Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Multi-lingual geoparsing based on machine translation

Xu Chen [a,b,*], Judith Gelernter [b], Han Zhang [b], Jin Liu [a,**]

[a] State Key Laboratory of Software Engineering, Computer School, Wuhan University, China
[b] Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

HIGHLIGHTS

- In traditional, multi-lingual Named Entity Recognition depends on variety local lexicon and tools.
- However, language resource in English are much more than other languages, and easier to acquire them from Internet.
- Our method for multi-lingual geoparsing uses monolingual tools and resources along with machine translation and alignment to return location words in many languages.
- Our main claim is that our LanguageBridge software can find location words in texts that are in the world's widely spoken languages.
- People who interested in multi-lingual Named Entity Recognition, will easy to build a native language geoparsers based on our LanguageBridge software, it is not only save the time and cost of developing geoparsers for each language separately, but also it allows the possibility of a wide range of having a wide range of language capabilities within a single interface.

ARTICLE INFO

Article history:
Received 31 August 2016
Received in revised form 26 April 2017
Accepted 23 July 2017
Available online xxxx

Keywords:
Named entities recognition
Location
Geoparse
Multi-lingual
Machine translation
Word Alignment

ABSTRACT

Our method for multi-lingual geoparsing uses monolingual tools and resources along with machine translation and alignment to return location words in many languages. Not only does our method save the time and cost of developing geoparsers for each language separately, but also it allows the possibility of a wide range of having a wide range of language capabilities within a single interface. We evaluated our method in our LanguageBridge prototype on location named entities using newswire, broadcast news and telephone conversations in English, Arabic and Chinese data from the Linguistic Data Consortium (LDC). Our results for geoparsing Chinese and Arabic text using our multi-lingual geoparsing method are comparable to our results for geoparsing English text with our English tools. Furthermore, our experiments using our tools on machine translation approach in accuracy results on results from the same data that was translated manually, further showing the robustness of locations to machine translation.

1. Introduction

Named Entity Recognition is central to many Natural Language Processing tasks, including information retrieval, question answering, data mining and text analysis. Often, finding named entities in different languages is approached by developing tools in each language separately. NLP tools for English are widely developed and used and can be downloaded easily on Internet. However, minority languages have little useful NLP tools, such as Mongol, Vietnamese and so on. In this paper, our method aims to reduce development time for Named Entity Recognition tools by processing in a single language via machine translation. We assume that our method extends to person and organization named entities, although our research focus is on named entities for location.

Named entities for location. Named Entity Recognition typically encompasses named entities for person, organization and location. Our focus for experimentation is on named entities for location, which we alternately refer to as toponym. That is because our ultimate goal is to produce not only the locations, but also the geographic coordinates for each location. Our results can be displayed on a geographic map, if desired.

Logic of method. The previous version of our English geoparser can find location named entities in high quality English text, as well as in English text produced by machine translation from other languages. Our method is based on a finding in our previous research that finding locations in Spanish tweets with a geoparser trained for Spanish was less accurate than geoparsing an English translation of the same Spanish tweets with a geoparser trained for English [1]. Similar results were found when using machine

[*] Corresponding author at: State Key Laboratory of Software Engineering, Computer School, Wuhan University, China.
[**] Corresponding author.
E-mail addresses: xuchen@whu.edu.cn (X. Chen), jinliu@whu.edu.cn (J. Liu).

http://dx.doi.org/10.1016/j.future.2017.07.057
0167-739X/© 2017 Elsevier B.V. All rights reserved.

translation and English tools to find named entities in source texts in Swahili and Arabic [2]. In fact, statistical machine translation is often used for cross-language information retrieval [3].

Novelty and contribution. Our goal in designing this multilingual system was to build a low-resource language Named Entity Recognition based on machine translation efficiently, while supporting a wide range of languages. To attain this goal, we used off-the-shelf machine translation and alignment tools from Google and Microsoft. This meant that we sacrificed the accuracy we could have achieved, had we trained our own machine translation and word alignment models for each language that we wanted our LanguageBridge geoparser to support.

In our method, Named Entity Recognition capabilities are inherited by the languages that can be translated into the English of our parser. Microsoft Translator supports more than 40 languages at the time of our writing,[1] whereas Google Translate supports more than 60.[2] Space limits our including of additional experiments to show the language-extensibility of our method. Even so, soundness of results for languages as linguistically diverse as Chinese and Arabic, coupled with reliable Machine Translations from Google or Microsoft, implies our method's extensibility.

Robustness of solution. Our LanguageBridge geoparser has been shown effective in informal testing of Russian, Ukrainian, Bahasa Indonesia, and Farsi, as well as formal testing of Chinese, Arabic and English. Adding another language requires adding to our program only a few lines of code, and preliminary testing to see whether Microsoft or Google Translate produces a better quality translation. Both translation algorithms are available in our interface, so user selection at the time of the data analysis is simple. In this paper, we use Automatic Content Extraction [4] Multilingual Training Data v6.0 to evaluate results based on finding locations in Chinese and Arabic, as well as English.[3]

Usability/Human Factors. For those not conversant in a source language, results in our LanguageBridge prototype may be set to display in both the source language and the English target, along with a confidence value that suggests the probability that the result is correct. System controls could be set to any language. The English tools for machine translation and Named Entity Recognition can be hidden from users uncomfortable in English or who are uninterested in steps preceding the display of results, so that the entire procedure might take place in a source language, even though the equipment is mostly in English.

Research questions

- How can we improve word alignment so as to improve the accuracy of the output?
- What is the main source of the geoparsing error: Machine translation? Word alignment? Our Geoparser?
- How does our precision and recall in geoparsing without machine translation (in English) compare generally with precision and recall in geoparsing that relies on machine translation (in Chinese and Arabic)?
- To what extent does translation quality influence geoparsing result?

Section 2 describes related work, and Section 3 describes the architecture for our multi-lingual geoparser, that we call LanguageBridge and details the sub-processes required for each step of our implementation. Data for the experiment data is described in Section 4. Evaluation experiments for the LanguageBridge appears in Section 5. The paper concludes in Section 6 with potential research directions and a summary of our contributions.

2. Related work

Our end-to-end solution in finding locations in text consists of translating that language into English while maintaining the word alignment with English. We can then use English Named Entity Recognition tools in that translation, before outputting the source and target languages. Our solution is described more fully in Section 3 — here, we review subtasks for our solution. We conclude here by comparing to similar end-to-end solutions for finding named entities in text.

2.1. History and importance of named entity research

Named entities are generally defined as nouns, in categories such as person, organization and location. Gey [5] found that 30% of content-bearing words, and Friburger and Maurel [6] found that 10% of words in their document set are proper nouns. Friburger and Maurel found that of the proper nouns, 43.9% are locations. Named entity solutions have been accomplished in various ways, some with machine learning, with or without a match list, and using language processing cues of the word (such as capitalization) as well as cues in the sentence. Fifteen years of named entity solutions up to 2006 are reviewed in [7].

2.2. Sub-tasks for the solution: External resources and knowledge engineering

Some of the more language-independent approaches to finding named entities use language-specific resources. This may be accomplished with statistics and some heuristics, but without necessarily using part-of-speech or grammar tagging for individual languages. However, for the identification of named places, at least, the named entities have been recognized by using gazetteers with entities for locations in many languages [8]. This method is similar to that used by Kumar et al. [9], who found named entities in Hindi and Marathi by using bi-lingual dictionaries, and using a dictionary built from Wikipedia to find entities in English.

2.3. Sub-tasks for the solution: Machine learning techniques

Täckström showed that it is possible to use labels in one language to train another language, and that also a system trained on one language can be used for data in another language. He used this approach when no annotated resources were available in a target language. He found that using multi-lingual word clusters can improve performance [10]. Maximum Entropy Models and Hidden Markov Models were used in the Conference on Computational Natural Language Learning, 2003. Features used to find the named entities varied from system to system [11].

2.4. Sub-tasks for the solution: Machine translation and word alignment

Some have determined that not every word in a source text requires translation, only the named entities [12,13]. If an entire sentence is wanted but the exact translation is not apparent, a translation could be selected among a set of word possibilities from the web [14], or synonyms for those phrases could be used instead [15], or else the named entity might be taken into the target language in transliteration. While collecting high quality labeled data for multi-lingual named entity recognition is time-consuming, unsupervised word clustering has been attempted with some success [16].

When Named Entity Recognition is performed on data that was machine translated from the original language, the alignment between source and target language becomes critical. Languages

[1] Microsoft Translator API list of languages: http://msdn.microsoft.com/en-us/library/hh456380.aspx.

[2] Google Translate API list of languages: https://developers.google.com/translate/v2/using_rest.

[3] Linguistic Data Consortium catalog number ACE2005E18.

Please cite this article in press as: X. Chen, et al., Multi-lingual geoparsing based on machine translation, Future Generation Computer Systems (2017), http://dx.doi.org/10.1016/j.future.2017.07.057.

might be aligned in groups of two words to one word, or no words [17], or alignment of one-to-one might produce better results [18]. What has been called the “classical approach” to word alignment by Hidden Markov Models (HMMs) in the 1990s has since been updated [19]. This approach is employed in the Microsoft alignment used in our Language Bridge.

2.5. Similar solutions: Crosslingual named entity recognition

Bilingual texts, also called parallel corpora, have been used to strengthen mono-lingual Named Entity Recognition algorithms [20,21], and create named entity annotations [22]. Our objective, by contrast, is to find location named entities in texts in many languages immediately by adding machine translation tools to an already strong Named Entity Recognizer—without additional training when possible.

The Cross-Language Retrieval Forum, CLEF, ran geo-tracks in 2005 (a pilot year) 2006, 2007 and 2008, in order to test the ability of a system to find location information in multiple languages [23]. The CLEF experiments differ from ours in that participating systems were expected to answer questions regarding location to express geographical relationships (proximity, inclusion and exclusion), rather than just to identify locations, as ours does. Furthermore, our method using Google or Microsoft Translator permits a wide range of language capabilities with very little additional coding required.

Thus, our method allows us to find location expressions in dozens of languages due to the translation range of Google and Microsoft. This is in comparison to the GeoCLEF experiments in 2008 that were in European languages only, and Rosette NER[4] from BasisTech can find locations (and in fact, standard named entities) in 16 languages at the time of this writing. However, our use of our own Geoparser that has Stanford NER, our own CRF-trained classifier and some heuristics, makes our algorithm more robust [24].

Some similar solutions has been proposed for crosslingual Named Entity Recognition, especially using a resource fortunate language to aid a resource deprived language. A neural network based architecture is proposed for crosslingual Named Entity Recognition, which allows sharing of various parameters between the two languages [25]. Dandapat and Way [26] propose a technique to improve named entity recognition in a resource-poor language (Hindi) by using cross-lingual information, this work is similar with our work, however, our work provide crosslingual Named Entity Recognition method for many more languages. A word embedding–based named entity recognition (NER) approach is used for low-resource languages without the presence of sufficiently large training Data [27]. Agerri and Rigau [28] present a multilingual Named Entity Recognition approach based on a robust and general set of features across languages and datasets, their system combines shallow local information with clustering semi-supervised features induced on large amounts of unlabeled text.

3. Our multi-lingual geoparser, LanguageBridge

This section provides an overview of our method for finding location terms in texts in different languages. We illustrate the architecture of our LanguageBridge prototype system, and then describe how we adjusted the word alignment and machine translation components for better results. Note that although we give most examples in Chinese and Arabic for consistency throughout the paper, similar errors in word alignment and machine translation as described here might arise in other languages.

3.1. English geoparser and cross-lingual geoparsing

Our group has already developed an English geoparser, based on supervised learning with Condition Random Fields [1]. The parser had been trained on part of the ACE Multi2005 data, as well as annotated tweets.

Pre-processing for both includes tokenization, lemmatization, part-of-speech tagging, and feature extraction. The English geoparser outputs toponyms, along with some widely recognized locations. The geoparser relies on English resources such as the GeoNames gazetteer for place names, Stanford NLP tools for tokenization and part of speech tagging (as well as CMU ARK Twitter for part of speech tagging), and Cybozu for language recognition. The output of the Geoparser is each location in the source language and in English, with geographic coordinates, and a confidence value reflecting the probability that the toponym was output accurately. The confidence value is based on the Conditional Random Fields model.

3.2. Architecture for multi-lingual geoparsing

Fig. 1 illustrates the procedure for multi-lingual geo-parsing. Word alignment is created between the input language and English when the machine translation algorithm is run. We have added scripts to improve the alignments, as discussed below, as well as a hashmap to store the alignment.

Next, our English geoparser finds locations in the English translation. The multi-lingual geoparser uses the alignment information to match the found locations with those in the original language.

The last step is for the location words identified in the English translation to be displayed both in English and in the original language. There is an option to display also the latitude and longitude coordinates for location.

Our prototype system called LanguageBridge was implemented using a component-based approach that includes our own geoparser, a Machine Translation component (whether Google or Microsoft), and word alignment (whether from Google or Microsoft) along with our alignment adjustment scripts. A developer might substitute another component for ours to alter the processing result.[5]

3.3. Google or Microsoft for machine translation

At the core of our multi-lingual geoparser are machine translation and word alignment algorithms, which we have adopted from either Microsoft or Google. We added word alignment improvement scripts to our framework. The point of building a choice of Google and Microsoft online translation services is that each uses a different translation and alignment algorithm, and these effect results. Sometimes either translation algorithm will work with a given language. However, we have found even in initial testing that the Microsoft algorithm works better with Chinese, for example. And Microsoft does not provide reliable word alignment from Arabic to English, so we use the Google translation and word alignment for Arabic.

[4] https://www.rosette.com/function/entity-extraction/.

[5] Our Java-based prototype runs on Windows, Mac or Linux. Because the index of the gazetteer is huge, we recommend that the hardware used to run it has at least 4 GB of memory. Write to gelern@cs.cmu.edu to inquire about access to the LanguageBridge.

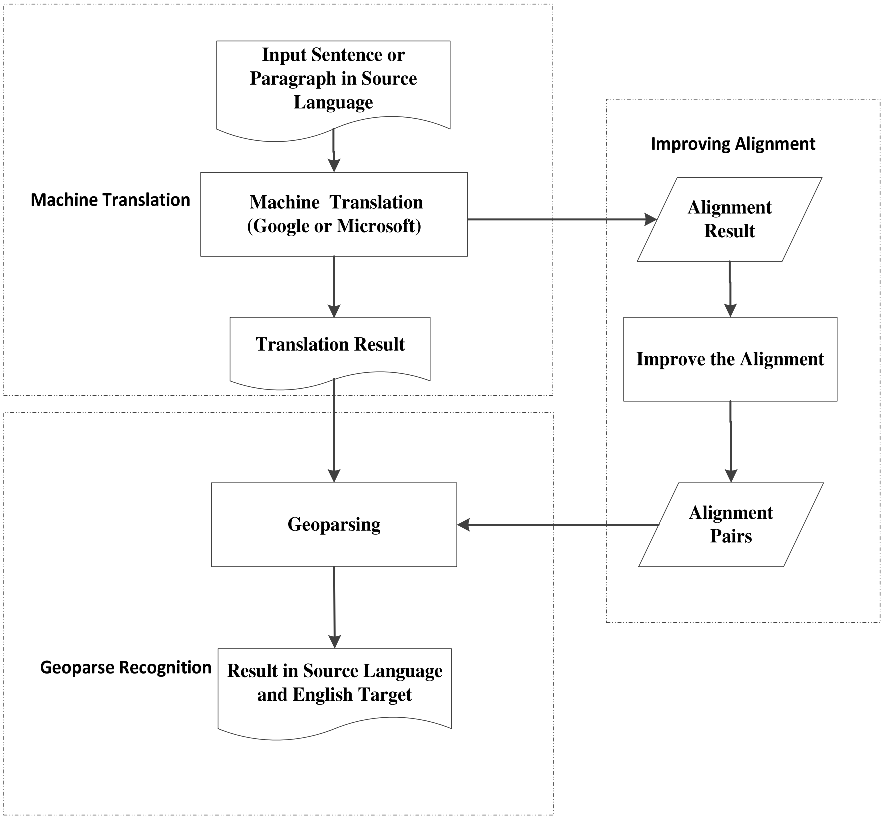


Fig. 1. Architecture for our multi-lingual geoparser, LanguageBridge.

3.4. Word alignment adjustments

Adjustments for locations found in the source language

The word alignment from machine translation often loses words, or combines function words with toponyms so as to complicate the geoparsing. Hence, we geoparse the coarse alignment before re-aligning with the source language.

1. Some non-English tokens bundle a preposition with a named entity. For example, in Microsoft alignment in the Russian language, "Крыма = of Crimea". Our algorithm must separate preposition from location in order to identify the location.
2. Synonymous Chinese words align with the same word in English. For example, even though "塞尔维亚共和国" translates as "Republic of Serbia", and "塞尔维亚" translates as "Serbia", when they appear in the same sentence, both "塞尔维亚共和国" and "塞尔维亚" align with "Serbia", and our algorithm chooses only one in the pair for the hash map.
3. Some words may lose alignment information during machine translation. However, if the word occurs two or more times in the same sentence, we can supplement the lost alignment information based on the same word in other places of sentence. For example, if we find "اسرائيل" aligns with "Israel", then the algorithm will use "Israel" in other parts of the translated sentence.

We propose Algorithm 1 to improve the alignment information from machine translation. The input is the Alignment String and output is a hash map which stores the pairs of word or phrases for the source language and English.

Errors are caused when the oldValue and newValue differ but share the same key. The algorithm compares the values in the sentence to see whether one is one a subset of the other. For example, both "塞尔维亚共和国" Republic of Serbia and "塞尔维亚" Serbia are in the original sentence, with the second as a subset of the first. We created a rule to output the longer version from the original. Alternatively, we could retain the alignment positions of both versions of the toponym from the original text and use different keys to store in the hash map, but this would make it take more time to process the alignment and output the result.

Adjustments for locations found in the English Machine Translation

When we find the locations that are output by our English geoparser, we should get the location words in source language too. However, sometimes we cannot find the alignment information directly.

1. Adjacent English words align with the same word in Chinese. For example, "United" maps to "美国", and also "States" maps to "美国".
2. Alignment information is not available. The alignment cannot find a phrase in Chinese to match "Carnegie Mellon University". Therefore, it finds the words one by one rather than in a phrase, and then combines them to give a result.

Our implementation takes all of these errors into account to improve source — target alignment of multi-word location expressions.

Algorithm 2 fixes errors after the location words in English are output from the geoparser. For the example of "Great Britain", the alignment algorithm cannot identify Great Britain as a phrase, so the hash map stores "Great" and "Britain" as separate keys. Algorithm 2 restores the values of the two keys in the output.

```
Algorithm1 alignment information improvement
Input: Alignment String S (sourceNumSeq$_1$-enNumSeq$_1$,…,sourceNumSeq$_i$-enNumSeq$_i$ ,…, sourceNumSeq$_n$-enNumSeq$_n$).
Output: HashMap pairs.
Scanner alignmentScanner = new Scanner(S);
paris = new HashMap();
While (alignmentScanner.hasnext){
  String wordAlignment = alignmentScanner.next();
  String newValue = FindW$_i$ (wordAlignment.getSourceNumSeq$_i$());
  newValue.removePreopostion();
  String Key = FindW'$_i$ (wordAlignment.getEnNumSeq$_i$());
  String oldValue = pairs.getValue (Key);
  If (oldValue==null||oldValue.equals(newValue))
  pairs.put (Key, newValue);
  else if
  String Value = Compare (oldValue, newVaule);
  paris.put (key,Value);    }
return pairs;
```

```
Algorithm2 find source toponym
Input : Toponym Set in English, enToponym(t$_1$, ...,t$_i$,...t$_n$)
Output: Toponym Pairs of Chinese and English toponymPair(t$_1$---t'$_1$, ..., t$_i$---t'$_i$, ..., t$_n$---t'$_n$);
for (topnymEntity t$_i$: enToponym){
  String key = t$_i$;
  String t'$_i$ = getOriginalWords(key);
  if (t'$_i$  == null){
StringTokenizer toponymTok = new StringTokenizer(key);
    while (toponymTok.hasMoreTokens()){
      String temporary = toponymTok.nextToken();
      if ((temporary != null)&&!temporary.equals(t'$_i$ )){
t'$_i$ += temporary;          }          }          }
addToponymPair(t$_i$,t'$_i$)
  }
return toponymPair;
```

3.5. Machine translation and alignment based on Google

The Google Translate API v2[6] does not provide alignment information. A Google employee advised us to approximate alignment information by doing HTML translation, where we will appropriately propagate HTML tags from the source to the target. But he warned that sometimes doing this can affect translation quality, as the system tries to preserve the HTML formatting, which might confuse location expressions.[7]

We used html tags to separate each word in the source sentence. For a source language in Arabic, the markup looks like this: <font font="1">إسرائيل</font>.

We assemble these words into an html file, which becomes our input for Google Translate. Then we get the translation result from Google. It is easy to align the word or phrase based on the <html > tag number. As shown in Table 1, we can get the Arabic–English pairs for the toponym, e.g. Israel إسرائيل. Finally, we store the alignment in a hash like this: {Israel = إسرائيل, in = في, I = أنا, live = أعيش}.

3.6. Machine translation and alignment based on Microsoft

At the core of our multi-lingual geoparser are machine translation and word alignment algorithms which we have adopted from either Microsoft or Google. We added word alignment improvement scripts to our framework. The point of building a choice of Google and Microsoft online translation services is that each uses a different translation and alignment algorithm, and these effect results. Sometimes either translation algorithm will work with a given language. But we have found in initial testing that the Microsoft algorithm works better with Chinese, and Microsoft does not provide reliable word alignment from Arabic to English, so we use the Google translation and word alignment for Arabic.

Microsoft's online statistical translation service, Microsoft Translator,[8] delivers automatic translation into the language specified. In our experiment, we use the SOAP API from Microsoft, because it provides alignment information as well as the translation. The Simplified Chinese sentences below have been geoparsed and aligned based on Microsoft. For example:

Source: 美国在加勒比海和太平洋还拥有多处领土和岛屿地区

[6] https://developers.google.com/translate/.

[7] Josh Estelle at Google, personal communication, May 23, 2014.

[8] http://www.microsoft.com/en-us/translator/developers.aspx.

Table 1
Translation of the Arabic أنا أعيش في إسرائيل based on Google Translation.

Arabic	English
<html lang="en-x-mtfrom-ar">	<html lang="en-x-mtfrom-ar">
<head></head>	<head></head>
<body>	<body style=";text-align:left;direction:ltr">
<doctype html="">	<doctype html="">
<title></title>	<title></title>
<font font="1">أنا</font>	<font font="1">I</font>
<font font="2">أعيش</font>	<font font="2">live</font>
<font font="3">في</font>	<font font="3">in</font>
<font font="4">إسرائيل</font>	<font font="4">Israel</font>
</doctype>	</doctype>
</body>	</body>
</html>	</html>

Translation from Microsoft: [The] United States [is] in the Caribbean and the Pacific, [and] also has a number of territories and insular areas.

Alignment information: 0:1–0:5 0:1–7:12 2:2–14:15 3:5–21:30 6:6–21:30 7:7–31:33 8:10–39:45 11:11–48:51 12:13–53:55 14:14–59:64 16:17–69:79 18:18–81:83 19:20–85:91 21:22–93:97.

HashMap that we create from the alignment: {territories = 领土, Pacific = 太平洋, Caribbean = 加勒比海, areas = 地区, number = 多, insular = 岛屿, also = 还, in = 在, has = 拥有, States = 美国, United = 美国, and = 和}.

English Geoparser: Caribbean, Pacific, United States.

LanguageBridge output: Caribbean—- 加勒比海, Pacific—- 太平洋United States—- 美国.

The alignment between any two languages is straightforward when it consists of one source token to one target token, or one source token to many target tokens, because a single concept may be expressed by multiple tokens. In the example above, aligning to English Caribbean and Pacific are likely correct. Algorithm 2 fixed the mechanical error that arises when the translation when two or more tokens align to two or more tokens. In the example above, the error is that the United States is repeated twice, with 美国美国.

3.7. More improvements to word alignment

We propose a method to improve alignment for any language based on Microsoft translation because we cannot predict whether Microsoft will align in either phrases or words. The open source alignment tool, Fast Align, uses an EM (Expectation Maximum) algorithm and implements IBM model 2 to produce token-to-token alignment.[9] Note that we use Russian here because the problem does not occur in Chinese, and we aimed to use Microsoft Translate which is not compatible with Arabic.

In cases where the Fast Align output differs from that of Microsoft Align, we prefer the output of Fast Align which has fewer word alignment errors and therefore fewer potential geoparsing errors. On the other hand, Fast Align requires a large quantity of training data, is slow to train, and distances us from our goal of language independent geoparsing. We thus offer Fast Align to supplement our pipeline when alignment from Microsoft or Google is inadequate.

This example demonstrates the problem with Microsoft alignment.

Source:
Поставщики: компаниииизИталии, Франции и Испании.

Translation from Microsoft: Suppliers: companies from Italy, France and Spain.

Alignment numerical information: 0:11–0:9 13:32–11:31 34:52–33:49.

Word pairs that we create from the Microsoft alignment:
{Suppliers: = Поставщики,
France and Spain = ФранциииИспании, companies from Italy = компаниииизИталии}.

Using Fast Align is able to improve geoparsing in some cases. In the above example, the locations in the Russian are unclear because the Microsoft Alignment is not for word, but in numerical phrases. So we use a hash map to create semantic pairs from the numerical pairs, however, we cannot get location results in Russian because of the lack of word alignment. Therefore, we use Fast Align [29] as a second alignment pass after the initial translation with Microsoft.

After we use the fast alignment algorithm, we get the word alignment result:
{France = Франции, Spain = Испании,
suppliers = Поставщики, :=: , Italy = Италии}.

Multi-lingual Geoparser output:
Italy- Италии.
France- Франции.
Spain- Испании.

The drawback of this approach is that Fast Align requires training. We need parallel text, and can also use whatever sentences were run in the first pass, to train the Fast Align before implementation.

4. Data

In order to test our multi-lingual geoparsing method, we used Automatic Content Extraction (ACE) 2005 Multilingual Training Corpus LDC2005E18. This includes three separate data sets for English, Chinese and Arabic. We relied on the ACE annotations for experiments reported here.

The data corpus includes three tags which we consider locations: GPE, LOC, and NAM. GPE stands for geopolitical entities, LOC for location, and NAM[10] for proper name references. We use texts in the corpus taken from Newswire, Broadcast News, Broadcast Conversation, Conversational Telephone Speech, and we randomly chose about 100 files for each language for testing.

We selected this number of files to roughly balance the number of unique locations among languages. In Table 2, our count of the number of words, the number of locations and unique locations are based on the annotations provided in the ACE Multi2005 data set.

[9] https://github.com/clab/fast_align.

[10] According to the ACE annotation guidelines, NAM = proper name reference to an entity, such as “American” which matches with the toponym, America. This is from the Linguistic Data Consortium, “ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13.

Table 2
Testing data from about 100 files from each language of ACE 2005Multilingual LDC2005E18.

	Number of words	Number of locations	Number of unique locations
Chinese	33,349	912	238
Arabic	20,087	1435	348
English	31,255	851	243

Table 3
Testing data from about 50 files for each language from the Parallel Corpora LDC2012T16 and LDC2014T05.

	Number of words	Number of locations	Number of unique locations
Chinese	20,357	469	113
Arabic	13,422	850	182

Table 4
Geoparsing for location words from LDC2005E18: Precision, Recall and F1 for Chinese, Arabic and English.

	Precision	Recall	F1
Chinese	0.821	0.737	0.777
Arabic	0.781	0.784	0.782
English	0.887	0.826	0.855

We selected two sets of LDC parallel corpora with Newswire text, 2012T16 (Chinese–English) and 2014T05 (Arabic–English), because they include high-quality English translations by bilingual speakers for Chinese and Arabic that we could use to compare with the machine translations. In Table 3, we created our own annotations in order to count the locations and unique locations.

5. Evaluation of our LanguageBridge prototype for multi-lingual geoparsing

5.1. Experimentation

We tested the accuracy of our method with three experiments.

Exp_1(a). How does finding locations in the same language as the tool (native English with an English tool) compare to finding locations in machine translation into English with an English tool.

Experimental procedure: We geoparse the English directly with our own geoparser [24], and also we are using our multi-lingual geoparsing methods that include machine translation (here, with Microsoft and Google Translator) to output locations in Chinese and Arabic. We selected files for testing from ACE 2005Multilingual LDC2005E18.

Experiment results: Precision in all three languages suffers due to the geo/non-geo disambiguation problem of non-geographic names (example: Jordan as a man's name) being mistaken for a toponym (Jordan, the country). Results in Table 4 show that the overall F1 of Chinese is close to Arabic, and both of them are comparable to the output of English.

Result analysis: Note that the precision in Chinese is higher than that in Arabic. Some of the variability can be explained by the fact that the Chinese → English path is easier than the Arabic → English path. It has been said that up to 75% of this variability can be explained by factors such as the amount of word reordering necessary, and the historical relatedness of the two languages [30].

Exp_1(b). How does finding locations in the same language as the tool (native Chinese with a Chinese tool, Rosette NER) compare to finding locations in machine translation into English with an English tool (Language Bridge).

Experimental procedure: We selected 106 test files in Chinese from ACE2005 Multilingual LDC2005E18. These files were run both in the Chinese version of the Rosette NER by BasisTech, and in our LanguageBridge (by our lab at Carnegie Mellon). The Rosette NER finds other named entities, but we scored only for location.

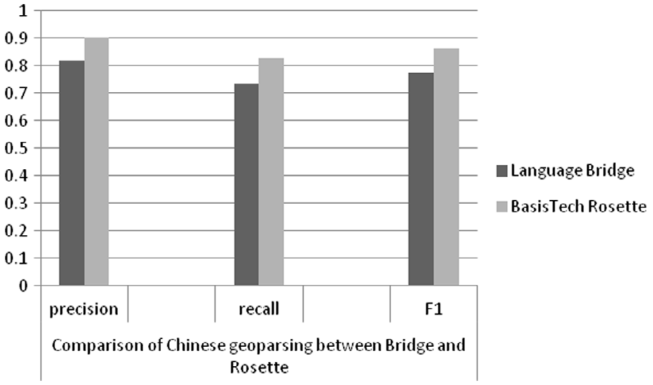


Fig. 2. Comparison of geoparsing from a translation with the Carnegie Mellon Language Bridge vs geoparsing in the original language with BasisTech Rosette NER on 106 test files in Chinese from ACE2005 Multilingual LDC2005E18.

Experimental results: Fig. 2 demonstrates that somewhat higher results were achieved in using the BasisTech software than in using our English-based tool.

Result analysis: As demonstrated by Fig. 2, quality geoparsing in the original language has the potential to achieve better results than in parsing via machine translation. Nevertheless, it has been found that named entities can be identified with success via Machine Translation for Arabic and Swahili [2], and also for Spanish [1]—and that, as shown in Fig. 2, the differential between the in Chinese parsing and the cross-language parsing is not high.

Significance. Significant for our argument is that the comparison cannot be performed in many languages (Rosette NER presently supports 16 languages only). Our system, by contrast, handles dozens of languages owing to the range of Google Translate and Microsoft Translate. This demonstrates the wide significance of the "black box" method using machine translation for location detection.

Exp_2. Given the results of Exp_1 that geoparsing translations (with Named Entity Recognizers) achieves solid results, to what extents does the translation quality matter?

Experimental procedure: We tested our LanguageBridge multilingual geoparser on data sets in which the same text is provided in two languages, and in both manual and machine translation: Arabic and English (parallel corpus LDC2014T05) and Chinese and English (parallel corpus LDC2012T16). We randomly selected 50 files from each data set and used the manual translations to annotate the original. Then we used the Microsoft Machine translation algorithm with LanguageBridge to find locations in Chinese, and the Google Machine translation algorithm with LanguageBridge to find locations in the Arabic.

Experimental results: Table 5 shows that the precision and recall of machine translation results approaches precision and recall

Table 5
Geoparsing for location words: Precision, Recall and F1 found through Machine translation vs. manual translation.

	Chinese		Arabic	
	Machine translation	Manual translation	Machine translation	Manual translation
Precision	0.796	0.810	0.708	0.758
Recall	0.776	0.783	0.923	0.942
F1	0.786	0.796	0.801	0.840

achieved by manual translation. According to Pearson's Product-Moment Correlation, we found no statistically significant difference in the geoparsing precision and recall between manual and machine translation for Chinese, and also no statistically significant difference in geoparsing precision and recall between manual and machine translations for Arabic.

Result analysis: Why does the translation quality for finding locations in Chinese and Arabic text seem insignificant? The Named Entity Recognition that is the basis for finding locations does not rest on subtle text understanding. Instead, Named Entity Recognition relies upon correct recognition of part of speech of words, some location-indicating phrases, and location-word matches with a gazetteer, all of which can be accomplished adequately from a good machine translation.

Exp_3. How robust is our cross-lingual geoparsing method?

Experimental procedure. Parallel corpora are used for statistical machine translation and other procedures for Natural Language Processing. The Linguistic Data Consortium includes parallel corpora (2014T05—Arabic/English) and (2012T16—Chinese/English) that include both machine and manual translations. We selected 50 files at random from each corpus, and annotated the locations found in those files. The Arabic set had 182 unique toponyms, and the Chinese set had 113 (see Table 3). We sent both manual and machine translations from each language through our own Geoparser, and through the Yahoo GeoMaker,[11] (1) to compare parsing tools. We were interested also (2) in comparing relative accuracy between Chinese and Arabic, and (3) in comparing relative accuracy with different translation quality.

Experimental results. Three results comes from this experiment. (1) Tools: The first is that our geoparser is comparable to Yahoo GeoMaker in precision and recall, both for the Chinese and for the Arabic data set. The GeoMaker outperforms our Geolocator in precision for both languages, but the Geolocator dominates the GeoMaker in recall for Arabic, making it outperform the GeoMaker in Arabic overall. (2) Languages: The files were chosen randomly from Arabic and Chinese. There are more unique toponyms in the Arabic data set than in the Chinese. Proportionally, however, the ability to find locations accurately in both languages is comparable for our Geolocator (whereas the GeoMaker performed better in Arabic than in Chinese). (3) Translation quality: Results demonstrate finally that the quality of the translation matters little in results – that the locations found in the machine translation approximate those in the manual translation for both languages. The Geomaker even found more locations accurately based on Chinese machine translation than on the manual.

Result analysis: The overall results of Figs. 3 and 4 demonstrate the effectiveness of our method in using machine translations of text with English geoparsing tools.

CMU Geolocator[12] and GeoMaker sometimes make the same parsing errors, for example:

Chinese:

这艘渔船的船东来自西班牙北部的巴斯克地区

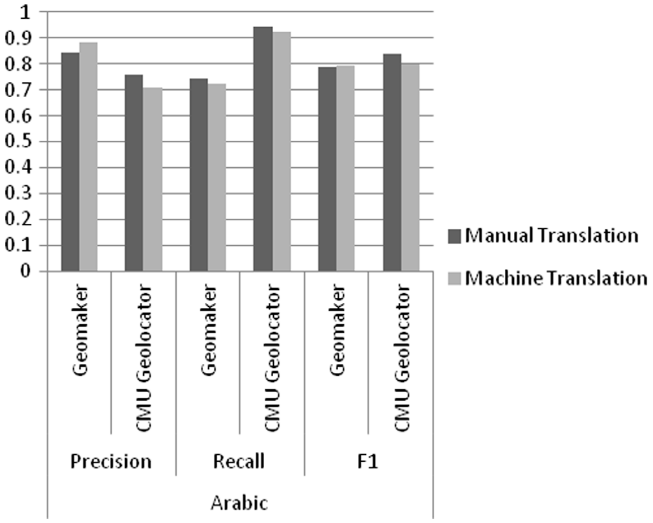


Fig. 3. Our geoparser compared to GeoMaker on 50 files in Arabic and English (parallel corpus LDC2014T05).

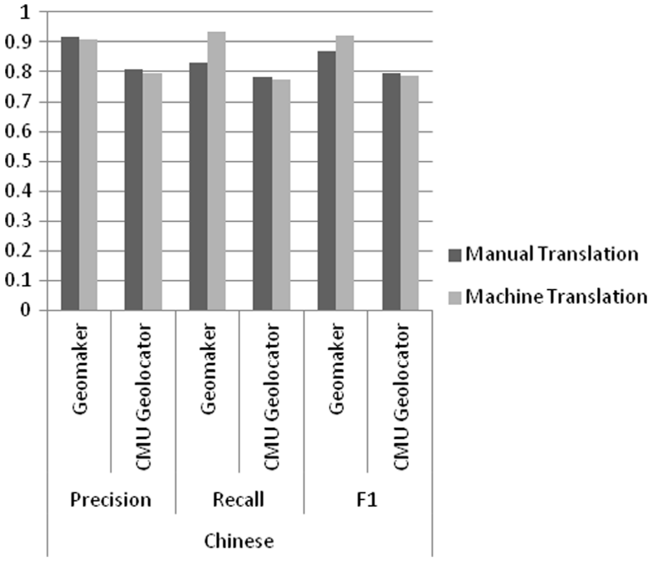


Fig. 4. Our geoparser compared to GeoMaker on 50 files in Chinese and English (parallel corpus LDC2012T16).

English:

This fishing boat's owner is from Spain's northern Basque region.

Algorithm output:

Spain *(Geoparsing error and Geomaker error, which is the same error)*

What the output should be:

Spain's northern Basque region.

From Chinese parallel LDC2012T16, Spain's northern Basque region is a location, because it is a chunking, both of two parsers

[11] https://developer.yahoo.com/geo/placemaker/.

[12] Our geo-parser has been available on GitHub at https://github.com/geoparser/geolocator.

Table 6
Examples of errors from LanguageBridge, with full errors charted in Fig. 5.

pID	Source	Machine Translation	Output actual	Output wanted	Error type
1	美国高科技纳斯达克指数爆跌，港股受到相当大的影响。	United States hi-tech NASDAQ burst down considerable impact on Hong Kong stocks	United States -美国 NASDAQ - 纳斯达克 Hong Kong -香港	United States -美国 Hong Kong -香港	Precision
2	月岛因为涉嫌违反日本核子安全法	Moon island on suspicion of violating Japan nuclear safety law	Moon island (mistranslation of person’s name) -月岛 Japan-日本	Japan-日本	Precision
3	وقال الحكيم، في كلمة له بمناسبة يوم الغدير، في مقر المجلس الاعلى للثورة الاسلامية في منطقة الجادرية ببغداد	the Supreme Council for the Islamic Revolution headquarters in Jadriya district of Baghdad	المجلس الاعلى- Supreme Council للثورة الاسلامية- Islamic Revolution Jadriya- منطقة الجادرية بغداد- Baghdad	Jadriya- منطقة الجادرية Baghdad- بغداد	Precision
4	另外一架向北飞往距离首都 450 公里的摩斯沃尔	Another plane flying 450 km from the capital north of Moss Wall	--	Moss Wall -摩斯沃尔	Recall
5	الدكتور مروان هندي نائب مدير مستشفى الشفا في مدينة غزة قوله إن إدارة	Dr. Marwan Indian deputy director of Shifa Hospital in Gaza City, saying that the management	Gaza City (but no Arabic found)	Gaza City- مدينة غزة	Recall
6	因为东海村油燃料加工厂去年 9 月发生核灾事故而被起诉的 6 名员工	Since Tokaimura fuel processing plant accused of nuclear disaster accident last September 6 staff	--	Tokaimura -东海村	Recall

cannot recognize the whole location, just find Spain. We need to be able to recognize when two or more locations next to each other.

5.2. Error analysis

We analyze the result from multi-lingual geoparsing experiment 1, above, in more detail in order to determine the cause of false positives. We collected the errors, and then classified them into two types: mono-lingual geoparsing error (finding English locations in English data) and cross-lingual parsing error which includes two subtypes: translation error (mistaken translation between source language and English), and alignment error (mis-aligned words or phrases between source language and English target). In Table 6, there are some examples of errors.

We calculated the proportion of errors as shown in Fig. 5. That the majority of the error is cause by finding English locations with an English tool suggests that it will be possible to improve multi-lingual results significantly by improving the capabilities of the English geoparser tool.

We found that much of the false positives geoparsing errors derive from geo/non-geo errors. These falsely-found locations tend to be names of people or organizations that the geoparser wrongly takes to be toponyms. That suggests that we will be able to improve the error greatly by improving the accuracy of our English geoparser. Geoparsing error from false positives (precision error) as well as from missing locations (recall error) is likely to stem from an intractable problem rather than an incomplete cross-lingual solution.

6. Conclusion

We propose a cost-efficient method to build a multi-lingual geoparser based on machine translation and word alignment adjustment. Our LanguageBridge prototype requires only a few lines of code to add the capability to geoparse other languages, provided that those languages are supported by one of the Machine Translation tools of Google or Microsoft.

We demonstrated the viability of our system by running our Language Bridge multi-lingual geoparser over Chinese and Arabic test data, as well as over English data. The experiment confirmed

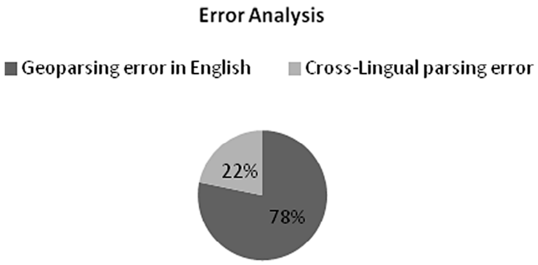


Fig. 5. Proportion of Errors in Arabic and English geoparsing (with data based on parallel corpus LDC 2005E18 corresponding to data in Table 4).

that results from geoparsing Arabic and Chinese were of comparable accuracy to results in English. Not surprisingly, therefore, we found that geoparsing the machine translation into English from Chinese and Arabic yields results comparable to geoparsing the high quality manual translations into English from Chinese and Arabic. The largest proportion of the error in finding location words, we found, comes from geoparsing English text with English tools (an English geoparser).

Validating our method, we found that results with English geoparsing tools other than ours were comparable to results with our own Language Bridge. Although there is variability in each language, for a data set of similar size in a language, which can be translated by Google or Microsoft, we expect that the location word output will be accurate.

In the future, we plan to use multi-lingual geoparsing for more minority languages and use GIS tools to show the geocoding result immediately. Because machine translation technology is developing quickly with the help of deep learning, we will integrate our tools with new machine translation algorithms and improve multilingual geoparsing result.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2016YFB0502601), the National Natural Science Foundation of China (Grant No. 41201405), China Scholarship Council (Grant No. 201308420300).

References

[1] J. Gelernter, W. Zhang, Cross-lingual geo-parsing for non-structured data, in: 7th Workshop on Geographic Information Retrieval, 2013. https://doi.org/10.1145/2533888.2533943.

[2] R. Shah, B. Lin, A. Gershman, R. Frederking, SYNERGY: A named entity recognition system for resource-scarce languages such as Swahili using online machine translation, 2011. http://www.cs.cmu.edu/~encore/synergy.pdf. (Accessed 13 July 2014).

[3] V. Nikoulina, S. Clinchant, Domain adaptation of statistical machine translation models with monolingual data for cross lingual information retrieval, in: Advances in Information Retrieval, in: Lecture Notes in Computer Science, vol. 7814, 2013, pp. 768–771.

[4] The ACE 2005 (ACE05) Evaluation Plan: evaluation of the detection and recognition of ACE entities, values, temporal expressions, relations, and events, 2005. http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v2a.pdf. (Accessed 13 July 2014).

[5] F. Gey, Research to improve cross-language retrieval — position paper for CLEF, in: Cross-Language Information Retrieval and Evaluation, in: Lecture Notes in Computer Science, vol. 2069, 2000, pp. 83–88.

[6] N. Friburger, D. Maurel, Textual similarity based on proper names, in: Proceedings of the 25th ACM SIGIR conference, 2002, pp. 155–167.

[7] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Linguist. Investig. 30 (1) (2007) 3–26.

[8] B. Pouliquen, R. Steinberger, C. Ignat, T. De Groeve, Geographical information recognition and visualization in text written in various languages, in: Proceedings of ACM Symposium on Applied Computing, 2004, pp. 1051–1058.

[9] N.K. Kumar, G.S.K. Santosh, V. Varma, A language-independent approach to identify the named entities in under-resourced languages and clustering multilingual documents, in: Multilingual and Multimodal Information Access Evaluation, in: Lecture Notes in Computer Science, vol. 6941, 2011, pp. 74–82.

[10] O. Täckström, Nudging the envelope of direct transfer methods for multilingual named entity recognition, in: Proceedings of NAACL-HLT Workshop on the Induction of Linguistic Structure, 2012, pp. 55–63.

[11] E.F.T. Sang, F. de Meulder, Introduction to the CoNLL-2003 shared task: Language independent named entity recognition, 2003. http://www.cnts.ua.ac.be/conll2003/pdf/14247tjo.sh.pdf. (Accessed 13 July 2014).

[12] C.J. Lee, C.H. Chen, S.H. Kao, P.J. Cheng, To translate or not to translate? in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 651–658.

[13] BasisTech.com, Entity Extractor, 2014. http://www.basistech.com/text-analytics/rosette/entity-extractor/. (Accessed 17 July 2014).

[14] D.B. Bracewell, F. Ren, S. Kuroiwa, A low cost machine translation method for cross-lingual information retrieval, Eng. Lett. 16 (1) (2008) 160–165.

[15] Y. Al-Onaizan, K. Knight, Translating named entities using monolingual and bilingual resources, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 404–408.

[16] M. Faruqui, "Translation can't change a name" using multilingual data for named entity recognition, 2014. http://arxiv.org/pdf/1405.0701v1.pdf. (Accessed 17 July 2014).

[17] H. Alshawi, S. Douglas, Learning dependency transduction models from unannotated examples, Phil. Trans. R. Soc. 358 (2000) 1357–1372.

[18] A. Lopez, M. Nossal, R. Hwa, P. Resnik, Word-level alignment for multilingual resource acquisition, in: Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data, 2002. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA458782. (Accessed 13 July 2014).

[19] X. He, Using word dependent transition models in HMM based word alignment for statistical machine translation, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2007, pp. 80–87.

[20] W. Che, M. Wang, C.D. Manning, T. Liu, Named entity recognition with bilingual constraints, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2013, Atlanta, GA, 2013.

[21] M. Wang, W. Che, C.D. Manning, Effective bilingual constraints for semi-supervised learning of named entity recognizers, in: The Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA, 2013.

[22] M. Ehrmann, M. Turchi, R. Steinberger, Building a multilingual named entity-annotated corpus using annotation projection, in: Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 12–14 September 2011, (2011) pp. 118–124.

[23] T. Mandl, P. Carvalho, G.M. Di Nunzio, F. Gey, R.R. Larson, D. Santos, C. Womser-Hacker, GeoCLEF 2008: The CLEF 2008 cross-language geographic information retrieval track overview, in: Evaluating Systems for Multilingual and Multimodal Information Access, in: Lecture Notes in Computer Science, vol. 5706, 2008, pp. 808–821.

[24] J. Gelernter, W. Zhang, Cross-lingual geo-parsing for non-structured data, in: Proceedings of 7th Workshop on Geographic Information Retrieval, 2013. https://doi.org/10.1145/2533888.2533943.

[25] V.R. Murthy, M. Khapra, P. Bhattacharyya, Sharing network parameters for crosslingual named entity recognition, 2016. arXiv preprint arXiv:1607.00198.

[26] S. Dandapat, A. Way, Improved named entity recognition using machine translation-based cross-lingual information, Comput. Sistemas 20 (3) (2016) 495–504.

[27] A. Das, D. Ganguly, U. Garain, Named entity recognition with word embeddings and wikipedia categories for a low-resource language, ACM Trans. Asian Low-Res. Lang. Inf. Process. 16 (3) (2017) 18.

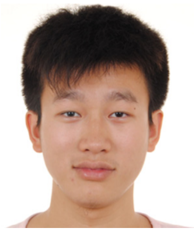
[28] R. Agerri, G. Rigau, Robust multilingual named entity recognition with shallow semi-supervised features, Artif. Intell. 238 (2016) 63–82.

[29] C. Dyer, V. Chahumeau, Noah A. Smith, (2013) A simple, fast, and effective re-parameterization of IBM model 2, in: Proceedings of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies, pp. 644–648.

[30] A. Birch, M. Osborne, P. Koehn, Predicting success in machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 745–754.

Dr. Xu Chen is an associate professor at Wuhan University, China. He focuses on Geographic Information Retrieval and Natural Language Processing. He holds a Ph.D. degree in Geographic Information Science from Wuhan University (2010, Hubei, China) and was a visiting scholar at Carnegie Mellon University (2013 Pennsylvania, USA).

Dr.Judith Gelernter has been a Research Scientist at National Institute of Standards and Technology (NIST) in Maryland Since 2015. Her work focuses on Natural Language Processing techniques and visualization for text-based work, and for mapping. Her Ph.D. degree in Information Science comes from Rutgers University, and she conducted research from 2008–2014 at the Language Technologies Institute of the School of Computer Science at Carnegie Mellon University.



Han Zhang is a machine learning scientist in Microsoft AI & Research. His work mainly focuses on Deep Learning and Natural Language Processing. He graduated from LTI, CMU with a master degree.

Jin Liu is a professor in the State Key Laboratory of Software Engineering, Computer School, Wuhan University. His research interests include Software Engineering and interactive collaboration on the Web. His work has been published in several international journals including CCPE, J SUPERCOMPUT and IEEE TSE.

Please cite this article in press as: X. Chen, et al., Multi-lingual geoparsing based on machine translation, Future Generation Computer Systems (2017), http://dx.doi.org/10.1016/j.future.2017.07.057.