Frame-Semantic Web: a Case Study for Korean*

Jungyeul Park[†‡], Sejin Nam[‡], Youngsik Kim[‡]
Younggyun Hahm[‡], Dosam Hwang[‡§], and Key-Sun Choi[‡]

[†]UMR 6074 IRISA, Université de Rennes 1, France
[‡]Semantic Web Research Center, KAIST, Republic of Korea
[§]Department of Computer Science, Yeungnam University, Republic of Korea
http://semanticweb.kaist.ac.kr

Abstract. FrameNet itself can become a resource for the Semantic Web. It can be represented in RDF. However, mapping FrameNet to other resources such as Wikipedia for building a knowledge base becomes more common practice. By such mapping, FrameNet can be considered to provide capability to describe the semantic relations between RDF data. Since the FrameNet resource has been proven very useful, multiple global projects for other languages have arisen over the years, parallel to the original English FrameNet. Accordingly, significant steps were made to further develop FrameNet for Korean. This paper presents how frame semantics becomes a frame-semantic web. We also provide the Wikipedia coverage by Korean FrameNet lexicons in the context of constructing a knowledge base from sentences in Wikipedia to show the usefulness of our work on frame semantics in the Semantic Web environment.

Keywords: Semantic Web, Frame Semantics, FrameNet, Korean FrameNet.

1 Introduction

FrameNet [1][1] is a both human- and machine-readable large-scale on-line lexical database, not only consists of thousands and thousands of words and sentences, but, moreover, an extensive and complex range of semantic information as well. Based on a theory of meaning called frame semantics, FrameNet strongly supports an idea that the meanings of words and sentences can be best understood on the basis of a semantic frame, a coherent conceptual structure of a word describing a type of event, relation, or entity and the participants in it. It is believed that semantic frames of related concepts are inseparable from each other, so that, one cannot have complete understanding of a word, without knowledge of all the semantic frames related to that word. FrameNet itself serves as a great example of such a principle, wherein 1,180 semantic frames closely link together by a system of semantic relations and provide a solid basis for reasoning about the meaning of the entire text.

* This work was supported by the IT R&D program of MSIP/KEIT. [10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform]

[1] https://framenet.icsi.berkeley.edu

FrameNet itself can become a resource for the Semantic Web as represented in RDF/OWL [2, 3]. Mapping FrameNet to other resources such as Wikipedia for building a knowledge base can also be considered to provide capability to describe the semantic relations between RDF data. Since the FrameNet resource has been proven useful in the development of a number of other NLP applications, even in the Semantic Web environment such as in [4], multiple global projects have arisen over the years, parallel to the original English FrameNet, for a wide variety of languages around the world. In addition to Brazilian Portuguese[2], French[3], German (the SALSA Project)[4], Japanese[5], Spanish[6], and Swedish[7], significant steps were made to further develop FrameNet for Korean, and the following sections of this paper present the process and mechanisms. By using FrameNet, it can become a frame-semantic web where frame semantics is enabled for the Semantic Web. We also provide the Wikipedia coverage by Korean FrameNet lexicons in the context of constructing a knowledge base from sentences in Wikipedia. It can show how the frame-semantic web would be useful in the Semantic Web environment.

2 Building a Database of Frame Semantic Information for Korean

We describe the manual construction of a FrameNet-style annotated corpus for Korean translated from the FrameNet corpus and its FE transfer based on English-Korean alignment using cross-linguistic projection proposed in [5, **?**]. We also explain this process by using the translated Korean FrameNet corpus and its counterpart English corpus as our bilingual parallel corpus. We propose a method for mapping a Korean LU to an existing FrameNet-defined frame to acquire a Korean frame semantic lexicon. Finally, we illustrate a self-training technique that can build a database of large-scale frame semantic information for Korean.

Manual Construction: The development of FrameNet for Korean has been the central goal of our project, and we have chosen to perform this task by starting off with "manually translating" the already-existing FrameNet from English to Korean language. Such decision was made on the grounds that, even though obtaining a large set of data through means of manual translation can be a difficult, costly and time-consuming process, its expected advantages indeed far outweigh the charge in the long run. The fact that only humans can really develop a true understanding and appreciation of the complexities of languages, subject knowledge and expertise, creativity and cultural sensitivity also makes manual translation the best option to adopt for our project. Expert translators

[2] `http://www.ufjf.br/framenetbr`
[3] `https://sites.google.com/site/anrasfalda`
[4] `http://www.coli.uni-saarland.de/projects/salsa`
[5] `http://jfn.st.hc.keio.ac.jp`
[6] `http://sfn.uab.es/SFN`
[7] `http://spraakbanken.gu.se/eng/swefn`

performed manual translation for all FrameNet full text annotated corpus with a word alignment recommendation system. A guideline manual for translating the FrameNet-style annotated corpus to Korean sentences was prepared for the clean transferring of English FrameNet annotated sentences to Korean.

Automatic Construction: We also extend previous approaches described in [5] using a bilingual English-Korean parallel corpus. Assuming that the same kinds of frame elements (FEs) exist for each frame for the English and Korean sentences, we achieve the cross-linguistic projection of English FE annotation to Korean via alignment of tokenized English and Korean sentences. English FE realization can be projected to its corresponding Korean sentences by transforming consecutive series of Korean tokens in the Korean translation of any given sentence. Since the alignment of English tokens to Korean tokens defines the transformation, the success of token alignment is crucial for the cross-linguistic projection process. For frame population to Korean lexical units (LUs), we present our method for the automatic creation of the Korean frame semantic lexicon for verbs in this section. We start by finding an appropriate translation for each verb to create a mapping between a Korean LU and an existing FrameNet-defined frame. In contrast to mapping from one sense to one frame, mapping to more than one frame requires using a further disambiguation process to select the most probable frame for a given verb. We use maximum likelihood estimation (MLE) for possible frames from the existing annotated corpora to select the correct frame. For the current work, we only used FrameNets lexicographic annotation to estimate MLE. We use the *Sejong* predicate dictionary[8] for frame semantic lexicon acquisition. We place 16,807 Korean verbs in FrameNet-defined frames, which constitute 12,764 distinctive orthographic units in Korean. We assume that FEs with respect to the assigned frame for Korean LUs are directly equivalent to the FEs in the corresponding English frames. Thus, we do not consider redefining FEs specifically for Korean.

Bootstrapping Frame-Semantic Information: Self-training for frame semantic role projection consists of annotating FrameNet-style semantic information, inducing word alignments between two languages, and projecting semantic information of the source language onto the target language. We used the bilingual parallel corpus for self-training, and a probabilistic frame-semantic parser [6] to annotate semantic information of the source language (English). Then, we induced an HMM word alignment model between English and Korean with a statistical machine translation toolkit. Finally, we projected semantic roles information from the English onto the Korean sentences. For the experiment, we employed a large bilingual English-Korean parallel corpus, which contains almost 100,000 bilingual parallel sentences to bootstrap the semantic information. During self-training, errors in the original model would be amplified in the new model; thus, we calibrate the results of the frame-semantic parser by using the confidence score of the frame-semantic parser as a threshold. As a result, 120,621 pairs of frames with their FEs are obtained and among them 30,149 are unique; 715 frames are used for 10,898 different lexica.

[8] `http://www.sejong.or.kr`

3 Linking FrameNet to Wikipedia

DBpedia[9] is a knowledge base constructed from Wikipedia based on DBpedia ontology (DBO). DBO can be viewed as a vocabulary to represent knowledge in Wikipedia. However, DBO is a Wikipedia-Infobox-driven ontology. That is, although DBO is suitable to represent essential information of Wikipedia, it does not guarantee enough to represent knowledge in Wikipedia written in a natural language. In overcoming such a problem, FrameNet has been considered useful in linguistic level as a language resource representing semantics. We calculate the Wikipedia coverage rate by DBO and FrameNets LUs to match the relation instantiation from DBpedia and FrameNet to Wikipedia. Before we calculate the Wikipedia coverage rate, we need to know which sentences within Wikipedia actually contain knowledge. We define that a typical sentence with *extractable knowledge* can be linked to DBpedia entities as a triple. From almost three millions sentences in Korean Wikipedia, we find over four millions predicates for cases where only a subject appears, only an object appears, or both of a subject and an object appear (2.11 predicates per sentence). We obtain 6.92% and 95.19% for DBO and FrameNets LUs, respectively. The shortage of DBO can be explained that DBO is too small to cover actual predicates in Wikipedia only by pre-defined predicates in DBO. However, FrameNet gives almost full coverage for sentences with extractable knowledge, which is very promising for extracting and representing knowledge in Wikipedia using FrameNet.

4 Discussion and Conclusion

Throughout this paper, by building a database of frame semantic information, we explained that FrameNet can become a resource for the Semantic Web and it can gather lexical linked data and knowledge patterns with almost full coverage for Wikipedia.

References

1. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. (2010)
2. Narayanan, S., Fillmore, C.J., Baker, C.F., Petruck, M.R.L.: FrameNet Meets the Semantic Web: A DAML+OIL Frame Representation. In: Proc. of AAAI-02.
3. Narayanan, S., Baker, C.F., Fillmore, C.J., Petruck, M.R.L.: FrameNet Meets the Semantic Web: Lexical Semantics for the Web. In: ISWC 2003.
4. Fossati, M., Tonelli, S., Giuliano, C.: Frame Semantics Annotation Made Easy with DBpedia. In: Proc. of CrowdSem2013. 69–78
5. Padó, S., Lapata, M.: Cross-lingual Bootstrapping of Semantic Lexicons: The Case of FrameNet. In: Proc. of AAAI-05.
6. Das, D., Schneider, N., Chen, D., Smith, N.A.: Probabilistic Frame-Semantic Parsing. In: Proc. of NAACL 2010.

[9] `http://dbpedia.org/About`