Subword-level Word Vector Representations for Korean

Sungjoon Park[1], Jeongmin Byun[1], Sion Baek[2], Yongseok Cho[3], Alice Oh[1]

Department of Computing, KAIST[1] , Program in Cognitive Science, Seoul National University[2], Natural Language Processing team, Adecco[3]
{sungjoon.park, jmbyun}@kaist.ac.kr, sioning1122@snu.ac.kr, yongseok.cho84@gmail.com, alice.oh@kaist.edu

Introduction

Background

- Research on distributed word representations is focused on widely-used languages such as English. Although the same methods can be used for other languages, language-specific knowledge can enhance the accuracy and richness of word vector representations.
- Despite their effectiveness in capturing syntactic features from subword features of diverse languages, decomposing a word into a set of n-grams and learning n-gram vectors does not consider the unique linguistic structures of various languages.

Contribution

- Our first contribution is the method to decompose the words into both character-level units and *jamo*-level units and train the subword vectors through the Skip-Gram model.
- Our second major contribution is the Korean evaluation datasets for word similarity and analogy tasks, a translation of the WS-353 with annotations by 14 Korean native speakers, and 10,000 items for semantic and syntactic analogies, developed with Korean linguistic expertise.
- Using these datasets, we show that our model improves performance over other baseline methods without relying on external resources for word decomposition.

Experiments

Dataset

	# of words	# of sentences	# of unique words
Wikipedia	43.4M	3.3M	299,528
Online News	47.1M	3.2M	282,955
Sejong Corpus	31.4M	2.2M	231,332
Total	121.9M	8.8M	638,708

- We aggregate three sources to make the corpus containing 0.12 billion word tokens with 0.6 million unique words.
- Our model and all of the comparison models for training word vectors are trained over the collected corpus.

Evaluation Tasks

1) Word Similarity & Analogy

- We develop the evaluation datasets.
- Similarity: Spearman's correlation coefficient between the human judgment and model's cosine similarity for the similarity of word pairs are reported.
- Analogy: Rank-based measures may not be an appropriate measure since the total number of unique n-grams/words over the same corpus largely differ from each other. For fair comparison, cosine distances between the vector a+b−c and d of each categories are reported.

2) Sentiment Analysis

- Given a sequence of words, a trained classifier should predict the binary sentiment from the inputs while maintaining the input word vectors fixed.
- Based on part of the Naver Sentiment Corpus, single layer RNN is trained as a classifier for the task.

Subword-level Word Vectors for Korean

Decomposition of Korean Words

- Decompose a word to *jamo* sequence
  *Jamo*s have names that reflect the position in a character:
  1) *chosung* (syllable onset), 2) *joongsung* (syllable nucleus), 3) *jongsung* (syllable coda)
- Add empty jongsung symbol e such that a character always has 3 (jamos)
- Add start/end symbol < / > in the sequence

먹었다(ate) → <, ㅁ, ㅓ, ㄱ, ㅇ, ㅓ, ㅆ, ㄷ, ㅏ, e , >

Extracting n-grams for jamo sequence

- Character-level n-grams, Gct (ㅁ, ㅓ, ㄱ) , (ㅇ, ㅓ, ㅆ) , (ㄷ, ㅏ, e) , (ㅇ, ㅓ, ㅆ, ㄷ, ㅏ, e) ...
- Inter-character *jamo*-level n-grams, Gjt (<, ㅁ, ㅓ) , (ㄱ , ㅇ, ㅓ) , (ㅆ , ㄷ, ㅏ) , (ㅏ, e , >) ...

Subword Information Skip-Gram (SISG, a.k.a FastText)

- Constructing word vector from subword vectors : $z_t = \frac{1}{|G_{ct} + G_{jt}|}(\sum_{g_{ct} \in G_{ct}}^{|G_{ct}|} z_{g_{ct}} + \sum_{g_{jt} \in G_{jt}}^{|G_{jt}|} z_{g_{jt}})$
- SISG Binary Logistic Loss : $\ell = \log\left(1 + e^{-s(w_t, w_{t+j})}\right) + \sum_{n=1}^{n_c} \log(1 + e^{s(w_t, w_{t+j})})$
- Scoring Function : $s(w_t, w_{t+j}) = \frac{1}{|G_{ct} + G_{jt}|}(\sum_{g_{ct} \in G_{ct}}^{|G_{ct}|} z_{g_{ct}}^T v_{t+j} + \sum_{g_{jt} \in G_{jt}}^{|G_{jt}|} z_{g_{jt}}^T v_{t+j})$

Developing Evaluations Sets

Word Similarity (WS-353) for Korean

- 2 native speakers translated the original item pairs.
- 14 other native speakers annotated similarity scores of the pairs.
- Correlation between the original scores and the annotated scores of the translated pairs is 0.82.

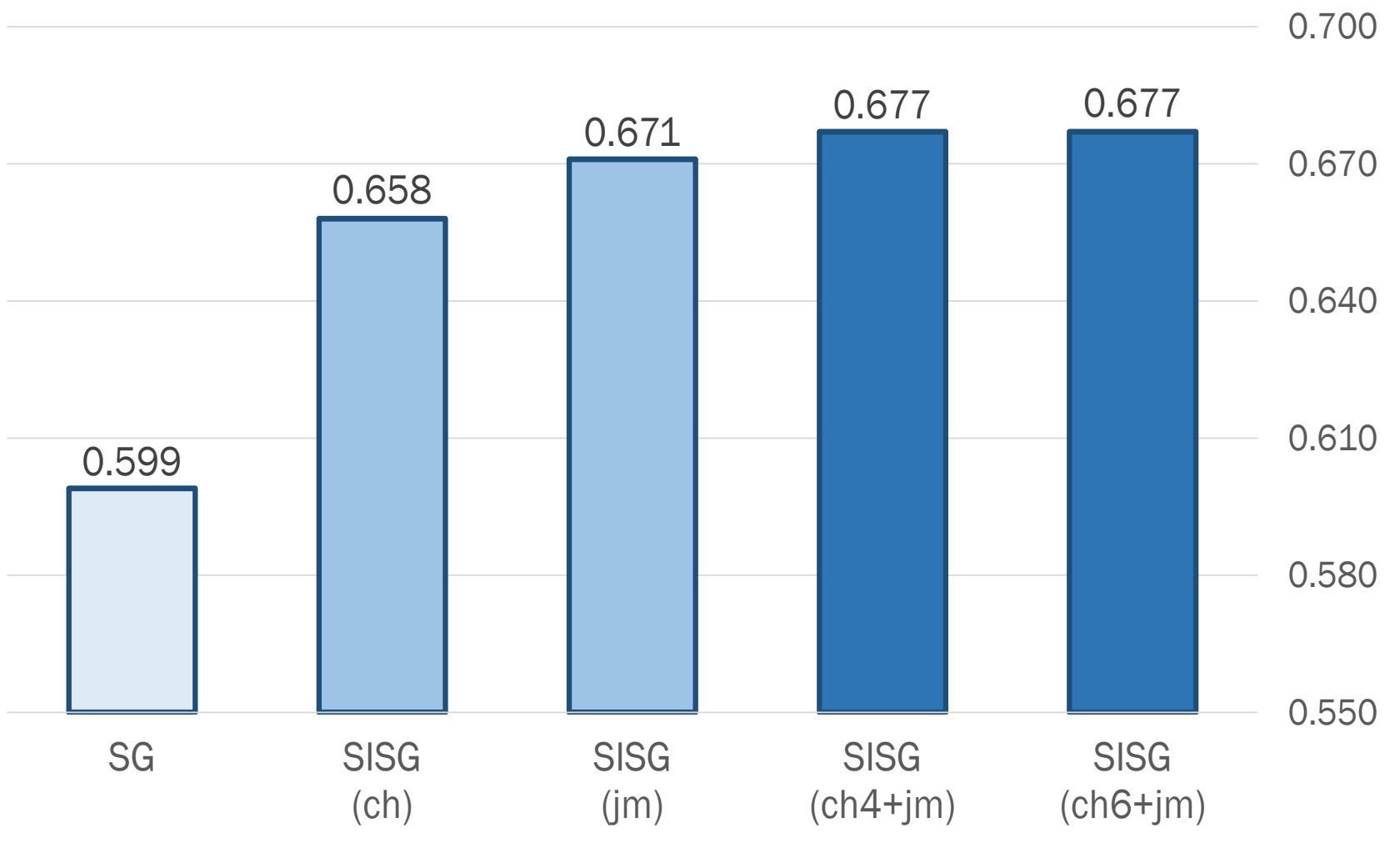
Word Analogy for Korean

- Semantic Features (5,000 items)
  - ○ *Capital-Country* : 아테네Athens : 그리스Greece = 바그다드Baghdad : 이라크Iraq
  - ○ *Male-Female* : 왕자prince:공주princess = 신사gentlemen:숙녀ladies
  - ○ *Name-Nationality* : 간디Gandhi : 인도India = 링컨Lincoln : 미국USA
  - ○ *Country-Language* : 아르헨티나Argentina : 스페인어Spanish = 미국USA : 영어English
  - ○ *Miscellaneous* : 개구리Frog : 올챙이tadpole = 말horse : 망아지pony
- Syntactic Features (5,000 items)
  - ○ *Case* : 교수Professor : 교수가Professor+case가 = 축구soccer : 축구가soccer+case가
  - ○ *Tense* : 싸우다fight : 싸웠다fought = 오다come : 왔다came
  - ○ *Voice* : 팔았다sold : 팔렸다be sold = 평가했다evaluated : 평가됐다was evaluated
  - ○ *Verb* : 가다go : 가고go+form고 = 쓰다write : 쓰고write+form고
  - ○ *Honorific* :도왔다helped : 도우셨다helped+honorific시 = 됐다done : 되셨다done+honorific시

* Publicly available at : https://github.com/SungjoonPark/KoreanWordVectors

Results

Word Similarity



- Decomposing words into *jamo*-level is helpful to learn good Korean word vectors.
- Spearman correlation with human evaluation is improved to 0.677.
  (1-4 characters n-grams / 3-5 jamo n-grams included)

Word Analogy

	Semantic					Syntactic				
	Capt	Gend	Name	Lang	Misc	Case	Tense	Voice	Form	Honr
SG	0.460	0.551	**0.537**	0.435	0.574	0.521	0.597	0.594	0.685	0.634
SISG(ch)	0.469	0.584	0.608	0.439	0.614	0.422	0.559	0.550	0.656	0.489
SISG(jm)	0.442	0.515	0.574	0.362	0.565	0.228	0.421	0.434	0.537	0.367
SISG(ch4+jm)	0.431	0.504	0.570	0.361	0.556	0.212	0.415	0.434	**0.501**	**0.364**
SISG(ch6+jm)	**0.425**	**0.498**	0.561	**0.354**	**0.554**	**0.210**	**0.414**	**0.426**	0.507	0.367

- Overall, decomposing words help to capture semantic/syntactic features.

Sentiment Analysis

	Acc.(%)	Prec.	Rec.	F1
SG	76.15	0.746	0.792	0.768
SISG(ch)	76.26	0.774	0.741	0.757
SISG(jm)	76.53	**0.790**	0.722	0.754
SISG(ch4+jm)	76.28	0.755	0.776	0.765
SISG(ch6+jm)	**76.54**	0.750	**0.795**	**0.772**

- Decomposing a word to 1-6 character n-grams and 3-5 jamo n-grams show slightly higher performance over comparable models.

Effect of n in n-grams

		# of chars			
		4	5	6	all
# of *jamos*	2-4	0.660	0.655	0.659	0.651
	3-4	0.660	0.650	0.652	0.660
	3-5	**0.677**	0.672	**0.677**	0.675
	3-6	0.665	0.663	0.664	0.669

- *n* of jamo-level *n*-grams, including n=5,6 of *n*-grams and excluding bigrams show higher performance.
- Including all of the character *n*-grams while decomposing a word does not guarantee performance improvement.

Conclusion and Discussion

- We demonstrated the effectiveness of the jamo- and character-level Korean word vectors in capturing the semantic and syntactic information by evaluating these vectors with newly developed word similarity and word analogy tasks.
- We plan to apply these vectors for various neural network based NLP models, and apply the same idea to other syntactic tasks such as POS tagging and parsing.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-01778, Development of Explainable Human-level Deep Machine Learning Inference Framework).