Enriching Entity Translation Discovery using Selective Temporality

Gae-won You, Young-rok Cha, Jinhan Kim, and Seung-won Hwang

Pohang University of Science and Technology, Republic of Korea
{gwyou, line0930, wlsgks08, swhwang}@postech.edu

Abstract

This paper studies named entity translation and proposes "selective temporality" as a new feature, as using temporal features may be harmful for translating "atemporal" entities. Our key contribution is building an automatic classifier to distinguish temporal and atemporal entities then align them in separate procedures to boost translation accuracy by 6.1%.

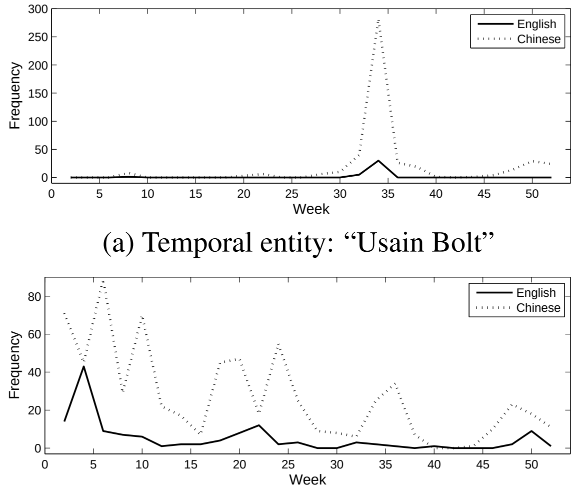
1 Introduction

Named entity translation discovery aims at mapping entity names for people, locations, *etc.* in source language into their corresponding names in target language. As many new named entities appear every day in newspapers and web sites, their translations are non-trivial yet essential.

Early efforts of named entity translation have focused on using *phonetic feature* (called PH) to estimate a phonetic similarity between two names (Knight and Graehl, 1998; Li et al., 2004; Virga and Khudanpur, 2003). In contrast, some approaches have focused on using *context feature* (called CX) which compares surrounding words of entities (Fung and Yee, 1998; Diab and Finch, 2000; Laroche and Langlais, 2010).

Recently, holistic approaches combining such similarities have been studied (Shao and Ng, 2004; You et al., 2010; Kim et al., 2011). (Shao and Ng, 2004) rank translation candidates using PH and CX independently and return results with the highest average rank. (You et al., 2010) compute initial translation scores using PH and iteratively update the scores using *relationship feature* (called R). (Kim et al., 2011) boost You's approach by additionally leveraging CX.

More recent approaches consider *temporal feature* (called T) of entities in two corpora (Klementiev and Roth, 2006; Tao et al., 2006; Sproat et al., 2006; Kim et al., 2012). T is computed using frequency vectors for entities and combined with PH (Klementiev and Roth, 2006; Tao et al., 2006). (Sproat et al., 2006) extend Tao's approach by iteratively updating overall similarities using R. (Kim et al., 2012) holistically combine all the features: PH, CX, T, and R.



(a) Temporal entity: "Usain Bolt"

(b) Atemporal entity: "Hillary Clinton"

Figure 1: Illustration on temporality

However, T used in previous approaches is a good feature only if temporal behaviors are "symmetric" across corpora. In contrast, Figure 1 illustrates asymmetry, by showing the frequencies of "Usain Bolt," a Jamaican sprinter, and "Hillary Clinton," an American politician, in comparable news articles during the year 2008. The former is mostly mentioned in the context of some temporal events, *e.g.*, Beijing Olympics, while the latter is not. In such case, as Hillary Clinton is a famous female leader, she may be associated with other Chinese female leaders in Chinese corpus, while such association is rarely observed in English corpus, which causes asymmetry. That is, Hillary Clinton is "atemporal," as Figure 1(b) shows, such that using such dissimilarity against deciding this pair as a correct translation would be harmful. In clear contrast, for Usain Bolt, similarity of temporal dis-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 201–205,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

tributions in Figure 1(a) is a good feature for concluding this pair as a correct one.

To overcome such problems, we propose a new notion of "selective temporality" (called this feature ST to distinguish from T) to automatically distinguish temporal and atemporal entities. Toward this goal, we design a classifier to distinguish temporal entities from atemporal entities, based on which we align temporal projections of entity graphs for the temporal ones and the entire entity graphs for the atemporal ones. We also propose a method to identify the optimal window size for temporal entities. We validate this "selective" use of temporal features boosts the accuracy by 6.1%.

2 Preliminaries

Our approach follows a graph alignment framework proposed in (You et al., 2010). Our graph alignment framework consists of 4 steps.

2.1 Step 1: Graph Construction

We first build a graph $G = (V, E)$ from each language corpus, where V is a set of entities (nodes) and E is a set of co-occurrence relationships (unweighted edges) between entities. We consider entities occurring more than η times as nodes and entity pairs co-occurring more than σ times as edges.

To identify entities, we use a CRF-based named entity tagger (Finkel et al., 2005) and a Chinese word breaker (Gao et al., 2003) for English and Chinese corpora, respectively.

2.2 Step 2: Initialization

Given two graphs $G_e = (V_e, E_e)$ and $G_c = (V_c, E_c)$, we initialize $|V_e|$-by-$|V_c|$ initial similarity matrix R^0 using PH and CX for every pair (e, c) where $e \in V_e$ and $c \in V_c$.

For PH, we use a variant of Edit-Distance (You et al., 2010) between English entity and a romanized representation of Chinese entity called Pinyin. For CX, the context similarity is computed based on entity context which is defined as a set of words near to the entity (we ignore some words such as stop words and other entities). We compute similarity of the most frequent 20 words for each entity using a variant of Jaccard index. To integrate two similarity scores, we adopt an average as a composite function.

We finally compute initial similarity scores for all pairs (e, c) where $e \in V_e$ and $c \in V_c$, and build the initial similarity matrix R^0.

2.3 Step 3: Reinforcement

We reinforce R^0 by leveraging R and obtain a converged matrix R^∞ using the following model:

$$R^{t+1}_{(i,j)} = \lambda R^0_{(i,j)} + (1-\lambda) \sum_{(u,v)_k \in B^t(i,j,\theta)} \frac{R^t_{(u,v)}}{2^k}$$

This model is a linear combination of (a) the initial similarity $R^0_{(i,j)}$ of entity pair $(i, j) \in V_e \times V_c$ and (b) the similarities $R^t_{(u,v)}$ of their matched neighbors $(u, v) \in V_e \times V_c$ where t indicates iteration, $B^t(i, j, \theta)$ is an ordered set of the matched neighbors, and k is the rank of the matched neighbors. λ is the coefficient for balancing two terms.

However, as we cannot assure the correctly matched neighbors (u, v), a chicken-and-egg dilemma, we take advantage of the current similarity R^t to estimate the next similarity R^{t+1}. Algorithm 1 describes the process of matching the neighbors where $N(i)$ and $N(j)$ are the sets of neighbor nodes of $i \in V_e$ and $j \in V_c$, respectively, and H is a priority queue sorting the matched pairs in non-increasing order of similarities. To guarantee that the neighbors are correctly matched, we use only the matches such that $R^t_{(u,v)} \geq \theta$.

Algorithm 1 $B^t(i, j, \theta)$

```
$M \leftarrow \{\}; H \leftarrow \{\}$
$\forall u \in N(i), \forall v \in N(j)$ $H$.push$(u, v)$ such that $R^t_{(u,v)} \geq \theta$
while $H$ is not empty do
    $(u, v) \leftarrow H$.pop()
    if neither $u$ nor $v$ are matched yet then
        $M \leftarrow M \cup \{(u, v)\}$
    end if
end while
return $M$
```

2.4 Step 4: Extraction

From R^∞, we finally extract one-to-one matches by using simple greedy approach of three steps: (1) choosing the pair with the highest similarity score; (2) removing the corresponding row and column from R^∞; (3) repeating (1) and (2) until the matching score is not less than a threshold δ.

3 Entity Translation Discovery using Selective Temporality

Overall Framework: We propose our framework by putting together two separate procedures for temporal and atemporal entities to compute the overall similarity matrix R

We first build two temporal graphs from the corpora within every time window, optimized in Section 3.1. We then compute the reinforced matrix R_s^∞ obtained from the window starting at the time-stamp s. To keep the best match scores among all windows, we update R using the best similarity among $\forall s, R_s^\infty$. we then extract the candidate translation pairs M_{ours} by running step 4.

As there can exist atemporal entities in M_{ours}, we classify them (Section 3.2). Specifically, we build two entire graphs and compute R^∞. We then distinguish temporal entities from atemporal ones using our proposed metric for each matched pair $(i, j) \in M_{ours}$ and, if the pair is atemporal, $R_{(i,j)}$ is updated as the atemporal similarity $R_{(i,j)}^\infty$.

From the final matrix R, we extract the matched pairs by running step 4 with R once again.

3.1 Projecting Graph for Temporal Entities

We first project graphs temporally to improve translation quality for temporal entities. As the optimal projection would differ across entities, we generate many projected graphs by shifting time window over all periods, and then identify the best window for each entity.

The rest of this section describes how we set the right window size w. Though each entity may have its own optimal w, we find optimizing for each entity may negatively influence on considering relationships with entities of different window sizes. Thus, we instead find the optimal window size $\hat{w}$ to maximize the global "symmetry" of the given two graphs.

We now define "symmetry" with respect to the truth translation pair M. We note it is infeasible to assume we have M during translation, and will later relax to consider how M can be approximated.

Given a set of graph pairs segmented by the shifted windows

$$\{(G_e^{(0,w)}, G_c^{(0,w)}), \cdots, (G_e^{(s,s+w)}, G_c^{(s,s+w)}),$$
$$(G_e^{(s+\Delta s, s+\Delta s+w)}, G_c^{(s+\Delta s, s+\Delta s+w)}), \cdots\},$$

where s is the time-stamp, our goal is to find the window size $\hat{w}$ maximizing the average symmetry S of graph pairs:

$$\hat{w} = \arg\max_{\forall w} \left(\frac{\sum_s S(G_e^{(s,s+w)}, G_c^{(s,s+w)}; M)}{N} \right)$$

Given M, symmetry S can be defined for (1) *node* and (2) *edge* respectively. We first define the *node symmetry* S_n as follows:

$$S_n(G_e, G_c; M) = \frac{\sum_{(e,c) \in V_e \times V_c} I(e, c; M)}{\max\{|V_e|, |V_c|\}}$$

where $I(u, v; M)$ to be 1 if $(u, v) \in M$, 0 otherwise. High node symmetry leads to accurate translation in R^0 (Initialization step). Similarly, we define the *edge symmetry* S_e as follows:

$$S_e(G_e, G_c; M) =$$
$$\frac{\sum_{(e_1,e_2) \in E_e} \sum_{(c_1,c_2) \in E_c} I(e_1, c_1; M) I(e_2, c_2; M)}{\max\{|E_e|, |E_c|\}}$$

In contrast, high edge symmetry leads to accurate translation in R^∞ (Reinforcement step).

We finally define the symmetry S as the weighted sum of S_n and S_e with parameter α (empirically tuned to 0.8 in our experiment).

$$S(G_e, G_c; M) =$$
$$\alpha S_n(G_e, G_c; M) + (1 - \alpha) S_e(G_e, G_c; M)$$

However, as it is infeasible to assume we have the truth translation pair M, we approximate M using intermediate translation results M_{ours} computed at step 4. To insert only true positive pairs in M_{ours}, we set threshold higher than the optimized value from the step 4. We found out that symmetry from M_{ours} closely estimates that from M:

$$S(G_e, G_c; M) \approx S(G_e, G_c; M_{\text{ours}})$$

Specifically, observe from Table 1 that, given a manually built ground-truth set $M_g \subset M$ as described in Section 4.1, $S(G_e, G_c; M_{\text{ours}})$ returns the best symmetry value in two weeks for person entities, which is expectedly the same as the result of $S(G_e, G_c; M_g)$. This suggests that we can use M_{ours} for optimizing window size.

Weeks	26	13	4	**2**	1
M_g	.0264	.0276	.0303	**.0318**	.0315
M_{ours}	.0077	.0084	.0102	**.0113**	.0107

Table 1: Symmetry of window size

3.2 Building Classifier

We then classify temporal/atemporal entities. As a first step, we observe their characteristics: **Temporal entities** have peaks in the frequency distribution of both corpora and these peaks are aligned, while such distribution of **atemporal entities** are more uniform and less aligned.

Based on these observations, we identify the following criteria for temporal entities: (1) Their two distributions $\mathbf{m}$ in English corpus and $\mathbf{n}$ in Chinese corpus should have aligned peaks. (2) Frequencies at the peaks are the higher the better.

For the *first criterion*, we first normalize the two vectors $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$ since two corpora have different scales, *i.e.*, different number of documents. We then calculate the inner product of the two vectors $\mathbf{x} = \langle \hat{\mathbf{m}}, \hat{\mathbf{n}} \rangle$, such that this aggregated distribution $\mathbf{x}$ peaks, only if both $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$ peak at the same time.

For the *second criterion*, we have a spectrum of option from taking the frequencies at all peaks in one extreme, to taking only the maximum frequency in another extreme. A metric representing such a spectrum is p-norm, which represents sum when $p = 1$ and maximum when $p = \infty$. We empirically tune the right balance to distinguish temporal and atemporal entities, which turns out to be $p = 2.2$.

Overall, we define a metric $d(\mathbf{m}, \mathbf{n})$ which satisfies both criteria as follow:

$$d(\mathbf{m}, \mathbf{n}) = \left(\sum_{i=1}^{n} (\hat{\mathbf{m}}_i \hat{\mathbf{n}}_i)^p \right)^{\frac{1}{p}}$$

For instance, this measure returns 0.50 and 0.03 for the distributions in Figure 1(a) and (b), respectively, from which we can determine the translation of Figure 1(a) is temporal and the one of Figure 1(b) is atemporal.

4 Experimental Evaluation

4.1 Experimental Settings

We obtained comparable corpora from English and Chinese Gigaword Corpora (LDC2009T13 and LDC2009T27) published by the Xinhua News Agency during the year 2008. From them, we extracted person entities and built two graphs, $G_e = (V_e, E_e)$ and $G_c = (V_c, E_c)$ by setting $\eta = 20$ which was used in (Kim et al., 2011).

Next, we built a ground truth translation pair set M_g for person entities. We first selected 500 person names randomly from English corpus. We then hired a Chinese annotator to translate them into their Chinese names. Among them, only 201 person names were matched to our Chinese corpus. We used all such pairs to identify the best parameters and compute the evaluation measures.

We implemented and compared the following approaches denoted as the naming convention of listing of the used features in a parenthesis ():

- (PH+R) in (You et al., 2010).
- (PH+CX+R) in (Kim et al., 2011).
- (PH+CX+R+T) in (Kim et al., 2012).
- (PH+CX+R+ST): This is our approach.

We evaluated the effectiveness of our new approach using four measures: MRR, precision, recall, and F1-score, where MRR (Voorhees, 2001) is the average of the reciprocal ranks of the query results defined as follows:

$$\text{MRR} = \frac{1}{|Q|} \sum_{(u,v) \in Q} \frac{1}{rank_{(u,v)}},$$

where Q is a set of ground-truth matched pairs (u, v) such that $u \in V_e$ and $v \in V_c$, and $rank_{(u,v)}$ is the rank of $R_{(u,v)}$ among all $R_{(u,w)}$'s such that $w \in V_c$. We performed a 5-fold cross validation by dividing ground truth into five groups. We used four groups to training the parameters to maximize F1-scores, used the remaining group as a test-set using trained parameters, and computed average of five results. (**bold numbers** indicate the best performance for each metric.)

4.2 Experimental Results

Effect of window size

We first validated the effectiveness of our approach for various window sizes (Table 2). Observe that it shows the best performance in two weeks for MRR and F1 measures. Interestingly, this result also corresponds to our optimization result $\hat{w}$ of Table 1 in Section 3.1.

Weeks	26	13	4	**2**	1
MRR	.7436	.8066	.8166	**.8233**	.8148
Precision	.7778	.7486	.8126	.8306	**.8333**
Recall	.6617	.6875	**.7320**	.7295	.7214
F1	.7151	.7165	.7701	**.7765**	.7733

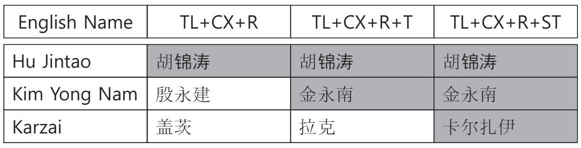
Table 2: Optimality of window size

Overall performance

Table 3 shows the results of four measures. Observe that (PH+CX+R+T) and (PH+CX+R+ST) outperform the others in all our settings. We can also observe the effect of selective temporality, which maximizes the symmetry between two graphs as shown in Table 1, *i.e.*, (PH+CX+R+ST)

Method	MRR	Precision	Recall	F1
(PH+R)	.6500	.7230	.4548	.5552
(PH+CX+R)	.7499	.7704	.6623	.7120
(PH+CX+R+T)	.7658	.8223	.6608	.7321
(PH+CX+R+ST)	**.8233**	**.8306**	**.7295**	**.7765**

Table 3: MRR, Precision, Recall, and F1-score



English Name	TL+CX+R	TL+CX+R+T	TL+CX+R+ST
Hu Jintao	胡锦涛	胡锦涛	胡锦涛
Kim Yong Nam	殷永建	金永南	金永南
Karzai	盖茨	拉克	卡尔扎伊

Figure 2: The translation examples where shaded cells indicate the correctly translated pairs.

outperforms (PH+CX+R+T) by 6.1%. These improvements were statistically significant according to the Student's t-test at $P < 0.05$ level.

Figure 2 shows representative translation examples. All approaches found famous entities such as "Hu Jintao," a former leader of China, but (PH+CX+R) failed to find translation of lesser known entities, such as "Kim Yong Nam." Using temporal features help both (PH+CX+R+T) and (PH+CX+R+ST) identify the right translation, as Kim's temporal occurrence is strong and symmetric in both corpora. In contrast, (PH+CX+R+T) failed to find the translation of "Karzai", the president of Afghanistan, as it only appears weakly and transiently during a short period time, for which only (PH+CX+R+ST) applying varying sizes of window per entity is effective.

5 Conclusion

This paper validated that considering temporality selectively is helpful for improving the translation quality. We developed a classifier to distinguish temporal/atemporal entities and our proposed method outperforms the state-of-the-art approach by 6.1%.

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

Mona Diab and Steve Finch. 2000. A statistical word level translation model for comparable corpora. In *RIAO '00*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel,Comparable Texts. In *COLING*.

Jianfeng Gao, Mu Li, and Chang-Ning Huang. 2003. Improved Source-channel Models for Chinese Word Segmentation. In *ACL*.

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*.

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 TKDE*.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *HLT-NAACL '06*.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *COLING*.

Haizhou Li, Zhang Min, and Su Jian. 2004. A Joint Source-Channel Model for Machine Transliteration. In *ACL*.

Li Shao and Hwee Tou Ng. 2004. Mining New Word Translations from Comparable Corpora. In *COLING*.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named Entity Transliteration with Comparable Corpora. In *ACL*.

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai. 2006. Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation. In *EMNLP*.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-language applications. In *SIGIR '03*.

Ellen M. Voorhees. 2001. The TREC Question Answering Track. *Natural Language Engineering*, 7(4):361–378.

Gae-won You, Seung-won Hwang, Young-In Song, Long Jiang, and Zaiqing Nie. 2010. Mining Name Translations from Entity Graph Mapping. In *Proceedings of EMNLP*, pages 430–439.