### 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
- **Wikipedia**: 43.4M words, 4.3M sentences, 299,528 unique words.
- **Online News**: 47.1M words, 282,955 unique words.
- **Sejong Corpus**: 31.4M words, 4.2M sentences, 231,332 unique words.
- **Total**: 121.9M unique words.

- 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 of word vectors is 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
- **Jamos** have names that reflect the position in a character: 1) chosung (syllabic onset), 2) jaosung (syllabic nucleus), 3) jongsung (syllabic coda)
- Add empty jongsung symbol e such that a character always has 3 jamos.
- Add start/end symbol / in the sequence.

#### Extracting n-grams for jamo sequence
- Character-level n-grams, $G_C$: (c, l), (c, l, m), (c, l, m, e, l, e, >)
- Inter-character jamo-level n-grams, $G_J$: (c, l, m), (c, l, m, e, l, e, >)

#### Subword Information Skip-Gram (SISG, a.k.a FastText)
- Constructing word vector from subword vectors: \[ \sum_{n=1}^{5} \frac{1}{\text{score}(v_{n,k,v_n,v_{n+1}})} \]
- SISG Binary Logistic Loss: \[ \log(1 + \exp(-z_n)) \]
- Scoring Function: \[ \frac{\text{score}(v_{n,k,v_n,v_{n+1}})}{\text{score}(v_{n,k,v_n,v_{n+1}}) + \text{score}(v_{n,k,v_n,v_{n+1}})} \]

### 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 Anomaly for Korean
- **Semantic Features (5,000 items)**
  - Capital-Country: 영국[UK], 러시아[Russia], 미국[USA]
  - Male-Female: 왕자[prince], 공주[princess], 신사[gentleman], 속녀[ladies]
  - Name-Nationality: 간디[Gandhi], 안도지안[andiozian], 잭슨[Lincoln], 미국[USA]
  - Country-Language: 아르헨티나[Argentina], 스페인[Spanish] = 영어[English]
  - Miscellaneous: 개구리[Frog], 물복[adapole], 말[horse], 말[horse]

- **Syntactic Features (5,000 items)**
  - Case: 고수[Professor], 교수[Professor] + case가[case가], 졌저[sooker], 졌저[sooker] + case가[case가]
  - Tenure: 쌓들[Chief], 쌓들[Chief] + 도거[come] + 도거[come] + 도거[come]
  - Voice: 쌓들[Chief], 쌓들[Chief] + 도거[come], 쌓들[Chief] + 도거[come], 쌓들[Chief] + 도거[come]

- Publicly available at: [https://github.com/SungjoonPark/KoreanWordVectors](https://github.com/SungjoonPark/KoreanWordVectors)

### Results

#### Word Similarity

| Semantic | Capt | Gend | Name | Lang | Misc | Syntactic | Case | Tense | Voice | Form | Hon
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>0.460</td>
<td>0.551</td>
<td>0.537</td>
<td>0.435</td>
<td>0.574</td>
<td>0.521</td>
<td>0.597</td>
<td>0.594</td>
<td>0.685</td>
<td>0.634</td>
<td></td>
</tr>
<tr>
<td>SISG(ch)</td>
<td>0.469</td>
<td>0.554</td>
<td>0.554</td>
<td>0.439</td>
<td>0.614</td>
<td>0.422</td>
<td>0.559</td>
<td>0.569</td>
<td>0.537</td>
<td>0.568</td>
<td></td>
</tr>
<tr>
<td>SISG(1m)</td>
<td>0.424</td>
<td>0.515</td>
<td>0.574</td>
<td>0.362</td>
<td>0.565</td>
<td>0.231</td>
<td>0.523</td>
<td>0.434</td>
<td>0.537</td>
<td>0.367</td>
<td></td>
</tr>
<tr>
<td>SISG(4+4m)</td>
<td>0.431</td>
<td>0.504</td>
<td>0.570</td>
<td>0.361</td>
<td>0.556</td>
<td>0.212</td>
<td>0.415</td>
<td>0.504</td>
<td>0.501</td>
<td>0.364</td>
<td></td>
</tr>
<tr>
<td>SISG(1m+4m)</td>
<td>0.425</td>
<td>0.509</td>
<td>0.573</td>
<td>0.394</td>
<td>0.544</td>
<td>0.210</td>
<td>0.414</td>
<td>0.426</td>
<td>0.507</td>
<td>0.367</td>
<td></td>
</tr>
</tbody>
</table>

#### Sentence Analysis

<table>
<thead>
<tr>
<th>Acc. (%)</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>76.15</td>
<td>0.734</td>
<td>76.02</td>
</tr>
<tr>
<td>SISG(ch)</td>
<td>76.26</td>
<td>0.744</td>
<td>74.74</td>
</tr>
<tr>
<td>SISG(1m)</td>
<td>76.53</td>
<td>0.790</td>
<td>77.01</td>
</tr>
<tr>
<td>SISG(4+4m)</td>
<td>76.28</td>
<td>0.755</td>
<td>77.76</td>
</tr>
<tr>
<td>SISG(1m+4m)</td>
<td>76.54</td>
<td>0.795</td>
<td>79.75</td>
</tr>
</tbody>
</table>

#### Effect of n in n-grams

<table>
<thead>
<tr>
<th># of chars</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td># of jamos</td>
<td>2-4</td>
<td>0.960</td>
<td>0.655</td>
<td>0.659</td>
</tr>
<tr>
<td>3-4</td>
<td>0.960</td>
<td>0.655</td>
<td>0.652</td>
<td>0.660</td>
</tr>
<tr>
<td>3-6</td>
<td>0.677</td>
<td>0.672</td>
<td>0.677</td>
<td>0.675</td>
</tr>
<tr>
<td>3-6</td>
<td>0.960</td>
<td>0.663</td>
<td>0.694</td>
<td>0.669</td>
</tr>
</tbody>
</table>

### 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.