Korean FrameNet Expansion Based on Projection of Japanese FrameNet

Jeong-uk Kim, Younggyun Hahm, and **Key-Sun Choi**
Machine Reading Lab, School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
{prismriver, hahmyg, kschoi}@kaist.ac.kr

Abstract

FrameNet project has begun from Berkeley in 1997, and is now supported in several countries reflecting characteristics of each language. The work for generating Korean FrameNet was already done by converting annotated English sentences into Korean with trained translators. However, high cost of frame-preservation and error revision was a huge burden on further expansion of FrameNet. This study makes use of linguistic similarity between Japanese and Korean to increase Korean FrameNet corpus with low cost. We also suggest adapting PubAnnotation and Korean-friendly valence patterns to FrameNet for increased accessibility.

1 Introduction

Growing demand for natural language processing (NLP) inevitably requires large scale of proper training dataset. In this sense, Berkeley proposed FrameNet project of semantically analyzing sentences with several 'frames.'Based on Frame Semantics (Fillmore, 1982), event invoking words from sentences are selected to from a frame, with core roles of the event as frame elements. Such frame annotated dataset in FrameNet can be widely used as Semantic Role Labeling training set for Machine Translation, Information Extraction, Event Recognition and etc.

Since characteristics of target language must be considered to apply on the other NLP tasks, FrameNets have been developed in several languages. Most FrameNets like Japanese annotated sentences one by one (Ohara et al., 2003), but this procedure requires much time and effort of frame experts. On the other hand, utilizing existing corpus to easily generate FrameNet was also researched. For example, projection algorithm of English frame semantic data into Italian (Tonelli and Pianta, 2008) was suggested.

However, comparing to the abundant corpus in English FrameNet, FrameNets in other languages are containing relatively small dataset, or even missing. There was no Korean FrameNet until importing 4025 English FrameNet sentences into Korean using trained translators (Park et al., 2014). The translated sentences with frame information might be used for NLP in Korean with secured quality, but the quantity is rather too small for machine learning training data.

Focusing on the lack of frame annotated sentence, this study proposes much cost efficient method for expanding size of Korean FrameNet utilizing structural similarities between Japanese and Korean – use of postposition and interchangeable word order. After dividing original sentences into a set of word chunks, they were translated keeping the order of chunks putting the original frame information aside. This procedure did not require translators to learn frame semantics or to continue revisions of all frames. The projections of word chunks to frame elements were held on the translated sentences, with the extracted frame information. Furthermore, Korean FrameNet website introduces visualization of frame annotated sentences using PubAnnotation and valence patterns including postpositions.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 175–179, Osaka, Japan, December 11-17 2016.

2 Expanding Korean FrameNet

For previous version of full-text annotations, Korean FrameNet contained 4025 sentences chosen from random categories of English FrameNet. They designed detailed guidelines to translate them into Korean (Park et al., 2014). The guidelines include keeping every frame elements, and preserving of frame element meanings even after the relocation. The translated sentences have high qualities with consecutive verifications with both expert translators and NLP majors, but the absolute size of the data is relatively small for NLP application.

Lexical units are the stemmed words that invoke frames in the full-text annotations. Part of speech tag for each lexical unit was chosen as the part of speech tag of the last morpheme of the stemmed words. In this way, 7130 lexical units were listed on the Korean FrameNet in the Korean alphabetic order.

Frame index contains brief definition of the frame, with its core or non-core frame elements for each lexical unit. Since the creation of Korean FrameNet used translation based approach, the frame index information is identical to that of English FrameNet. Therefore, Korean FrameNet makes use of 1019 frame index data from Berkeley as it is.

In order to expand the Korean FrameNet dataset, this study proposes machine aided projection approach from Japanese to Korean. The two languages both support flexible change of word order as role of each word is highly related to the attached preposition. With these advantages, new full-text annotated sentences can be achieved with low cost using the following approach.

2.1 Extract word chunks from Japanese FrameNet Corpus

Suppose a simple sentence “梅雨はすでに明け、九州地方は一気に夏模様である。” “The monsoon is already stopped, and summer seems to come in Kyusu area soon” from Japanese FrameNet. Its frame information is listed in Figure 1.

Figure 1: list of frames in an annotated sentence

Every start and end position of frame elements and lexical units become boundaries to split word chunks. Some positions can be used in different frames several times. Examples of sentences separated as word chunks shown in Figure 2. Frame information of the sentence is stored in a file for further use.

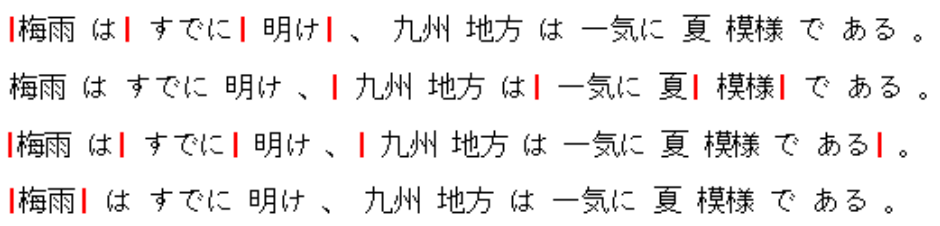

Figure 2: list of word chunks separated with frame data

2.2 Translate extracted word chunks

Every boundary in section 2.1 are merged to a single sentence to be sent to expert translators unfamiliar with NLP, just liken in Figure 3. Positions of every frame element and lexical unit are kept as index of boundaries in its start and end position.

|梅雨| は| すでに| 明け| 、| 九州 地方 は| 一気に 夏| 模様| で ある| 。

Figure 3: sentence to be translated

Guidelines for translation include – meaning of the full sentence as well as each word in a boundary must be unchanged, boundary must not be removed, added, or relocated, and that the translated text in Korean must be natural. Result of the above example is shown in Figure 4.

The basic concept behind such division is that both Korean and Japanese are flexible for word ordering. Instead, role of each noun largely depends on the postposition of the word. For instance, 'は'(ha) in Japanese and '는'(neun) in Korean are both postpositions that represents the former noun chunk is a subject of the current sentence. In this way, no matter where the original sentence has subject, we can always make a correct Korean sentence with subject in the same relative position.

|장마| 는| 이미| 끝나| .| 규슈 지방 은| 단숨에 여름| 이 올 듯한 모양| 이다| .

Figure 4: sentence translated with boundaries preserved

2.3 Retrieve annotated Korean sentences with frame information

The translated Korean sentence now requires to be reverted into frame annotated sentences. Stored frame information mapping word chunks indices to frame boundaries in the previous section is used to retrieve positions and frame index of frame elements and lexical units. An example of the result is shown in Figure 5. Frame annotated sentences in Korean are then added to the Korean FrameNet.

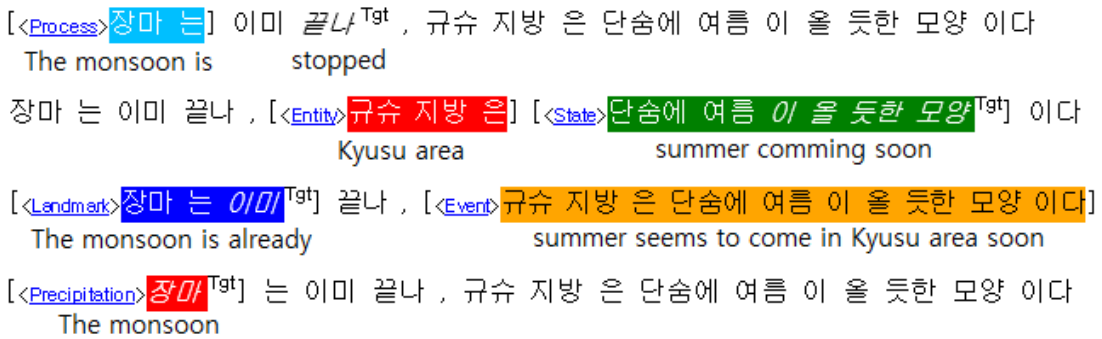

Figure 5: frame annotated sentence in Korean

2.4 Quality Insurance

Double-checking policy of translators secures the quality of translation. In addition, we ask for any problems of unnatural sentences translated in the original order. However, there was no such case among total 1795 sentences supporting the correctness of our approach.

3 Demo Website

This version of Korean FrameNet also changes visualization of annotated sentences. Most FrameNet web service shows each frame element with unique color for frame index. However, for non-experts in this field, users must find the corresponding frame index from the color table. To get rid of this unnecessary work, we focused on the PubAnnotation.

PubAnnotation (Kim and Wang, 2012) introduced open source interface for annotation sharing. The system supports annotating part of the given sentence and linking two annotations as a relation. Frame information in Korean FrameNet website became more intuitive with the interface. Comparison of a same annotation with the two different approaches is shown in Figure 6 and 7. Lexical unit 'fell' is more noticeable using PubAnnotation by focusing on the origin of every relation. Roles of frame elements in the sentence are also easily seen with the annotation instead of mapping colors of frame elements to its roles in the frame.

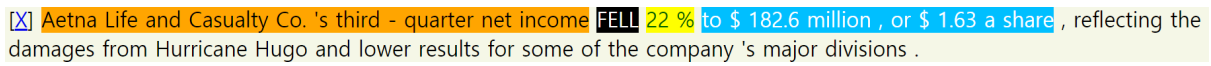

Figure 6: frame information in English FrameNet

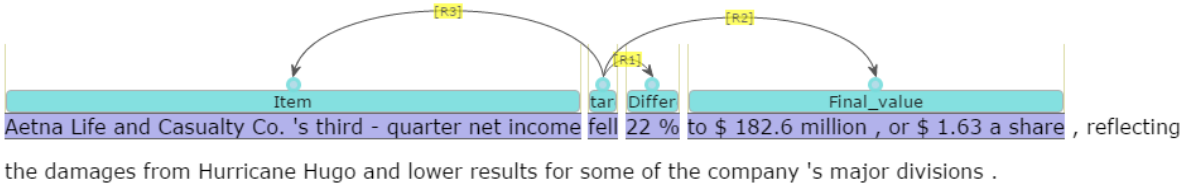

Figure 7: frame annotation using PubAnnotation

In addition, sentences with each lexical unit are categorized by their valence patterns. Originally, English FrameNet simply saves valences patterns as the order of frame elements. However, as mentioned above, postpositions play important roles in the meaning of a sentence. Valence patterns in Korean FrameNet are expressed with frame elements with part of speech tag of their prepositions like in Figure 8. For lexical unit '가' meaning 'go', the valence pattern shows a leading frame element 'theme' with a postposition for subject followed by 'time' without a postposition, 'goal' with a postposition for adverb, and the lexical unit. All FrameNet data is open to public in Korean FrameNet website[1].

[theme/JKS] + [time] + [goal/JKB] + 가

Figure 8: valence pattern with postposition

4 Conclusion

This study presents how to expand Korean FrameNet using Japanese FrameNet and to improve interface focused on Korean. Making use of linguistic characteristics dramatically reduces FrameNet transition costs with only few errors, and even more Japanese full-text annotations would be easily imported with reuse of parsing and error revision tools. Similar approaches can be applied to other language pair with the same characteristics. Combining FrameNet framework with PubAnnotation also lowers the accessibility of the public.

With richer annotation sets in Korean FrameNet and better visualization interface, Korean FrameNet has moved closer to the NLP researchers. Current Korean FrameNet might still be not enough for large scale NLP. However, Korean FrameNet would keep growing, and help researchers suffering from the lack of Korean dataset as a major resource of semantic role labeling.

[1] http://framenet.kaist.ac.kr

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R0101-16-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

This work was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP(2015M3A9A7029735)

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Charles Fillmore. 1982. Frame semantics. *Linguistics in the morning calm*, pages 111–137.

Jin-Dong Kim and Yue Wang. 2012. Pubannotation: a persistent and sharable corpus and annotation repository. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205. Association for Computational Linguistics.

Kyoko Hirose Ohara, Seiko Fujii, Hiroaki Saito, Shun Ishizaki, Toshio Ohori, and Ryoko Suzuki. 2003. The japanese framenet project: A preliminary report. In *Proceedings of pacific association for computational linguistics*, pages 249–254. Citeseer.

Jungyeul Park, Sejin Nam, Youngsik Kim, Younggyun Hahm, Dosam Hwang, and Key-Sun Choi. 2014. Frame-semantic web: a case study for korean. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 257–260. CEUR-WS. org.

Sara Tonelli and Emanuele Pianta. 2008. Frame information transfer from english to italian. In *LREC*.