Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1603–1610
Marseille, 11–16 May 2020

Korean-Specific Emotion Annotation Procedure Using N-Gram-Based Distant Supervision and Korean-Specific-Feature-Based Distant Supervision

Young-Jun Lee, Chae-Gyun Lim, Ho-Jin Choi
Korea Advanced Institute of Science and Technology, School of Computing
Daejeon, South Korea
{yj2961, rayote, hojinc}@kaist.ac.kr

Abstract

Detecting emotions from texts is considerably important in an NLP task, but it has the limitation of the scarcity of manually labeled data. To overcome this limitation, many researchers have annotated unlabeled data with certain frequently used annotation procedures. However, most of these studies are focused mainly on English and do not consider the characteristics of the Korean language. In this paper, we present a Korean-specific annotation procedure, which consists of two parts, namely n-gram-based distant supervision and Korean-specific-feature-based distant supervision. We leverage the distant supervision with the n-gram and Korean emotion lexicons. Then, we consider the Korean-specific emotion features. Through experiments, we showed the effectiveness of our procedure by comparing with the KTEA dataset. Additionally, we constructed a large-scale emotion-labeled dataset, Korean Movie Review Emotion (KMRE) Dataset, using our procedure. In order to construct our dataset, we used a large-scale sentiment movie review corpus as the unlabeled dataset. Moreover, we used a Korean emotion lexicon provided by KTEA. We also performed an emotion classification task and a human evaluation on the KMRE dataset.

Keywords: Korean-specific emotion annotation procedure, Korean emotion labeled dataset, distant supervision

1. Introduction

In recent years, the interest of detecting emotions in texts has grown in NLP, but there are still challenges with building desirable emotion detection models because of the lack of labeled data with emotions. Thus, many researchers have constructed fine-grained emotion labeled datasets, which are created in different domains, emotion models in psychology, and annotation procedures. Among these, the annotation procedure is divided into three categories: *expert-based*, *crowd-sourcing*, and *distant supervision*. Among these procedures, most of the previous studies have created annotated datasets by using standard methods such as *expert-based* and *crowd-sourcing*. Through these methods, datasets can be annotated manually by experts or platforms (e.g. Amazon Mechanical Turk). However, these methods are considerably costly in terms of both time and money. Therefore, recent studies have attempted to annotate datasets by using *distant supervision* (called *self-labeling*), as this procedure is quicker and cheaper than the others. The existing emotion labeled datasets built using these procedures can be described in (Klinger and others, 2018).

However, most of the publicly available emotion labeled datasets are in English. Therefore, it is difficult to detect emotions from non-English texts, particularly Korean texts. However, in some studies, researchers have constructed datasets annotated with sentiments or emotions in Korean (Shin et al., 2012; Do and Choi, 2015). In contrast, previous datasets are mainly labeled with sentiments. Moreover, the dataset labeled with emotions is not sufficiently large to build a better emotion detection model. Therefore, it is desirable to construct a large-scale dataset labeled with emotions in Korean.

In this paper, we present a novel annotation procedure that can construct a large-scale emotion labeled dataset by using n-gram-based distant supervision and the Korean emotion lexicon. More specifically, in this study, we analyzed the Korean emotion lexicon and a large amount of unlabeled data as morpheme units using an explicit morpheme analyzer. Then, we annotated the unlabeled dataset with seven types of emotions through n-gram-based distant supervision. Finally, we exploited Korean-specific features (e.g., emoticons and emotion letters) to better understand emotions by using distant supervision. We also constructed an emotion-labeled dataset, Korean Movie Review Emotion (KMRE) Dataset, by applying our presented procedure to the Naver Sentiment Movie Corpus (NSMC) in Korean.

Our first contribution is our Korean-specific annotation procedure that automatically annotates a large-scale unlabeled dataset with emotions by exploiting n-gram-based distant supervision and Korean-specific-features-based distant supervision. Our second contribution is that we provide a large-scale emotion labeled dataset, called the KMRE dataset. To the best of our knowledge, this is the first work that automatically constructs the Korean-specific emotion-labeled dataset by distant supervision with the use of an emotion lexicon. Moreover, this is the first large-scale textual Korean-specific emotion-labeled dataset, which is publicly available.

2. Related Work

There are three types of representative annotation procedures: *expert-based*, *crowd-sourcing*, and *distant supervision*. As a standard method, the *expert-based* method is the most commonly used procedure, allowing some experts who understand a domain of the dataset to annotate the dataset with emotions (Li et al., 2017). The *crowd-sourcing based* method is also a manual procedure similar to the former one, except that this allows some workers to annotate datasets by using the platforms (e.g., Amazon Mechanical

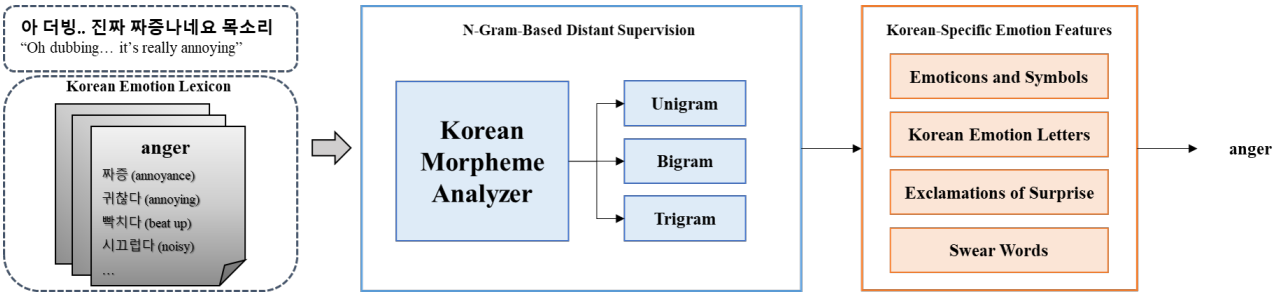

Figure 1: Overview of the annotation procedure using Korean emotion lexicon

Turk[1] or CrowdFlower[2]) (Milnea et al., 2015; Lapitan et al., 2016). However, these two procedures incur a considerable cost to build a large annotated emotion dataset manually, as doing so requires human time, money, and effort. Therefore, recent studies have used *distant supervision* (Go et al., 2009) to get a large amount of data annotated with emotions. *Distant supervision* is called *self-labeling* or *weak labeling*. Previous studies have leveraged this method to create emotion-labeled datasets with the use of emoticons (Tang et al., 2014; Deriu et al., 2016), hashtags (Mohammad, 2012; Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017), or emojis (Eisner et al., 2016; Felbo et al., 2017) as noisy labels. Another research used lexicons, which have the word-emotion association, to create large amounts of annotated emotion data for obtaining an emotion-enriched word representation (Agrawal et al., 2018). In this study, they add emotion vectors associated with words in a sentence. These emotion vectors are derived from lexicons and they get an emotion label from the calculated emotion vectors.

There are not many publicly available datasets labeled with emotions. Most of the previous studies created sentiment-labeled datasets, such as KOSAC (Shin et al., 2012; Jang et al., 2013) and NSMC. To the best of our knowledge, the only publicly available emotion labeled dataset is a Korean Twitter Emotion Analysis (KTEA[3]) dataset (Do and Choi, 2015). The KTEA contains various resources to help analyze Korean emotions in the Twitter domain. To build an emotion-labeled dataset, three annotators manually annotated each tweet with seven types of emotions. As a result, the KTEA contains 5,706 tweets labeled with seven types of emotions. Additionally, the KTEA provides two different Korean-specific emotion lexicons, which were made of the *Weighted Tweet Frequency* (TwF) approach and the *Thesaurus-Based + Translation-Based* (TT) approach. The TwF emotion lexicon has a weighted TwF value which represents the strength of the corresponding emotion. In other words, the higher the value, the more it is associated with emotion. Each lexicon has words associated with Ekman's six emotions (Ekman, 1999). In this work, we used a combination of two emotion lexicons as a form of distant supervision, as it achieved the best performance in (Do and Choi, 2015).

[1] https://www.mturk.com/
[2] https://www.figure-eight.com/
[3] goo.gl/Gu0GNw

3. Method

In this section, we introduce how to construct the large-scale emotion-labeled dataset in Korean automatically. First, we analyzed an unlabeled dataset and Korean-specific emotion lexicons at the morpheme-level. Second, we annotated the unlabeled dataset with seven types of emotions through n-gram-based distant supervision. Lastly, we additionally annotated the unlabeled dataset with distant supervision by using Korean-specific features. An overview of our annotation procedure is illustrated in Figure 1.

3.1. Morphological Analysis of Korean Words

As one of an agglutinative languages, Korean is a morphologically rich language, where each word is composed of a set of morphemes. More specifically, the Korean word (called *Eojeol*) is formed by combining postposition morphemes (*Eomi* and *Josa*) based on the root morpheme (*Eogan*). Because of this variation, we can produce various different forms that have a similar meaning, which leads to make an increase in the size of the vocabulary. Therefore, it is appropriate to analyze each word into morphemes. The morpheme is the smallest unit of meaning in linguistics. Recently, many studies have improved word representations by decomposing each word into *syllable-level* and *jamo-level* parts (Choi et al., 2017; Park et al., 2018). However, in this work, we assumed that the morpheme in each word was the smallest unit that expressed emotions well. To prepare the annotation, thus, we first analyze emotion lexicons and the unlabeled dataset at the *morpheme-level*. We exploited the `Open API` for the morpheme analyzer provided by the `ETRI`, which consists of 47 parts of speech (POS) tags[4]. Each tag is described in the Appendix section.

In general, morphemes are divided into two types according to their semantic characteristics (functions): *Actual morpheme* and *Formal morpheme*. In particular, the actual morpheme, called lexical morpheme, is a morpheme that represents concrete objects or abstract concepts. The formal morpheme, called grammatical morpheme, represents grammatical relationships between the actual morphemes by combining with the actual morpheme. In other words, the actual morphemes contain more semantic information than the formal morphemes. Thus, we did not consider the formal morphemes during the annotation procedure, as the actual morphemes have a considerable effect on the expression of the emotions of the sentence. We denote which

[4] http://aiopen.etri.re.kr/index.php

POS tags belonged to each type (i.e., the actual morpheme and the formal morpheme) in Table 6. In fact, some of the formal morphemes (e.g., the auxiliary particle, prefix, and suffix) may have meanings under certain conditions, but we excluded the very small details. Moerover, this part is described in the National Institute of Korean Language[5]. Empirically, we removed certain morphemes that would not have a significant effect on emotions among the formal morphemes (i.e. EP, EF, EC, ETN, ETM, XPN, XSN, XSV, and XSA), which belong to the Markers.

3.2. N-Gram-Based Distant Supervision

As aforementioned, we leveraged the distant supervision with the analyzed unlabeled dataset and the Korean emotion lexicons (i.e., a combination of TwF and TT) from the ETRI morpheme analyzer. Previous work (Agrawal et al., 2018) using distant supervision with lexicons only considered one word for labeling with emotions. However, in Korean, it is not appropriate to only take care of a unigram because of the Korean characteristics. In other words, the unigram can convey different emotions depending on what postposition morphemes are after the root morphemes. For example, there are the "좋아하/VV, 지/EC, 않/VX, 다/EF" of the dataset and the "좋아하/VV, 다/EF" of the emotion lexicons, which are analyzed from the sentence "좋아하지 않다" and the element of lexicons "좋아하다", respectively. According to a previous study, the result of the sentence can be happiness more than anger or disgust as the previous study determined emotions only by considering a unigram such as the "좋아하/VV" morpheme. However, this sentence "좋아하지 않다" is closer to anger or disgust. Thus, we utilized not only unigrams but also bigrams and trigrams.

More specifically, let $D = \{d_1, d_2, ..., d_N\}$ be a set of unlabeled sentences and $D' = \{d'_1, d'_2, ..., d'_N\}$ be a set of the analyzed unlabeled sentences. The difference between D and D' was only whether each sentence was analyzed. Each unlabeled sentence in D' is represented as follows:

$$d'_i = \{w_{i,1}, w_{i,2}, ..., w_{i,|d'_i|}\}, i = 1, ..., N \quad (1)$$

where $w_{i,j}$ denotes the analyzed morpheme unit and N is the size of the dataset. Let L_{fuse} denote the combination of TwF and TT. Each emotion lexicon has words or phrases associated with emotions. We adopted Ekman's model of six emotions (Ekman, 1999), namely *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Further, we added *neutral* to the six types of emotions, where *neutral* represented no emotion. Note that we generated six emotion labels based on Ekman's model for constructing our KMRE dataset. The reason for doing so is explained later.

In the case of a unigram, for each $w_{i,j}$, we computed an emotion vector of $w_{i,j}$ (i.e., $emo(w_{i,j})$), which is represented as one-hot encoded vector with seven dimensions. When we used the L_{fuse} emotion lexicon, we extracted only the top-*K* of each emotion list in the L_{TwF} on the basis of TwF value. We were free to set the value of *K*. If $w_{i,j}$ was in the emotion lexicons, then the corresponding emotion could be assigned to $emo(w_{i,j})$. For example, if $w_{i,j}$ was in the lists corresponding to the anger of the L_{fuse} lexicon, then a vector $(1, 0, 0, 0, 0, 0, 0)$ was assigned to $emo(w_{i,j})$. If not, then a zero vector is assigned to $emo(w_{i,j})$, because we thought that morphemes in the emotion lexicons had a considerable effect on the expression of the emotions of the sentence than morphemes not in the emotion lexicons. After the assignment, we added the emotion vector of all the elements of d'_i in order to get an emotion vector of d'_i (i.e., $emo(d'_i)$) as follows:

$$emo(d'_i) = \sum_{n=1}^{3} \sum_{j=1}^{|d'_i|-(n-1)} emo([w_{i,j}, ..., w_{i,j+n-1}]) \quad (2)$$

In the case of n-gram (n > 1), the range for computing the emotion vector could be different, where we computed an emotion vector $emo([w_{i,j}, ..., w_{i,j+n-1}])$ to get an emotion vector of d'_i as shown in Equation 2. In Algorithm 1, L can be any emotion lexicon such as only TwF or TT lexicons, but we uses L_{fuse} (i.e., a combination of the TwF and TT lexicons) in this work.

Algorithm 1 N-Gram-Based Distant Supervision

procedure NGRAM DISTANT SUPERVISION(d'_i, L)
  $emo(d'_i) \leftarrow \vec{0}$
  for $n \in [1, 2, 3]$ **do**
    while $j \neq |d'_i| - (n-1)$ **do**
      $ngram \leftarrow [w_{i,j}, ..., w_{i,j+n-1}]$
      if $ngram \in L$ **then**
        $emo(d'_i) \leftarrow emo(d'_i) + emo(ngram)$
      else
        $emo(d'_i) \leftarrow emo(d'_i) + \vec{0}$
      end if
      $j \leftarrow j + 1$
    end while
  end for
  return $emo(d'_i)$

3.3. Korean-Specific-Features-Based Distant Supervision

We annotated each unlabeled sentence of D' with seven types of emotions by using our N-gram-based distant supervision. According to our annotation procedure, we could capture not only an emotion contained in one morpheme, but an emotion of a combination of n-gram morphemes. However, it was difficult to consider more Korean-specific features (e.g., emoticons, symbols, and emotion letters), as emotion lexicons do not contain these features. Therefore, we proceeded with additional distant supervision using the Korean-specific features provided by KTEA (Do and Choi, 2015). In detail, let $EF = \{EF_1, EF_2, EF_3, EF_4, EF_5\}$ be a set of Korean-specific emotion features, where each element of EF represents a group of emoticons, symbols, Korean emotion letters, exclamations of surprise, and swear words as follows:

$$EF_k = \{EF_{k,1}, EF_{k,2}, ..., EF_{k,|EF_k|}\}, k = 1, ..., 5 \quad (3)$$

where $|EF_k|$ denotes the number of features in EF_k. To consider the Korean-specific features for labeling, we took

[5] https://www.korean.go.kr/front_eng/main.do

Algorithm 2 Korean-Specific-Features-Based Distant Supervision

procedure FT DISTANT SUPERVISION($d_i, emo(d'_i)$)
for $k \in [1, 2, 3, 4, 5]$ **do**
  $l \leftarrow 1$
  while $l \neq |EF_k|$ **do**
    if $EF_{k,l} \in d_i$ **then**
      $emo(d'_i) \leftarrow emo(d'_i) + emo(EF_{k,l})$
    end if
    $l \leftarrow l + 1$
  end while
end for
$label \leftarrow \arg\max emo(d'_i)$
if $|label| > 1$ **then**
  $idx \sim U(0, |label|)$
  $label_i \leftarrow idx$
else
  $label_i \leftarrow label$
end if
return $label_i$

an original sentence d_i and the emotion vector $emo(d'_i)$, which was calculated in Algorithm 1. If any feature of the group was in sentence d_i, then we added an emotion vector corresponding to this feature to $emo(d'_i)$. Then, we regarded the highest index of the final emotion vector $emo(d'_i)$ as the emotion label $label$ by executing the $\arg\max$ operation. If multiple indexes have the same maximum value, then we randomly sampled one emotion label by using a uniform distribution.

4. Data Resources

In this section, we introduce the data resources that we used for our annotation procedure.

4.1. Korean Twitter Emotion Analysis (KTEA) Dataset

As mentioned earlier, this dataset has various resources, such as an emotion-labeled dataset, an emotion lexicon, and Korean-specific features. The emotion-labeled dataset was constructed manually by using three annotators and contained 5,706 valid tweets labeled with seven types of emotions. For the n-gram-based distant supervision, we exploited a combination of TwF and TT. Moreover, we used Korean-specific emotion features for additional distant supervision. Moreover, we performed an experiment in which our method could annotate six emotions well, except the emotion *neutral*, using the KTEA dataset.

4.2. Naver Sentiment Movie Corpus

We took the Naver Sentiment Movie Corpus as our unlabeled dataset, called NSMC[6]. This dataset consists of reviews scraped from Naver Movies[7]. Moreover, they constructed the dataset according to the method described in (Maas et al., 2011). This dataset contains 150,000 sentences for training and 50,000 sentences for testing. The sentiment classes are balanced, and they exclude neutral reviews. We removed some sentences from the corpus for our annotation procedure. These removed sentences did not have content or lead to the errors of the morpheme analyzer. The statistics of NSMC are described in Table 1.

	NSMC	
	Training	Testing
Total	150,000	50,000
Sample	149,994	49,997

Table 1: Statistics of NSMC

5. Experiments

5.1. Experimental Setting

To evaluate whether an appropriate label is generated, we compared an emotion label produced by our annotation procedure with a ground-truth emotion label of KTEA dataset. Before the experiment, we used the KTEA dataset, which contains tweets that the three annotators all agreed on in terms of the emotions except neutral. We adopted the accuracy and the weighted f1 score as our metrics for each emotion. As mentioned earlier, to the best of our knowledge, this is the first work to construct a large-scale Korean emotion-labeled dataset by using distant supervision with emotion lexicons. However, we did not find an appropriate baseline. Thus, we chose the annotation procedure in (Agrawal et al., 2018) as our baseline. The results are shown in Table 2. For the experiments, we used a combination of the Twf and TT lexicons, because it achieved the best performance in (Do and Choi, 2015). Moreover, we fixed the value of *K* as 60, as this led to the best performance between 10 and 100, as shown in Figure 2.

5.2. Experimental Result

Table 2 shows that our method performed better than the baseline in terms of the overall accuracy and the weighted f1 score. This might be attributed to the fact that our method considered the characteristics of the Korean language by analyzing each sentence as morpheme units. In contrast, the baseline analyzed each sentence as word (i.e. *Eojeol*) units, because it used `word tokenize` from the `NLTK` toolkit. We compared the performance of each annotation procedure with increasing *n*-gram, applying post-processing, or removing specific morphemes or not. The average accuracy was affected by the *n-gram*, where the *3-gram* showed a higher performance of 72.7% than the *1-gram* and the *2-gram*, as we could consider the characteristics of Korean by using the n-gram morphemes. Moreover, we observed that the feature-based distant supervision captured emotions well because of the improvement of the average accuracy from 65.2% to 72.7%, in the same environment of *3-gram* and *w ex morp*. Moreover, the results showed that the removal of certain specific morphemes improved the accuracy from 70.4% to 72.7%. As a result, we concluded that our annotation procedure could produce an appropriate emotion-labeled dataset. Furthermore, we constructed a large-scale emotion-labeled dataset from a large amount of unlabeled data, by using our annotation procedure whose accuracy was 72.7%.

[6] `https://github.com/e9t/nsmc`
[7] `https://movie.naver.com/movie/point/af/list.nhn`

	Korean Twitter Emotion Analysis Dataset							
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.	Weighted f1
Base	43.6	56.9	34.3	50.0	47.9	48.0	48.0	49.9
Base + *ft*	57.9	52.3	31.4	55.8	60.5	54.7	56.6	57.9
Our: *K*-60 + *1-gram* (w ex morp) + *ft*	66.9	73.8	65.7	73.3	71.3	53.3	69.7	70.0
Our: *K*-60 + *2-gram* (w ex morp) + *ft*	66.9	72.3	62.9	74.0	71.6	62.7	70.6	71.1
Our: *K*-60 + *3-gram* (w ex morp) + *ft*	68.4	75.4	65.7	74.8	74.9	64.0	**72.7**	**73.0**
Our: *K*-60 + *3-gram* (w ex morp)	59.4	78.5	80.0	71.3	63.8	42.7	65.2	65.7
Our: *K*-60 + *3-gram* (w/o ex morp) + *ft*	69.9	73.8	51.4	69.4	72.8	70.7	70.4	70.7

Table 2: Accuracy(%) per emotion on Korean Twitter Emotion Analysis dataset. ***ft*** stands for the Korean-specific-feature-based distant supervision in our procedure. **ex morp** stands for removing specific morphemes in the n-gram based distant supervision.

	Korean Movie Review Emotion Dataset						
	# of sentence	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Training	119,995	29.88	9.84	8.42	20.36	23.93	7.57
Development	29,999	29.86	9.6	8.44	20.35	24.17	7.58
Testing	49,997	29.82	9.93	8.32	20.35	24.0	7.58

Table 3: Emotion distributions(%) of Korean Movie Review Emotion dataset

6. Data Construction

In this section, we introduced the information of our constructed dataset and performed an emotion classification task. Moreover, we measured the inter-annotation agreement.

6.1. Korean Movie Review Emotion (KMRE) Dataset

We constructed a Korean Movie Review Emotion (KMRE) dataset annotated with six types of emotions in the NSMC dataset by following our annotation procedure, which achieved the best performance. Here, we excluded the emotion *neutral* because there were no neutral reviews in the NSMC. After the construction, we split the training dataset into 60% for training, 15% for development, and 25% for testing. We found that specific emotions (i.e., anger, happiness, and sadness) were more frequently annotated in the KMRE dataset. Furthermore, the training, development, and testing of the KMRE dataset tended to have similar emotion label distributions. Table 3 shows the emotion label distribution of the KMRE dataset.

6.2. Emotion Classification on KMRE

We performed the emotion classification task on our KMRE dataset with the GRU model and the bidirectional GRU (Bi-GRU) model. We implemented these models using Tensorflow 2.0[8]. The word embedding size was set to 300. The vocabulary size was set to 32,771, and all the OOV (out-of-vocabulary) tokens were mapped to a special token $< unk >$. Each model had 1-layer-GRU with 300 hidden units. Moreover, we set the batch size as 128. We used the Adam optimizer with a fixed learning rate of 0.002. We stopped training each model when the validation loss failed to improve compared with the best validation loss for five epochs. Furthermore, we applied 30% dropout to prevent the overfitting of our model. In addition, we removed sentences whose length was less than 20 on the KMRE dataset. Table 4 shows the results. Overall, the Bi-GRU model achieved slightly better performance than the GRU model. Overall, the accuracy was high for certain emotions, such as anger, happiness, and sadness. This might be attributed to the fact that these emotions appeared mostly in our KMRE dataset, as shown in Table 3. We expected that the performance could be increased, if we used the subword-level (e.g., *syllable* or *jamo*) word vector representations for Korean.

6.3. Human Evaluation

We conducted a human evaluation to measure the quality of the generated emotion-labeled dataset. More specifically, first, we randomly sampled 100 sentences per emotion (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) from the test set. Then, we presented these sampled sentences to two human annotators. Finally, we asked two annotators to score whether the labeled emotion was appropriate for each sentence on a rating scale of 0 (Disagree), 1 (Neither agree nor disagree), and 2 (Agree). The reason why we use this rating scale is that each sentence may have different emotions depending on the context or situations. Thus, we use this rating scale rather than a binary agreement. Next, we calculated Cohen's kappa (Cohen, 1960) to measure the inter-annotator agreement. The Cohen's kappa score for the labels of our dataset was 0.560, indicating "Moderate Agreement".

7. Discussion

We constructed the KMRE dataset by annotating the NSMC dataset with emotions, using Korean emotion lexicons and Korean-specific emotion features, provided by KTEA. The domains of NSMC and KTEA were different: one was movie reviews, and the other was Twitter. Therefore, we were concerned that the domain inconsistency might degrade the quality of the KMRE dataset. However,

[8] https://github.com/tensorflow/tensorflow

	Korean Movie Review Emotion Dataset							
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.	Weighted f1
GRU	69.9	45.5	45.7	66.2	61.5	40.2	60.4	60.2
Bi-GRU	72.7	41.7	54.5	66.5	58.9	43.3	61.3	61.1

Table 4: Accuracy(%) per emotion on Korean Movie Review Emotion dataset

the quality did not degrade considerable, as observed in the experiments. Had the domains been similar, we would have expected a better quality emotion-labeled dataset.

As mentioned earlier, we empirically removed certain morphemes (i.e., EP, EF, EC, ETN, ETM, XPN, XSN, XSV, and XSA) from the formal morphemes. In fact, the formal morphemes consisted of two categories (i.e., Particles and Markers without the XR tag), as specified in Table 6. In addition to the formal morphemes, we considered the symbol morphemes. To understand which category in the formal morphemes was important for annotating emotions, we examined the ablation study by considering all of the combinations. Table 5 shows the results of the ablation study conducted on *K*-60 and, *3*-gram, and with the Korean-specific-feature-based distant supervision. From this ablation study, we found that the removal of only Marker's morphemes was effective for annotating emotions in Korean. Moreover, there were some meaningful morphemes in Particles because the removal of the Particle's morphemes lead to a lower performance than that observed otherwise. Additionally, the Symbol's morphemes had no significant effect on the annotation of emotions, as the performances between with and without the Symbol's morphemes were the same or slightly lower. However, the removal of all of the formal morphemes and symbols led to a slight degradation of the performance. Figure 2 illustrate the results of the ablation study.

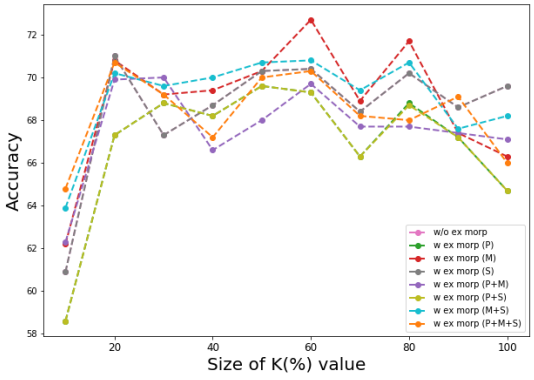

Figure 2: Average accuracy depending on the *K*(%) value

8. Conclusion and Future Work

In this paper, we presented a Korean-specific emotion annotation procedure using n-gram-based distant supervision and Korean-specific-feature-based distant supervision. In the experiments, we showed that our annotation procedure could generate appropriate emotion labels. Moreover, we constructed the Korean Emotion Movie Review (KMRE) dataset that contains six types of emotions, using the annotation procedure; it exhibited the best performance in the experiments.

In our future work, we will study a more advanced annotation procedure that can capture the contextual information of each sentence for more precise emotions. KMRE is publicly available at `https://github.com/passing2961/KMRE`. We hope that our dataset will be used for various emotion-related tasks. Furthermore, we hope to help researchers construct emotion-labeled datasets by using our annotation procedure for the Korean language.

9. Acknowledgements

This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services). This research was supported by Korea Electric Power Corporation. (Grant number:R18XA05)

10. References

Abdul-Mageed, M. and Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 718–728.

Agrawal, A., An, A., and Papagelis, M. (2018). Learning emotion-enriched word representations. In Proceedings of the 27th International Conference on Computational Linguistics, pages 950–961.

Choi, J. D. and Palmer, M. (2011). Statistical dependency parsing in korean: From corpus generation to automatic parsing. In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, pages 1–11. Association for Computational Linguistics.

Choi, S., Kim, T., Seol, J., and Lee, S.-g. (2017). A syllable-based technique for word embeddings of korean words. *arXiv preprint arXiv:1708.01766*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., Luca, V. D., and Jaggi, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In Proceedings of the 10th international workshop on semantic evaluation, number CONF, pages 1124–1128.

Do, H. J. and Choi, H.-J. (2015). Korean twitter emotion classification using automatically built emotion lexicons and fine-grained features. In Proceedings of the 29th

	Korean Twitter Emotion Analysis Dataset							
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.	Weighted f1
w/o ex morp	69.9	73.8	51.4	69.4	72.8	70.7	70.4	70.7
w ex morp (P)	60.2	69.2	42.9	68.6	75.7	72.0	69.3	69.4
w ex morp (M)	68.4	75.4	65.7	74.8	74.9	64.0	**72.7**	**73.0**
w ex morp (S)	69.9	73.8	51.4	69.4	72.8	70.7	70.4	70.7
w ex morp (P+M)	66.2	73.8	57.1	69.8	75.4	52.0	69.7	69.8
w ex morp (P+S)	60.2	69.2	42.9	68.6	75.7	72.0	69.3	69.4
w ex morp (M+S)	70.7	72.3	60.0	73.6	72.5	57.3	70.8	71.2
w ex morp (P+M+S)	66.9	75.4	48.6	67.4	79.3	52.0	70.3	70.3

Table 5: Performance on Korean Twitter Emotion Analysis dataset according to certain morphemes. **ex morp** stands for removing specific morphemes in n-gram-based distant supervision. (P: Particles, M: Markers, S: Symbols)

Pacific Asia Conference on Language, Information and Computation: Posters, pages 142–150.

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., and Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009.

Jang, H., Kim, M., and Shin, H. (2013). Kosac: A full-fledged korean sentiment analysis corpus. In Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), pages 366–373.

Klinger, R. et al. (2018). An analysis of annotated corpora for emotion classification in text. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2104–2119.

Lapitan, F. R., Batista-Navarro, R. T., and Albacea, E. (2016). Crowdsourcing-based annotation of emotions in filipino and english tweets. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), pages 74–82.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, pages 142–150. Association for Computational Linguistics.

Milnea, D., Parisb, C., Christensenc, H., Batterhamc, P., and O'Deac, B. (2015). We feel: Taking the emotional pulse of the world. In Proceedings 19th Triennial Congress of the IEA, volume 9, page 14.

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. (2012). # emotional tweets. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 246–255. Association for Computational Linguistics.

Park, S., Byun, J., Baek, S., Cho, Y., and Oh, A. (2018). Subword-level word vector representations for korean. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2429–2438.

Shin, H., Kim, M., Jang, H., and Cattle, A. (2012). Annotation scheme for constructing sentiment corpus in korean. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pages 181–190.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1555–1565.

Appendix: ETRI POS Tagset

Table 6 displays the ETRI POS tags including brief descriptions and whether it is the actual or formal morphemes. For the description of each POS tag, we referred to (Choi and Palmer, 2011).

Category	ETRI POS Tags	Description	Actual/Formal
Nouns	NNG	General noun	A
	NNP	Proper noun	A
	NNB	Bound noun	A
	NP	Pronoun	A
	NR	Numeral	A
Verbs	VV	Verb	A
	VA	Adjective	A
	VX	Auxiliary predicate	A
	VCP	Copula	A
	VCN	Negation adjective	A
Modifiers	MMA, MMD, MMN	Determiner	A
	MAG	General adverb	A
	MAJ	Conjunctive adverb	A
Interjection	IC	Interjection	A
Particles	JKS	Subjective case particle	F
	JKC	Complemental case particle	F
	JKG	Adnomial case particle	F
	JKO	Objective case particle	F
	JKB	Adverbial case particle	F
	JKV	Vocative case particle	F
	JKQ	Quotative case particle	F
	JX	Auxiliary particle	F
	JC	Conjunctive particle	F
Markers	EP	Prefinal ending marker	F
	EF	Final ending marker	F
	EC	Conjunctive ending marker	F
	ETN	Nominalizing ending marker	F
	ETM	Adnominalizing ending marker	F
	XPN	Noun prefix	F
	XSN	Noun derivational suffix	F
	XSV	Verb derivational suffix	F
	XSA	Adjective derivational suffix	F
	XR	Base morpheme	A
Symbols	SF, SP, SS, SE, SO	Punctuation marks	-
	SW	Special word	-
	SL	Foreign word	-
	SH	Chinese word	-
	SN	Number	-
	NA	Unknown word	-

Table 6: POS tags in the ETRI morpheme analyzer based on the Sejong POS tag sets (A: Actual morphemes, F: Formal morphemes)