Watch Drama, Learn Language

Liling Tan
Rakuten Institute of Technology (Singapore)
10 Jan 2018 @ Global WordNet Conference
Overview

• Demos

• Generate Quiz Automatically

• Numbers and Graphs

• WordNet and Language Learning
Demos
Drama Vocab Quiz
Watch dramas, learn languages

Learn Korean and Chinese words while watching your favorite dramas on Rakuten Viki!

How many drama words do you already know? Take the quiz to find out!

KOREAN QUIZ  CHINESE QUIZ
전화
jeon-hwa

anticipation

signal

telephone

I don't know
낭만

nang-man

logical

romance

soldier

I don't know >
Your results

Hey,

Heo! (Omg!) You know about 11% of the common drama words. It’s a start, but you need to try harder to impress Jisung.

SEE DETAILED RESULTS

Share your result with friends. See who scores better!
Results Breakdown

30  7  17
Total  Correct  Wrong

Out of the 30 words in the quiz, there were 10 easy, 10 medium, and 10 hard words. You answered 2 easy words, 2 medium words and 3 hard words correctly, and 17 words incorrectly. Based on your answers, we estimate that you know about 1% of the common drama words.

Please note that getting answers wrong decreases your score. If you are unsure of the answer, you can select the "I don't know" option - this will not impact your score negatively.

You think you can do better? Try again! Every time you take our quiz, you see a new set of words.

Curious to know the science behind your score? Find out [here](#).
Watch and Learn

Here are some of the words you missed. With Learn mode, you can watch dramas and learn Korean at the same time. Check it out!

유서 yu-seo History
This word appears in
EP 1: The Legend of the Blue Sea

낭만 nang-man Romance
This word appears in
EP 9: Healer
Generate Quiz Automatically
자격증

Select a list of words and translation

Certificate

Document

Church

I don’t know

Select distractors
How to generate quiz automatically?

- **Select a list of words**
  - Selected sample needs to be representative of the corpus

- **Find the translation**
  - Translation has to be accurate and most prototypical / least ambiguous

- **Generate distractors**
  - Distractors MUST NOT have the same meaning as the answer
  - Distractors should not be too easy to eliminate
Select a list of words

1. **Create a Corpus**: Collect all captions from Korean dramas
2. **Tokenize**: Extract lemmas from surface words.

<table>
<thead>
<tr>
<th>미남</th>
<th>- 이(다)</th>
<th>- (으)시-</th>
<th>- (이)라구</th>
<th>- 요</th>
</tr>
</thead>
<tbody>
<tr>
<td>minam</td>
<td>- is</td>
<td>- si-</td>
<td>- lagu</td>
<td>- yo</td>
</tr>
<tr>
<td>handsome guy</td>
<td>(he) is</td>
<td>(honorific)</td>
<td>Someone said</td>
<td>(polite marker)</td>
</tr>
</tbody>
</table>

- The word 미남 [minam] “handsome guy” can be followed by multiple suffixes at once -이시라구요 [-issilaguyo] to form a single word meaning “Someone said that he is handsome”.

- Extract the root word 미남 [minam], and count it as a unique word type
Select a list of words

1. Create a Corpus
2. Tokenize
3. Filter Rare Words
   - Term Frequency > 20
   - Drama Frequency > 5

<table>
<thead>
<tr>
<th>Word</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>송전</td>
<td>power transmission</td>
</tr>
<tr>
<td>림프절</td>
<td>lymph nodes</td>
</tr>
<tr>
<td>감가상각</td>
<td>depreciation</td>
</tr>
<tr>
<td>달맞이꽃</td>
<td>evening primrose</td>
</tr>
<tr>
<td>보리밭</td>
<td>barley field</td>
</tr>
<tr>
<td>열화학</td>
<td>thermal chemistry</td>
</tr>
<tr>
<td>식기세척기</td>
<td>dishwasher</td>
</tr>
<tr>
<td>테라포밍</td>
<td>terraforming</td>
</tr>
<tr>
<td>삼각함수</td>
<td>trigonometric function</td>
</tr>
<tr>
<td>지방산</td>
<td>fatty acid</td>
</tr>
</tbody>
</table>
Select a list of words

1. Create a Corpus
2. Tokenize
3. Filter Rare Words
4. Translate
   - Find translations from **in-house** dictionary and MT engines.
   - **Dictionaries**: More info on slides 12 and 15-38
     [https://www.slideshare.net/rakutentech/ai-based-language-learning-tools](https://www.slideshare.net/rakutentech/ai-based-language-learning-tools)
   - **Machine Translation**: Come talk to me =)
Select a list of words

1. Create a Corpus
2. Tokenize
3. Filter Rare Words
4. Translate
5. Filter More Words
   • Konglish words are too easy to guess
   • Swear words are inappropriate
   • https://github.com/alvations/expletives
<table>
<thead>
<tr>
<th>Korean</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>그린</td>
<td>green</td>
</tr>
<tr>
<td>배드민턴</td>
<td>badminton</td>
</tr>
<tr>
<td>스트라이크</td>
<td>strike</td>
</tr>
<tr>
<td>홍혜정</td>
<td>hong hye jung</td>
</tr>
<tr>
<td>햄버거</td>
<td>hamburger</td>
</tr>
<tr>
<td>보디가드</td>
<td>bodyguard</td>
</tr>
<tr>
<td>김태평</td>
<td>kim tae pyung</td>
</tr>
<tr>
<td>설렁탕</td>
<td>seolleongtang</td>
</tr>
<tr>
<td>카피라이터</td>
<td>copywriter</td>
</tr>
<tr>
<td>패션모델</td>
<td>fashion model</td>
</tr>
</tbody>
</table>
## Filter Konglish Words

<table>
<thead>
<tr>
<th>Korean</th>
<th>Romanization</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>그린</td>
<td>geu-rin</td>
<td>green</td>
</tr>
<tr>
<td>배드민턴</td>
<td>bae-deu-min-teon</td>
<td>badminton</td>
</tr>
<tr>
<td>스트라이크</td>
<td>seu-teu-ra-i-keu</td>
<td>strike</td>
</tr>
<tr>
<td>홍혜정</td>
<td>hong-hye-jeong</td>
<td>hong hye jung</td>
</tr>
<tr>
<td>햄버거</td>
<td>haem-beo-geo</td>
<td>hamburger</td>
</tr>
<tr>
<td>보디가드</td>
<td>bo-di-ga-deu</td>
<td>bodyguard</td>
</tr>
<tr>
<td>김태평</td>
<td>gim-tae-pyeong</td>
<td>kim tae pyung</td>
</tr>
<tr>
<td>설렁탕</td>
<td>seor-reong-tang</td>
<td>seolleongtang</td>
</tr>
<tr>
<td>카피라이터</td>
<td>ka-pi-ra-i-teo</td>
<td>copywriter</td>
</tr>
<tr>
<td>패션모델</td>
<td>pae-syeon-moder</td>
<td>fashion model</td>
</tr>
<tr>
<td>Korean</td>
<td>Romanization</td>
<td>Romanization (No vowel)</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>그린</td>
<td>geu-rin</td>
<td>grn</td>
</tr>
<tr>
<td>배드민턴</td>
<td>bae-deu-min-teon</td>
<td>bdmntn</td>
</tr>
<tr>
<td>스트라이크</td>
<td>seu-teu-ra-i-keu</td>
<td>strk</td>
</tr>
<tr>
<td>홍혜정</td>
<td>hong-hye-jeong</td>
<td>hnghyjng</td>
</tr>
<tr>
<td>햄버거</td>
<td>haem-beo-geo</td>
<td>hmbg</td>
</tr>
<tr>
<td>보디가드</td>
<td>bo-di-ga-deu</td>
<td>bdgd</td>
</tr>
<tr>
<td>김태평</td>
<td>gim-tae-pyeong</td>
<td>gmtpyng</td>
</tr>
<tr>
<td>설렁탕</td>
<td>seor-reong-tang</td>
<td>srrngtng</td>
</tr>
<tr>
<td>카피라이터</td>
<td>ka-pi-ra-i-teo</td>
<td>kprt</td>
</tr>
<tr>
<td>패션모델</td>
<td>pae-syeon-moder</td>
<td>psynmldr</td>
</tr>
<tr>
<td>Korean</td>
<td>Romanization</td>
<td>Romanization (No vowel)</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>그린</td>
<td>geu-rin</td>
<td>grn</td>
</tr>
<tr>
<td>배드민턴</td>
<td>bae-deu-min-teon</td>
<td>bdmntn</td>
</tr>
<tr>
<td>스트라이크</td>
<td>seu-teu-ra-i-keu</td>
<td>strk</td>
</tr>
<tr>
<td>홍혜정</td>
<td>hong-hye-jeong</td>
<td>hnghyjng</td>
</tr>
<tr>
<td>햄버거</td>
<td>haem-beo-geo</td>
<td>hmbg</td>
</tr>
<tr>
<td>보디가드</td>
<td>bo-di-ga-deu</td>
<td>bdgd</td>
</tr>
<tr>
<td>김태평</td>
<td>gim-tae-pyeong</td>
<td>gmtpyng</td>
</tr>
<tr>
<td>설렁탕</td>
<td>seor-reong-tang</td>
<td>srrngtng</td>
</tr>
<tr>
<td>카피라이터</td>
<td>ka-pi-ra-i-teo</td>
<td>kprt</td>
</tr>
<tr>
<td>패션모델</td>
<td>pae-syeon-mo-der</td>
<td>psynmdr</td>
</tr>
</tbody>
</table>
Generate the Distractors

• Create a Word2Vec model on the source language

• For each word:

  • **Distractor 1:** Select the top 5\textsuperscript{th} to 20\textsuperscript{th} closest words (cosine)

  • **Distractor 2:** Use Distractor 1 as negative and question word as positive, select 1\textsuperscript{st} to 20\textsuperscript{th} closest word (cosmul)
Generate the Distractors

Question word

남자 (man)

소녀 (girl)

일요일 (Sunday)

Distractor 1

Distractor 2
Splitting Word List into 3 Difficulty Levels

• **Word list after Step 1-5:**
  - Chinese: 14,000 common drama words
  - Korea: 8,000 common drama words

• **Using frequency as difficulty proxy:**
  - Split vocab into 3 difficulty bins;
  - ↑ frequency = simpler
  - Percentile ranges: (0-25], (25,75), [75,100]
  - Randomly select 10 words from each bin
(A little) Smarter than Absolute Score
Hey,

Heol! (Omg!) You know about 11% of the common drama words. It's a start, but you need to try harder to impress Jisung.

SEE DETAILED RESULTS

Share your result with friends. See who scores better!
How to Estimate learners’ Vocab Size?

• Treat the no. of correct answers per level as observed mean of the test parameter from t-distribution.

• Confidence interval gives range of values parameterized by t-distribution such that observed mean is statistically probable.

• Find the lower & upper bound of confidence interval from one-tail t-test (p < 0.1)

• Multiply the lower/upper bound of CI by no. of words
How to Estimate learners’ Vocab Size?

• Let’s say you answered 5 out of 10 correctly

• input = [1, 1, 0, 1, 0, 0, 1, 0, 1, 0]
• stdev(input) = 0.527
• mean(input) = 5

• One-tail t-test (p<0.1), so cumulative frequency = 0.9
• Degree of freedom = len(input) -1 = 9
• Critical value of t = percent-point-function(0.9, 9) = 1.38

• CI = mean ± (1.38 * 0.527) / sqrt(len(input)) = 5 ± 0.229 = (4.771, 5.229)
Numbers and Graphs
Language Leaners Like Quizzes!!

- 60,000+ quizzes taken in 2 weeks
- 35,000+ unique users completed quiz
- 16% of the users repeated quiz
Is Word Frequency a Good Proxy to Word Difficulty?
Word Frequency is a Good Indicator for Difficulty Level

No. of Easy Words
Users got correct

No. of Medium Words
Users got correct

No. of Hard Words
Users got correct
But it’s not the Perfect Indicator...

Infrequent words can be simple
Does watching K-drama more often help you improve Korean?
Conclusion
Conclusion

• How to generate quiz automatically?

• Frequency is a good indicator of lexical complexity
  • but we can do better

• Advance learners benefit from watching dramas
  • but not too much
Resources

Learn Mode: https://viki.com

Quiz: https://languagequiz.viki.com

Techblog: https://techblog.rakuten.co.jp/2017/05/26/lang-quiz/
WordNets and CALL
WordNet for Language Learning

자격증:
{
Translations / Crosslingual WSD / Annotated Corpus

자격증 [ja-gyeok-jeung]

Certificate
Document
Church
I don’t know
WordNet for Language Learning

자격증:
{
  xling: {certificate: (sentid_1, sentid_2, ...),
          license: (sentid_34, sentid_56, ...)
  },

[ja-gyeok-jeung]
Certificate
Document
Church
I don’t know
WordNet for Language Learning

자격증:

Translations / Crosslingual WSD / Annotated Corpus

Lemma frequency information

Certificate

Document

Church

I don’t know

[ja-gyeok-jeung]
WordNet for Language Learning

자격증:
{xling: {certificate: (sentid_1, sentid_2, ...),
 license: (sentid_34, sentid_56, ...)
},
counts: {sejong: (432, 0.03),
 chosun: (2342, 0.008),
 wiki: (952, 0.045),
 ...},
}
WordNet for Language Learning

자격증:
{xling: {certificate: (sentid_1, sentid_2, ...),
license: (sentid_34, sentid_56, ...)}
},
counts: {DoTS: (432, 0.03),
LBS: (2342, 0.008),
BoF: (952, 0.045),
...},

[ja-gyeok-jeung]
Certificate
Document
Church
I don’t know
WordNet for Language Learning

자가격증:

{ Translations / Crosslingual WSD / Annotated Corpus

Lemma frequency information

Extra-modality: Video, Romanization, Speech

자가격증 [ja-gyeok-jeung]

Certificate
Document
Church
I don’t know
WordNet for Language Learning

자격증:
{xling: {certificate: (sentid_1, sentid_2, ...),
  license: (sentid_34, sentid_56, ...),
},
counts: {sejong: (432, 0.03),
  chosun: (2342, 0.008),
  wiki:   (952, 0.045),
  ...},
sound: {male:   (jageokjeung-m1.wav,
  jageokjeung-m2.wav),
  female: (jageokjeung-w1.wav,
  jageokjeung-w2.wav}),

[ja-gyeok-jeung]  
Certificate
Document
Church
I don’t know
WordNet for Language Learning

자격증:
{xling: {certificate: (sentid_1, sentid_2, ...),
  license: (sentid_34, sentid_56, ...)}},

counts: {sejong: (432, 0.03),
  chosun: (2342, 0.008),
  wiki: (952, 0.045),
  ...},

sound: {male: (jageokjeung-m1.wav, jageokjeung-m2.wav),
  female: (jageokjeung-w1.wav, jageokjeung-w2.wav),

ontology: {synonym: {...}, related_to: {...}}

forms: {plural: {-을: (sentid_345, sentid_789, ...)},
  singular: {∅: (sentid_85, sentid_29, ...)}

collocations: {교사 자격증: (sent_523, ...), ... }}}}
WordNet for Language Learning

• Can contextualization (frequency, example sentences, crosslingual mappings) be scaled to 100 M / 1B sentences at “user-able” quality?

• Moving towards Multi-Modality (Video, Speech, Romanization)

• Should Grammatical Knowledge (morphology, syntax, collocations) be Lexicalized?
What Makes a (Good) Language Learning Quiz Application?
What Makes a Language Learning Quiz Application?

• Generate Quiz Automatically
What Makes a Language Learning Quiz Application?

• Generate Quiz Automatically

• Give Meaningful Results
What Makes a Language Learning Quiz Application?

- Generate Quiz Automatically
- Give Meaningful Results
What Makes a Language Learning Quiz Application?

• Generate Quiz Automatically
• Give Meaningful Results
• Reinforce Learning Experience
Hey,

*Heoll!* (Omg!) You know about 11% of the common drama words. It's a start, but you need to try harder to impress jisung.

SEE DETAILED RESULTS

Share your result with friends. See who scores better!  

Hey,

*Daebak!* (Awesome!) You know about 4% of the common drama words. You will survive a noraebang date with a pop star!

SEE DETAILED RESULTS

Share your result with friends. See who scores better!

Hey,

*jjong!* (Great!) You know about 4% of the common drama words. You can probably order an Americano at a coffee shop in Hongdae.

SEE DETAILED RESULTS

Share your result with friends. See who scores better!