Bootstrapping Entity Translation on Weakly Comparable Corpora

Taesung Lee and **Seung-won Hwang**
Department of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
{elca4u, swhwang}@postech.edu

Abstract

This paper studies the problem of mining named entity translations from comparable corpora with some "asymmetry". Unlike the previous approaches relying on the "symmetry" found in parallel corpora, the proposed method is tolerant to asymmetry often found in comparable corpora, by distinguishing different semantics of relations of entity pairs to selectively propagate seed entity translations on weakly comparable corpora. Our experimental results on English-Chinese corpora show that our selective propagation approach outperforms the previous approaches in named entity translation in terms of the mean reciprocal rank by up to 0.16 for organization names, and 0.14 in a low comparability case.

1 Introduction

Identifying and understanding entities is a crucial step in understanding text. This task is more challenging in the presence of multilingual text, because translating named entities (NEs), such as persons, locations, or organizations, is a non-trivial task. Early research on NE translation used phonetic similarities, for example, to mine the translation 'Mandelson'→'曼德尔森'[ManDeErSen] with similar sounds. However, not all NE translations are based on transliterations, as shown in Table 1—Some translations, especially the names of most organizations, are based on semantic equivalences. Furthermore, names can be abbreviated in one or both languages, e.g., the 'World Trade Organization' (世界贸易组织) can be called the 'WTO' (世贸组织). Another challenging example is that, a translation can be arbitrary, e.g., 'Jackie Chan' → '成龙' [ChengLong]. There are many approaches that deal with some of these challenges (Lam et al., 2007; Yang et al., 2009), e.g., by combining phonetic similarity and a dictionary. However, arbitrary translations still cannot be handled by examining the NE pair itself. Corpus-based approaches (Kupiec, 1993; Feng, 2004), by mining external signals from a large corpus, such as parenthetical translation "成龙 (Jackie Chan)", complement the problem of transliteration-based approaches, but the coverage of this approach is limited to popular entities with such evidence.

English	Chinese
World Trade Organization	世界贸易组织 [ShiJieMaoYiZuZhi]
WTO	世贸组织 [ShiMaoZuZhi]
Jackie Chan	成龙 [ChengLong]

Table 1: Examples of non-phonetic translations.

The most effective known approach to NE translation has been a holistic framework (You et al., 2010; Kim et al., 2011; You et al., 2012) combining transliteration- and corpus-based methods. In these approaches, both 1) arbitrary translations and 2) lesser-known entities can be handled, by propagating the translation scores of known entities to lesser-known entities if they co-occur frequently in both corpora. For example, a lesser-known entity Tom Watson can be translated if Mandelson and Tom Watson co-occur frequently in an English corpus, and their Chinese translations also co-occur frequently in a Chinese corpus, i.e., if the co-occurrences in the two corpora are "symmetric".

A research question we ask in this paper is: What if comparable corpora are not comparable enough to support this *symmetry* assumption? We found that this is indeed the case. For example, even English and Chinese news from the same publisher may have different focus– the Chinese version focuses more on Chinese Olympic

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 631–640,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

teams and Chinese local news. In the presence of such asymmetry, all previous approaches, building upon symmetry, quickly deteriorate by propagating false positives. For example, co-occurrence of Mandelson and Tom Watson may not appear in a Chinese corpus, which may lead to the translation of Tom Watson into another Chinese entity Gordon Brown which happens to co-occur with the Chinese translation of Mandelson.

Our key contribution is to avoid such false propagation, by discerning the semantics of relations. For example, relations between Mandelson and Tom Watson, should be semantically different from Chinese relations between '戈登·布朗' (Gordon Brown) and '曼德尔森' (Mandelson). A naive approach would be finding documents with a similar *topic* such as politics, and scientific discovery, and allowing propagation only when the topic agrees. However, we found that a topic is a unit that is too coarse for this task because most articles on Mandelson will invariably fall into the same topic[1]. In clear contrast, we *selectively propagate* seed translations, only when the relations in the two corpora share the same semantics.

This selective propagation can be especially effective for translating challenging types of entities such as *organizations* including the WTO used with and without abbreviation in both languages. Applying a holistic approach (You et al., 2012) on organizations leads to poor results, 0.06 in terms of the F1-score. A naive approach to increase the precision would be to consider multi-type co-occurrences, hoping that highly precise translations of some type, e.g., persons with an F1-score of 0.69 (You et al., 2012), can be propagated to boost the precision on organizations. In our experiments, this naive multi-type propagation still leads to an unsatisfactory F1-score of 0.12. Such a low score can be explained by the following example. When translating 'WTO' using the co-occurrence with 'Mandelson', other co-occurrences such as (London, Mandelson) and (EU, Mandelson) produce a lot of noise because the right translation of WTO does not share much phonetic/semantic similarity. Our understanding of relation semantics, can distinguish "Mandelson *was born in* London" from "Mandelson *visited* the WTO", to stop false propagations, which generates an F1-score 0.25 higher than the existing approaches.

More formally, we enable *selective propagation* of seed translations on weakly comparable corpora, by 1) clarifying the detailed meaning of relational information of co-occurring entities, and 2) identifying the contexts of the relational information using statement-level context comparison. In other words, we propagate the translation score of a known translation pair to a neighbor pair *if* the semantics of their relations in English and Chinese corpora are *equivalent* to accurately propagate the scores. For example, if we know 'Russia'→'俄罗斯'$_{(1)}$ and *join*→加入$_{(2)}$, then from a pair of statements "Russia$_{(1)}$ joins$_{(2)}$ the WTO$_{(3)}$" and "俄罗斯$_{(1)}$ 加入$_{(2)}$ 世贸组织$_{(3)}$", we can propagate the translation score of (Russia, 俄罗斯)$_{(1)}$ to (WTO, 世贸组织)$_{(3)}$. However, we do not exploit a pair of statements "Russia joined the WTO" and "俄罗斯 谴责$_{(2')}$ 摩洛哥" because 谴责$_{(2')}$ does not mean *join*$_{(2)}$. Furthermore, we mine a similar English-Chinese document pair that can be found by comparing the entity relationships, such as "Mandelson visited Moscow" and "Mandelson met Alexei Kudrin", within the English document and the Chinese document to leverage similar contexts to assure that we use symmetric parts.

For this goal, we first extract *relations* among entities in documents, such as *visit* and *join*, and mine semantically equivalent relations across the languages, e.g., English and Chinese, such as *join*→加入. Once these relation translations are mined, similar document pairs can be identified by comparing each constituent relationship among entities using their relations. Knowing document similarity improves NE translation, and improved NE translation can boost the accuracy of document and relationship similarity. This iterative process can continue until convergence.

To the best of our knowledge, our approach is the first to translate a broad range of multilingual relations and exploit them to enhance NE translation. In particular, our approach leverages semantically similar document pairs to exclude incomparable parts that appear in one language only. Our method outperforms the previous approaches in translating NE up to 0.16 in terms of the mean reciprocal rank (MRR) for organization names. Moreover, our method shows robustness, with 0.14 higher MRR than seed translations, on less comparable corpora.

[1] The MRR for organization names achieved by a topic model-based approach was 0.15 lower than our best.

2 Related Work

This work is related to two research streams: NE translation and semantically equivalent relation mining.

Entity translation

Existing approaches on NE translation can be categorized into 1) transliteration-based, 2) corpus-based, and 3) hybrid approaches.

Transliteration-based approaches (Wan and Verspoor, 1998; Knight and Graehl, 1998) are the foundations of many decent methods, but they alone suffer from ambiguity (e.g., 史蒂夫 and 始第夫 have the same sounds) and cannot handle non-transliterated cases such as 'Jackie Chan (成龙[ChengLong])'. Some methods (Lam et al., 2007; Yang et al., 2009) rely on meanings of constituent letters or words to handle organization name translation such as 'Bank of China (中国银行)', whose translation is derived from 'China (中国)', and 'a bank (银行)'. However, many names often originate from abbreviation (such as 'WTO'); hence we cannot always leverage meanings.

Corpus-based approaches (Kupiec, 1993; Lin et al., 2008; Jiang et al., 2009) exploit high-quality bilingual evidence such as parenthetical translation, e.g., "成龙 (Jackie Chan)", (Lin et al., 2008), semi-structural patterns (Jiang et al., 2009), and parallel corpus (Kupiec, 1993). However, the coverage of the corpus-based approaches is limited to popular entities with such bilingual evidences. On the other hand, our method can cover entities with monolingual occurrences in corpora, which significantly improves the coverage.

The most effective known approach is a holistic framework that combines those two approaches (You et al., 2012; You et al., 2010; Kim et al., 2011). You et al. (2010; 2012) leverage two graphs of entities in each language, that are generated from a pair of corpora, with edge weights quantified as the strength of the relatedness of entities. Then, two graphs are iteratively aligned using the common neighbors of two entities. Kim et al. (2011) build such graphs using the context similarity, measured with a bag of words approach, of entities in news corpora to translate NEs. However, these approaches assume the symmetry of the two graphs. This assumption holds if two corpora are parallel, but such resources are scarce. But our approach exploits comparable parts from corpora.

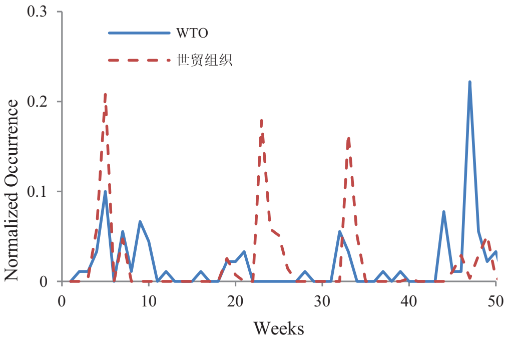


Figure 1: Dissimilarity of temporal distributions of 'WTO' in English and Chinese corpora.

Other interesting approaches such as (Klementiev and Roth, 2006; Kim et al., 2012) rely on temporal distributions of entities. That is, two entities are considered to be similar if the two entities in different languages have similar occurrence distributions over time. However, the effectiveness of this feature also depends on the comparability of entity occurrences in time-stamped corpora, which may not hold as shown in Figure 1. In clear contrast, our method can find and compare articles, on different dates, describing the same NE. Moreover, our method does not require time stamps.

Semantically similar relation mining

Recently, similar relation mining in one language has been studied actively as a key part of automatic knowledge base construction. In automatically constructed knowledge bases, finding semantically similar relations can improve understanding of the Web describing content with many different expressions. As such an effort, PATTY (Nakashole et al., 2012) finds similar relations with almost the same support sets–the sets of NE pairs that co-occur with the relations. However, because of the regional locality of information, bilingual corpora contain many NE pairs that appear in only one of the support sets of the semantically identical relations. NELL (Mohamed et al., 2011) finds related relations using seed pairs of one given relation; then, using K-means clustering, it finds relations that are semantically similar to the given relation. Unfortunately, this method requires that we set K manually, and extract relations for each given relation. Therefore, this is unsuitable to support general relations.

There are only few works on translating relations or obtaining multi-lingual similar relations. Schone et al. (2011) try to find relation patterns in multiple languages for given seed pairs of a relation. Because this approach finds seed pairs in Wikipedia infoboxes, the number of retrievable relations is restricted to five. Kim et al. (2010) seek more diverse types of relations, but it requires parallel corpora, which are scarce.

3 Framework Overview

In this section, we provide an overview of our framework for translating NEs, using news corpora in English and Chinese as a running example. Because such corpora contain asymmetric parts, the goal of our framework is to overcome asymmetry by distinguishing the semantics of relations, and leveraging document context defined by the relations of entities.

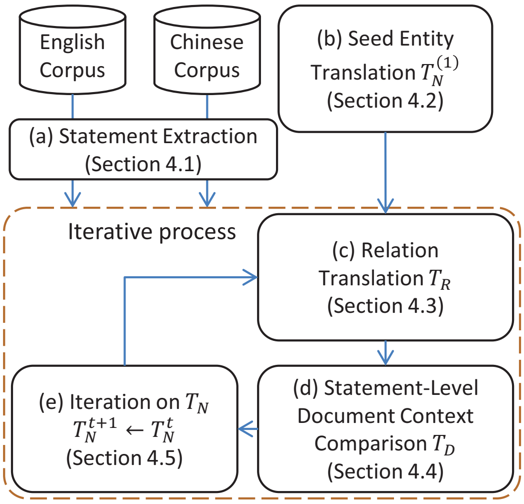

English Corpus
Chinese Corpus
(b) Seed Entity Translation $T_N^{(1)}$ (Section 4.2)
(a) Statement Extraction (Section 4.1)
Iterative process
(c) Relation Translation T_R (Section 4.3)
(e) Iteration on T_N $T_N^{t+1} \leftarrow T_N^t$ (Section 4.5)
(d) Statement-Level Document Context Comparison T_D (Section 4.4)

Figure 2: Framework overview.

For this purpose, we build a mutual bootstrapping framework (Figure 2), between entity translation and relation translation using extracted relationships of entities (Figure 2 (a), Section 4.1). More formally, we use the following process:

1. Base condition (Figure 2 (a), Section 4.2): Initializing $T_N^{(1)}(e_E, e_C)$, a seed entity translation score, where e_E is an English entity, and e_C is a Chinese entity. $T_N^{(1)}$ can be initialized by phonetic similarity or other NE translation methods.

2. Iteration: Obtaining T_N^{t+1} using T_N^t.

   1) Using T_N^t, we obtain a set of relation translations with a semantic similarity score, $T_R^t(r_E, r_C)$, for an English relation r_E and a Chinese relation r_C (Figure 2 (b), Section 4.3) (e.g., r_E =*visit* and r_C =访问).

   2) Using T_N^t and T_R^t, we identify a set of semantically similar document pairs that describe the same event with a similarity score $T_D^t(d_E, d_C)$ where d_E is an English document and d_C is a Chinese document (Figure 2 (c), Section 4.4).

   3) Using T_N^t, T_R^t and T_D^t, we compute T_N^{t+1}, an improved entity translation score (Figure 2 (d), Section 4.5).

Each sub-goal reinforces the result of others in the $(t+1)$-th iteration, and by iteratively running them, we can improve the quality of translations. Note that, hereinafter, we omit (t) for readability when there is no ambiguity.

4 Methods

In this section, we describe our method in detail. First, we explain how we extract statements, which are units of relational information, from documents in Section 4.1, and how we obtain seed name translations in Section 4.2. Next, we present our method for discovering relation translations across languages in Section 4.3. In Section 4.4, we use the name translations and the relation translations to compare document contexts which can boost the precision of NE translation. In Section 4.5, we describe how we use the resources obtained so far to improve NE translation.

4.1 Statement Extraction

We extract relational statements, which we exploit to propagate translation scores, from an English news corpus and a Chinese news corpus. A *relational statement*, or simply a *statement* is a triple (x, r, y), representing a relationship between two names, x and y. For example, from "Mandelson recently visited Moscow," we obtain this statement: (Mandelson, visit, Moscow). We follow a standard procedure to extract statements, as similarly adopted by Nakashole et al. (2012), using Stanford CoreNLP (Klein and Manning, 2003) to lemmatize and parse sentences. Here, we refer readers to existing work for further details because this is not our key contribution.

4.2 Seed Entity Translation

We need a few seed translation pairs to initiate the framework. We build a seed translation score $T_N^{(1)}(e_E, e_C)$ indicating the similarity of an English entity e_E and a Chinese entity e_C using an existing method. For example, most methods would give high value for

$T_N^{(1)}$(Mandelson, 曼德尔森 [ManDeErSen]). In this work, we adopted (You et al., 2012) with (Lam et al., 2007) as a base translation matrix to build the seed translation function. We also use a dictionary to obtain non-NE translations such as 'government'. We use an English-Chinese general word dictionary containing approximately 80,000 English-Chinese translation word pairs that was also used by Kim et al. (2011) to measure the similarity of context words of entities.

4.3 Relation Translation

We need to identify relations that have the equivalent semantics across languages, (e.g., *visit*→ 访问), to enable selective propagation of translation scores. Formally, our goal is to measure a pairwise relation translation score $T_R(r_E, r_C)$ for an English relation $r_E \in R_E$ and a Chinese relation $r_C \in R_C$ where R_E is a set of all English relations and R_C is a set of all Chinese relations.

We first explain a basic feature to measure the similarity of two relations, its limitations, and how we address the problems. A basic clue is that relations of the same meaning are likely to be mentioned with the same entity pairs. For example, if we have (Mandelson, visit, Moscow) as well as (Mandelson, head to, Moscow) in the corpus, this is a positive signal that the two relations may share the same meaning. Such NE pairs are called *support pairs* of the two relations.

We formally define this clue for relations in the same language, and then describe that in the bilingual setting. A *support intersection* $H_m(r^i, r^j)$, a set of support pairs, for monolingual relations r^i and r^j is defined as

$$H_m(r^i, r^j) = H(r^i) \cap H(r^j) \tag{1}$$

where $H(r)$ is the *support set* of a relation r defined as $H(r) = \{(x, y) | (x, r, y) \in \mathbf{S}\}$, and $\mathbf{S}$ is either $\mathbf{S}_E$, a set of all English statements, or $\mathbf{S}_C$, a set of all Chinese statements that we extracted in Section 4.1.

Likewise, we can define a support intersection for relations in the different languages using the translation score $T_N(e_E, e_C)$. For an English relation r_E and a Chinese relation r_C,

$$\begin{aligned} H_b(r_E, r_C) = \{ & (x_E, x_C, y_E, y_C) | \\ & T_N(x_E, x_C) \geq \theta \\ & \text{and } T_N(y_E, y_C) \geq \theta \\ & \text{for } (x_E, r_E, y_E) \in \mathbf{S}_E \\ & \text{and } (x_C, r_C, y_C) \in \mathbf{S}_C \} \end{aligned} \tag{2}$$

where $\theta = 0.6$ is a harsh threshold to exclude most of the false translations by T_N.

Finally, we define a support intersection, a set of support pairs between two relations r^i and r^j of any languages,

$$H(r^i, r^j) = \begin{cases} H_b(r^i, r^j) & \text{if } r^i \in R_E \text{ and } r^j \in R_C \\ H_b(r^j, r^i) & \text{if } r^j \in R_E \text{ and } r^i \in R_C \\ H_m(r^i, r^j) & \text{otherwise} \end{cases} \tag{3}$$

Intuitively, $|H(r^i, r^j)|$ indicates the strength of the semantic similarity of two relations r^i and r^j of any languages. However, as shown in Table 2, we cannot use this value directly to measure the similarity because the support intersection of semantically similar *bilingual* relations (e.g., $|H(\text{head to}, \text{访问})| = 2$) is generally very low, and normalization cannot remedy this problem as we can see from $|H(\text{visit}, \text{访问})| = 27$ and $|H(\text{visit})| = 1617$.

Set	Cardinality
H(visit)	1617
H(访问)	2788
H(visit, 访问)	27
H(head to, 访问)	2

Table 2: Evidence cardinality in the corpora.

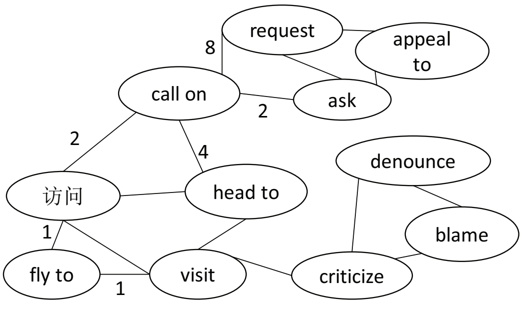


Figure 3: Network of relations. Edges indicate that the relations have a non-empty support intersection, and edge labels show the size of the intersection.

We found that the connectivity among similar relations is more important than the strength of the similarity. For example, as shown in Figure 3, *visit* is connected to most of the *visit*-relations such as *head to*, *访问*. Although *visit* is connected to *criticize*, *visit* is not connected to other criticize-relations such as *denounce* and *blame*, whereas *criticize*, *denounce*, and *blame* are inter-

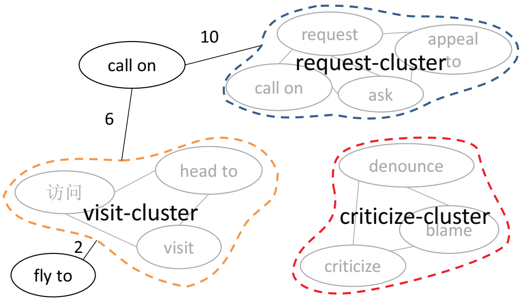


Figure 4: Relation clusters and a few individual relations. Edge labels show the size of the intersection.

connected. To exploit this feature, we use a random walk-based graph clustering method.

Formally, we use Markov clustering (Van Dongen, 2000) on a graph $G = (V, E)$ of relations, where $V = R_E \cup R_C$ is a set of all English and Chinese relations. An edge (r^i, r^j) indicates that two relations in any languages are similar, and its weight is quantified by a sigmoid function on a linear transformation of $|H(r^i, r^j)|$ that was empirically found to produce good results.

Each resultant cluster forms a set of bilingual similar relations, $c = \{r^{c_1}, ..., r^{c_M}\}$, such as *visit*-cluster, which consists of *visit*, *head to*, and *访问* in Figure 4. However, this cluster may not contain all similar relations. A relation may have multiple meanings (e.g., *call on*) so it can be clustered to another cluster, or a relation might not be clustered when its support set is too small (e.g., *fly to*). For such relations, rather than assigning zero similarity to *visit*-relations, we compute a cluster membership function based on support pairs of the cluster members and the target relation, and then formulate a pairwise relation translation score.

Formally, we learn the membership function of a relation r to a cluster c using support vector regression (Joachims, 1999) with the following features based on the support set of cluster c, $H(c) = \bigcup_{r \in c} H(r)$, and the support intersection of r and c, $H(r, c) = \bigcup_{r^* \in c} H(r, r^*)$.

- $f^1(r, c) = |H(r, c)|/|H(r)|$: This quantifies the degree of inclusion, $H(c) \in H(r)$.
- $f^2(r, c) = |H(r, c)|/|H(c)|$: This quantifies the degree of inclusion, $H(r) \in H(c)$.
- $f^3(r, c) = |H_{\text{within}}(r, c)|/|H_{\text{within}}(c)|$: This is a variation of f^2 that considers only noun phrase pairs shared at least once by relations in c.
- $f^4(r, c) = |H_{\text{within}}(r, c)|/|H_{\text{shared}}(c)|$: This is a variation of f^2 that considers only noun phrase pairs shared at least once by any pair of relations.
- $f^5(r, c) = |\{r^* \in c | H(r, r^*) > 0\}|/|c|$: This is the degree of connectivity to the cluster members.

where $H_{\text{within}}(r, c) = \bigcup_{r^* \in c} H(r, c) \cap H(r, r^*)$, the intersection, considering translation, of $H(r)$ and noun phrase pairs shared at once by relations in c, $H_{\text{within}}(c) = \bigcup_{r^* \in c} H(r^*, c - \{r^*\})$, and $H_{\text{shared}}(c) = \bigcup_{r^* \in R_E \cup R_C} H(r^*, c)$, the noun phrase pairs shared at once by any relations. The use of H_{within} and H_{shared} is based on the observation that a noun phrase pair that appear in only one relation tends to be an incorrectly chunked entity such as 'World Trade' from the 'World Trade Organization'.

Based on this membership function $S(r, c)$, we compute pairwise relation similarity. We consider that two relations are similar if they have at least one cluster that the both relations belong to, which can be measured with $S(r, c)$. More formally, pairwise similarity of relations r^i and r^j is defined as

$$T_R(r^i, r^j) = \max_{c \in \mathcal{C}} S(r^i, c) \cdot S(r^j, c) \qquad (4)$$

where $\mathcal{C}$ is a set of all clusters.

4.4 Statement-level Document Context Comparison

A brute-force statement matching approach often fails due to ambiguity created by ignoring context, and missing information in T_N or T_R. Therefore, we detect similar document pairs to boost the statement matching process. Unlike the previous approaches (e.g., bag-of-words), we focus on the relationships of entities within documents using the extracted statements.

Formally, we compute the similarity of two statements $s_E = (x_E, r_E, y_E)$ and $s_C = (x_C, r_C, y_C)$ in different languages as follows:

$$T_S(s_E, s_C) = T_N(x_E, x_C) T_R(r_E, r_C) T_N(y_E, y_C) \qquad (5)$$

With this definition, we can find similar statements described with different vocabularies in different languages.

To compare a document pair, we use the following equation to measure the similarity of an

English document d_E^i and a Chinese document d_C^j based on their statements S_E^i and S_C^j, respectively:

$$T_D(d_E^i, d_C^j) = \frac{\sum_{(s_E,s_C)\in B} T_S(s_E^{i,r}, s_C^{j,r})}{|S_E^i| + |S_E^i| - |B|} \tag{6}$$

where $B \subset S_E^i \times S_C^j$ is a greedy approximate solution of maximum bipartite matching (West, 1999) on a bipartite graph $G_B = (V_B = (S_E^i, S_C^j), E_B)$ with edge weights that are defined by T_S. The maximum bipartite matching finds a subset of edges in $S_E^i \times S_C^j$ that maximize the sum of the selected edge weights and that do not share a node as their anchor point.

4.5 Iteration on T_N

In this section, we describe how we use the statement similarity function T_S, and the document similarity function T_D to improve and derive the next generation entity translation function $T_N^{(t+1)}$. We consider that a pair of an English entity e_E and a Chinese entity e_C are likely to indicate the same real world entity if they have 1) semantically similar relations to the same entity 2) under the same context. Formally, we define an increment function as follows.

$$\Delta T_N(e_E,e_C)=\sum_{d_E^i}\sum_{d_C^j} T_D(d^i, d^j) \max_{(s_E,s_C)\in B^*} T_S(s_E, s_C) \tag{7}$$

where B^* is a subset of $B \subset S_E^i \times S_C^j$ such that the connected statements mention e_E and e_C, and B is the greedy approximate solution of maximum bipartite matching for the set S_E^i of statements of d_E^i and the set S_C^j of statements of d_C^j. In other words, B^* is a set of matching statement pairs mentioning the translation target e_E and e_C in the document pair. Then, we use the following equation to improve the original entity translation function.

$$\begin{aligned} T_N^{(t+1)}(e_E, e_C) = (1-\lambda)\frac{\Delta T_N(e_E, e_C)}{\sum_{e_C^*} \Delta T_N(e_E, e_C^*)} \\ + \lambda T_N(e_E, e_C) \end{aligned} \tag{8}$$

where λ is a mixing parameter in $[0, 1]$. We set $\lambda = 0.6$ in our experiments.

With this update, we obtain the improved NE translations considering the relations that an entity has to other entities under the same context to achieve higher precision.

5 Experiments

In this section, we present experimental settings and results of translating entity names using our methods compared with several baselines.

5.1 Data and Evaluation

We processed news articles for an entire year in 2008 by Xinhua news who publishes news in both English and Chinese, which were also used by Kim et al. (2011) and Shao and Ng (2004). The English corpus consists of 100,746 news articles, and the Chinese corpus consists of 88,031 news articles. The news corpora are not *parallel* but *comparable* corpora, with asymmetry of entities and relationship as the asymmetry in the number of documents also suggest. Examples of such locality in Xinhua news include the more extensive coverage of Chinese teams in the Olympics and domestic sports in the Chinese news. Our framework finds and leverages comparable parts from the corpora without document-content-external information such as time stamps. We also show that, under the decreasing comparability, our method retains higher MRR than the baselines.

We follow the evaluation procedures used by You et al. (2012) and Kim et al. (2011) to fairly and precisely compare the effectiveness of our methods with baselines. To measure performance, we use mean reciprocal rank (MRR) to evaluate a translation function T:

$$MRR(T) = \frac{1}{|Q|} \sum_{(u,v)\in Q} \frac{1}{rank_T(u,v)} \tag{9}$$

where Q is the set of gold English-Chinese translation pairs (u, v) and $rank_T(u, v)$ is the rank of $T(u, v)$ in $\{T(u, w)|w$ is a Chinese entity$\}$. In addition, we use precision, recall, and F1-score.

As gold translation pairs, we use the evaluation data used by You et al. (2012) with additional labels, especially for organizations. The labeling task is done by randomly selecting English entities and finding their Chinese translation from the Chinese corpus. We only use entities with translations that appear in the Chinese corpus. We present the evaluation results for persons and organizations to show the robustness of the methods. In total, we identified 490 English entities in the English news with Chinese translations in the Chinese news. Among the 490 entities, 221 NEs are persons and 52 NEs are organizations.

	Person				Organization			
	MRR	P.	R.	F1	MRR	P.	R.	F1
$T_N^{(2)}$	**0.80**	**0.81**	**0.79**	**0.80**	**0.53**	**0.56**	**0.52**	**0.54**
$T_N^{(1)}$	0.77	0.80	0.77	0.78	0.44	0.49	0.44	0.46
T_{PH+P}^S	0.73	0.70	0.67	0.69	0.14	0.17	0.04	0.06
T_{PH+P}^M	0.68	0.70	0.68	0.69	0.08	0.31	0.08	0.12
T_{HB}	0.71	0.59	0.59	0.59	0.37	0.29	0.29	0.29
T_{Dict}	0.09	1.00	0.09	0.17	0.17	1.00	0.17	0.30

Table 3: Evaluation results of the methods.

5.2 Baselines

We compare our methods with the following baselines.

- T_{PH+P}^S (You et al., 2012) is a holistic method that uses a transliteration method as base translations, and then reinforces them to achieve higher quality. This method uses only a single type of entities to propagate the translation scores.

- T_{PH+P}^M is the holistic method revised to use naive multi-type propagation that uses multiple types of entities to reinforce the translation scores.

- T_{HB} is a linear combination of transliteration and semantic translation methods (Lam et al., 2007) tuned to achieve the highest MRR.

- T_{Dict} is a dictionary-only method. This dictionary is used by both T_{HB} and T_N.

Only the translation pairs of scores above 0.35 are used for T_{PH+P} to maximize the F1-score to measure precision, recall and F1-score. For our method $T_N^{(t)}$, we use the result with $(t) = 1$, the seed translations, and $(t) = 2$, which means that only one pass of the whole framework is performed to improve the seed translation function. In addition, we use translation pairs with scores above 0.05 to measure precision, recall, and F1-score. Note that these thresholds do not affect MRRs.

5.3 NE Translation Results

We show the result of the quantitative evaluation in Table 3, where the highest values are boldfaced, except T_{Dict} which shows 1.00 precision because it is a manually created dictionary. For both the person and organization cases, our method $T_N^{(2)}$ outperforms the state-of-the-art methods in terms of precision, recall, F1-score and MRR. With only one iteration of selective propagation, the seed translation is improved to achieve the 0.09 higher MRR.

The baselines show lower, but comparable MRRs and F1-scores for persons that mostly consist of transliterated cases. However, not all translations have phonetic similarity, especially organization names, as the low F1-score of T_{PH+P}^S, 0.06, for organizations suggests. The naive multi-type propagation T_{PH+P}^M shows decreased MRR for both persons and organizations compared to the single-type propagation T_{PH+P}^S, which shows a negative influence of diverse relation semantics of entities of different types. T_{HB} achieves a better MRR than T_{PH+P} due to the semantic translation of organization names. However, despite the increased recall of T_{HB} over that of T_{Dict}, the precision of T_{HB} is unsatisfactory because T_{HB} maps abbreviated names such as 'WTO' with other NEs. On the other hand, our method achieves the highest MRR and precision in both the person and organization categories.

As shown in Table 4, T_{HB} translates 'WTO' inaccurately, linking it to an incorrect organization '巴解组织' (Palestine Liberation Organization).

English name	$T_N^{(2)}$	$T_N^{(1)}$	T_{HB}
Mandelson	**曼德尔森** [ManDeErSen]	**曼德尔森** [ManDeErSen]	**曼德尔森** [ManDeErSen]
WTO	**世贸组织** [ShiMaoZuZhi]	上合组织 [ShangHeZuZhi]	巴解组织 [BaJieZuZhi]
White House	**白宫** [BaiGong]	加州 [JiaZhou]	加州 [JiaZhou]
Microsoft	**微软公司** [WeiRuanGongSi]	美国司法部 [MeiGuoSiFaBu]	米罗诺夫 [MiLuoNuoFu]

Table 4: Example translations from the different methods. Boldface indicates correct translations.

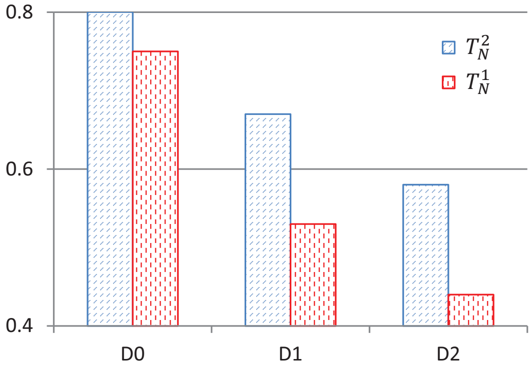
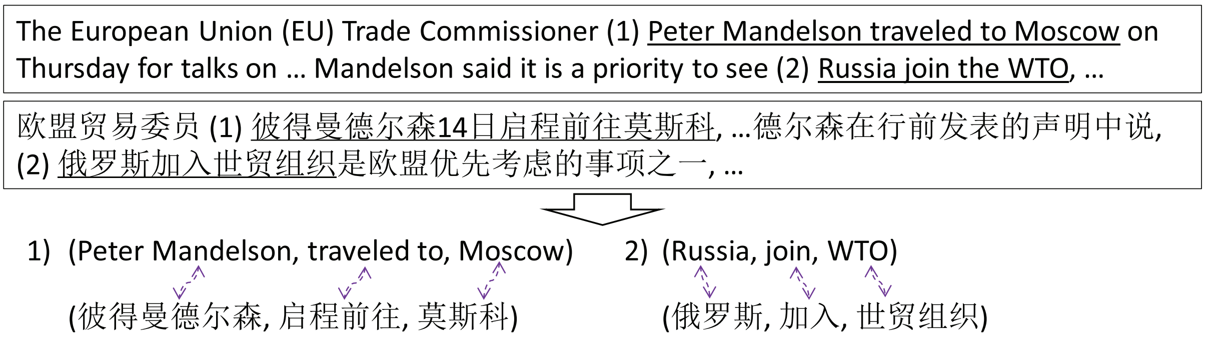


Figure 6: MRR with decreasing comparability.



The European Union (EU) Trade Commissioner (1) Peter Mandelson traveled to Moscow on Thursday for talks on … Mandelson said it is a priority to see (2) Russia join the WTO, …

欧盟贸易委员 (1) 彼得曼德尔森14日启程前往莫斯科, …德尔森在行前发表的声明中说, (2) 俄罗斯加入世贸组织是欧盟优先考虑的事项之一, …

1) (Peter Mandelson, traveled to, Moscow)

(彼得曼德尔森, 启程前往, 莫斯科)

2) (Russia, join, WTO)

(俄罗斯, 加入, 世贸组织)

Figure 5: Example of similar document pairs.

Moreover, the use of the corpora by $T_N^{(1)}$ could not fix this problem, and it finds another organization related to trade, '上合组织' (Shanghai Cooperation Organization). In contrast, our selective propagation method $T_N^{(2)}$, which uses the wrong seed translation by $T_N^{(1)}$, '上合组织' (Shanghai Cooperation Organization), successfully translates the WTO using statements such as (Russia, join, WTO), and its corresponding Chinese statement (俄罗斯, 加入, 世贸组织). Similarly, both the baseline T_{HB} and the seed translation $T_N^{(1)}$ matched *Microsoft* to incorrect Chinese entities that are phonetically similar as indicated by the underlined text. In contrast, $T_N^{(2)}$ finds the correct translation despite the phonetic dissimilarity.

5.4 NE Translation Results with Low Corpus Comparability

We tested the methods using less comparable data to evaluate the robustness with the following derived datasets:

- D0: All news articles are used.
- D1: January-December English and July-December Chinese articles are used.
- D2: April-September English and July-December Chinese articles are used.

Figure 6 shows the MRR comparisons of our method $T_N^{(2)}$ and $T_N^{(1)}$ on all test entities. Because the commonly appearing NEs are decreasing, the performance decline is inevitable. However, we can see that the MRR of the seed translation method drops significantly on D1 and D2, whereas our method shows 0.14 higher MRR for both cases.

5.5 Similar Documents

In this section, we show an example of similar documents in Figure 5. Both articles describe the same event about the visit of Mandelson to Moscow for the discussion on the joining of Russia to the WTO. The extracted statements are the exact translations of each corresponding part as indicated by the arrows. We stress this is an extreme case for illustration, where the two sentences are almost an exact translation, except for a minor asymmetry involving the date (Thursday in English, and 14th in Chinese). In most similar documents, the asymmetry is more significant. The seed translation score T_N^1(WTO, 世贸组织) is not enough to match the entities. However, the context similarity, due to other similar statements such as (1), allows us to match (2). This match helps translation of 'WTO' by inspecting the organization that Russia considers to join in both documents.

6 Conclusions

This paper proposed a bootstrapping approach for entity translation using multilingual relational clustering. Further, the proposed method could finds similar document pairs by comparing statements to enable us to focus on comparable parts of evidence. We validated the quality of our approach using real-life English and Chinese corpora, and its performance significantly exceeds that of previous approaches.

Acknowledgment

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency). (NIPA-2012-H0503-12-1036).

References

Donghui Feng. 2004. A new approach for english-chinese named entity alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 372–379.

Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining bilingual data from the web with adaptively learnt patterns. In *Joint Conference of the ACL and the IJCNLP*, pages 870–878, Stroudsburg, PA, USA.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. A cross-lingual annotation projection approach for relation detection. In *COLING*, pages 564–571, Stroudsburg, PA, USA.

Jinhan Kim, Long Jiang, Seung-won Hwang, Young-In Song, and Ming Zhou. 2011. Mining entity translations from comparable corpora: a holistic graph mapping approach. In *CIKM*, pages 1295–1304, New York, NY, USA.

Jinhan Kim, Seung won Hwang, Long Jiang, Young-In Song, and Ming Zhou. 2012. Entity translation mining from comparable corpora: Combining graph mapping with corpus latent features. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints).

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 82–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Comput. Linguist.*, 24(4):599–612, December.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *ACL*, pages 17–22, Stroudsburg, PA, USA.

Wai Lam, Shing-Kit Chan, and Ruizhang Huang. 2007. Named entity translation matching and learning: With application for mining unseen translations. *ACM Trans. Inf. Syst.*, 25(1), February.

Dekang Lin, Shaojun Zhao, Benjamin Van Durme, and Marius Pasca. 2008. Mining parenthetical translations from the web by word alignment. In *ACL*.

Thahir Mohamed, Estevam Hruschka, and Tom Mitchell. 2011. Discovering relations between noun categories. In *EMNLP*, pages 1447–1455, Edinburgh, Scotland, UK., July.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *EMNLP*.

Patrick Schone, Tim Allison, Chris Giannella, and Craig Pfeifer. 2011. Bootstrapping multilingual relation discovery using english wikipedia and wikimedia-induced entity extraction. In *ICTAI*, pages 944–951, Washington, DC, USA.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *COLING*, Stroudsburg, PA, USA.

S. Van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht, The Netherlands.

Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic english-chinese name transliteration for development of multilingual resources. In *ACL*, pages 1352–1356, Stroudsburg, PA, USA.

Douglas Brent West. 1999. *Introduction to graph theory (2nd edition)*. Prentice Hall.

Fan Yang, Jun Zhao, and Kang Liu. 2009. A chinese-english organization name translation system using heuristic web mining and asymmetric alignment. In *Joint Conference of the ACL and the IJCNLP*, pages 387–395, Stroudsburg, PA, USA.

Gae-won You, Seung-won Hwang, Young-In Song, Long Jiang, and Zaiqing Nie. 2010. Mining name translations from entity graph mapping. In *EMNLP*, pages 430–439, Stroudsburg, PA, USA.

Gae-Won You, Seung-Won Hwang, Young-In Song, Long Jiang, and Zaiqing Nie. 2012. Efficient entity translation mining: A parallelized graph alignment approach. *ACM Trans. Inf. Syst.*, 30(4):25:1–25:23, November.