds.

From Table 2, we can see that in the top-20 ranked results, the “Camera” domain have high precision. “Camera” domain has only two incorrect result, the top-20 results for “Car” domain, however, includes some noisy entities that are incorrect, such as product workshop names (“大众汽车” and “Audi company”), and the similarity concept name (“Honda car” and “福特汽车公司”).

Camera	Car
富士-FUJIFILM	奥迪-audi
卡西欧-Casio	宝马-BMW
徕卡-Leica	别克-Bulk
柯达-Kodak	福特-Ford
理光-Ricoh	**福克斯-Focus**
索尼-SONY	本田-Honda
奥林巴斯-OLYMPUS	马自达-Mazda
松下-Panasonic	**丰田-Toyota**
佳能-Canon	尼桑-Nissan
尼康-Nikon	丰田皇冠-Toyota Crown
宾得-Pentax	textbf沃尔沃-Volvo
康佳-Konka	大众-Volkswagen
柯尼卡-Konica	马自达6-Mazda6
尼康S2-Nikon S2	日制汽车-Honda
佳能 VTD-Canon VTD	奔驰-Benz
尼康-Konka	本田雅阁-Honda
三星-SAMSUNG	雷克萨斯-Lexus
佳能-Nikon	现代-Hyundai
美能达-Minolta	通用-GM
柯尼卡-Konica	雪铁龙-Citroen

Table 2: Top -20 results by two stage method

6.3 Our Methods VS. State-of-art Methods

To prove the effectiveness of our method, we select the following state-of-art methods as baseline for comparison.

1) Rapp is a typical context-similarity-based method (Rapp, 1999). Context words are words in a window (window size is 10) and are treated separately for each position. Associations with context words are computed using the log-likelihood ratio. The similarity measure between context vectors is the city-block metric.

2) Andrade is a sophisticated method in context-similarity-based methods (Andrade et al., 2010). Context is a set of words with a positive association in a window (window size is 10). The association is calculated using the PMI estimated by a Bayesian method, and a similarity between contexts is estimated based on the number of overlapping words.

3)Tamura proposes a method for lexicon extraction that extracts translation pairs from comparable corpora by using graph-based label propagation (Tamura et al., 2012). They utilize indirect relations with the bilingual seeds together with direct relations, in which each word is represented by a distribution of translated seeds. The seed distributions are propagated over a graph representing relations among words, and translation pairs are extracted by identifying word pairs with a high similarity in the seed distributions.

6.4 Experiments Results

Table 3 and Table 4 show the performance of each method using Car and Camera review dataset. Table 3 and Table 4 show that the proposed methods outperform the baselines on both datasets. The results show that expansion of bilingual product name by using two stage algorithm is effective .

Rapp's method computed associations with context words using the log-likelihood ratio. The city-block metric is used to compute similarity between context vector. Andrade define context as a set of words with a positive association in a window, Pointwise Mutual Information estimated by a Bayesian method is used to calculate. The similarity between contexts is estimated based on the number of overlapping words. Tamura's method utilize indirect relations with the bilingual seeds together with direct relations, in which each word

Methods	Acc_1	Acc_{10}	Acc_{20}
Rapp	1.6%	2.5%	3.9%
Andrade	1.8%	3.2%	4.1%
Tamura	2.5%	5.8%	7.5%
Ours	4.5 %	8.6%	12.4%

Table 3: Performance statistics on Camera domain by using Top N accuracy (Acc_N).N is 1,10,20 respectively.

Methods	Acc_1	Acc_{10}	Acc_{20}
Rapp	1.5%	2.3%	4.5%
Andrade	1.7%	3.6%	5.1%
Tamura	2.3%	6.2%	8.5%
Ours	4.3 %	9.6%	13.8%

Table 4: Performance statistics on Car domain by using Top N accuracy (Acc_N).N is 1,10,20 respectively.

is represented by a distribution of translated seeds. Then they extracts translation pairs from comparable corpora by using graph-based label propagation. The parameter setting in these three baselines are the same as the original papers. The overall performance results are shown in Table 3 and 4. From these results, we can make the following observations.

1) Ours achieves performance improvement over other methods. This indicates that our method is effective for bilingual product name extraction.

2) Our two stage method outperform Rapp's method, Andrade's method and Tamura's method. The reason is that two stage-based method extract bilingual entity name in a flexible way, we first consider entity set expansion, then find bilingual entity pair by using machine translation methods from comparable corpora, which is not only find the same semantic entity, but also can find entity translation pair, so we can extract bilingual product name on specific domain. but Rapp's method, Andrade's method and Tamura's method only build a general bilingual lexicon.

3) Our method construct context association by utilizing machine translation model between bilingual entity name. Machine translation model have the characteristic of accurate and interpretation, which favor our problems. Our test data, on the other hand, includes many low-frequency words. It is generally true that translation of high-frequency words is much easier than that of low frequency words. The accuracies of the baselines in Table 3 and 4 are worse than the previous reports: 14% Acc_1 and 46% Acc_{10} (Andrade et al., 2010), and 72% Acc_1 (Rapp, 1999).

4) Our methods expand entity name of the same semantic concept by using the bootstrapping algorithm, which is weak-supervised learning algorithm. The algorihtm need not labeled dateset to train model, meanwhile which is easier to implement it, it exceeds Tamura's method,which only considers distribution of translated seeds, then each word is represented by seeds distribution. The seed distributions are propagated over a graph representing relations among words, but constructing a graph is consuming lot of forces, its effect is very low.

6.5 Effect of Seeds Size

In this subsection, we aim to prove the effectiveness and robustness of our algorithms for bilingual entity extraction. We vary the number of input seeds and report the corresponding bilingual entity extraction performance. Specifically, given the 4, 6 and 8 seeds for each of the two domains in the experiments,we aim to test the performance of our two stage algorithm. The results are reported in Table 5, Table 6. The overall trend stands out that the performance of our algorithm with 6 seeds is in general much better and more stable than the case where only 4 or 8 seeds are used as input. We consider three kinds of the characters that the entity seed set have. The seed must be first the most representative of a semantic class, and polysemy of a seed should be avoided, we also consider the coverage of a seed set. This suggests that our algorithm is more robust when a reasonable number of seeds are given, and the performance may fluctuate with very few number of seeds, largely depending on the quality of the seeds given.

6.6 Effect of Translation Model

We can find entity of similar pattern by using GIZA++ model, but the alignment model result in some errors, there are two central reasons. Our test data includes words whose translation equivalence inherently cannot be found. The first of these types are words whose equivalence does not

Number	Acc_1	Acc_{10}	Acc_{20}
4	1.6%	2.5%	4.9%
6	2.7%	4.3%	6.5%
8	2.3%	3.9%	5.5%

Table 5: Performance statistics on Car domain by using Top N accuracy (Acc_N).The number of seeds choose 4,6 and 8 respectively.

Number	Acc_1	Acc_{10}	Acc_{20}
4	2.0%	3.5%	4.9%
6	2.7%	4.5%	7.4%
8	2.5%	4.1%	5.9%

Table 6: Performance statistics on Camera domain by using Top N accuracy (Acc_N).The number of seeds choose 4,6 and 8 respectively.

exist in the English corpus, which is an unavoidable problem for our methods based on comparable corpora. The second reason of errors is word sense ambiguity, which is different in every language, the Chinese word “宝马” means either “horse” or “car” in English, the proposed methods could not identify correct translation pairs. We will leave this word sense disambiguation problem for future work.

7 Conclusions

This paper proposes a novel two-stage method for product name dictionary construction from comparable corpora. The bootstrapping algorithm is used to expand bilingual product name in the first stage, Then in the second stage we find bilingual product name pair by calculating context similarity. The alignment model is used to Calculate alignment probability of every entity's context in different languages. Evaluations using English and Chinese comparable corpora outperforms conventional methods.

In future work, we are planning to investigate the following open problems : word sense disambiguation and translation of compound words in bilingual entity extraction. We are also planning an end-to-end evaluation, for instance, by employing the extracted bilingual product name into an machine translation system.

References

Nasreen AbdulJaleel and Leah S Larkey. 2003. Statistical transliteration for english-arabic cross language information retrieval. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 139–146. ACM.

Daniel Andrade, Tetsuya Nasukawa, and Jun'ichi Tsujii. 2010. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Daniel Andrade, Takuya Matsuzaki, and Jun'ichi Tsujii. 2011. Learning the optimal use of dependency-parsing information for finding translations with comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 10–18. Association for Computational Linguistics.

Li Bo, Eric Gaussier, Akiko N Aizawa, et al. 2011. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 473–478.

Béatrice Daille and Emmanuel Morin. 2005. French-english terminology extraction from comparable corpora. In *Natural Language Processing–IJCNLP 2005*, pages 707–718. Springer.

Darja Fišer, Špela Vintar, Nikola Ljubešić, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26. Association for Computational Linguistics.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Wei Gao, Kam-Fai Wong, and Wai Lam. 2005. Phoneme-based transliteration of foreign names for oov problem. In *Natural Language Processing–IJCNLP 2004*, pages 110–119. Springer.

Zoubin Ghahramani and Katherine Heller. 2005. Bayesian sets.

Yeye He and Dong Xin. 2011. Seisa: set expansion by iterative similarity aggregation. In *Proceedings of the 20th international conference on World wide web*, pages 427–436. ACM.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*, pages 1044–1049.

Partha Pratim Talukdar, Thorsten Brants, Mark Liberman, and Fernando Pereira. 2006. A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 141–148. Association for Computational Linguistics.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36. Association for Computational Linguistics.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 214–221. Association for Computational Linguistics.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-language applications. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 365–366. ACM.

Ivan Vulic, Wim De Smet, Marie-Francine Moens, and KU Leuven. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pages 479–484.

Richard C Wang and William W Cohen. 2007. Language-independent set expansion of named entities using the web. In *Data Mining,2007.ICDM 2007. Seventh IEEE International Conference on*, pages 342–350. IEEE.

Richard C Wang and William W Cohen. 2008. Iterative set expansion of named entities using the web. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 1091–1096. IEEE.

Richard C Wang and William W Cohen. 2009. Character-level analysis of semi-structured documents for set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1503–1512. Association for Computational Linguistics.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics.