Multilingual Open Relation Extraction Using Cross-lingual Projection

Manaal Faruqui
Carnegie Mellon University
Pittsburgh, PA 15213
mfaruqui@cs.cmu.edu

Shankar Kumar
Google Inc.
New York, NY 10011
shankarkumar@google.com

Abstract

Open domain relation extraction systems identify relation and argument phrases in a sentence without relying on any underlying schema. However, current state-of-the-art relation extraction systems are available only for English because of their heavy reliance on linguistic tools such as part-of-speech taggers and dependency parsers. We present a cross-lingual annotation projection method for language independent relation extraction. We evaluate our method on a manually annotated test set and present results on three typologically different languages. We release these manual annotations and extracted relations in ten languages from Wikipedia.

1 Introduction

Relation extraction (RE) is the task of assigning a semantic relationship between a pair of arguments. The two major types of RE are closed domain and open domain RE. While closed-domain RE systems (Bunescu and Mooney, 2005; Bunescu, 2007; Mintz et al., 2009; Yao and Van Durme, 2014; Berant and Liang, 2014) consider only a closed set of relationships between two arguments, open domain systems (Yates et al., 2007; Carlson et al., 2010; Fader et al., 2011; Mausam et al., 2012) use an arbitrary phrase to specify a relationship. In this paper, we focus on open-domain RE for multiple languages. Although there are advantages to closed domain RE (Banko and Etzioni, 2008), it is expensive to construct a closed set of relation types which would be meaningful across multiple languages.

Open RE systems extract patterns from sentences in a given language to identify relations. For learning these patterns, the sentences are analyzed using a part of speech tagger, a dependency parser and possibly a named-entity recognizer. In languages other than English, these tools are either unavailable or not accurate enough to be used. In comparison, it is easier to obtain parallel bilingual corpora which can be used to build machine translation systems (Resnik and Smith, 2003; Smith et al., 2013).

In this paper, we present a system that performs RE on a sentence in a source language by first translating the sentence to English, performing RE in English, and finally projecting the relation phrase back to the source language sentence. Our system assumes the availability of a machine translation system from a source language to English and an open RE system in English but no any other analysis tool in the source language. The main contributions of this work are:

- A pipeline to develop relation extraction system for any source language.
- Extracted open relations in ten languages based on Wikipedia corpus.
- Manual judgements for the projected relations in three languages.

We first describe our methodology for language independent cross-lingual projection of extracted relations (§2) followed by the relation annotation procedure and the results (§3). The manually annotated relations in 3 languages and the automatically extracted relations in 61 languages are available at: http://cs.cmu.edu/~mfaruqui/soft.html.

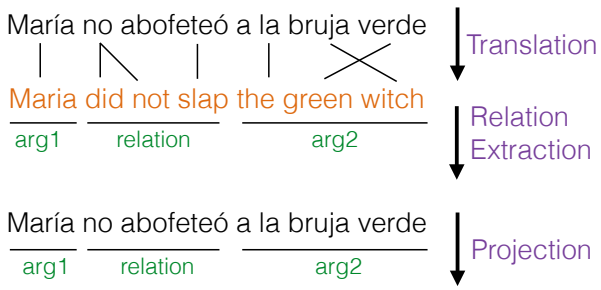

Figure 1: RE in a Spanish sentence using the cross-lingual relation extraction pipeline.

2 Multilingual Relation Extraction

Our method of RE for a sentence $\mathbf{s} = \langle s_1, s_2, \ldots s_N \rangle$ in a non-English language consists of three steps: (1) Translation of **s** into English, that generates a sentence $\mathbf{t} = \langle t_1, t_2, \ldots t_M \rangle$ with word alignments **a** relative to **s**, (2) Open RE on **t**, and (3) Relation projection from **t** to **s**. Figure 1 shows an example of RE in Spanish using our proposed pipeline.[1] We employ OLLIE[2] (Mausam et al., 2012) for RE in English and GOOGLE TRANSLATE[3] API for translation from the source language to English, although in principle, we could use any translation system to translate the language to English. We next describe each of these components.

2.1 Relation Extraction in English

Suppose $\mathbf{t} = \langle t_1, t_2, \ldots, t_M \rangle$ is a tokenized English sentence. Open relation extraction computes triples of non-overlapping phrases (**arg1**; **rel**; **arg2**) from the sentence **t**. The two arguments **arg1** and **arg2** are connected by the relation phrase **rel**.

We utilized OLLIE (Mausam et al., 2012) to extract the relation tuples for every English sentence. We chose OLLIE because it has been shown to give a higher yield at comparable precision relative to other open RE systems such as REVERB and WOE$^{\text{parse}}$ (Mausam et al., 2012). OLLIE was trained by extracting dependency path patterns on annotated training data. This training data was bootstrapped from a set of high precision seed tuples extracted from a simpler RE system REVERB (Fader et al., 2011). In *Godse killed Gandhi*, the extracted relation (Godse; killed; Gandhi) can be expressed by the dependency pattern: **arg1** ↑ nsubj ↑ **rel**:postag=VBD ↓ dobj ↓ **arg2**.[4] OLLIE also normalizes the relation phrase for some of the phrases, for example *is president of* is normalized to *be president of*. [5]

Data: **s**, **t**, **a**, p_t
Result: p_s
$P \leftarrow$ PhraseExtract(**s**, **t**, **a**)
$p_s = \emptyset$, $score = -\infty$, $overlap = 0$
for $(phr_s, phr_t) \in P$ **do**
  if *BLEU*$(phr_t, p_t) > score$ **then**
    if $phr_t \cap p_t \neq \emptyset$ **then**
      $p_t \leftarrow phr_t$
      $score \leftarrow$ BLEU(phr_t, p_t)
      $overlap \leftarrow phr_t \cap p_t$
if $overlap \neq 0$ **then**
  $length = \infty$
  for $(phr_s, p_t) \in P$ **do**
    if $len(phr_s) < length$ **then**
      $length \leftarrow len(phr_s)$
      $p_s \leftarrow phr_s$;
else
  $p_s \leftarrow$ WordAlignmentProj(**s**, **t**, **a**, p_t);

Algorithm 1: Cross-lingual projection of phrase p_t from a target sentence **t** to a source sentence **s** using word alignments **a** and parallel phrases P.

2.2 Cross-lingual Relation Projection

We next describe an algorithm to project the extracted relation tuples in English back to the source language sentence. Given a source sentence, the GOOGLE TRANSLATE API provides us its translation along with the word-to-word alignments relative to the source. If $\mathbf{s} = s_1^N$ and $\mathbf{t} = t_1^M$ denote the source and its English translation, then the alignment $\mathbf{a} = \{a_{ij} : 1 \leq i \leq N; 1 \leq j \leq M\}$ where,

[1] This is a sample sentence and is not taken from Wikipedia.
[2] http://knowitall.github.io/ollie/
[3] https://developers.google.com/translate/
[4] Example borrowed from Mausam et al. (2012)
[5] For sentences where the veracity of a relation depends on a clause, OLLIE also outputs the clause. For example, in *Early astronomers believed that Earth is the center of the universe*, the relation (Earth; be center of; universe) is supplemented by an (*AttributedTo:* believe; Early astronomers) clause. We ignore this clausal information.

$a_{ij} = 1$ if s_i is aligned to t_j, and is 0 otherwise. A naive word-alignment based projection would map every word from a phrase extracted in English to the source sentence. This algorithm has two drawbacks: first, since the word alignments are many-to-many, each English word can be possibly mapped to more than one source word which leads to ambiguity in its projection; second, a word level mapping can produce non-contiguous phrases in the source sentence, which are hard to interpret semantically.

To tackle these problems, we introduce a novel algorithm that incorporates a BLEU score (Papineni et al., 2002) based phrase similarity metric to perform cross-lingual projection of relations. Given a source sentence, its translation, and the word-to-word alignment, we first extract phrase-pairs P using the phrase-extract algorithm (Och and Ney, 2004). In each extracted phrase pair $(phr_s, phr_t) \in P$, phr_s and phr_t are contiguous word sequences in **s** and **t** respectively. We next determine the translations of **arg1**, **rel** and **arg2** from the extracted phrase-pairs.

For each English phrase $p \in \{\text{arg1}, \text{rel}, \text{arg2}\}$, we first obtain the phrase-pair $(phr_s, phr_t) \in P$ such that phr_t has the highest BLEU score relative to p subject to the condition that $p \cap phr_t \neq \emptyset$ i.e, there is at least one word overlap between the two phrases. This condition is necessary since we use BLEU score with smoothing and may obtain a non-zero BLEU score even with zero word overlap. If there are multiple phrase-pairs in P that correspond to the same target phrase phr_t, we select the shortest source phrase (phr_s). However, if there is no word overlap between the target phrase p and any of the target phrases in P, we project the phrase using the word-alignment based projection. The cross-lingual projection method is presented in Algorithm 1.

3 Experiments

Evaluation for open relations is a difficult task with no standard evaluation datasets. We first describe the construction of our multilingual relation extraction dataset and then present the experiments.

Annotation. The current approach to evaluation for open relations (Fader et al., 2011; Mausam et al., 2012) is to extract relations from a sentence and manually annotate each relation as either valid (1) or invalid (0) for the sentence. For example, in the sentence: *"Michelle Obama, wife of Barack Obama was born in Chicago"*, the following are possible annotations: a) (Michelle Obama; born in; Chicago): 1, b) (Barack Obama; born in; Chicago): 0. Such binary annotations are not available for languages apart from English. Furthermore, a binary 1/0 label is a coarse annotation that could unfairly penalize an extracted relation which has the correct semantics but is slightly ungrammatical. This could occur either when prepositions are dropped from the relation phrase or when there is an ambiguity in the boundary of the relation phrase.

Therefore to evaluate our multilingual relation extraction framework, we obtained annotations from professional linguists for three typologically different languages: French, Hindi, and Russian. The annotation task is as follows: *Given a sentence and a pair of arguments (extracted automatically from the sentence), the annotator identifies the most relevant contiguous relation phrase from the sentence that establishes a plausible connection between the two arguments*. If there is no meaningful contiguous relation phrase between the two arguments, the arguments are considered invalid and hence, the extracted relation tuple from the sentence is considered incorrect.

Given the human annotated relation phrase and the automatically extracted relation phrase, we can measure the similarity between the two, thus alleviating the problem of coarse annotation in binary judgments. For evaluation, we first report the percentage of valid arguments. Then for sentences with valid arguments, we use smoothed sentence-level BLEU score (max n-gram order = 3) to measure the similarity of the automatically extracted relation relative to the human annotated relation.[6]

Results. We extracted relations from the entire Wikipedia[7] corpus in Russian, French and Hindi from all sentences whose lengths are in the range of $10 - 30$ words. We randomly selected $1,000$ relations for each of these languages and annotated them. The results are shown in table 1. The percent-

[6] We obtained two annotations for ≈ 300 Russian sentences. Between the two annotations, the perfect agreement rate was 74.5% and the average BLEU score was 0.85.

[7] www.wikipedia.org

Language	Argument 1	Relation phrase	Argument 2
French	Il *He*	fut enrôlé de force au *was conscripted to*	RAD *RAD*
Hindi	bahut se log *Many people*	aaye *came to*	cailifornia *California*
Russian	Автокатастрофа *Crash*	произошла *occured*	Черногории *Montenegro*

Table 3: Examples of extracted relations in different languages with English translations (Hindi is transliterated).

Language	% valid	BLEU	Relation length	
			Gold	Auto
French	81.6%	0.47	3.6	2.5
Hindi	64.9%	0.38	4.1	2.8
Russian	63.5%	0.62	1.8	1.7

Table 1: % of valid relations and BLEU score of the extracted relations across languages with the average relation phrase length (in words).

Language	Size	Language	Size
French	6,743	Georgian	497
Hindi	367	Latvian	491
Russian	7,532	Tagalog	102
Chinese	2,876	Swahili	114
Arabic	707	Indonesian	1,876

Table 2: Number of extracted relations (in thousands) from Wikipedia in 10 languages out of a total of 61.

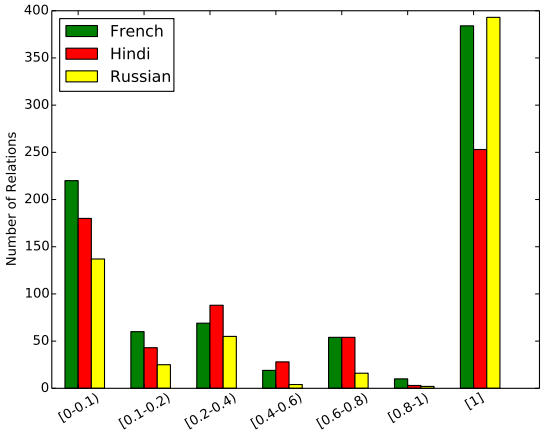

Figure 2: Number of automatically extracted relations binned by their BLEU scores computed relative to the manually annotated relations.

age of valid extractions is highest in French (81.6%) followed by Hindi and Russian (64.0%). Surprisingly, Russian obtains the lowest percentage of valid relations but has the highest BLEU score between the automatic and the human extracted relations. This could be attributed to the fact that the average relation length (in number of words) is the shortest for Russian. From table 1, we observe that the length of the relation phrase is inversely correlated with the BLEU score.

Figure 2 shows the distribution of the number of extracted relations across bins of similar BLEU scores. Interestingly, the highest BLEU score bin (1) contains the maximum number of relations in all three languages. This is an encouraging result since it implies that the majority of the extracted relation phrases are identical to the manually annotated relations. Table 2 lists the sizes of automatically extracted relations on 61 different languages from Wikipedia that we are going to make publicly available. These were selected to include a mixture of high-resource, low-resource, and typologically different languages. Table 3 shows examples of randomly selected relations in different languages along with their English translations.

4 Related Work

Cross-lingual projection has been used for transfer of syntactic (Yarowsky and Ngai, 2001; Hwa et al., 2005) and semantic information (Riloff et al., 2002; Padó and Lapata, 2009). There has been a growing interest in RE for languages other than English. Gamallo et al. (2012) present a dependency-parser based open RE system for Spanish, Portuguese and Galician. RE systems for Korean have been developed for both open-domain (Kim et al., 2011) and closed-domain (Kim and Lee, 2012; Kim et al., 2014) using annotation projection. These approaches use a Korean-English parallel corpus to project relations extracted in English to Korean. Following projection, a Korean POS-tagger and a dependency parser are employed to learn a RE system

for Korean.

Tseng et al. (2014) describe an open RE for Chinese that employs word segmentation, POS-tagging, dependency parsing. Lewis and Steedman (2013) learn clusters of semantically equivalent relations across French and English by creating a semantic signature of relations by entity-typing. These relations are extracted using CCG parsing in English and dependency parsing in French. Blessing and Schütze (2012) use inter-wiki links to map relations from a relation database in a pivot language to the target language and use these instances for learning in a distant supervision setting. Gerber and Ngomo (2012) describe a multilingual pattern extraction system for RDF predicates that uses pre-existing knowledge bases for different languages.

5 Conclusion

We have presented a language independent open domain relation extraction pipeline and have evaluated its performance on three typologically different languages: French, Hindi and Russian. Our cross-lingual projection method utilizes OLLIE and GOOGLE TRANSLATE to extract relations in the language of interest. Our approach does not rely on the availability of linguistic resources such as POS-taggers or dependency parsers in the target language and can thus be extended to multiple languages supported by a machine translation system. We are releasing the manually annotated judgements for open relations in the three languages and the open relations extracted over the entire Wikipedia corpus in ten languages. The resources are available at: `http://cs.cmu.edu/~mfaruqui/soft.html`.

Acknowledgment

This work was performed when the first author was an intern at Google. We thank Richard Sproat for providing comments on an earlier draft of this paper. We thank Hao Zhang for helping us with the relation extraction framework, and Richard Zens and Kishore Papineni for their feedback on this work. We are grateful to Bruno Cartoni, Vitaly Nikolaev and their teams for providing us annotations of multilingual relations.

References

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL*.

J. Berant and P. Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of ACL*.

Andre Blessing and Hinrich Schütze. 2012. Crosslingual distant supervision for extracting relations of different complexity. In *Proceedings of CIKM*.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of EMNLP*.

Razvan C. Bunescu. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of ACL*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of AAAI*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP*.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of ROBUS-UNSUP*.

Daniel Gerber and Axel-Cyrille Ngonga Ngomo. 2012. Extracting multilingual natural-language patterns for rdf predicates. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:11–311.

Seokhwan Kim and Gary Geunbae Lee. 2012. A graph-based cross-lingual projection approach for weakly supervised relation extraction. In *Proceedings of ACL*.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2011. A cross-lingual annotation projection-based self-supervision approach for open information extraction. In *Proceedings of IJCNLP*.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Trans. Asian Lang. Inf. Process.*, pages 3–3.

Mike Lewis and Mark Steedman. 2013. Unsupervised induction of cross-lingual semantic relations. In *Proceedings of EMNLP*.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of EMNLP-CoNLL*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, pages 417–449.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*.

Ellen Riloff, Charles Schafer, and David Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of COLING*.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of ACL*.

Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, Bo-Shun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Oren Etzioni, and Anthony Fader. 2014. Chinese open relation extraction for knowledge acquisition. In *Proceedings of EACL*.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of ACL*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*.

Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. 2007. Textrunner: Open information extraction on the web. In *Proceedings of NAACL: Demonstrations*.