한국어 쓰기 능력 자동 분석

IAKLE국제학술대회

박정열

Department of Linguistics
University at Buffalo

jungyeul@buffalo.edu

2019년 7월5-6일

Overview

1. 한국어 학습자 말뭉치 자질

2. 한국어 학습자 말뭉치 쓰기 능력 자동 처리
   - 한국어 학습자 말뭉치 쓰기 능력 자동 채점
   - 한국어 학습자 말뭉치 쓰기 능력 자동 분류

한국어 학습자 말뭉치 자질

학습자 말뭉치 자질의 종류

1. 어휘자질: 문장 길이, 단어 길이/갯수, 타입/토큰 비율
2. 통사자질: 동사 갯수, 술어-논항 구조, PCFG 규칙, 수형도 깊이
3. Fluency 자질

어휘자질

(1) a. 하지만 빌리씨하고 나오코씨는 모두 사진기가 없었어요.
(어절 토큰 갯수 = 6)

b. 하지만 빌리 씨 하고 나오코 씨 는 모두 사진
기 가 없 었 어요 .
(형태소 토큰 갯수 = 14) (Park and Lee, 2016)

문장길이 (평균)	단어길이 (평균)	단어갯수	타입/토큰 비율*
14	1.714	14	0.928

*Type/Token ratio: it is calculated with respect to lemma

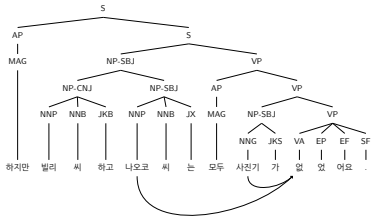
하지만	하지만/MAJ
빌리씨하고	빌리/NNP+씨/NNB+하고/JKB
나오코씨는	나오코/NNP+씨/NNB+**는/JX**
모두	모두/MAG
사진기가	사진기/NNG+**가/JKS**
없었어요.	없/VA+었/EP+어요/EF+./SF

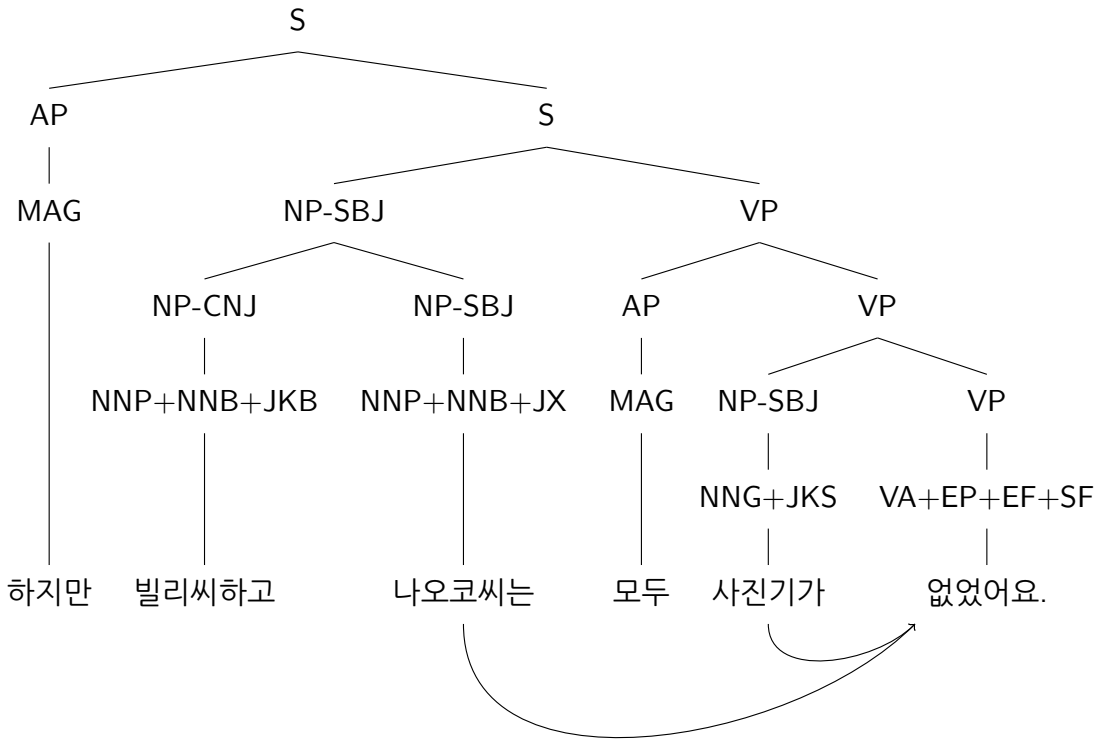
Figure: Example of POS tagging

	word form	lemma	
verbal ending	ㄴ	은	
	ㄹ지	을지	
case marker	가	이	('NOM')
	를	을	('ACC')
	는	은	('TOP')

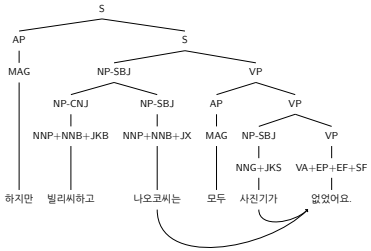
Figure: Suffix normalisation examples (Park and Tyers, 2019)

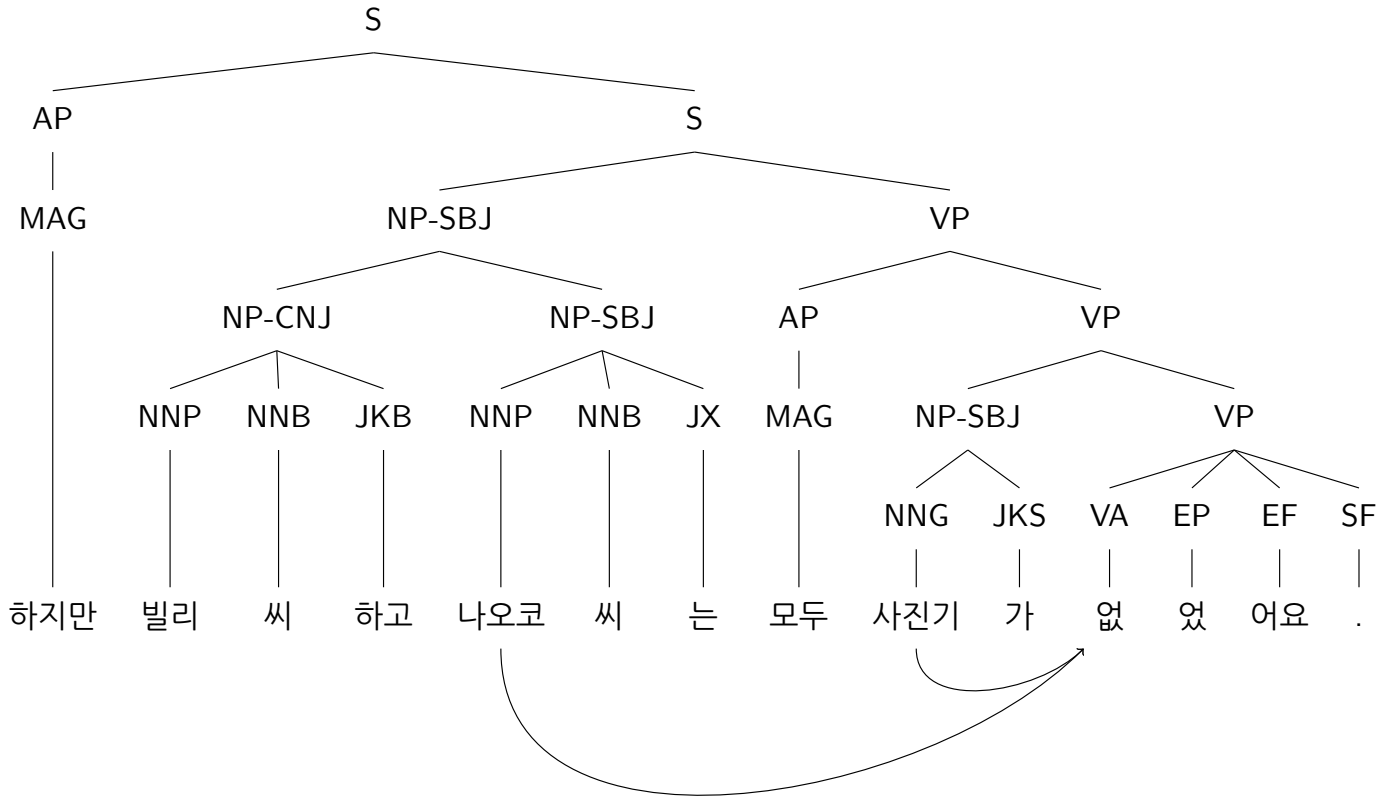
통사자질



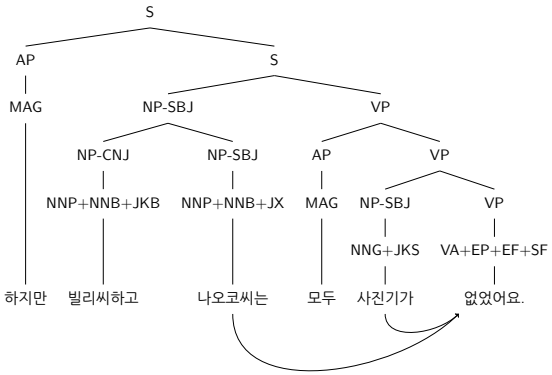


의존구조분석 (Park et al., 2013)





구구조분석 (Choi et al., 2012; Park et al., 2016)



동사 갯수	수형도 깊이	술어-논항 구조	PCFG 규칙
1	6	np-sbj np-sbj adj	s→ap s
			s→np-sbj vp
			np-sbj→np-sbj vp
			...

Fluency 자질

1. The perplexity is the inverse probability of the sentence, normalized by the number of words:

$$f_1(h) = \frac{ppl}{|h|}$$

2. Fluency score $S_F(h)$ by Asano et al. (2017):

$$f_2(h) = \frac{\log P_m(h) - \log P_u(h)}{|h|}$$

3. Fluency score $f_3(x)$ by Ge et al. (2018):

$$f_3(h) = \frac{1}{1 + H(x)}$$

$$H(x) = -\frac{\log P_m(h)}{|h|}$$

where P_m is the probability of the sentences given by language model, P_u is the unigram probability of the sentences, and $|h|$ is total number of words.

언어모델:

unigram

bos	하지만	빌리	씨	하고	나오코	씨	는	모두	사진기	가	없	었	어요	.	eos
$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

bigram

bos 하지만	하지만 빌리	빌리 씨	씨 하고	하고 나오코	...	어요 .	. eos
$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	...	$\frac{1}{15}$	$\frac{1}{15}$

SRILM (Stolcke, 2002) 및 세종형태소분석 코퍼스[1]을 사용한 한국어 언어 모델 구축

[1] 676,951 문장, 19,014,530 토큰 (9M 어절)

학습자 말뭉치 자질 분류

1. complexity: 어휘 및 통사 자질
2. fluency: 언어 모델
3. accuracy: requires the error annotated corpus

accuracy

S Despite of it is an industrial city , there are many shops and department stores .
A 0 1|||R:PREP|||Although|||REQUIRED|||-NONE-|||0
A 1 2|||U:PREP||||||REQUIRED|||-NONE-|||0

	grammatical error correction 예:
input	Despite of it is an industrial city , there are many shops and department stores .
output	Although it is an industrial city , there are many shops and department stores .

한국어 오류 주석 말뭉치

```
<SENTENCE to="83" from="60">
 <s>수업이 끝난 후에 친구하고 약속 있어요. </s>
 <LearnerErrorAnnotations>
  <word>
   <w>친구하고</w>
   <morph from="70" to="74" subsequence="1" wordStart="Start">
    <Preserved>친구</Preserved>
   </morph>
   <morph from="70" to="74" subsequence="2" wordStart="None">
    <Proofread pos="JKB">와</Proofread>
    <ErrorArea type="FAP" />
    <ErrorPattern type="REP" />
    <ErrorLevel type="DS" />
   </morph>
  </word>
  <word>
   <w>약속</w>
   <morph from="75" to="77" subsequence="1" wordStart="Start">
    <Preserved>약속</Preserved>
   </morph>
   <morph from="75" to="77" subsequence="2" wordStart="None">
    <Proofread pos="JKS">이</Proofread>
    <ErrorArea type="FNP" />
    <ErrorPattern type="OM" />
   </morph>
  </word>
 </LearnerErrorAnnotations>
</SENTENCE>
```

한국어 학습자 말뭉치 쓰기 능력 자동 처리

한국어 학습자 말뭉치 쓰기 능력 자동 처리

1. 쓰기 능력 자동 채점
   - 점수 1-10,
   - continuous values,
   - linear regression
2. 쓰기 능력 자동 분류:
   - 레벨 1-6,
   - discrete values,
   - logistic regression

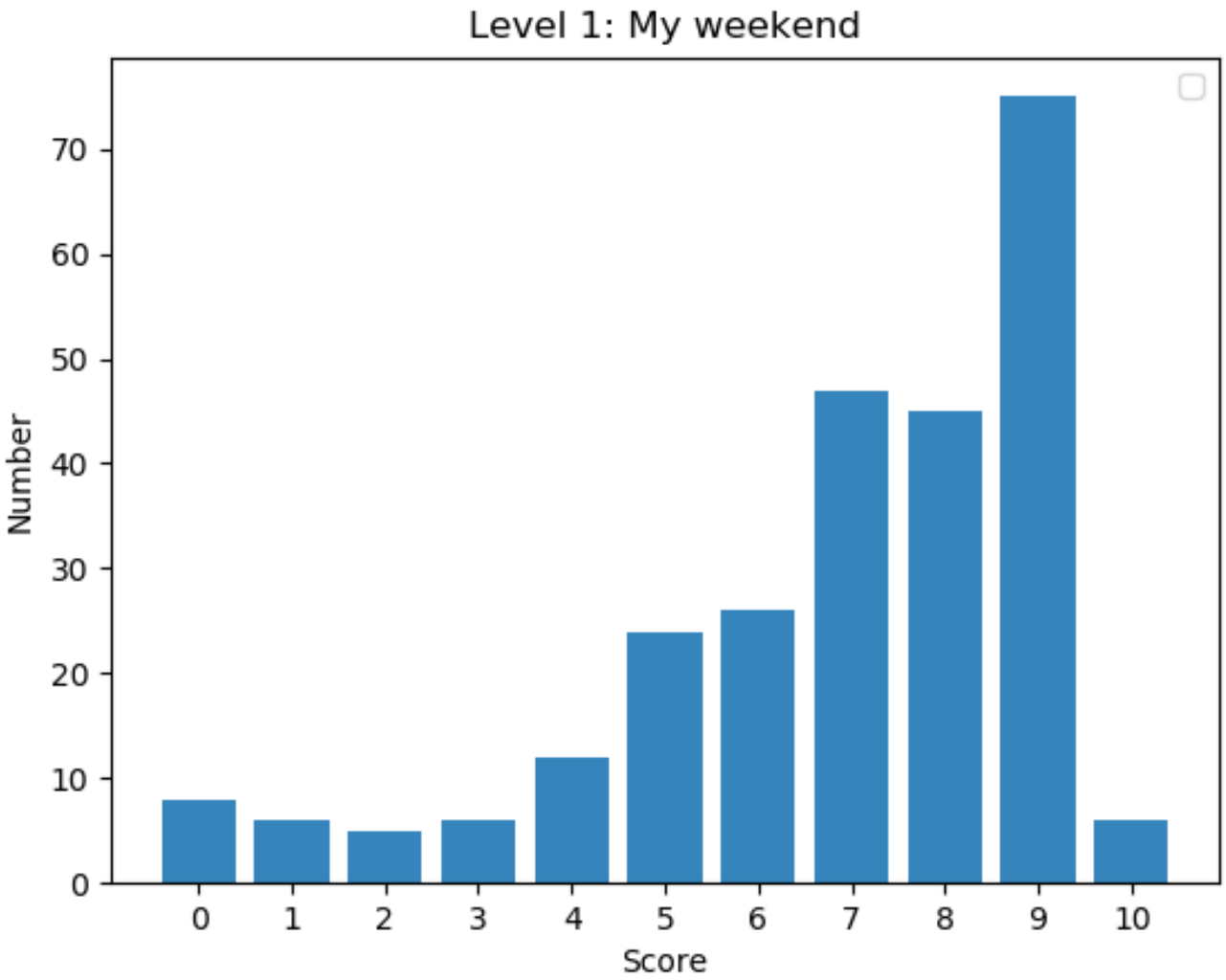
한국어 학습자 말뭉치 쓰기 능력 자동 채점

주말 이야기, 260 examples for Level 1

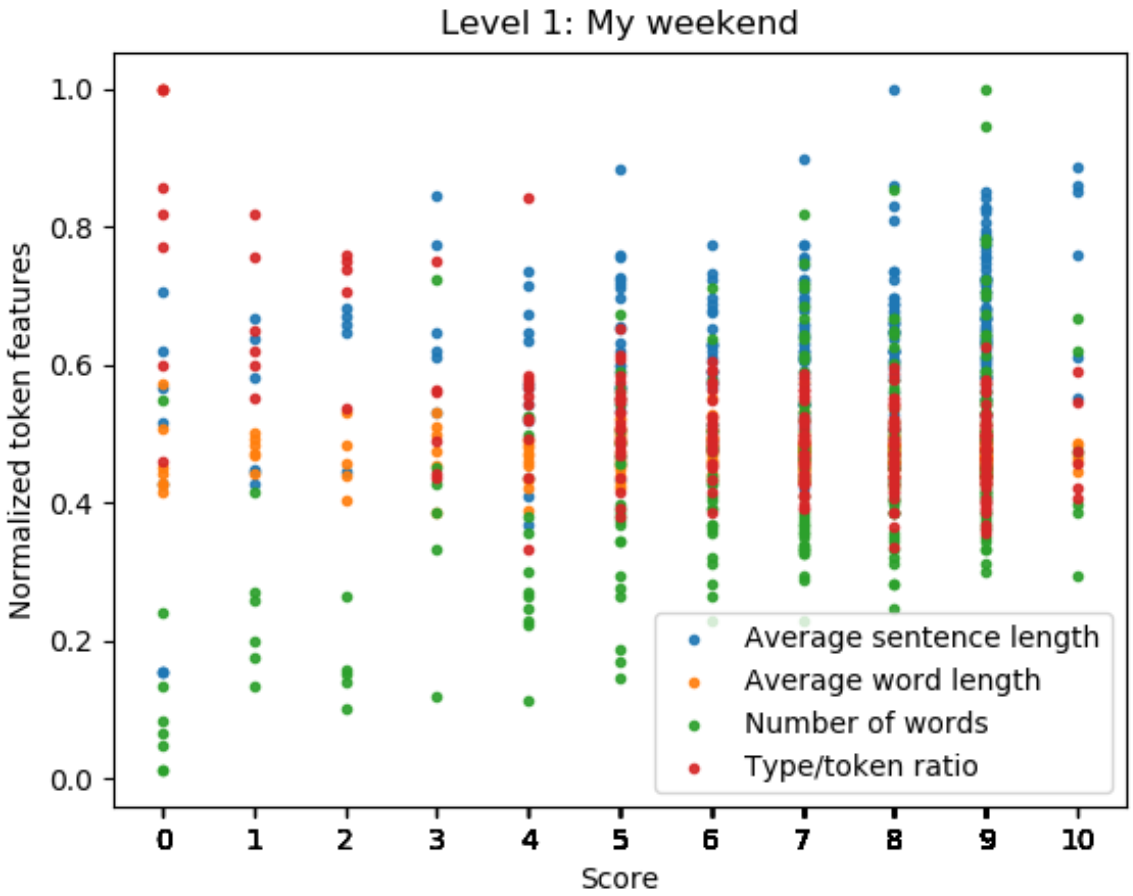
A100003_v01
<topic>주말 이야기</topic>
<score>80</score>
<s>우리 집에 갔습니다.</s> <s>강아지하고 제 집에 갔습니다.</s> <s>우리가 놀았습니다.</s> <s>저는 줌을 자지 않았습니다.</s> <s>집에 제침대를 있습니다.</s> <s>집에 강아지 침대도 있습니다.</s> <s>하지만 우리가 줌을 안 잡니다.</s> <s>그리고 강아지하고 밥을 먹었습니다.</s> <s>저는 빵을 먹었습니다.</s> <s>제 강아지가 동물 음식을 먹었습니다.</s> <s>모두 맛있었습니다.</s> <s>저는 재미있었습니다.</s>

A100305_v02
<topic>주말 이야기</topic>
<score>60</score>
<s>저는 서울숍에 갑니다.</s> <s>나하고 리사가 갑니다.</s> <s>서울숖에 사람 많습니다.</s> <s>저는 군구를 합니다.</s> <s>리사도 공부합니다.</s> <s>재는 옷습니다.</s>

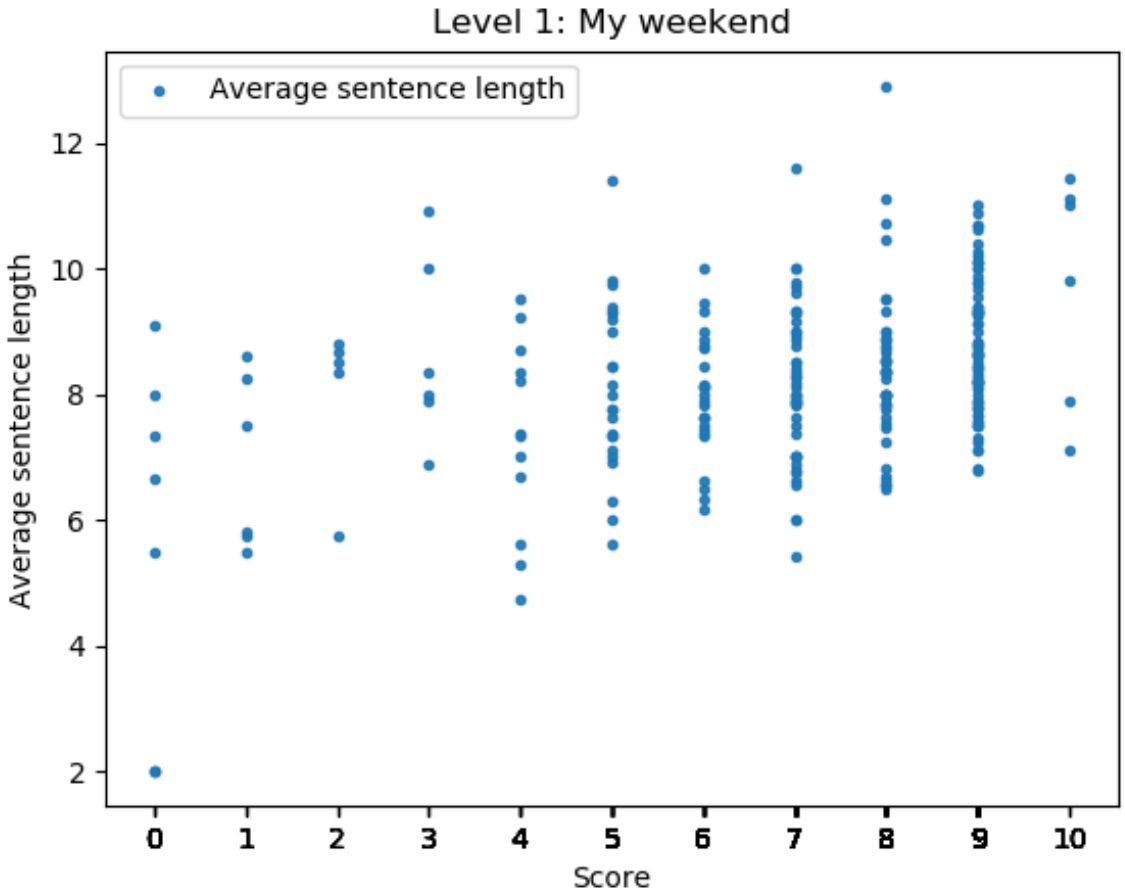
학습자 점수 분포


Level 1: My weekend
Number
Score
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6
7
8
9
10

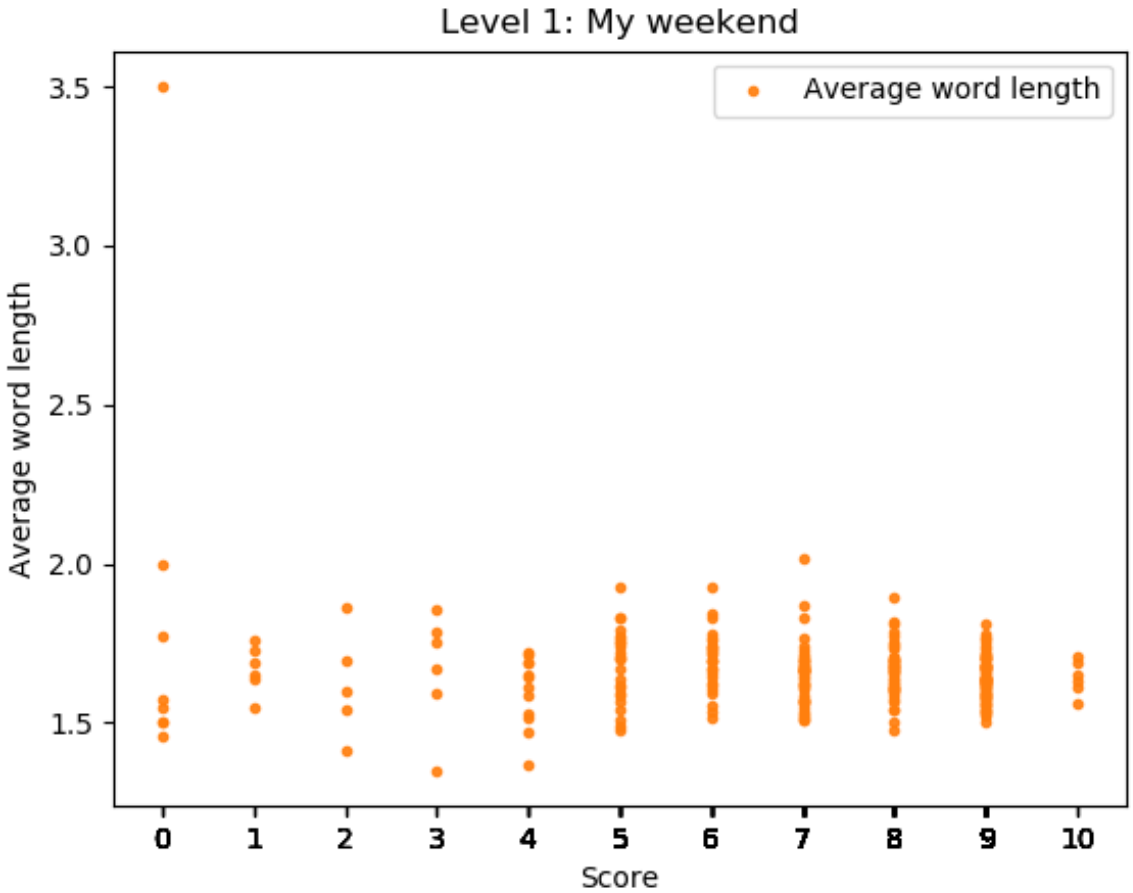
어휘자질 분포


Level 1: My weekend
Normalized token features
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
Score
Average sentence length
Average word length
Number of words
Type/token ratio

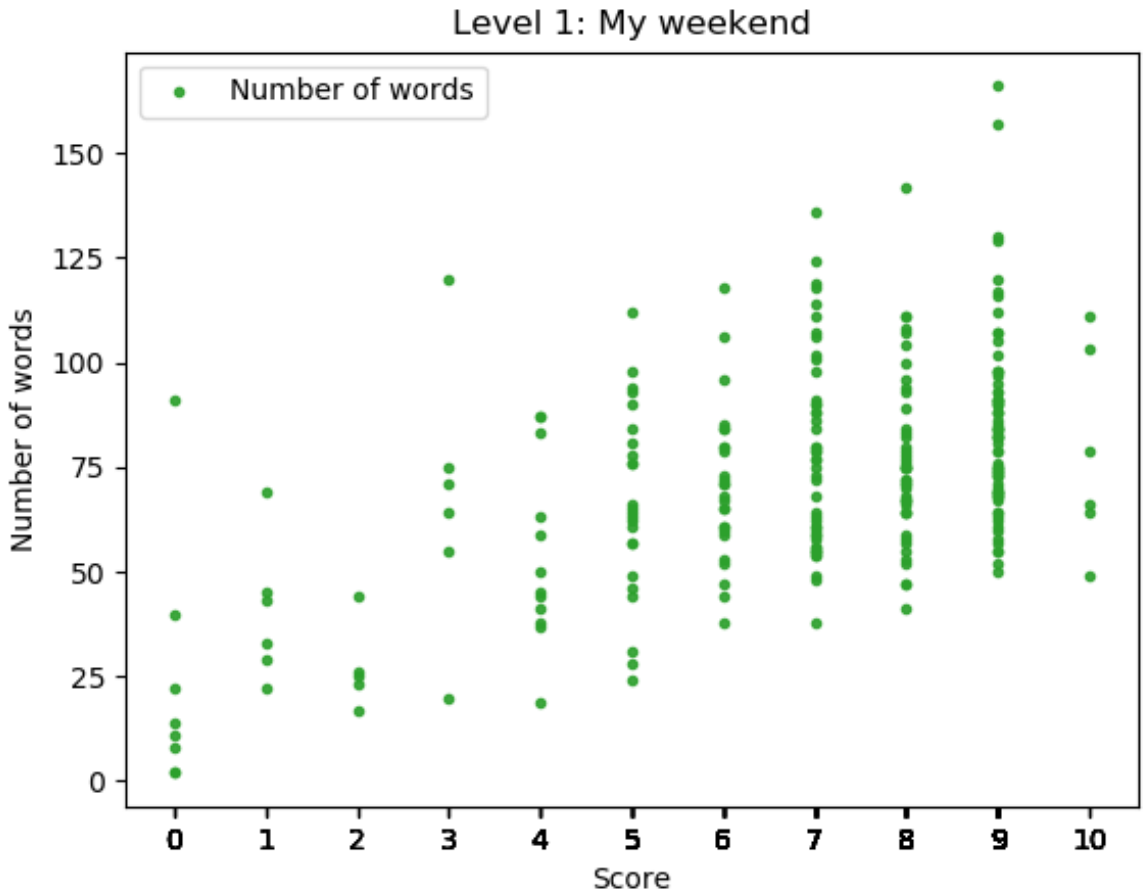
average sentence length:



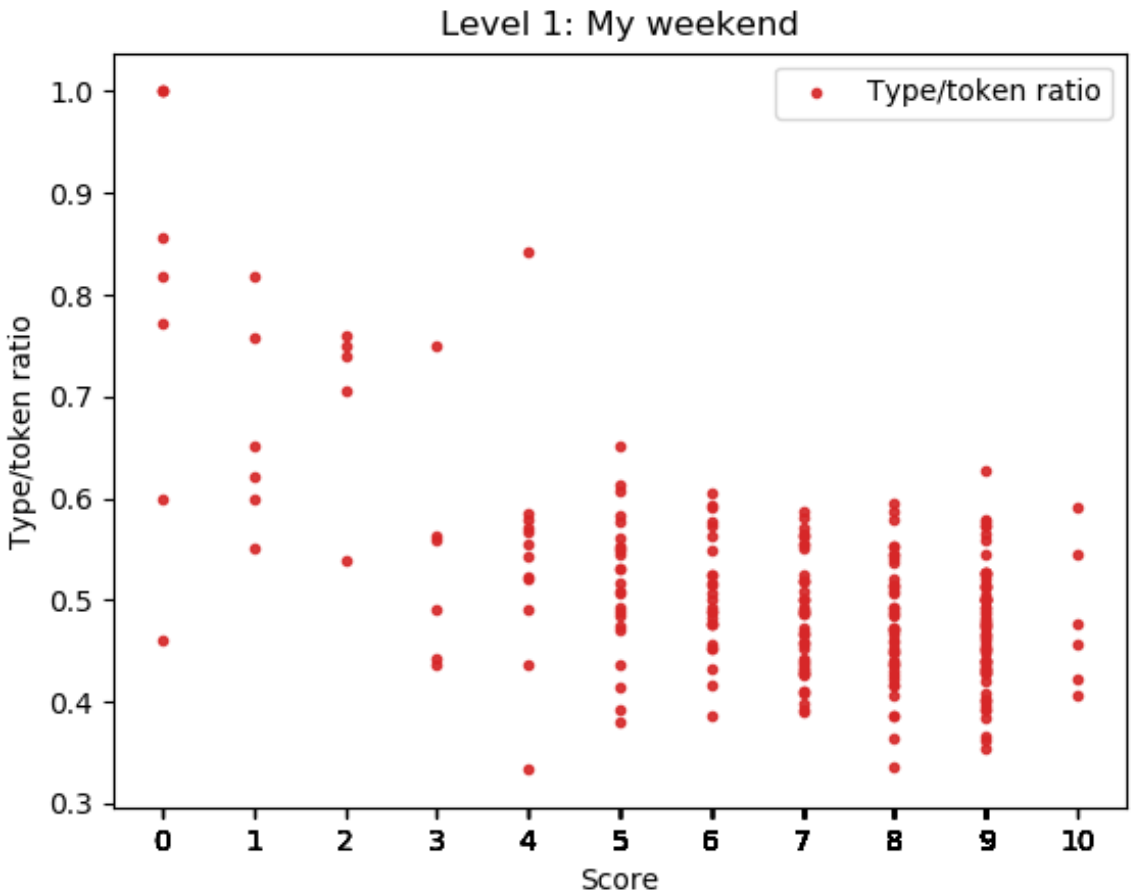
average word length:



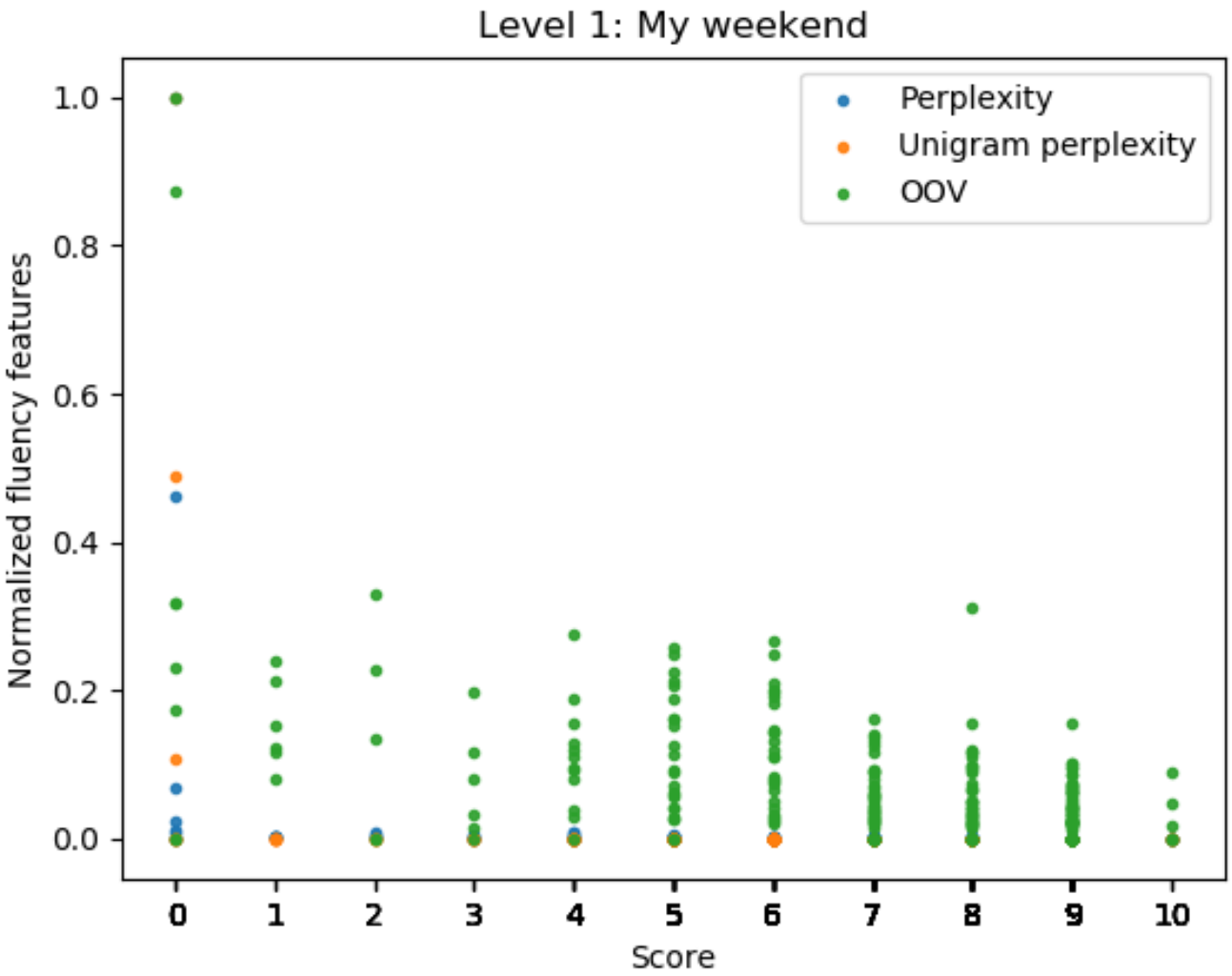
number of words:



type/token ratio:



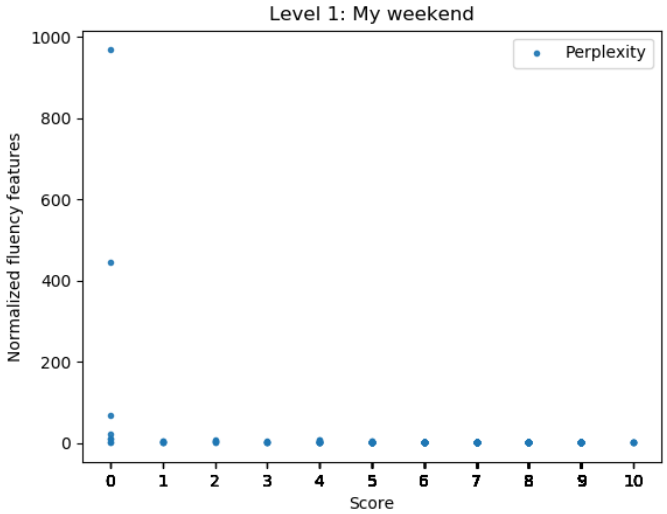
Fluency 자질 분포

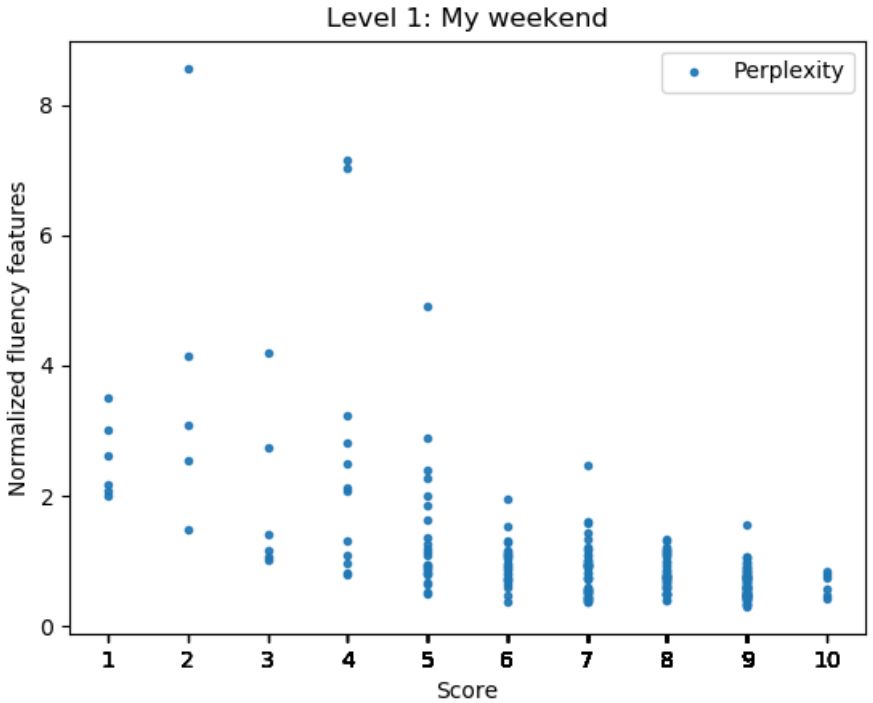


Perplexity 자질 예제

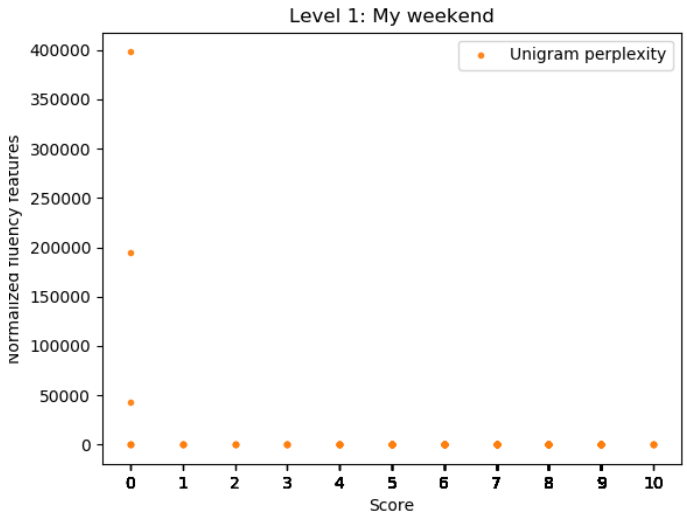
```
==> A100007_v01.sent.tok.ppl <==
file A100007_v01.sent.tok: 9 sentences, 103 words, 1 OOVs
0 zeroprobs, logprob= -211.9383 ppl= 81.16233 ppl1= 119.6264

Score: 10

==> A100007_v02.sent.tok.ppl <==
file A100007_v02.sent.tok: 6 sentences, 49 words, 0 OOVs
0 zeroprobs, logprob= -94.92311 ppl= 53.19547 ppl1= 86.53788

Score: 7
```

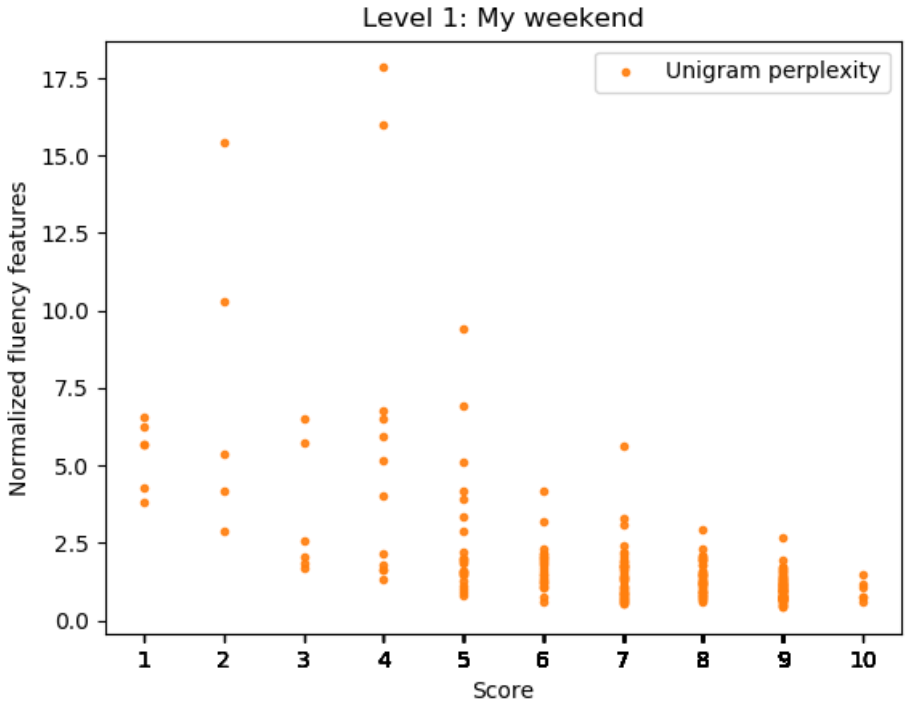
Perplexity:


Level 1: My weekend
Perplexity
Normalized fluency features
Score
0
200
400
600
800
1000
0
1
2
3
4
5
6
7
8
9
10

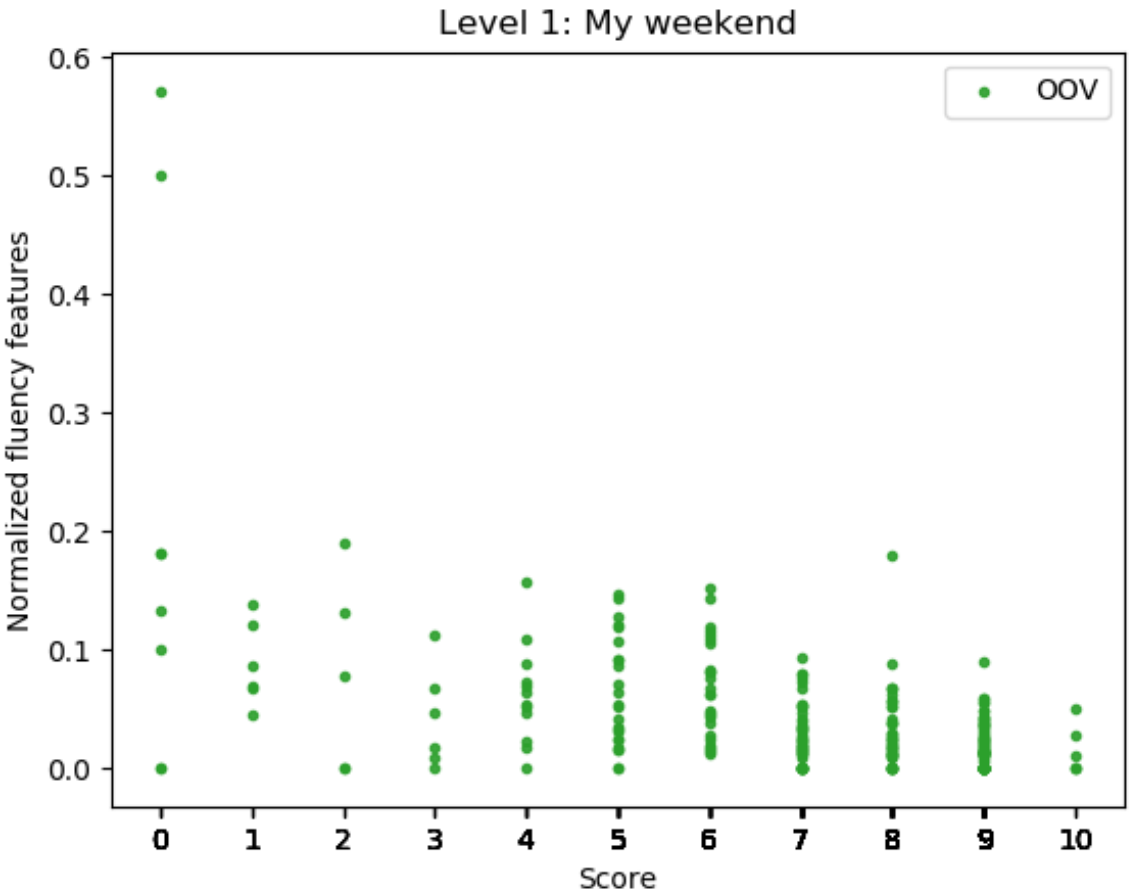

Level 1: My weekend
Perplexity
Normalized fluency features
Score
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10

Unigram perplexity:





OOV:


Level 1: My weekend
OOV
Normalized fluency features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
7
8
9
10
Score

구구조 자질 예제

```
==> A100303_v02.sent.berkeley.out.cfg <==
6 (VP NP VP)
3 (VP VA EF SF)
3 (VP VV EP EF SF)
2 (VP VA EP EF SF)
2 (VP NNG XSV EC)
1 (VP AP VP)
1 (VP VV ETM)
1 (VP VP VP)
```

```
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 3
1 0 0 0 0 0 0 0 2 0 1 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 3 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 1 0 0 1 0 4 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 6 0 0 1 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 3 0 1 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
```

232-dimension

의존구조 자질 예제

```
==> A100303_v02.sent.malt.out.pred <==
1 VA+EF+SF NNG+JX
1 VA+EF+SF NNP+JKB NNG+JKS
1 VA+EF+SF PRON+JKB+JX NNG+NNG+JKS
1 VA+EP+EF+SF NNG
1 VA+EP+EF+SF NNP+JX MAG
1 VV+EP+EF+SF NNG+JKO
1 VV+EP+EF+SF PRON+JKS NNG+JKO
1 VV+EP+EF+SF PRON+JX NNG+JKB NNG+JKB
```

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 ...
```

846-dimension

한국어 학습자 말뭉치 쓰기 능력 자동 분류

나의 미래 계획, 24/96 examples for Level 1/2

A100064_v02
<topic>나의 미래 계획</topic>
<score>30</score>
<s>제가 한국에 좋아하세요.</s> <s>그래서 한국에 왔습니다.</s> <s>한국 음식을 좋아하세요.</s> <s>비빔밥하고 나명도 좋아하세요.</s> <s>친구 같이 공부 하고 싶습니다.</s>

A200000_v03
<topic>나의 미래 계획</topic>
<score>60</score>
<s>저는 2월 25일에 한국에 왔습니다.</s> <s>저는 지금 한국어를 배우고 있습니다.</s> <s>1년동안 한국어를 배운 후에 저는 한국에서 있는 대학교에 갈 겁니다.</s> <s>대학교에 갈 때 저는 한국 대학생들과 같이 배워서 지금 한국어를 열심히 공부합니다.</s> <s>저는 서울대학교에 가겠습니다.</s> <s>저는 아마 역사학과나 국어국문을 전공하려고 합니다.</s> <s>한국에서 공부가 아주 어려울 것 같습니다.</s> <s>그래서 시간이 있으면 책을 많이 읽습니다.</s> <s>많은 준비하기 때문에 대학교 공부가 더 쉴 것 같습니다.</s> <s>감사합니다.</s>

결과

5-fold cross validation:

자동채점 linear regression
Mean squared: -6.05 (+/- 1.19)[2]

자동분류 linear logistic: classification
Accuracy: 0.97 (+/- 0.08)[3]

[2] [-6.25144946 -5.84226331 -5.88071923 -7.05126297 -5.24455246]
[3] [1. 0.91666667 0.91666667 1. 1.]

끝

Asano, H., Mizumoto, T., and Inui, K. (2017). Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Choi, D., Park, J., and Choi, K.-S. (2012). Korean Treebank Transformation for Parser Training. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 78–88, Jeju, Republic of Korea. Association for Computational Linguistics.

Ge, T., Wei, F., and Zhou, M. (2018). Fluency Boost Learning and Inference for Neural Grammatical Error Correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1055–1065, Melbourne, Australia. Association for Computational Linguistics.

Park, J., Hong, J.-P., and Cha, J.-W. (2016). Korean Language Resources for Everyone. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers (PACLIC 30)*, pages 49–58, Seoul, Korea. Pacific Asia Conference on Language, Information and Computation.

Park, J., Kawahara, D., Kurohashi, S., and Choi, K.-S. (2013). Towards Fully Lexicalized Dependency Parsing for Korean. In *Proceedings of The 13th International Conference on Parsing Technologies (IWPT 2013)*, Nara, Japan. International Conference on Parsing Technologies (IWPT 2013).

Park, J. and Lee, J. H. (2016). A Korean Learner Corpus and its Features. *Journal of the Linguistic Society of Korea*, 75:69–85.

Park, J. and Tyers, F. (2019). A New Annotation Scheme for the Sejong Part-of-speech Tagged Corpus. In *Proceedings of the 13th Linguistic Annotation Workshop (The LAW XIII)*.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002*, pages 901–904, Denver, Colorado. International Conference on Spoken Language Processing.