SinK-DaT: Reducing age gap through Sino-Korean Detector and Translator

Mina Huh
20160697_MinaHuh

SeokJun Kim
20160793_SeokJunKim

Jeongeon Park
20160811_JeongeonPark

Juhoon Lee
20160832_JuhoonLee

Hyunchang Oh
20170410_HyunchangOh

Abstract

We introduce a new Korean language model called **SinK-DaT**, the **Sin**o-**K**orean (Sin-Ko) **D**etector **a**nd **T**ranslator. SinK-DaT provides the correct meaning and Chinese character representations of the "difficult" Sin-Ko word based on context from a given Korean text. Experiments on old Korean news articles show that the model can not only detect most Sin-Ko words, but also distinguish difficult ones among the detected Sin-Ko words and thus, supply the most suitable definitions to the users. As a result, Sink-DaT helps the young Korean generation in comprehending difficult corpora.

1 Introduction

Sino-Korean (Sin-Ko) words are Hanja-based words incorporated to the Korean language. Such words are widely used in many Korean corpora for their ability to contain complex meanings in just few letters. Despite its wide usage, especially among the older generation, Hanja is seldom taught in the modern curriculum, resulting in an unbridgeable linguistic gap between the younger and the older generations. Understanding each other's language is vital in reducing the generational and domain language gap. Thus, we constructed a model that detects and defines difficult Sin-Ko words to aid reading comprehension and to expedite language acquisition for the younger generation.

2 Approach

There are three main steps for developing SinK-DaT: *1) Detecting Sin-Ko words*, *2) Scoring their difficulty levels*, and *3) Showing their appropriate definitions in the context*.

To detect the Sin-Ko words within a sentence, we searched each word in an online dictionary, and treated it as a Sin-Ko word if a match was found. The underlying heuristic assumes Sin-Ko and pure Korean semes rarely share the same form. Even if a pure Korean word was misidentified, it would be eliminated in later steps as the difficulty and the context would not be mapped correctly.

To construct the 'difficult' vocabulary set in a sentence, we used a rule-based approach. As the degree of difficulty is subjective to reader's proficiency levels, we set the target group as TOPIK L3-L4 learners, whose evaluation criteria include the ability to understand easy parts of news broadcasts and newspapers. We used three different metrics to measure the difficulty of each word. First, we counted the number of Synsets using KorLex. The number signifies different meanings a word can have, with a greater number implying higher complexity. Second, we analyzed the level of each Hanja character in a word by referring to the official reading level classification from the Hanja Education Research Institute. If the levels were low, meaning they are difficult and rarely used, we scored the difficulty as high. Lastly, we compared the number of occurrences of Sin-Ko words within basic and advanced corpora, which were elementary school textbooks and newspaper articles respectively. In addition to the corpora comparison, TOPIK vocabulary guide was used to filter easy words. The weighted scores from the three metrics were combined to rank the difficulties of the Sin-Ko words.

Sin-Ko words tend to have many homonyms, making it difficult to identify the correct definition. To perform word sense disambiguation, we constructed an n-gram language model to compare the contexts of each Sin-Ko word in the sentence and its definitions. To compare the context of each word, its n-gram probability scores were calculated with the n-grams created from each of the possible definitions to identify the definition with the highest probability of being the correct meaning.

Precision	Recall	Accuracy	F-score
0.538	0.560	0.439	0.549

Table 1: Model's Sin-Ko Detection Result.

3 Experiment

For the training dataset, more than 7K sentences were retrieved from Chosun Ilbo News Library using random articles from 1990. The model first tokenized raw text with Kkma PoS tagging function of KoNLPy, with auxiliary components removed based on the PoS tags. Then, each word was searched on NAVER's Hanja Dictionary. Tokens matching the Korean form of Sin-Ko word in the dictionary were regarded as potential Sin-Ko words. Using the newspaper sentences, we built an n-gram (n=3) model using NLTK to compare the context of words with n-grams created from possible definitions from NAVER's Hanja Dictionary. To represent an easy corpora, public elementary textbooks provided by the Korean textbook publication Kyohaksa were used along with other tools and classifications as noted in section 2. To represent the advanced corpora, news article sentences from above was used.

The evaluation consists of three parts - *1) Detection of Sin-Ko words*, *2) Detection of difficult Sin-Ko words*, and *3) Mapping of difficult words to their meanings*.

For 1) Detection of Sin-Ko words and 3) Mapping of words to meanings in context, we used 500 sentences from the Chosun Ilbo news articles as the test set. The model-detected words were compared with manually tagged words to calculate the performance, and for finding meanings, we calculated the accuracy, also using manual tagging.

We used *crowdsourcing* to evaluate 2) Detection of difficult Sin-Ko words. We surveyed 12 middle school students for their proficiency on the Sin-Ko words. Four paragraphs with roughly 25 sentences each from the *Chosun Ilbo* news articles were used for the survey. The students were asked to rate the given Sin-Ko word on a 3 point scale – 1) Never heard/seen, 2) Seen but don't know the meaning, 3) Seen and can explain the meaning. We then calculated the average score of each word.

4 Result

Sin-Ko Word Detection The results are shown in Table 1. Recall was higher than the precision, which shows that the existing Sin-Ko words were correctly identified, even if some pure Korean words were misidentified. Considering the fact that each sentence contained on average ten Sin-Ko words and that a sentence was marked to be incorrect for even one wrong detection, the results are promising.

Difficult Sin-Ko Word Detection The average rated score of the Sin-Ko words from the survey was *2.560*, while that of the detected difficult words was *2.243*. This shows that the model-detected difficult words were perceived to be more difficult than other Sin-Ko words.

Context-wise Definition Mapping The accuracy of the mapped definition of the difficult Sin-Ko words was *0.686*. The majority of the difficult Sin-Ko word definitions were correctly identified given their context, but a room for improvement exists.

5 Discussion

Discussion & Limitation We have demonstrated that our model using a rule-based approach together with the data-driven model is effective in detecting and mapping definitions of difficult Sin-Ko words. As our project solves NLP problems in Korean, the accuracy of KoNLPy module influenced the performance of pre-processing. When PoS tags were mismatched, certain Sino-Koreans were undetected or annotated with wrong Sin-Ko words. One of the scoring criteria in detecting difficult words is the difficulty of constituent Hanjas. While this is true for most of the cases, some easy words (e.g. 여유: 餘裕) have difficult Hanjas while some difficult words (e.g. 소실: 消失) have easy Hanjas.

Future Work We plan to add more criteria to detecting the difficult words and allow users to specifically choose the difficulty level of Sin-Ko words. This will develop our model into an essential tool for Sin-Ko learning for all generations.

6 Conclusion

We present **SinK-DaT**, a model that detects and translates difficult Sin-Ko words with appropriate context. Our major contribution is the analysis of the characteristic of Sin-Ko words distinguished from the pure Korean words and proposing various measures to score the level of difficulty of a word. By yielding context-aware definition of difficult Sin-Ko words, SinK-DaT helps the younger generation to advance to a wider range of corpora.

Acknowledgements

We thank Prof. Jong Chul Park for teaching us how to tackle NLP problems as well as encouraging us during the whole process. We also thank the TAs of CS372, Huijee Lee, Soyeong Jeong, Hoyun Song, Eugene Jang, Euijun Hwang, Wonsuk Yang and Junseop Ji for assisting us and doing all the hard works. We also thank Chosun News for providing available archived resources.

Bibliographical References

Dongjun Lee, Yubin Lim, and Ted "Taekyoung" Kwon. 2018. Morpheme-based Efficient Korean Word Embedding. Journal of KIISE, 45(5):444–450.

Eunjeong L. Park, Sungzoon Cho. "KoNLPy: Korean natural language processing in Python", Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea, Oct 2014.

Gil-Ja So, Seung-Hee Lee, and Hyuk-Chul Kwon. 2011. Generalization of error decision rules in a grammar checker using Korean WordNet, KorLex. The KIPS Transactions:PartB, 18B(6):405–414.

Hye-Jeong Song, Ji-Eun Chi, Yoon-Kyoung Lee, Ji Hye Yoon, Jong-Dae Kim, Chan-Young Park and Yu-Seop Kim. 2019. A Web Service for Evaluating the Level of Speech in Korean Applied Sciences. February.

Kevin P. Yancey, Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) 438-445, May.

News.chosun.com. 2020. [한자 문맹(漢字文盲) 벗어나자] "안중근 의사(義士), 어떤 과목 진료했죠?"… 젊은층 '독해不能(한자어 뜻 몰라 문장 해독 못하는 현상)' 심각. [online] Available at: <https://news.chosun.com/site/data/htmldir/2014/\04/22/2014042200992.html>

Sanghyuk Choi, Taeuk Kim, Jinseok Seol, and Sang-Goo Lee. 2017. A Syllable-based Technique for Word Embeddings of Korean Words. Proceedings of the First Workshop on Subword and Character Level Models in NLP.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level Word Vector Representations for Korean. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

T. J. Ostrand and M. J. Balcer. 1988. The category-partition method for specifying and generating fuctional tests. Communications of the ACM, 31(6):676–686, January.

Language References

KorLex. http://korlex.pusan.ac.kr/

KoNLPy (Python package for natural language processing of the Korean language). http://konlpy.org/

Test Of Proficiency in Korean. https://www.topik.go.kr/

Corporation of Hanja Education Research Institute http://new.hanja.net/

The Chosun Ilbo. https//www.news.chosun.com/

교학사 교과서. https://www.kyohak.co.kr/

네이버 한자사전 (NAVER Hanja dictionary). https://hanja.dict.naver.com/

조선일보-라이브러리. https://newslibrary.chosun.com/

한자검증능력시험. https://www.hanja.ne.kr/

A Appendix

0. Raw Text

오전 10시7분 증인선서 낭독으로 시작된 전 전대통령의 증언은 5공비리, 광주문제에 관한 답변 속에 저녁까지 계속됐다. 그러나 증언도중 증언내용이 부실하고 위증이 있다는 야당의원들의 고함과 의사진행 발언요구 등속에 소란이 끊임없이 벌어져 모두 7여 차례의 정회가 거듭되다. 결국 밤 12시를 넘김으로써 전씨가 의사당을 떠나, 청문회는 전씨의 구두 답변을 다 듣지 못한 채 끝나고 말았다.

1. Sin-Ko Detection

오전(Sin-Ko) 10시7분 **증인**(Sin-Ko)**선서**(Sin-Ko) **낭독**(Sin-Ko)으로 **시작**(Sin-Ko) 된 전 **전대통령**(Sin-Ko)의 **증언**(Sin-Ko)은 5공비리, **광주**(Sin-Ko)**문제**(Sin-Ko)에 관한 **답변**(Sin-Ko) 속에 저녁까지 **계속**(Sin-Ko)됐다. 그러나 **증언**(Sin-Ko)도중 **증언**(Sin-Ko)**내용**(Sin-Ko)이 **부실**(Sin-Ko)하고 **위증**(Sin-Ko)이 있다는 **야당**(Sin-Ko)**의원**(Sin-Ko)들의 **고함**(Sin-Ko)과 **의사**(Sin-Ko)**진행**(Sin-Ko) **발언**(Sin-Ko)**요구**(Sin-Ko) 등속에 **소란**(Sin-Ko)이 끊임없이 벌어져 모두 7여 차례(Sin-Ko)의 **정회**(Sin-Ko)가 거듭되다. **결국**(Sin-Ko) 밤 12시를 넘김으로써 전씨가 **의사당**(Sin-Ko)을 떠나, **청문회**(Sin-Ko)는 전씨의 **구두**(Sin-Ko) **답변**(Sin-Ko)을 다 듣지 못한 채 끝나고 말았다.

2. Difficult Sin-Ko Word Identification

오전(Easy) 10시7분 증인(Easy)선서(Easy) **낭독**(Hard)으로 시작(Easy) 된 전 전대통령(Easy)의 증언(Easy)은 5공비리, 광주(Easy)문제(Easy)에 관한 답변(Easy) 속에 저녁까지 계속(Easy)됐다. 그러나 증언(Easy)도중 증언(Easy) 내용(Easy)이 부실(Easy)하고 **위증**(Hard)이 있다는 야당(Easy) 의원(Easy) 들의 **고함**(Hard) 과 의사(Easy) 진행(Easy) 발언(Easy) 요구(Easy) 등속에 **소란**(Hard)이 끊임없이 벌어져 모두 7여 차례(Easy)의 정회(Easy)가 거듭되다. 결국(Easy) 밤 12시를 넘김으로써 전씨가 의사당(Easy)을 떠나, 청문회(Easy)는 전씨의 구두(Easy) 답변(Easy)을 다 듣지 못한 채 끝나고 말았다.

3. Difficult Sin-Ko Word Mapping

오전(Easy) 10시7분 증인(Easy)선서(Easy) **낭독**(朗讀, 소리를 높이어 밝게 읽음)으로 시작(Easy) 된 전 전대통령(Easy)의 증언(Easy)은 5공비리, 광주(Easy)문제(Easy)에 관한 답변(Easy) 속에 저녁까지 계속(Easy)됐다. 그러나 증언(Easy)도중 증언(Easy)내용(Easy)이 부실(Easy)하고 **위증**(僞證, 거짓으로 증명함. 또는 그런 증거)이 있다는 야당(Easy)의원(Easy)들의 **고함**(叩喊, 북을 치면서 여러 사람이 함께 큰 소리를 지름과 의사(Easy) 진행(Easy) 발언(Easy) 요구(Easy) 등속에 **소란**(騷亂)이 끊임없이 벌어져 모두 7여 차례(Easy)의 정회(Easy)가 거듭되다. 결국(Easy) 밤 12시를 넘김으로써 전씨가 의사당(Easy)을 떠나, 청문회(Easy)는 전씨의 구두(Easy) 답변(Easy)을 다 듣지 못한 채 끝나고 말았다.

Figure 1. The process of Sino-Korean detection and definition mapping with a sample paragraph.