A Cross-lingual Annotation Projection-based Self-supervision Approach for Open Information Extraction

Seokhwan Kim, Minwoo Jeong[†], Jonghoon Lee, Gary Geunbae Lee
Department of Computer Science and Engineering,
Pohang University of Science and Technology
{megaup, stardust, jh21983, gblee}@postech.ac.kr

Abstract

Open information extraction (IE) is a weakly supervised IE paradigm that aims to extract relation-independent information from large-scale natural language documents without significant annotation efforts. A key challenge for Open IE is to achieve self-supervision, in which the training examples are automatically obtained. Although the feasibility of Open IE systems has been demonstrated for English, utilizing such techniques to build the systems for other languages is problematic because previous self-supervision approaches require language-specific knowledge. To improve the cross-language portability of Open IE systems, this paper presents a self-supervision approach that exploits parallel corpora to obtain training examples for the target language by projecting the annotations onto the source language. The merit of our method is demonstrated using a Korean Open IE system developed without any language-specific knowledge.

1 Introduction

The objective of information extraction (IE) is to generate structured information representing semantic relationships among a set of arguments from natural language documents. Although many supervised machine learning approaches have been successfully applied to IE tasks, applications of these approaches are still limited because large amounts of training data are required to achieve good extraction results. Because manual annotation for training examples is very expensive, weakly-supervised techniques to learn the IE system without significant annotation efforts have been sought (Zhang, 2004; Chen et al., 2006).

Open IE is an alternative weakly-supervised IE paradigm (Banko et al., 2007). The goal of Open IE is to yield both domain-independent and relation-independent extractions from a large amount of natural language text without requiring hand-crafted rules or hand-annotated training examples. A key challenge to implementing Open IE is to learn extractors without manually annotated training examples. Self-supervised learning approaches have allowed Open IE systems such as TextRunner (Banko et al., 2007) and WOE (Wu and Weld, 2010) to extract relations from large-scale English text with automatically annotated training examples obtained using external knowledge.

However, applying the self-supervision approaches adopted by previously reported Open IE systems to build a new system is problematic in languages other than English, because these approaches mainly depend on language-specific knowledge for English. For example, TextRunner obtains training examples from the English Penn Treebank by triggering a set of hand-written heuristics denoting syntactic structural constraints to decide whether or not a given instance has a semantic relationship. To learn an extractor for a new language, this approach requires a syntactically annotated corpus and language-specific heuristics for the target language. WOE achieves self-supervised learning of Open IE by using heuristic matches between attribute values in Wikipedia infoboxes and their corresponding sentences. This method can reduce the cost of

[†] Now at Microsoft Bing

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 741–748,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

building an Open IE system for a new language, because Wikipedia articles and their infoboxes are available not only for English, but also for most other languages. But differences among languages in the amount of available resources from Wikipedia are still severe; for example, English Wikipedia includes about 3.5 million articles, but Korean Wikipedia includes only about 150,000 articles as of January 2011.

In this paper, we propose a cross-lingual annotation projection-based self-supervision approach to improve the cross-language portability of Open IE systems. This method exploits parallel corpora to obtain training examples in the target language by projecting the annotations generated by the Open IE system for the source language. The goal is to determine whether a semantic relationship in a pair of noun phrases in the target language L_T is the same as in the corresponding pair of noun phrases in the source language L_S; this process is called cross-lingual annotation projection. Using our self-supervision approach, we developed the first English-to-Korean Open IE system that does not require any language-specific knowledge. We use an English-Korean parallel corpus to project the results of an English Open IE system onto training examples for the target Korean system.

We present the definition of Open IE problem in Section 2, describe our cross-lingual annotation projection-based self-supervision approach for Open IE in Section 3, present details about implementation of the Korean Open IE system developed based on our proposed approach in Section 4, report the evaluation result of the system in Section 5, present related work in Section 6, and conclude this paper in Section 7.

2 Open Information Extraction

The problem of Open IE is to learn a function $f : D \rightarrow \{\langle e_i, r_{i,j}, e_j \rangle | 1 \leq i, j \leq N\}$, where D is a given natural language document, e_i and e_j are entities which have a semantic relationship that is explicitly expressed in a contextual subtext $r_{i,j}$, and N is the total number of entities in D. For example, the output of an Open IE system for an input sentence *"Obama was born in Hawaii."* will be a tuple $\langle$ *Obama, was born in, Hawaii* $\rangle$. Whereas traditional relation extraction problems such as ACE RDC have attempted to process both explicit and implicit relationships, Open IE aims to only extract explicit relationships $r_{i,j}$ in the context (Banko et al., 2007). Following Banko (2007), this paper concerns semantic relationships between entity pairs within a single sentence and considers each base noun phrase as an entity candidate.

Because the goal of Open IE paradigm is to eliminate direct human supervision, an extractor should be learned from the training examples obtained automatically without requiring handcrafted rules or hand-labeled annotations: this process is called self-supervised learning. Self-supervised learning for Open IE is performed in two steps: (1) self-supervision and (2) extractor learning. In the self-supervision step, the training examples to learn an extractor are generated for each instance, i.e., pair of noun phrases in the given sentence. Next, self-supervised learning determines whether or not each instance is semantically related. The key to achieving self-supervision is to determine how to automatically identify the existence of a semantic relationship between noun phrases. Whereas previously reported Open IE systems have performed this determination based on syntactic structural heuristics or structured information from Wikipedia, our proposed self-supervision approach utilizes the projected annotations from the results of Open IE system developed for another language. Details about our self-supervision approach are provided in Section 3.

In the learning step, a set of training examples obtained from self-supervision is utilized to learn an extractor f. The extractor has been successfully implemented using statistical models such as the Naive Bayes classifier (Banko et al., 2007) and conditional random fields (CRF) (Banko et al., 2008).

3 Cross-Lingual Annotation Projection-Based Self-Supervision

Cross-lingual annotation projection is an approach to obtain training examples for L_T by projecting the annotations for L_S using parallel corpora between L_T and L_S. This approach has been applied for several natural language processing tasks which have differences in the amounts of available resources among target languages (Yarowsky and Ngai, 2001; Yarowsky et al., 2001; Merlo et al., 2002; Hwa et al., 2002; Zitouni and Florian, 2008; Pado and Lapata, 2009). A premise of our method is that parallel corpora between L_T and L_S are much easier to obtain than is a task-specific training dataset for L_T: this premise is generally reasonable because large numbers of parallel corpora for various language pairs are available.

We consider the Open IE as a task with an imbalance problem in resource according to the target language, because most reported systems for Open IE were developed only for English and because they depend on language-specific knowledge. We propose a cross-lingual annotation projection-based self-supervision method of obtaining training examples for Open IE. The cross-lingual annotation projection for self-supervision can be performed for each bi-sentence pair $\langle S_S^i, S_T^i \rangle$ in parallel corpora between L_T and L_S as follows:

1) **Annotation:** Given an input sentence S_S^i, a set of extracted tuples O_S^i is yielded by the extractor f_s for the source language L_S.

2) **Projection:** The annotations O_T^i for the sentence S_T^i are generated by projecting from O_S^i based on word alignment between S_S^i and S_T^i.

3.1 Annotation

The first step in projecting annotations from L_S onto L_T is to obtain annotations for the sentences in L_S, as follows:

1) A set of entities $\{e_S^1, \cdots, e_S^N\}$ in the given sentence S_S^i is identified using a base phrase chunker in L_S. Each base noun phrase is considered as an entity candidate.

2) Each instance is composed of a pair of entities $\langle e_S^l, e_S^m \rangle$ in S_S^i, where $1 \leq l < m \leq N$.

3) For each instance $\langle e_S^l, e_S^m \rangle$, the extractor f_s for the source language L_S outputs the existence of semantic relation between e_S^l and e_S^m and the textual fragment $r_S^{i,j}$ indicating the detected relationship.

As an example of annotation projection for self-supervision of Korean Open IE with a bi-text in an L_T Korean and an L_S English (Figure 1), the annotation of the sentence in English shows that the pair of entities "Barack Obama" and "Honolulu" has a semantic relationship and "was born in" indicates the relationship between two entities.

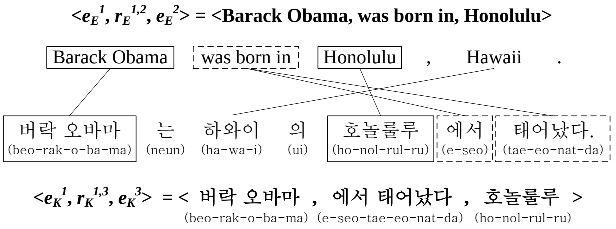
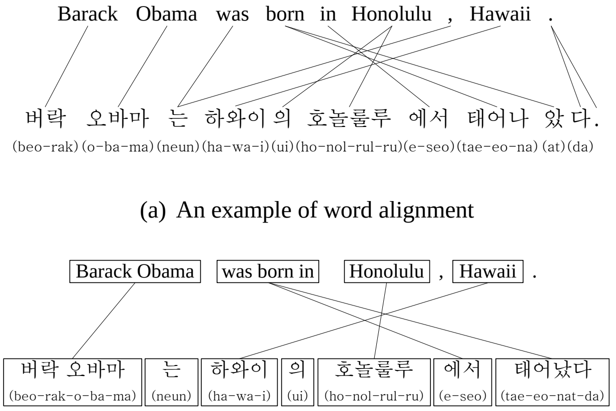

$<e_E^1, r_E^{1,2}, e_E^2>$ = <Barack Obama, was born in, Honolulu>
Barack Obama
was born in
Honolulu
,
Hawaii
.
버락 오바마 (beo-rak-o-ba-ma)
는 (neun)
하와이 (ha-wa-i)
의 (ui)
호놀룰루 (ho-nol-rul-ru)
에서 (e-seo)
태어났다. (tae-eo-nat-da)
$<e_K^1, r_K^{1,3}, e_K^3>$ = < 버락 오바마 , 에서 태어났다 , 호놀룰루 >
(beo-rak-o-ba-ma) (e-seo-tae-eo-nat-da) (ho-nol-rul-ru)

Figure 1: An example of cross-lingual annotation projection for Open IE of a bitext in English and Korean


Barack Obama was born in Honolulu , Hawaii .
버락 오바마 는 하와이 의 호놀룰루 에서 태어나 았 다.
(beo-rak) (o-ba-ma) (neun) (ha-wa-i) (ui) (ho-nol-rul-ru) (e-seo) (tae-eo-na) (at) (da)
Barack Obama
was born in
Honolulu
,
Hawaii
.
버락 오바마 (beo-rak-o-ba-ma)
는 (neun)
하와이 (ha-wa-i)
의 (ui)
호놀룰루 (ho-nol-rul-ru)
에서 (e-seo)
태어났다 (tae-eo-nat-da)

(a) An example of word alignment

(b) An example of chunk alignment

Figure 2: Comparision between word and chunk alignments

3.2 Projection

To use cross-lingual annotation projection to project the annotations from the sentences in L_S onto the sentences in L_T, we utilize word alignment information, which is an important component of statistical machine translation techniques. The objective of the word alignment task is to identify translational relationships among the words in a bi-text, and to produce a bipartite graph with a set of edges between words with translational relationships (Figure 2(a)). However, the results of automatic word alignment may include incorrect alignments because of technical difficulties. For example, the alignments (Figure 2(a)) have some errors such as ⟨Honolulu, ui⟩, ⟨COMMA, neun⟩ and ⟨PERIOD, da⟩.

The success of annotation projection is highly dependent on the quality of word alignment, to obtain quality results, the efforts to minimize harmful effects of erroneous word alignments should be minimized. In this work, we use alignments (Fig-

```
$A_P \leftarrow \vec{C}_S \times \vec{C}_T$
$A_C \leftarrow \emptyset$
for all $\left(C_S^i, C_T^j\right) \in A_P$ do
    $M(i,j) \leftarrow$ # of aligned words among $C_S^i$ and $C_T^j$
end for
while $A_P \neq \emptyset$ do
    $(i,j) \leftarrow \underset{(i',j')}{\operatorname{argmax}} \left(M(i',j') | (C_S^{i'}, C_T^{j'}) \in A_P\right)$
    if $(i,*) \notin A_C$ and $(*,j) \notin A_C$ then
        $A_C \leftarrow A_C \bigcup \{(i,j)\}$
    else if $(i, \vec{j_i}) \in A_C$ and $j$ is adjacent to $\vec{j_i}$ then
        $A_C \leftarrow \left(A_C - (i, \vec{j_i})\right) \bigcup \left\{ \left(i, \vec{j_i} \bigcup \{j\}\right) \right\}$
    else if $(\vec{i_j}, j) \in A_C$ and $i$ is adjacent to $\vec{i_j}$ then
        $A_C \leftarrow \left(A_C - (\vec{i_j}, j)\right) \bigcup \left\{ \left(\vec{i_j} \bigcup \{i\}, j\right) \right\}$
    end if
    $A_P \leftarrow A_P - (C_S^i, C_T^j)$
end while
return $A_C$
```

Figure 3: A chunk alignment algorithm

ure 3) between pairs of base phrase chunks instead of between pairs of words. For a given bi-text $\langle S_S^i, S_T^i \rangle$, a base phrase chunker for corresponding language produces chunk lists $\vec{C}_S$ for the source language and $\vec{C}_T$ for the target language. To identify the translational alignment between each pair of chunks C_S^i and C_T^j, the algorithm is performed in a simple greedy manner, i.e., a chunk pair that includes more word alignments is aligned before a chunk pair with few alignments, and a series of adjacent chunks aligned with the same counterpart can be merged. Chunk-based reorganization (Figure 3) of the word alignment in Figure 2(a) reduced the number of erroneous word alignments (Figure 2(b)).

Using chunk alignment, the annotations in the target language sentence S_T^i are projected from the annotations in the source language sentence S_S^i as follows:

1) As in the annotation phase, each instance is composed of a pair of base noun phrases $\langle e_T^l, e_T^m \rangle$ in S_T^i, where $1 \leq l < m \leq N$.

2) For each instance $\langle e_T^l, e_T^m \rangle$, its translational instance $\langle e_S^o, e_S^p \rangle$ in S_S^i is explored based on the result of chunk alignment.

3) The existence of semantic relationship in $\langle e_T^l, e_T^m \rangle$ is determined by projection.

4) If $\langle e_T^l, e_T^m \rangle$ is projected as a positive instance, the contextual subtext in S_T^i aligned with $r_S^{o,p}$ in S_S^i is extracted as $r_T^{l,m}$, and the final annotation $\langle e_T^l, r_T^{l,m}, e_T^m \rangle$ is produced.

In the Figure 1, an instance $\langle e_K^1, e_K^3 \rangle = \langle$ beo-rak-o-ba-ma, ho-nol-rul-ul $\rangle$ in the Korean sentence is aligned with the instance $\langle e_E^1, e_E^2 \rangle = \langle$ Barack Obama, Honolulu $\rangle$ in the English sentence. Because $\langle e_E^1, e_E^2 \rangle$ is predicted as a positive instance in the annotation phase, $\langle e_K^1, e_K^3 \rangle$ can be also considered to be a semantically related instance. Then, "e-seo-tae-eo-nat-da" in the Korean sentence is identified as $r_K^{1,3}$ which is aligned to $r_E^{1,2} =$ "was born in" in S_E^i, and finally, $\langle e_K^1, r_K^1, 3, e_K^3 \rangle = \langle$beo-rak-o-ba-ma, e-seo-tae-eo-nat-da, ho-nol-rul-ul$\rangle$ is yielded.

4 Implementation

We developed a Korean Open IE system (Figure 4) based on our proposed cross-lingual annotation projection-based self-supervised learning. Our system is operated with no language-specific knowledge or resource for the target language, Korean. It requires only an Open IE system for another source language and a parallel corpus between source and target languages. In this system, we have used English as the source language, because most reported techniques for Open IE were developed for English. According to the advantages of English Open IE systems, the objective of our system is to perform domain-independent and relation-independent extraction. Furthermore, the fact that manual annotations are not needed to obtain training examples is also valid for applying the system to a new language. The system consists of three parts: self-supervision, learning and extraction.

4.1 Self-supervision

The sole input of our self-supervision method is a parallel corpus of L_S and L_T. We used an English-Korean parallel corpus [1] which consists of 266,892 bi-sentence pairs in English and Korean. Each sentence in the corpus was processed for POS tagging and base phrase chunking using OpenNLP [2]

[1] The parallel corpus collected is available in our website http://isoft.postech.ac.kr/~megaup/ijcnlp/datasets

[2] http://incubator.apache.org/opennlp/

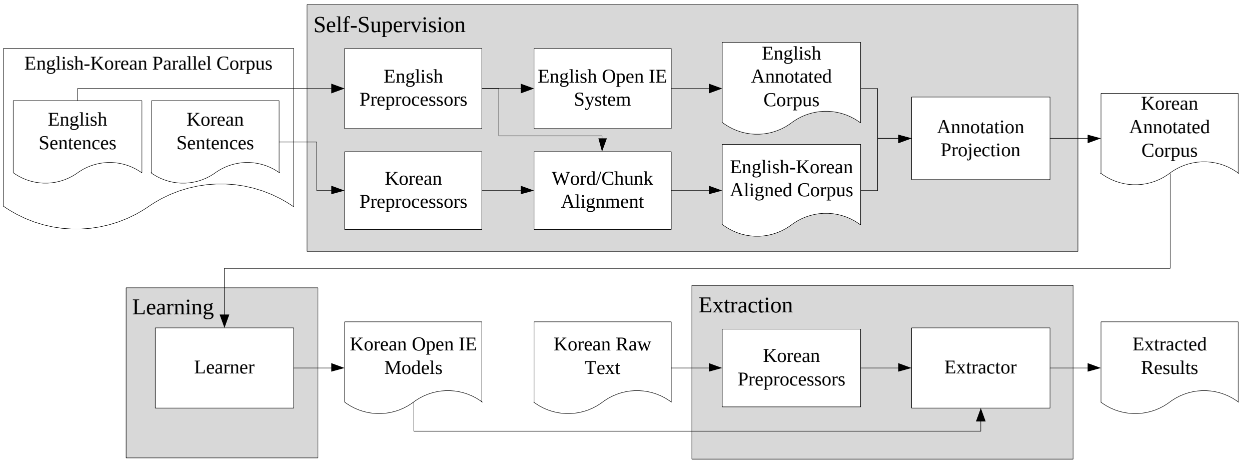


Figure 4: Overall architecture of the Korean Open IE system

for English sentences and Espresso [3] for Korean sentences.

For each preprocessed bi-sentence, word alignment was performed using GIZA++ software [4] (Och and Ney, 2003) in the standard configuration in both English-Korean and Korean-English directions. The bi-directional alignments were joined using the grow-diag-final algorithm (Koehn et al., 2003). The results of word alignment were reorganized by the chunk alignment algorithm 3.

The other prerequisite of the self-supervision in our system is that the Open IE system for the source language should obtain the annotations for source language sentences in the parallel corpus. We used our own implementation of the English Open IE system (Banko et al., 2007). We obtained a set of training examples to learn the extractor by applying a series of heuristics to the WSJ part of the Penn Treebank. From 49,208 sentences, 1,028,361 instances were generated; 9.0% of them were determined to be positive instances by the heuristic-based self-supervision. Using these instances, lexical and POS tag features were used to learn a CRF model. The CRF++ toolkit [5] was used to learn the extractor for English.

For the given preprocessed parallel corpus and Open IE system for the source sentences, annotation projection was performed. First, each English sentence in the parallel corpus was analyzed using the English Open IE system. Of 598,115 acquired instances, 169,771 positive annotations were produced by the annotation phase. These annotations were projected to the corresponding instances in the Korean part of the parallel corpus. This operation was performed based on the information obtained from chunk alignment. Finally, a set of training examples for the Korean Open IE system was projected. The projected dataset included 278,730 instances; 89,743 were positive.

4.2 Learning

Using training examples obtained by self-supervision, an extractor for Korean Open IE was generated. The extractor is composed of two statistical models. One is a maximum entropy (ME) classifier to detect whether or not each given instance is positive; the other is a CRF model to identify the contextual subtext indicating the semantic relationship for each positive instance. Both models utilized lexical and POS tag features in the node sequence of the dependency path between two entities organizing a given instance. The dependency path for each instance was generated using MSTParser (McDonald et al., 2005) [6] with a model trained on the Sejong corpus (Kim, 2006). The extractor was implemented using CRF++ and Maximum Entropy Modeling toolkits [7].

4.3 Extraction

During execution, the input of the system is raw text in Korean and the output is a set of extractions

[3] http://air.changwon.ac.kr/AIR/entry/Espresso_POS_K

[4] http://code.google.com/p/giza-pp/

[5] http://crfpp.sourceforge.net/

[6] http://sourceforge.net/projects/mstparser/

[7] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Model	P	R	F
Heuristic	47.7	20.1	28.3
Projection	33.6	49.0	39.8
Heuristic + Projection	41.9	46.4	44.1

Table 1: Comparison of performances among heuristic-based, projection-based and the merged models.

for the given text. The input text should be preprocessed by the analyzers for Korean sentences in the previously mentioned parts of the system. Then the instances and their features are extracted for each preprocessed sentence. The two models (Section 4.2) are operated in a cascaded manner for a given instance and its features: first the ME model identifies the existence of semantic relationship in a given instance, then the CRF model explores the context indicator only for instances determined to be positive by the ME model. Based on the results of two cascaded models, the system outputs the extracted results in the form of a triple, $\langle e_i, r_{i,j}, e_j \rangle$.

5 Evaluation

To evaluate our Korean Open IE system introducing cross-lingual annotation projection-based self-supervision, extractions were performed for two types of datasets. One dataset was built by annotating the semantic relationships denoted in a small number of sentences randomly selected from Korean Wikipedia articles. The dataset consists of 250 sentences and 1,434 instances, 308 of which were annotated to be positive instances. To compare with our system, we built a heuristic-based Korean Open IE system considered as a baseline. The baseline model was trained on the corpus automatically obtained from Sejong treebank corpus using a set of heuristics which were utilized for the English Open IE system except language-specific rules. On the test dataset, we measured the performances of three models: heuristic-based model, projection-based model, and the merged model trained on the mixture of both training datasets. Precision, Recall and F-measure were adopted for our evaluation.

Table 1 compares the performances of three models. The baseline model using only language-independent heuristics achieves poor performance, especially in recall. On the other hand, our proposed projection-based model outperforms the baseline model, due to largely increased recall. Moreover, the projected instances helps to improve the performance of the heuristic-based approach by merging the training datasets. The results show that our proposed projection-based method is more effective than the previous approach to build an Open IE system for a new language.

Type	Newswire		Wikipedia	
	prec.	# of extr.	prec.	# of extr.
Birth Place	65.2	256	69.1	971
Won Award	57.4	824	63.3	286
Acquisition	67.0	1112	50.3	143
Invent Of	53.1	32	47.6	103

Table 2: Evaluation results for four relation types

Error Type	# of errors
Chunking Error	364 (26.9%)
Dependency Parsing Error	461 (34.1%)
Extracting Error	527 (39.0%)

Table 3: Distribution of the errors

The second evaluation was performed on the extractions of our system for the large amount of documents. We used two datasets: one dataset consists of 2,565,487 sentences in 302,276 documents obtained from Korean Newswire Second Edition published by LDC; the other contains 1,342,003 sentences in 123,000 articles from Korean Wikipedia.

The evaluation was performed manually for the extracted results annotated by four relation types $\{$BIRTH_PLACE, WON_AWARD, ACQUISITION, INVENT_OF$\}$. The relation type of each extracted result was determined by manual clustering based on its contextual indicator $r_{i,j}$. Our system output 3,727 extractions with an average precision of 63.7% for four relation types (Table 2).

To investigate the reason for erroneous extractions, a qualitative analysis of 1,352 errors was performed (Table 3). Errors were classified into three categories: chunking errors and dependency parsing errors (both caused by the preprocessors), and extracting errors (caused by the extractor for well-preprocessed instances). About 60% of the errors were caused by preprocessors including base phrase chunking and dependency parsing. Because our system is highly dependent on the result of preprocessors, the performance of the extractor can be increased by reducing its sensitivity to preprocessor errors; this is a topic for future work.

6 Related Work

Many supervised machine learning approaches have been successfully applied to solve traditional relation extraction tasks (Kambhatla, 2004; Zhou et al., 2005; Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006), but these approaches require a large number of training examples to achieve high performance. To reduce the annotation cost, weakly-supervised techniques have been designed (Zhang, 2004; Chen et al., 2006).

Open IE pioneered by TextRunner (Banko et al., 2007) is an alternative weakly-supervised IE paradigm. TextRunner aims to perform relation-independent extraction by introducing the self-supervision approach based on a small set of heuristics about syntactic structural constraints. The performance of TextRunner was further improved using O-CRF and casting the Open IE task as a kind of sequential labeling problem (Banko et al., 2008). Wu and Weld (2010) presented another Open IE system WOE which utilizes an alternative self-supervision approach based on Wikipedia infoboxes. The main difference between our work and previous Open IE approaches is that we did not use language-dependent knowledge or resources for self-supervision, but implemented it using cross-lingual annotation projection techniques.

Early studies of cross-lingual annotation projection considered lexically-based tasks, e.g., part-of-speech tagging (Yarowsky and Ngai, 2001), named-entity tagging (Yarowsky et al., 2001), and verb classification (Merlo et al., 2002). Recently, applications of annotation projection such as dependency parsing (Hwa et al., 2002), mention detection (Zitouni and Florian, 2008), and semantic role labeling (Pado and Lapata, 2009) have been studied. To the best of our knowledge, no work has reported on the Open IE task.

7 Conclusions

This paper presented a novel self-supervision approach for Open IE. Our approach uses cross-lingual annotation projection to automatically obtain training examples for a target language by propagating annotations generated by an existing Open IE system for a source language via a parallel corpus between two languages. The main advantage of our method is that no language-dependent knowledge is required to learn the extractor. Our method can contribute to improving the cross-language portability of the Open IE paradigm.

The feasibility of our approach was demonstrated by our Korean Open IE system. The system was developed using only an English Open IE system and an English-Korean parallel corpus; the system never utilized any language specific knowledge or resources for the target language Korean. Furthermore, the system operated in fully unsupervised manner, because all components including prerequisites do not require hand-labeled annotations or hand-crafted rules for the target task. The system outperformed the baseline system based on the language-independent heuristics. For large amount of documents, the system produced 3,727 extractions with a precision of 63.7% for four relation types.

However, our method can still be improved. Many erroneous extractions were caused by errors committed by preprocessors. To reduce sensitivity to these kinds of errors, we plan to introduce assessment techniques which are not included in this work. Another direction of our future work is to investigate a hybrid approach to self-supervision considering not only cross-lingual projected annotations, but also various external knowledge source such as Wikipedia and WordNet. We expect that this fusion approach can help to improve the quality of extracted results, because the effectiveness of each approach has been demonstrated for IE tasks.

Acknowledgments

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2011-(C1090-1121-0008))

References

M. Banko, M. J Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. In *Proceedings of the IJCAI-07*, pages 2670–2676.

M. Banko, O. Etzioni, and T. Center. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of the ACL-08:HLT*, pages 28–36.

R. C Bunescu and R. J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the HLT/EMNLP 2005*, pages 724–731.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the COLING/ACL 2006*, pages 129–136.

A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the ACL 2004*, pages 423–429.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the ACL 2002*, pages 392–399.

N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL 2004*, pages 178–181.

H. Kim. 2006. Korean national corpus in the 21st century sejong project. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT-NAACL 2003*, volume 1, pages 48–54.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the HLT/EMNLP 2005*, pages 523–530.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the ACL 2002*, pages 207–214.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

S. Pado and M. Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

F. Wu and D. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the ACL 2010*, pages 118–127.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the NAACL 2001*, pages 1–8.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the HLT 2001*, pages 1–8.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.

Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of the COLING/ACL 2006*, pages 825–832.

Zhu Zhang. 2004. Weakly-supervised relation classification for information extraction. In *Proceedings of the CIKM 2004*, pages 581–588.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the ACL 2005*, page 434.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the EMNLP 2008*, pages 600–609.