International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.9 (2014), pp.143-150
http://dx.doi.org/10.14257/ijmue.2014.9.9.16

Finding and Typing New Named Entities in Tibetan from Chinese-Tibetan Parallel Corpora

Lirong Qiu

School of Information Engineering, Minzu University of China
Beijing, China
qiu_lirong@126.com

Abstract

Currently there is much interest in the automatic acquisition of entities, with the goal of Named Entity Recognition (NER). However previous work has focused primarily on major languages, with the large, structured, and semantically rich knowledge bases and using the large corpus with annotated NER tags. In this paper, we describe a method for Chinese-Tibetan bilingual named entity recognition using easily obtainable bilingual dictionary and parallel political corpora. We present two distinct steps for NER, one step identifying entity candidates in Tibetan, and the second step typing the entity into the semantic class. We then test the approach on the dataset and give the analysis of NE type errors.

Keywords: *named entity recognition, bilingual NE alignment, semantic dictionary, knowledge base*

1. Introduction

Extracting and acquiring entities and their relations from text automatically is a long-standing issue in Natural language processing. Named-entity recognition (NER) seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. NER technologies are important for various applications including question answering system, translation system and information retrieval system [1].

Over the last decade or so, the enormous abundance of information and data that has become available as made it possible to extract huge amounts of patterns and named entities. Most research on NER systems has been structured as taking an unannotated block of text, and producing an annotated block of text that highlights where the named entities are.

However, when interpreting unrestricted, domain-independent text, it is difficult to determine in advance what kind of entities will be focused and how it will be expressed, especially without the supports of the large, structured, and semantically rich knowledge bases. Building a NER system for a language is still crucially depends on the existence of manually annotated texts as training data. As the annotation is costly, one would like to leverage existing resources to minimize the human effort to construct corpora for a new language. While previous work has focused primarily on major languages, how to extend these results to other languages is the way to avoid working start from scratch.

In this line of consideration, we choose to focus on bilingual political corpora because for many less widely spoken languages and for political domains where entities such as person names, locations, organizations etc. are constantly being introduced. And the digitization of news often broadcasted in multiple languages, which has provided the opportunity for systematic study of bilingual NER. Since the political news are translated by government, the Tibetan translations were reviewed thoroughly to correct any wrong misunderstanding translations.

Our goal in this paper is to discover new entities in a target language that not registered in KB. To avoid working start from scratch, our main approach is to map lexico-syntactic named

ISSN: 1975-0080 IJMUE

entity projection from those resource-rich source languages (English, Chinese) to a resource-poor target language (Tibetan) using bilingual political corpora. The expectation is that this will boost the known entity names and the grouping of new entities in target language (*e.g.* Tibetan), and will also construct a named entity bilingual translation lexicon automatically.

The remainder of this paper is organized as follows. In Section 2 we outline the methodology of the work. Section 3 describes the name entity alignment method. In Section 4 we evaluate the proposed method and present experimental data used in our evaluations and discuss our results. In Section 5 we reviewed related work on bilingual named entity recognition. Section 6 draws the conclusions and mentions future work.

2. Motivation

Considering the bilingual entity pairs should share the same entity type and the detected NEs in source language should provide further information while selecting NEs in the target language, this work is proposed to abstract the entities away from the particular source language and apply it to a new one.

To select a large bilingual corpus annotated with NER tags as training data is essential for our work. Considering the fact that named entities, especially person names, location names, and organization names, are constantly being introduced in news corpus. On the other hand, some government report and news report are often published in multiple languages (e.g. Chinese, English, Tibetan, Uighur) in China. Thus, we select political corpus with tagged name entities as the training data.

Our method works by first pairing each entity extracted from a source language documents set *S* with each entity extracted from a target language document sets *T* aligned with *S* in the bilingual political corpus. The method we present below for bilingual named entity recognition is a bilingual alignment approach, i.e. it assumes a method exists for monolingual named entity recognition in source language.

Our aim is for all aligned newly discovered entities to be semantically interpretable by connecting them to KB classes according to the bilingual political corpus, which contains two main sub-tasks:

Entity candidate identification in target language: For each entity defined in source language, discover a list of entity candidates $E = \{e_1, e_2, \cdots, e_n\}$ in target language. First, relevant comparable data sets that include *m* are retrieved. In this paper, we collect parallel data from the report on the work of the government and government web documents such as news articles both in Chinese and in Tibetan. Then we apply various technologies such as word segmentation, part-of-speech tagging, noun phrase chunking and name tagging in both source language and target language. It is worth mentioning that since there are not much available natural language processing technologies in Tibetan language, the annotations in Tibetan language is far from useful. Therefore, the alignment from the source language to the target language is important for identifying the entity candidates.

Typing the new entity to semantic named entity classes: Rank the entity candidates in target language set *E*. Then, we should choose the named entity types from the definitions of NER hierarchy. Finally, the top ranked candidates should be typed into the chosen entity types.

In the NER hierarchy, we adopt two hierarchies of named entity types, BBN categories and Sekine's extended hierarchy. BBN categories are used for question answering and consist of 29 types and 64 subtypes [6]. Sekine's extended hierarchy is made of 200 subtypes [7]. Since which entity type should be chosen has been one of the considerations, we should abstract the entity types with respect to the definition of BBN and Sekine's extended hierarchy. To deduce types for new entities we propose to align new entities along the type of patterns they occur with.

3. The Proposed Bilingual NER Method

The NE identified in source language can be used to identify its less reliable counterpart in the target language, with respect to the NE pairs abstracted from the bilingual political corpus. This benefit can be largely used in those languages that NEs are initially incorrectly recognized with monolingual NER method.

3.1. Large-Scale Harvesting of Entities

The general goal of the first step is to identify a list of candidate entities in target language for each entity in source language acquired from the bilingual parallel political corpus including the report on the work of the government and web documents.

For a given sentence in source language, we extract all its possible entities. For each entity, there are two possibilities: (1) the entity is a known entity that can be directly mapped to the knowledge base; (2) the entity is a new entity not known to the knowledge base at all. We use an extensive dictionary for in-KB entities to determine whether refers to a known entity.

In this work, four universally accepted entity categories, person, location, organization and GPE are taken into consideration of the bilingual NER method. The types of entities are defined as PER, LOC, ORG and GPE, and many patterns are defined for each type.

The task is selecting Chinese entities and finding their Tibetan translations from the Chinese-Tibetan sentence pairs.

Given a Chinese-Tibetan sentence pair (s_c, s_t), a pair of a Chinese entity e_c (from s_c) and a Tibetan entity e_t (from s_t) are likely to indicate the same real world entity if the two conditions are satisfied: (1) under the same entity type, (2) semantically similar to the same entity.

In this study, we will apply bilingual dictionary and semantic similarity calculating method to acquire named entities.

3.1.1 Lexical Information: We have a NER dictionary in Chinese and Tibetan language, which is listed several examples as follows:

Table 1. Chinese-Tibetan Dictionary with NER Labels

Label	Chinese	Tibetan	English	Label	Chinese	Tibetan	English
LOC	林芝地区	ཉིན་ཁི་ས་ཁུལ་	Nyingchi	ORG	联合国	མཉམ་འབྲེལ་རྒྱལ་ཁབ་	The united Nations
LOC	昌都地区	ཆབ་མདོ་ས་ཁུལ་	Qamdo	ORG	教育部	སློབ་གསོ་སྡུད་	the Ministry of Education
LOC	青海	མཚོ་སྔོན་	Qinghai	ORG	国务院	རྒྱལ་སིད་སྡི་ཁུལ་ཁང་	the State Department
LOC	布拉格	པོ་ལན་གེའེ་	Prague	ORG	农业部	ཞིང་ལས་སྡུད་	the Department of Agriculture
PER	国家主席	རྒྱལ་ཁབ་ཀྱི་གྲུབ་ཞེ་	Chinese President	PER	军委主席	དམག་དོན་ཞུ་ཡོན་ཚན་ཁང་གི་གྲུབ་ཞེ་	Chairman of the Military Commission
PER	江泽民	ཇང་རྩེའེ་མིན་	Jiang Zemin	PER	胡锦涛	ཧུའུ་རྗིན་ཏོའོ་	Hu Jintao

The first option is simply to build a bilingual dictionary derived from a word-aligned parallel corpus. However there are many different transliterating names from one Chinese word to Tibetan language, mainly due to the differences in their sound and writing system. Especially in political domain, many new entity names are transliteration, and not to be found in bilingual dictionaries.

Given a Chinese-Tibetan sentence pair (s_c, s_t), entity e_c are found in the sentence s_c by applying monolingual NER method.

Definition 1 (bilingual entity pairs): Given a new entity e_c in source language, we abstract entity triples $<e_c, e_{ti}, dct_i>$, in which e_c is a source word, e_{ti} is the i^{th} translation of e_c, and dct_i ($dct_i \in [0,1]$) denotes the word pair relationship ($e_c \rightarrow e_{ti}$).

The definition of word pair relationship dct_i will be introduced in the next section.

3.1.2. Semantic Similarity: There is no guarantee that each of the words in the source-language is present in our dictionary, so our method would benefit from the ability to generalize closely related words by semantic knowledge base.

Given an entity e_c in source language, if there are no translations in target languages, other entities e_{c1}, e_{c2}… are taken into consideration, which share the same semantic meaning of e_c. Then we add the translations of e_{c1}, e_{c2}… into the target word set e_{t1}, e_{t2}, etc.

When comes to named entities, the traditional similarity calculation method is no longer fulfill the desire. For example, LOC type comprises patterns of "北京市" ("Beijing city"), "海淀区"("Haidian district"), and the synonym is hardly found in a dictionary. Whether the two words belong to same entity pattern , such as "* 市" ("* city"), is important to judge the similarity of two named entities.

To accomplish this goal in our work, we leverage the common situation that the sentence in source language only contains one novel or unknown entity not registered in the KB. The similarity between two words is calculated based on a semantic dictionary HowNet and the method is introduced in our work [12].

After the combining translations of the word and the corpus-based and semantic similarity calculating methods, the approach can, both 1) locate the target translations of a source word e_c and 2) give the similarity and entity type of target word for reference.

3.1.3. Target Candidate Ranking: The Chinese named entity candidate selection is using semantic dictionary and knowledge base as described above. Candidate Tibetan translations are generated by consulting a list of factors that can be acquired by the bilingual corpus. For Tibetan, there is no open-source corpus available with annotated NER tags. It is necessary to utilize a method to map named entity projection from Chinese to Tibetan.

Definition 2 (Word Pair Relationship): Given a new entity e_c in source language, search sentence pairs and select (s_{c1}, s_{t1}), (s_{c2}, s_{t2}), …, (s_{cm}, s_{tm}) if e_c appears in the source language sentences s_{ci}. Then the translations (e_{t1}, e_{t2}, …, e_{tn}) of e_c are acquired in sentence s_{ti} in target language. Then we add the e_{ti} (<e_c, e_{ti} >) into the candidate entity set.

To accomplish goals in our work, we leverage the common situation that there only contains one novel or unknown entity in the target sentence s_{ti} not registered in the KB from all the (s_{ci}, s_{ti}) sentence pairs. Also the word in Tibetan may be inflected for person, tense, and mood, especially the variants of the same mention of the entity, but the different forms and variants are not taken into consideration in this paper.

Given a set of translations (e_{t1}, e_{t2}, …, e_{tn}) in target language, dct_i is the value calculated by word pair relationship ($e_c \rightarrow e_{ti}$), which is calculated as follows:

Inputs:
Given word pairs ($e_c \rightarrow e_{ti}$) and Chinese-Tibetan sentence corpora
Begin:
For the given word e_c, the sentences set S_c[1..m]: S_{c1}, S_{c2}, and S_{cm} uses the word e_c
Locate the sentence pair (S_{c1}, S_{t1}), (S_{c2}, S_{t2}), and (S_{cm}, S_{tm})
Define the target sentence set S_t[1..m]: S_{t1}, S_{t2} and S_{tm} are the translation sentences of S_{c1}, S_{c2}, and S_{cm}
Given the translation word set of (e_{t1}, e_{t2}…, e_{tn}), for each target word e_{ti} is a possible translation of e_c
Calculate the numbers whether the translation word appears in the target sentences
{for each e_{ti} (i=1,2…, n): $num_i=0$
if e_{ti} is found in sentence s_{tj} (j=1,2,..., m) then $num_i=num_i+1$
$dct_i=\sum_{i=1}^{m} num_i/m$
Output the entity pairs < e_c, e_{ti}, dct_i >}

Rank the e_{ti} according to the value of dct_i, and put the words into the bilingual dictionary after the double check by linguists.

3.2. Joint Disambiguation

The goal of disambiguation step is to compute the newly discovered Tibetan candidate named entities onto KB classes, such that each entity is assigned to at most one item. Armed with the Chinese word e_c and Tibetan candidate lists e_{ti}, we then consider type the entities into different class, by enforcing that both Chinese entity and Tibetan entity jointly in the same entity type.

The joint disambiguation method assume that the entity in source and target language being selected are drawn from the same semantic class and fail to capture the corresponding

sets if this is not the case. Names of the same entity that occur in different languages often have correlated frequency patterns due to common events in the parallel corpus [9]. However, the use of distinct semantic feature is not the only problem when trying to type the source and target named entities.

This task is twofold: first, generate candidate types separately for Chinese entity and Tibetan entity; second, find the best type for both two entities among its candidate types. For a given entity, candidate types are defined by the method introduced in [10].

Definition 3 (Co-occurring Type): The Type-Pattern co-occurring type for Chinese entity e_c and Tibetan entity e_t is given by $P[(t)|(t_c, t_t)]$, where t_c and t_t are the types of the entities obtained with e_c and e_t separately. $P[(t)|(t_c, t_t)]$ is expanded as follows:

$P[(t)|(t_c, t_t)]=[P(t_c, e_c), P(t_t, e_t)]|P(t)$, $P(t_c, e_c)$ or $P(t_t, e_t)$ is the relative occurrence frequency of the typed pattern among all the matches in Chinese or Tibetan corpus.

If we observe where a Chinese entity can be easily disambiguated to a KB classes as the value of $P(t_c, e_c)$ is nearly approaching 1, and the other Tibetan entity is considered to be the same class.

4. Evaluation

The evaluation of the bilingual NER method requires a rather specific kind of sentence set. Namely, the sentence in both languages has to be annotated with the same tags following the same guidelines.

We only use entities with translations that appear in the Tibetan corpus. After discarding sentences with no aligned counterpart, a total of 210 documents and 6,127 parallel sentence pairs were used for evaluation. We choose 2000 sentence pairs as training data, and other sentence pairs as test data set.

We tested the method on the Chinese-Tibetan bilingual dataset, which contains 4127 sentence pairs. For building knowledge resources, all Chinese sentences are POS tagged and chunked with in-house tools. Our goal is to utilize the NER tagger outcome of resource-rich source languages (English, Chinese) to a resource-poor target language (Tibetan). Thus the results of NER tagger in Chinese are vital to our method. We use the Stanford CRF-based NER tagger as the monolingual component, which serves as a state-of-the-art monolingual baseline. We train the Chinese CRF models on corpus that are annotated with named entity tags.

We use a dictionary with location and organization names to verify the outcomes. Within the 6127 sentence pairs, there are 8860 entities found in Chinese sentence.

For Tibetan, there are no open-source NER taggers, so we only apply the word segmentation and part-of- speech tagging to the Tibetan sentences.

For example:

ཚན་རྩལ་ཡུན་[ORG][*the Ministry of Science and Technology*]/ཡིས་/སྐྱེ་ཁྲིད་ /བྱས་/ཤིང་།/ འཛེལ་རྒྱས་བཅོས་སྐྱར་ཡུ་ཡོན་ལྟན་ཁང་/[ORG] [*the National Development and Reform Commission*] དང་/ རྩོར་སྡིད་ཡུན་[ORG][*The Treasury*] /ཞིང་ལས་ཡུན་[ORG] [*the Department of Agriculture*] /སོགས་/ ཡུན་ཁུས་[ORG] [*department*]/བཅུ་བདུན་ /ཀྱིས་/མཉམ་དུ་/གྲོས་ཞིབ་

由/科技部/牵头/，发改委/、财政部/、农业部/等/17/个/部委/的/专家/联席/评审

Led by the ministry of Science and Technology, followed by the National Development and Reform Commission, the Treasury, the Department of Agriculture and so on, experts from 17 departments organize a combined review meeting.

5. Related Work

Previous work explored the use of bilingual corpora to improve existing monolingual analyzers. Huang et al. improved parsing performance using a bilingual parallel corpus by extracting the NE translation dictionary from the bilingual corpus [4]. Chen *et al.* proposed an integrated model to jointly identify and align bilingual NEs between Chinese and English, and constructed transformation rules [2]. Aker *et al.* presented a method for extracting bilingual terminologies from comparable corpora [8], which treated term alignment as a classification problem.

However, most commonly used NER models are trained on manually annotated data, and Tibetan language NER problem should be resolved through other ways. Yu *et al.* proposed a

NER method based on the syntactic rules in Tibetan language, using auxiliary word, lexicon and boundary information labels to improve the performance of the person name recognition system in Tibetan language [5].

Another promising for improving performance of NER systems in Tibetan language is in enforcing different treatment methods in different levels of sentences within Tibetan context based on form logic case, semantic logic case and phonological tendency studies [15].

Our bilingual NER alignment method is inspired by prior works, such as bilingual named entity recognition and word alignment using dual decomposition method [3], which yield improvements by combining English and Chinese tagging models with alignment model.

As the method for the initial building of a Tibetan thesaurus of named entities without much help of knowledge resources, the method is utilizing the alignment method to Chinese-Tibetan bilingual named entity recognition. Two distinct steps are defined: one is identifying entity candidates in Tibetan, and the other is typing the entity into the semantic class.

6. Conclusion and Future Work

The method of named entity recognition is mainly based on rules, statistics and machine learning method for the current mainstream language, while the research on Tibetan named entity recognition is in the early stage.

In this paper, we presented an approach for identifying bilingual named entities based on NER dictionary and political corpus.

We hope that our work of bilingual NER will enable attention to Tibetan language processing area, which is in urgent need of those methods and application systems. Various experiments and applications have been conducting in the current research. Future work will include more precise and accurate simulations as well as a complete description of the bilingual alignment algorithms of the NER recognition. Another work will look at the possibility of the performance gain to corpus size and variation. More semantic resources and dictionaries in Tibetan are will be built and be available online.

Acknowledgements

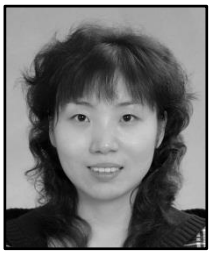
Our work is supported by the National nature science foundation of China (No. 61103161), the Program for New Century Excellent Talents in University (NCET-12-0579) and the “985” special funds in School of information engineering, Minzu university of China.

References

[1] Y. Al-Onaizan and K. Knight, “Translating named entities using monolingual and bilingual resources”, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, **(2002)**, pp. 400-408.
[2] Y. Chen, C. Zong and K.-y. Su, “A Joint Model to Identify and Align Bilingual Named Entities”, Journal of Computational Linguistics, vol. 39, no. 2, (**2012**), pp. 229-266.
[3] M. Wang, W. nxiang Che and C. D. Manning, “Joint Word Alignment and Bilingual Named Entity Recognition Using Dual Decomposition”, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **(2013)**.
[4] L. Huang, W. Jiang and Q. Liu, “Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing”, Association for Computational Linguistics, vol. 3, **(2009),** pp. 1222-1231.
[5] H. Yu, T. Jiang and N. Ma, “Named entity recognition for Tibetan texts using case-auxiliary grammars”, Proceedings of International Multi-conference of Engineers and Computer Scientists, **(2010)**.
[6] Bbn’S Proposed Answer Categories for Question Answering. Retrieved on 2013-07-21.
[7] Sekine's Extended Named Entity Hierarchy, http://nlp.cs.nyu.edu/ene/.
[8] A. Aker, M. Paramita and R. Gaizauskas, “Extracting bilingual terminologies from comparable corpora”, ACL, **(2013).**
[9] R. Sproat, T. Tao and C. Zhai, “Named Entity Transliteration with Comparable Corpora. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics”, Sydney, **(2006**), pp. 73-80.
[10] N. Nakashole, T. Tylenda and G. Weikum, “Fine-grained Semantic Typing of Emerging Entities”, ACL **(2013).**

[11] X. Jiang and L. Qiu, "A Tibetan ontology concept acquisition method based on HowNet and Chinese Tibetan dictionary", Proceedings of International Conference on Asian language, **(2013),** pp. 189-192.
[12] G. Zhou and S. Jian, "Machine Learning-based Named Entity Recognition via Effective Integration of Various Evidence. Natural Language Engineering, vol. 11, no. 2, (**2005**), pp. 189-206.
[13] B. O'Connor, B. M. Stewart and N. A. Smith, "Learning to Extract International Relations from Political Context", ACL, **(2013).**
[14] O. Franz Josef and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, vol. 29, no. 1, (**2003**), pp. 19-51.
[15] K. Qi, "On Tibetan Automatic Participate Research with the Aid of Information Treatment", Journal of Northwest University for Nationalities, vol. 16, no. 4, (**2006**), pp. 31-35.

Author



Lirong Qiu, She received her Ph.D. degree in Computer Science from Chinese Academy of Science (2007). Now she is an associate professor of Information Engineering Department, Minzu University of China. Her current research interests include natural language processing, artificial intelligence and distributed systems.