Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks

Heng Ji
Computer Science Department, Queens College and the Graduate Center
The City University of New York, New York, NY, 11367, USA
hengji@cs.qc.cuny.edu

Abstract

This paper describes a new task to extract and align information networks from comparable corpora. As a case study we demonstrate the effectiveness of this task on automatically mining name translation pairs. Starting from a small set of seeds, we design a novel approach to acquire name translation pairs in a bootstrapping framework. The experimental results show this approach can generate highly accurate name translation pairs for persons, geopolitical and organization entities.

1 Introduction

Accurate name translation is crucial to many cross-lingual information processing tasks such as information retrieval (e.g. Ji et al., 2008). Recently there has been heightened interest in discovering name pairs from comparable corpora (e.g. Sproat et al., 2006; Klementiev and Roth, 2006). By comparable corpora we mean texts that are about similar topics, but are not in general translations of each other. These corpora are naturally available, for example, many news agencies release multi-lingual news articles on the same day. There are no document-level or sentence-level alignments across languages, but important facts such as names, relations and events in one language in such corpora tend to co-occur with their counterparts in the other.

However, most of the previous approaches used a phonetic similarity based name transliteration module as baseline to generate translation hypotheses, and then exploit the distribution evidence from comparable corpora to re-score these hypotheses. As a result, these approaches are limited to names which are phonetically transliterated (e.g. translate Chinese name “*尤申科* (*You shen ke)*” to “*Yushchenko*” in English). But many other types of names such as organizations are often rendered semantically, for example, the Chinese name “*解放之虎 (jie fang zhi hu)*” is translated into “*Liberation Tiger*” in English. Furthermore, many name translations are context dependent. For example, a person name “*亚西尔·阿拉法特*” should be translated into “*Yasser Arafat (PLO Chairman)*” or “*Yasir Arafat (Cricketer)*” based on different contexts.

Information extraction (IE) techniques – identifying important entities, relations and events – are currently available for some non-English languages. In this paper we define a new notion ‘*bilingual information networks*’ which can be extracted from comparable corpora. An information network is a set of directed graphs, in which each node is a named entity and the nodes are linked by various ‘attributes’ such as hometown, employer, spouse etc. Then we align the information networks in two languages automatically in a bootstrapping way to discover name translation pairs. For example, after we extract bilingual

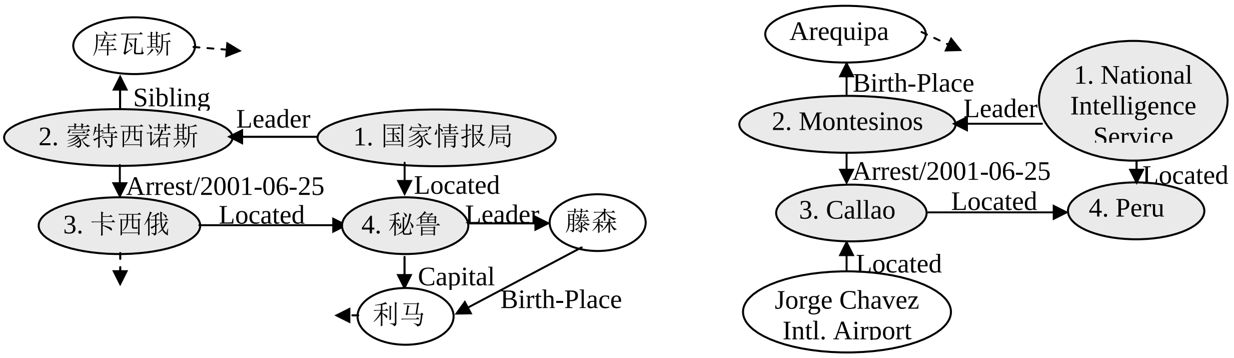

Figure 1. An example for Bilingual Information Networks

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 34–37,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

information networks as shown in Figure 1, we can start from a common name translation "国家情报局-National Intelligence Service (1)", to align its *leader* as "蒙特西诺斯- Montesinos (2)", align the *arrest place* of Montesinos as "卡西俄-Callao (3)", and then align the *location* of Callao as "秘鲁-Peru (4)". Using this approach we can discover name pairs of various types (person, organization and location) while minimizing using supervised name transliteration techniques. At the same time, we can provide links among names for entity disambiguation.

2 General Approach

Figure 2 depicts the general procedure of our approach. The language pair that we are considering in this paper is Chinese and English. We apply IE techniques to extract information networks (more details in section 3), then use a bootstrapping algorithm to align them and discover name pairs (section 4).

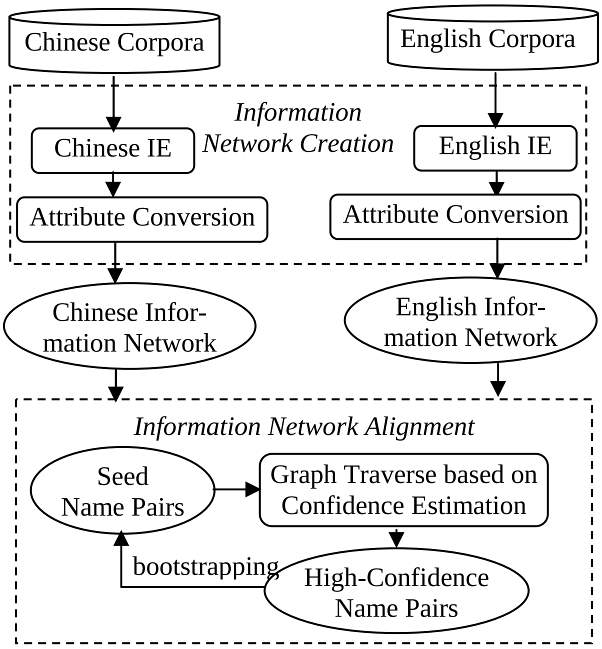
Chinese Corpora
English Corpora
Information Network Creation
Chinese IE
English IE
Attribute Conversion
Attribute Conversion
Chinese Information Network
English Information Network
Information Network Alignment
Seed Name Pairs
Graph Traverse based on Confidence Estimation
bootstrapping
High-Confidence Name Pairs

Figure 2. Name Translation Mining Overview

3 Information Network Creation

3.1 Bilingual Information Extraction

We apply a state-of-the-art bilingual information extraction system (Chen and Ji, 2009; Ji and Grishman, 2008) to extract ACE[1] types of entities, relations and events from the comparable corpora. Both systems include name tagging, nominal mention tagging, coreference resolution, time expression extraction and normalization, relation extraction and event extraction. Entities include persons, geo-political (GPE) and organizations; Relations include 18 types (e.g. "*a town some 50 miles south of Salzburg*" indicates a *located* relation.); Events include the 33 distinct event types defined in ACE05 (e.g. *"Barry Diller on Wednesday quit as chief of Vivendi"* indicates that *"Barry Diller"* is the *person argument* of a *quit* event occurred on *Wednesday*). The relation extraction and event extraction components produce confidence values.

3.2 Attribute Conversion

Then we construct a set of directed graphs for each language $G = \{G_i(V_i, E_i)\}$, where V_i is the collection of named entities, and E_i is the edges linking one name to the other, labeled by the attributes derived from the following two sources: (1) We select the relations with more static types to form specific attributes in Table 2[2], according to the entity types of a linked name pair. (2) For each extracted event we compose an attribute by combining its type and time argument (e.g. the "Arrest/2001-06-25" link in Figure 1). As we will see in the next section, these attributes are the key to discover name translations from the information networks because they are language-independent.

4 Information Network Alignment

After creating the information networks from each language, we automatically align them to discover name translation pairs. The general idea is that starting from a small seed set of common name pairs, we can rely on the link attributes to align their related names. Then the new name translations are added to the seed set for the next iteration. We repeat this bootstrapping procedure until no new translations are produced. We start from names which are frequently linked to others so that we can traverse through the information networks efficiently. For example, the seed set in processing ACE newswire data includes famous names such as "Indonesia", "China", "Palestine", "Sharon" and "Yugoslavia".

For each name pair <*CHName*, *EName*>, we search for all its related pairs <*CHName'*,

[1] http://www.itl.nist.gov/iad/mig//tests/ace/

[2] Many of these attributes are consistent with the definitions in NIST TAC-KBP task: http://apl.jhu.edu/~paulmac/kbp/090220-KBPTaskGuidelines.pdf

Name’ Name	Person	Geo-political	Organization
Person	Spouse, Parent, Child, Sibling	Birth-Place, Death-Place, Resides-Place, Nationality	Schools-Attended, Employer
Geo-political	Leader	Located-Country, Capital	-
Organization	Leader	Location	-

Table 2. Relation-driven Attributes (Name → Name’) in Information Network

Language Corpus	Chinese	English
ACE	CHSet1: XIN Oct-Dec 2000: 150 documents	ENSet1: APW Oct-Dec 2000: 150 documents ENSet2: AFP&APW Mar-June 2003: 150 documents
TDT-5	CHSet3: XIN Apr-Aug 2003: 30,000 documents	ENSet3: XIN Apr-Aug 2003: 30,000 documents ENSet4: AFP Apr-Aug 2003: 30,000 documents

Table 3. Number of Documents

ENName’>. Assuming *CHName* is linked to *CHName*’ by an edge *CHEdge*, and *ENName* is linked to *ENName’* by *ENEdge*, then if the following conditions are satisfied, we align *CHName’* and *ENName’* and add them as seeds for the next iteration:

- *CHEdge* and *ENEdge* are generated by IE systems with confidence values higher than thresholds;
- *CHEdge* and *ENEdge* have the same attributes;
- *CHName’* and *ENName’* have the same entity type;
- If *CHName’* and *ENName’* are persons, the Damerau–Levenshtein edit distance between the pinyin form of *CHName’* and *ENName’* is lower than a threshold.

It’s worth noting that although we exploit the pinyin information as essential constraints, this approach differs from the standard transliteration models which convert pinyin into English by adding/deleting/replacing certain phonemes.

5 Experimental Results

5.1 Data

We use some documents from the ACE (2004, 2005) training corpora and TDT-5 corpora to manually evaluate our approach. Table 3 shows the number of documents from different news agencies and time frames. We hold out 20 ACE texts from each language to optimize the thresholds of confidence values in section 4. A name pair <*CHName*, *EName*> is judged as correct if both of them are correctly extracted and one is the correct translation of the other in the certain contexts of the original documents.

5.2 Overall Performance

Table 4 shows the number and accuracy of name translation pairs discovered from CH-Set3 and EN-Set3, using 100 name pairs as seeds. After four iterations we discovered 968 new name translation pairs with accuracy 82.9%. Among them there are 361 persons (accuracy 76.4%), 384 geo-political names (accuracy 87.5%) and 223 organization names (accuracy 85.2%).

Iteration	1	2	3	4
Number of Name Pairs	205	533	787	968
Accuracy (%)	91.8	88.5	85.8	82.9

Table 4. Overall Performance

5.3 Impact of Time Frame and News Source Similarity

One major evidence exploited in the prior work is that the bilingual comparable corpora should be weakly temporally aligned. For example, Klementiev and Roth (2006) used the time distribution of names to re-score name transliteration. In order to verify this observation, we investigated how well our new approach can perform on comparable corpora with different time frames. Table 5 presents the performance of two combinations: CHSet1-ENSet1 (from the same time frame) and CHSet1-ENSet2 (from different time frames) with a seed set of 10 name pairs after 5 iterations.

Corpora	CHSet1-ENSet1	CHSet1-ENSet2
Number of Name Pairs	42	17
Accuracy (%)	81.0	76.5

Table 5. Impact of Time Frame Similarity

In addition, in order to measure the impact of news source similarity, we apply our approach to the combination of CHSet3 and ENSet4 which are from different news agencies. In total 815 name pairs are discovered after 4 iterations with overall accuracy 78.7%, which is worse than the results from the corpora of the same news source as shown in Table 4. Therefore we can clearly see that time and news source similarities are important to the performance of name translation pair mining.

5.4 Impact of IE Errors

Since in our approach we used the fully automatic IE pipeline to create the information networks, the errors from each component will be propagated into the alignment step and thus limit the performance of name translation discovery. For example, Chinese name boundary detection errors caused about 30% of the incorrect name pairs. As a diagnostic analysis, we tried to discover name pairs from CHSet1 and ENSet1 but with perfect IE annotations. We obtained 63 name pairs with a much higher accuracy 90.5%.

6 Related Work

Most of the previous name translation work combined supervised transliteration approaches with Language Model based re-scoring (e.g. Al-Onaizan and Knight, 2002; Huang et al., 2004). Ji et al. (2009) described various approaches to automatically mine name translation pairs from aligned phrases (e.g. cross-lingual Wikipedia title links) or aligned sentences (bi-texts). Our approach of extracting and aligning information network from comparable corpora is related to some prior work using comparable corpora to re-score name transliterations (Sproat et al., 2006; Klementiev and Roth, 2006).

In this paper we extend the target names from persons to geo-political and organization names, and extract relations links among names simultaneously. And we use a bootstrapping approach to discover name translations from the bilingual information networks of comparable corpora. In this way we don't need to have a name transliteration module to serve as baseline, or compute document-wise temporal distributions.

7 Conclusion and Future Work

We have described a simple approach to create bilingual information networks and then discover name pairs from comparable corpora. The experiments on Chinese and English have shown that this method can generate name translation pairs with high accuracy by using a small seed set. In the short term, our approach will provide a framework for many byproducts and directly benefit other NLP tasks. For example, the aligned sub-graphs with names, relations and events can be used to improve information redundancy in cross-lingual question answering; the outlier (mis-aligned) sub-graphs can be used to detect the novel or local information described in one language but not in the other.

In the future we plan to import more efficient graph mining and alignment algorithms which have been widely used for protein-protein interaction detection (Kelley et al., 2003). In addition, we will attempt using unsupervised relation extraction based on lexical semantics to replace the supervised IE pipeline. More importantly, we will investigate the tradeoff between coverage and accuracy by applying the generated name pairs to cross-lingual name search and machine translation tasks.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023 via 27-001022, and the CUNY Research Enhancement Program and GRTI Program.

References

Y. Al-Onaizan and K. Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. *Proc. ACL.*

Z. Chen and H. Ji. 2009. Language Specific Issue and Feature Exploration in Chinese Event Extraction. *Proc. HLT-NAACL.*

F. Huang, S. Vogel and A. Waibel. 2004. Improving Named Entity Translation Combining Phonetic and Semantic Similarities. *Proc. HLT/NAACL.*

H. Ji, R. Grishman, D. Freitag, M. Blume, J. Wang, S. Khadivi, R. Zens and H. Ney. 2009. Name Translation for Distillation. *Global Automatic Language Exploitation*.

H. Ji R. Grishman. 2008. Refining Event Extraction Through Cross-document Inference. *Proc. ACL.*

H. Ji, R. Grishman and W. Wang. 2008. Phonetic Name Matching for Cross-lingual Spoken Sentence Retrieval. *Proc. IEEE-ACL SLT*.

B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D.E. Root, B. R. Stockwell and T. Ideker. 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *The National Academy of Sciences of the United States of America*.

A. Klementiev and D. Roth. 2006. Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. *Proc. HLT-NAACL.*

R. Sproat, T. Tao and C. Zhai. 2006. Named Entity Transliteration with Comparable Corpora. *Proc. ACL.*