EM-based Hybrid Model for Bilingual Terminology Extraction from Comparable Corpora

Lianhau Lee, Aiti Aw, Min Zhang, Haizhou Li
Institute for Inforcomm Research
{lhlee, aaiti, mzhang, hli}@i2r.a-star.edu.sg

Abstract

In this paper, we present an unsupervised hybrid model which combines statistical, lexical, linguistic, contextual, and temporal features in a generic EM-based framework to harvest bilingual terminology from comparable corpora through comparable document alignment constraint. The model is configurable for any language and is extensible for additional features. In overall, it produces considerable improvement in performance over the baseline method. On top of that, our model has shown promising capability to discover new bilingual terminology with limited usage of dictionaries.

1 Introduction

Bilingual terminology extraction or term alignment has been well studied in parallel corpora. Due to the coherent nature of parallel corpora, various statistical methods, like EM algorithm (Brown et. al., 1993) have been proven to be effective and have achieved excellent performance in term of precision and recall. The limitation of parallel corpora in all domains and languages has led some researchers to explore ways to automate the parallel sentence extraction process from non-parallel corpora (Munteanu and Marcu, 2005; Fung and Cheung, 2004) before proceeding to the usual term alignment extraction using the existing techniques for parallel corpora. Nevertheless, the coverage is limited since parallel sentences in non-parallel corpora are minimal.

Meanwhile, some researchers have started to exploit comparable corpora directly in a new manner. The motivations for such an approach are obvious: comparable corpora are abundantly available, from encyclopedia to daily newspapers, and the human effort is reduced in either generating or collecting these corpora. If bilingual terminology can be extracted directly from these corpora, evolving or emerging terminologies can be captured much faster than lexicography and this would facilitate many tasks and applications in accessing cross-lingual information.

There remain challenges in term alignment for comparable corpora. The structures of texts, paragraphs and sentences can be very different. The similarity of content in two documents varies through they talk about the same subject matter. Recent research in using transliteration (Udupa et. al., 2008; Knight and Graehl, 1998), context information (Morin et. al., 2007; Cao and Li, 2002; Fung, 1998), part-of-speech tagging, frequency distribution (Tao and Zhai, 2005) or some hybrid methods (Klementiev and Roth, 2006; Sadat et. al., 2003) have shone some light in dealing with comparable corpora. In particular, context information seems to be popular since it is ubiquitous and can be retrieved from corpora easily.

In this paper, we propose an EM-based hybrid model for term alignment to address the issue. Through this model, we hope to discover new bilingual terminology from comparable corpora without supervision. In the following sections, the model will be explained in details.

Coling 2010: Poster Volume, pages 639–646,
Beijing, August 2010

2 System Architecture

It is expensive and challenging to extract bilingual terminologies from a given set of comparable corpora if they are noisy with very diverse topics. Thus the first thing we do is to derive the document association relationship between two corpora of different languages. To do this, we adopt the document alignment approach proposed by Vu et. al. (2009) to harvest comparable news document pairs. Their approach is relying on 3 feature scores, namely Title-n-Content (TNC), Linguistic Independent Unit (LIU), and Monolingual Term Distribution (MTD). In the nutshell, they exploit common words, numbers and identical strings in titles and contents as well as their distribution in time domain. Their method is shown to be superior to Tao and Zai (2005) which simply make use of frequency correlation of words.

After we have retrieved comparable document pairs, we tokenize these documents with prominent monolingual noun terms found within. We are interested only in noun terms since they are more informative and more importantly they are more likely not to be covered by dictionary and we hope to find their translations through comparable bilingual corpora. We adopt the approach developed by Vu et. al. (2008). They first use the state-of-the-art C/NC-Value method (Frantzi and Ananiadou, 1998) to extract terms based on the global context of the corpus, follow by refining the local terms for each document with a term re-extraction process (TREM) using Viterbi algorithm.

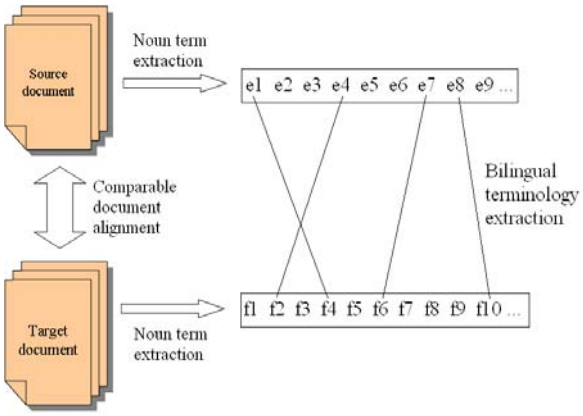

Figure 1. The procedure of bilingual terminology extraction from comparable documents.

After these preprocesses, we have a set of comparable bilingual document pairs and a set of prominent monolingual noun terms for each monolingual document. The aim of our term alignment model is to discover new bilingual terminology formed from these monolingual terms across aligned document pairs (Figure.1).

Like other approaches to comparable corpora, there exist many challenges in aligning bilingual terms due to the presence of noises and the significant text-structure disparity across the comparable bilingual documents. To overcome this, we propose using both corpus-driven and non-corpus-driven information, from which we draw various features and derive our hybrid model. These features are used to make initial guess on the alignment score of term pair candidates. Figure 2 shows the overall process of our term alignment model on comparable corpora. This model is language independent and it comprises several main components:

- EM algorithm
- Term alignment initialization
- Mutual information (MI) & TScore rescoring

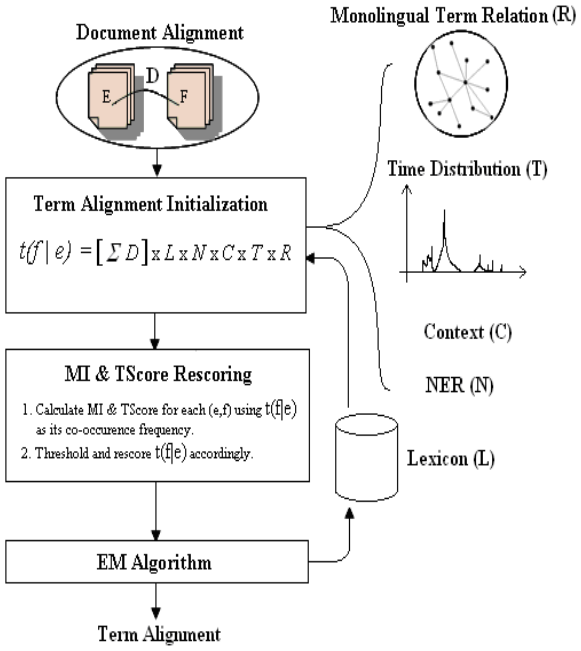

Figure 2. Term alignment model. D = document alignment score, L = lexical similarity, N = named entity similarity, C = context similarity, T = temporal similarity, R = related term similarity.

3 EM Algorithm

We make two assumptions on the preprocesses that the extracted monolingual terms are good representatives of their source documents, and the document alignment scores derived from document alignment process are good indicators of how well the contents of various documents align. Hence, the logical implication suggests that the extracted terms from both well aligned documents could well be candidates of aligned term pairs.

By reformulating the state-of-the-art EM-based word alignment framework IBM model 1 (Brown et. al., 1993), we can derive a term alignment model easily. In IBM word alignment model 1, the task is to find word alignment by using parallel sentences. In the reformulated model for term alignment, parallel sentences are replaced by comparable documents, characterized by document alignment score and their representative monolingual terms.

The significant advantage over the original IBM model 1 is the relaxation of parallel sentences or parallel corpora, by incorporating an additional feature of document alignment score. We initialize the term alignment score of the corresponding term pair candidates with the document alignment score to reflect the confidence level of document alignment. Other than that, we also employ a collection of feature similarity score: lexical similarity, named entity similarity, context similarity, temporal similarity, and related term similarity, to term alignment initialization. We will explain this further in the next section.

As we know, IBM model 1 will converge to the global maximum regardless of the initial assignment. This is truly good news for parallel corpora, but not for comparable corpora which contains a lot of noises. To prevent IBM model 1 from overfitting, we choose to run ten iterations (each iteration consists of one E-step and one M-step) for each cycle of EM in both e-f and f-e directions.

After each cycle of EM process, we simply filter off the weak term alignment pairs of both directions with a high threshold (0.8) and populate the lexicon database with the remaining pairs and use it to start another cycle of EM. The process repeats until no new term alignment pair is found. The EM algorithm for term alignment is shown as follow:

Initialize t(f|e).
for (iteration = 1 to 10)
E step

$$a[i,j,k] = \frac{t(f[k,j] \mid e[k,i])}{\sum_{i} t(f[k,j] \mid e[k,i])}, \; for \quad all \quad i,j,k$$

M step

$$tcount(e,f) = \sum_{\substack{i,j,k: \\ e[k,i]=e, \\ f[k,j]=f}} a[i,j,k], \; for \quad all \quad (e,f)$$

$$t(f \mid e) = \frac{tcount(e,f)}{\sum_{f} tcount(e,f)}, \; for \quad all \quad (e,f)$$

End for.

Figure 3. EM algorithm for e-f direction, where e[k] = k-th aligned source document, f[k] = k-th aligned target document, e[k,i] = i-th term in e[k], f[k,j] = j-th term in f[k], a[i,j,k] = probability of alignment from f[k,j] to e[k,i], t(f|e) = probability of alignment from term e to term f.

4 Term Alignment Initialization

We retrieve term alignment candidates by pairing all possible combinations of extracted monolingual source terms and target terms across the aligned document pairs. Before each cycle of EM, we assign an initial term alignment score, t(f|e) to each of these term pair candidates. Basically, we initialize the term alignment score t(f|e) based on document alignment score (D), lexical similarity (L), named entity similarity (N), context similarity (C), temporal similarity (T), and related term similarity (R). The similarity calculations of the corpus-driven features (D, C, T, R) are derived directly from the corpus and require limited lexical resource. The non-corpus-driven features (L, N) make use of a small word based bilingual dictionary to measure their lexical relevancy. That makes our model not resource-demanding and it shows that our model can work under limited resource condition.

All the above features contribute to the term alignment score t(f|e) independently, and we formulate their cumulative contributions as the following:

$$t(f|e)=\left[\sum_{(E,F):e\in E,\, f\in F} D(F|E)\right]\times L(f|e) \times N(f|e)\times C(f|e)\times T(f|e)\times R(f|e) \quad (1)$$

where,

e = source term
f = target term
E = source document
F = target document
D = document alignment score
L = lexical similarity
N = named entity similarity
C = context similarity
T = temporal similarity
R = related term similarity

This formula allows us to extend the model with additional features without affecting the existing configuration.

4.1 Document Alignment Score (D)

As explained in the Section 3, the relaxation on the requirement of parallel corpora in the new EM model leads to the incorporation of document alignment score. To indicate the confidence level of document alignment, we credit every aligned term pair candidate formed across the aligned documents with the corresponding document alignment score. Although it is not necessary, document alignment score is first normalized to the range of [0,1], with 1 indicates parallel alignment.

4.2 Lexical Similarity (L)

We design a simple lexical similarity measurement of two terms based on word translation. Term pairs that share more than 50% of word translation pairs will be credited with lexical similarity of L_0, where L_0 is configurable contribution weightage of lexical similarity. This provides us a primitive hint on term alignment without resorting to exhaustive dictionary lookup.

$$L(f|e)=\begin{cases} L_0, & if \quad T_W(f|e)\geq 0.5 \\ 1, & otherwise \end{cases} \quad (2)$$

where $L_0 > 1$ and $T_W(f|e)$ is word translation score.

4.3 Named Entity Similarity (N)

Named entity similarity is a measure of predefined category membership likelihood, such as person, location and organization. Term pairs that belong to the same NE categories will be credited with named entity similarity of N_0, where N_0 is a configurable weightage of named entity similarity. We use this similarity score to discover bilingual terms of same NE categories, yet not covered by bilingual dictionary.

$$N(f|e)=\begin{cases} N_0, if \quad NE \quad categories \quad match \\ 1, otherwise \end{cases} \quad (3)$$

where $N_0 > 1$.

4.4 Context Similarity (C)

We assume that terms with similar contexts are likely to have similar meaning. Thus, we make use of context similarity to measure semantic similarity. Here, only k nearest content words (verbs, nouns, adjectives and adverbs) before or after the terms within the sentence boundary are considered as its contexts. The following shows the calculation of context similarity of two terms based on cosine similarity between their context frequency vectors before scaling to the range of [1, C_0], where C_0 is a configurable contribution weightage of context similarity. As shown in the formula, the t(f'|e') accounts for the translation probability from the source context word to the target context word, hence the cosine similarity calculation is carried out in the target language domain.

$$C(f|e)=1+(C_0-1)\frac{\sum_{\substack{e'\in context(e)\\ f'\in context(f)}} freq(e')\,freq(f')\,t(f'|e')}{\sqrt{\sum_{e'\in context(e)} freq(e')^2}\sqrt{\sum_{f'\in context(f)} freq(f')^2}} \quad (4)$$

where $C_0 > 1$.

4.5 Temporal Similarity (T)

In temporal similarity, we make use of date information which is available in some corpus (e.g. news). We assume aligned terms are synchronous in time, this is especially true for comparable news corpora (Tao and Zai, 2005). We

use Discrete Fourier Transform (DFT) to transform the distribution function of a term in discrete time domain to a representative function in discrete frequency domain, which is usually known as "spectrum". We then calculate the power spectrum, which is defined as magnitude square of a spectrum. Power spectrum is sensitive to the relative spacing in time (or frequency component), yet invariant to the shifting in time, thus it is most suitably to be used for pattern matching of time distribution. The temporal similarity is calculated based on cosine similarity between the power spectrums of the two terms before scaling to the range of [1, T_0], where T_0 is a configurable contribution weightage of temporal similarity.

$$T(f \mid e) = (T_0 - 1)\cos ine(P_e(k), P_f(k)) + 1 \quad (5)$$

where $T_0 > 1$ and

$$\cos ine(u(k), v(k)) = \frac{\sum_k u(k)v(k)}{\sqrt{\sum_k u(k)^2}\sqrt{\sum_k v(k)^2}}$$

$$P_x(k) = |\, DFT\{DistributionFunction_x(n)\}\,|^2$$

$$= \left| \sum_{n=0}^{N-1} DistributionFunction_x(n) \times e^{-\frac{2\pi i}{N}kn} \right|^2$$

4.6 Related Term Similarity (R)

Related terms are terms that correlate statistically in the same documents and they can be found by using mutual information or t-test in the monolingual corpus. Basically, related term similarity is a measure of related term likelihood. Aligned terms are assumed to have similar related terms, hence related term similarity contributes to semantic similarity. The related term similarity is calculated based on weighted contribution from the related terms of the source term before scaling to the range of [1, R_0], where R_0 is a configurable contribution weightage of related terms similarity.

$$R(f \mid e) = (R_0 - 1)Rsimilarity(f \mid e) + 1 \quad (6)$$

where $R_0 > 1$ and

$$Rsimilarity(f \mid e) = \frac{\sum_{e' \in R(e)} vote(f \mid e')}{\sum_{f \in F}\sum_{e' \in R(e)} vote(f \mid e')}$$

$$vote(f \mid e') = \frac{\sum_{\substack{e'' \in R(e) \cap \\ [R(e') \cup \{e'\}]}} w(e'', f) \times MI(e, e'') \times vote(f \mid e'')}{\sum_{f \in F}\sum_{\substack{e'' \in R(e) \cap \\ [R(e') \cup \{e'\}]}} w(e'', f) \times MI(e, e'') \times vote(f \mid e'')}$$

$$w(e'', f) = \begin{cases} 1.5, & if \quad Tr(e'') \cap R(f) \neq \varnothing \\ 1, & otherwise \end{cases}$$

$$MI(e, e'') = \log\left(\frac{p(e, e'')}{p(e)p(e'')}\right)$$

vote(f|e') is initialized to 1 before it is computed iteratively until it converges. R(e) is the set of related term of e and Tr(e) is the set of translated term of e.

5 MI & TScore Rescoring

We design the MI & TScore rescoring process to enhance the alignment score t(f|e) of e-f term pairs that have significant co-occurrence frequencies in aligned document pairs, based on pointwise mutual information and TScore (or commonly known as t-test) of the terms. By using both measures concurrently, the association relationship of a term pair can be assumed with higher confidence. On top of that, the association of a term pair can also be suggested by a much higher TScore value alone. In this rescoring process, we scale up the alignment score t(f|e) of any term pair which is strongly associated by a constant factor. The following shows the mathematical expressions of what has been described, with M_0 as the configurable scaling factor.

Rescoring condition:

$$if \quad \{[TScore(e, f) \geq 2.5 \quad and \quad MI(e, f) \geq 0.6 \times \max_{\substack{(e', f'): freq(e') = freq(e) \\ or \quad freq(f') = freq(f)}} MI(e', f')] \quad (7)$$

$$or \quad TScore(e, f) \geq 5\} \quad then$$

$$T(f \mid e) = T(f \mid e) \times M_0$$

where $M_0 > 1$ and

$$TScore\ (e, f) = \frac{p(e, f) - p(e)p(f)}{\sqrt{\frac{p(e, f)}{2N}}}$$

$$N = NumberOfPair(e, f)$$

6 Experiment and Evaluation

We conduct the experiment on articles from three newspapers of different languages published by Singapore Press Holding (SPH), namely Straits Times[1] (English), ZaoBao[2] (Chinese) and Berita Harian[3] (Malay), in June 2006. There are 3187 English articles, 4316 Chinese articles and 1115 Malay articles. English is chosen to be the source language and the remaining two languages as target languages. To analyze the effect of the quality of comparable document in our term alignment model, we prepare two different input sets of document alignment, namely golden document alignment and automated document alignment for each source-target language pair. The former is retrieved by linguistic experts who are requested to read the contents of the articles in the source and the target languages, and then match the articles with similar contents (e.g. news coverage on same story), while the latter is generated using unsupervised method proposed by Vu et. al. (2009), mentioned in Section 2.

In both cases of document alignments, only monolingual noun terms extracted automatically by program (Vu et. al., 2008) will be used as basic semantic unit. There are 23,107 unique English noun terms, 31,944 unique Chinese noun terms and 8,938 unique Malay noun terms extracted in overall. In average, there are 17.3 noun term tokens extracted for each English document, 16.9 for Chinese document and 13.0 for Malay document. Also note that the term alignment reference list is constructed based on these extracted monolingual terms under the constraints of document alignment. In other words, the linguistic experts are requested to match the extracted terms across aligned document pairs (for both golden document alignment and automated document alignment sets respectively). The numbers of comparable document pairs and the corresponding unique term alignment reference pairs are shown in Table 2.

[1] http://www.straitstimes.com/ an English news agency in Singapore. Source © Singapore Press Holdings Ltd.

[2] http://www.zaobao.com/ a Chinese news agency in Singapore. Source © Singapore Press Holdings Ltd.

[3] http://cyberita.asia1.com.sg/ a Malay news agency in Singapore. Source © Singapore Press Holdings Ltd.

In the experiment, we will conduct the named entity recognition (NER) by using the developed system from the Stanford NLP Group, for English, and an in-house engine, for Chinese. Currently, there is no available NER engine for Malay.

Dictionary	E-C	C-E	E-M	M-E
Entry	23,979	71,287	28,496	18,935

Table 1. Statistics of dictionaries, where E = English, C = Chinese, M = Malay.

Corpus	GoldenDocAlign		AutomatedDocAlign	
	Doc Align	Term Align Ref	Doc Align	Term Align Ref
ST-ZB	90	313	899	777
ST-BH	42	113	475	358

Table 2. Statistics of comparable document alignment pairs and term alignment reference pairs.

For baseline, we make use of IBM model 1, modified in the same way which has been described in the section 3, except that we treat all comparable documents as parallel sentences, i.e. document alignment score is 1. Precision and recall are used to evaluate the performance of the system. To achieve high precision, high thresholds are used in the system and they are kept constant throughout the experiments for consistency. To evaluate the capability of discovering new bilingual terminology, we design a novelty metric, which is the ratio of the number of correct out-of-dictionary term alignment over the total number of correct term alignment.

$$Precision = \frac{C}{T} \qquad Recall = \frac{C}{G} \qquad Novelty = \frac{N}{C} \tag{8}$$

where,

C = total number of correct term alignment result.
T = total number of term alignment result.
G = total number of term alignment reference.
N = total number of correct term alignment result that are out-of-dictionary.

Table 3 shows the evaluation result of term alignment using EM algorithm with incremental feature setting. The particular order of setting is due to the implementation sequences and it is not expected to affect the result of analysis.

We observe that the precision, recall and novelty of the system are comparatively higher when the golden document alignment is used instead of the automated document alignment.

corpora	Setting	GoldenDocAlign			AutomatedDocAlign		
		Precision	Recall	Novelty	Precision	Recall	Novelty
ST-ZB	IBM 1	75.0%	1.92%	50.0%	22.2%	0.26%	50.0%
	(D)	75.0%	1.92%	50.0%	22.2%	0.26%	50.0%
	(D,L)	81.8%	2.88%	55.6%	33.3%	0.52%	25.0%
	(D,L,R)	81.8%	2.88%	55.6%	33.3%	0.52%	25.0%
	(D,L,R,M)	78.6%	3.51%	63.6%	35.7%	0.64%	40.0%
	(D,L,R,M,N)	88.2%	4.79%	53.3%	35.7%	0.64%	40.0%
	(D,L,R,M,N,C)	89.5%	5.43%	52.9%	33.3%	0.64%	40.0%
	(D,L,R,M,N,C,T)	89.5% (17/19)	5.43% (17/313)	52.9% (9/17)	37.5% (6/16)	0.77% (6/777)	16.7% (1/6)
ST-BH	IBM 1	33.3%	0.89%	0.00%	33.3%	0.78%	0.00%
	(D)	33.3%	0.89%	0.00%	33.3%	0.78%	0.00%
	(D,L)	75.0%	5.31%	50.0%	50.0%	1.94%	0.00%
	(D,L,R)	75.0%	5.31%	50.0%	50.0%	1.94%	0.00%
	(D,L,R,M)	75.0%	5.31%	50.0%	54.5%	2.33%	0.00%
	(D,L,R,M,N)	75.0%	5.31%	50.0%	54.5%	2.33%	0.00%
	(D,L,R,M,N,C)	83.3%	8.85%	60.0%	50.0%	1.94%	0.00%
	(D,L,R,M,N,C,T)	83.3% (10/12)	8.85% (10/113)	60.0% (6/10)	50.0% (5/10)	1.94% (5/258)	0.00% (0/5)

Table 3. Performance of term alignment using EM algorithm with incremental feature setting, where D = document alignment, L = lexical similarity, R = related term similarity, M = MI & TScore rescoring, N = named entity similarity, C = context similarity, T = temporal similarity.

This is expected since the golden document alignment provides document pairs with stronger semantic bonding. This also suggests that improving on the document alignment would further improve the term alignment result.

It is noteworthy observation that the implemented features improve the system precision and recall under various scenarios, although the degree of improvement varies from case to case. This shows the effectiveness of these features in the model.

On the other hand, the novelty of the system is around 40%+ and 50%+ for ST-ZB and ST-BH respectively (except for the automated document alignment in ST-BH scenarios). This suggests that the system can discover quite a large percentage of the correct bilingual terminologies that do not exist in the lexicon initially.

Compared with the baseline IBM model 1, there is an increase of 14.5% in precision, 3.51% in recall and 2.9% in novelty for ST-ZB, using the golden document alignment. For ST-BH, there is an even larger increase: 50% in precision, 7.96% in recall and 60% in novelty.

7 Conclusion

We have proposed an unsupervised EM-based hybrid model to extract bilingual terminology from comparable corpora through document alignment constraint. Our strategy is to make use of various information (corpus-driven and non-corpus-driven) to make initial guess on the semantic bonding of the term alignment candidates before subjecting them to document alignment constraint through EM algorithm. The hybrid model allows inclusion of additional features without reconfigurations on existing features, this make it practically attractive. Moreover, the proposed system can be easily deployed in any language with minimal configurations.

We have successfully conducted the experiments in English-Chinese and English-Malay comparable news corpora. The features employed in the model have shown incremental improvement in performance over the baseline method. In particular, the system shows improvement in the capability to discover new bilingual terminology from comparable corpora even with limited usage of dictionaries.

From the experiments, we have found that the quality of comparable bilingual documents is a

major limiting factor to achieve good performance. In future, we want to explore ways to improve on this.

References

R. Agrawal, C. Faloutsos, and A. Swami. 1993. *Efficient similarity search in sequence databases*. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. Chicago, United States.

P. F. Brown, V. S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. *The mathematics of statistical machine translation: Parameter estimation*. Computational Linguistics, 19(2): 263-312.

Yunbo Cao and Hang Li. 2002. *Base Noun Phrase Translation Using Wed Data and the EM Algorithm*, Computational Linguistics, pp.1-7.

Pascale Fung, 1998. *A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora.* Proceedings of AMTA, pp.1-17.

Pascale Fung and Percy Cheung. 2004. *Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM*, Proceedings of EMNLP, pp.57-63.

Alexandre Klementiev and Dan Roth, 2006. *Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora*. Computational Linguistics, pp. 817-824.

K. Knight and J. Graehl. 1998. *Machine transliteration*, Computational Linguistics, 24(4): 599-612.

E. Morin, B. Daille, K. Takeuchi, K. Kageura. 2007. *Bilingual Terminology Mining – Using Brain, not brawn comparable corpora*, Proceedings of ACL.

Dragos Stefan Munteanu and Daniel Marcu. 2005. *Improving Machine Translation Performance by Exploiting Non-Parallel Corpora.* Computational Linguistics, 31(4): 477-504.

Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura, 2003. *Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach.* Proceedings of ACL, vol.11, pp.57-64.

Tao Tao and Chengxiang Zhai. 2005. *Mining comparable bilingual text corpora for cross-language information integration*, Proceedings of ACM.

Raghavendra Udupa, K. Saravanan, A. Kumaran, Jagadeesh Jagarlamudi. 2008. *Mining named entity transliteration equivalents from comparable corpora*, Proceedings of ACM.

Thuy Vu, Aiti Aw, Min Zhang, 2008. *Term extraction through unithood and termhood unification.* Proceedings of IJCNLP-08, Hyderabad, India.

Thuy Vu, Aiti Aw, Min Zhang, 2009. *Feature-based Method for Document Alignment in Comparable News Corpora*. Proceedings of EACL-09, Athens, Greece.