Analyzing the Impact of Spelling Errors on POS-Tagging and Chunking in Learner English

Tomoya Mizumoto Ryo Nagata

Background

- are used on NLP tasks that target learner English
  - 10 of the 12 teams used POS-tagging in the CoNLL ST
- also are used for linguistic analysis of learner English
  - explored characteristic patterns in learner English
  - POS sequences can be used to distinguish between mother tongue interferences
- Detailed investigation would improve related tasks
  - none of studies described the root cause of POS-tagging errors in detail

♠Summary♠

- have investigated performance of POS-tagging
  - focused our investigation on spelling errors

Extent of performance degradation due to spelling errors

Performance of POS-tagging: 0.23%↓
Spelling errors do not influence accuracy of estimating POS of their surrounding words

Types of spelling errors

No DIFF on performance between known and unknown

Effects of spell checker

Improvement: 0.06% → spell checker is not required

Performance Analysis of Spelling Errors

1. Extent of performance degradation due to spelling errors

- Learner English includes 3.4% spelling errors
  - assuming that POS-tagging fails for all unknown words: performance 3.4%↓ ✕
- Effect of misspelled words have on them or their surrounding words

It is very ineresting/*interesting game .

Final seen/*scene is very good .

2. Types of spelling errors

- Various types of spelling errors

e.g. Unknown word error:
typographical (studing/*studying)
Known word error:
homophones (sea/*see)、
derivations (smell/*smelly)

- Some spelling errors have effective information that helps determine POSs

e.g. affix information (e.g. ed, ing)

3. Effect of a spell checker

- Accuracy of spell checker is not 100%
  - can correct unknown errors
  - difficult to correct known word errors
  - correct unknown errors to different words

e.g. movile → **movie** or **mobile**

- Does ideal spell checker have positive effect on POS-tagging?

Experiments

Experimental Setup

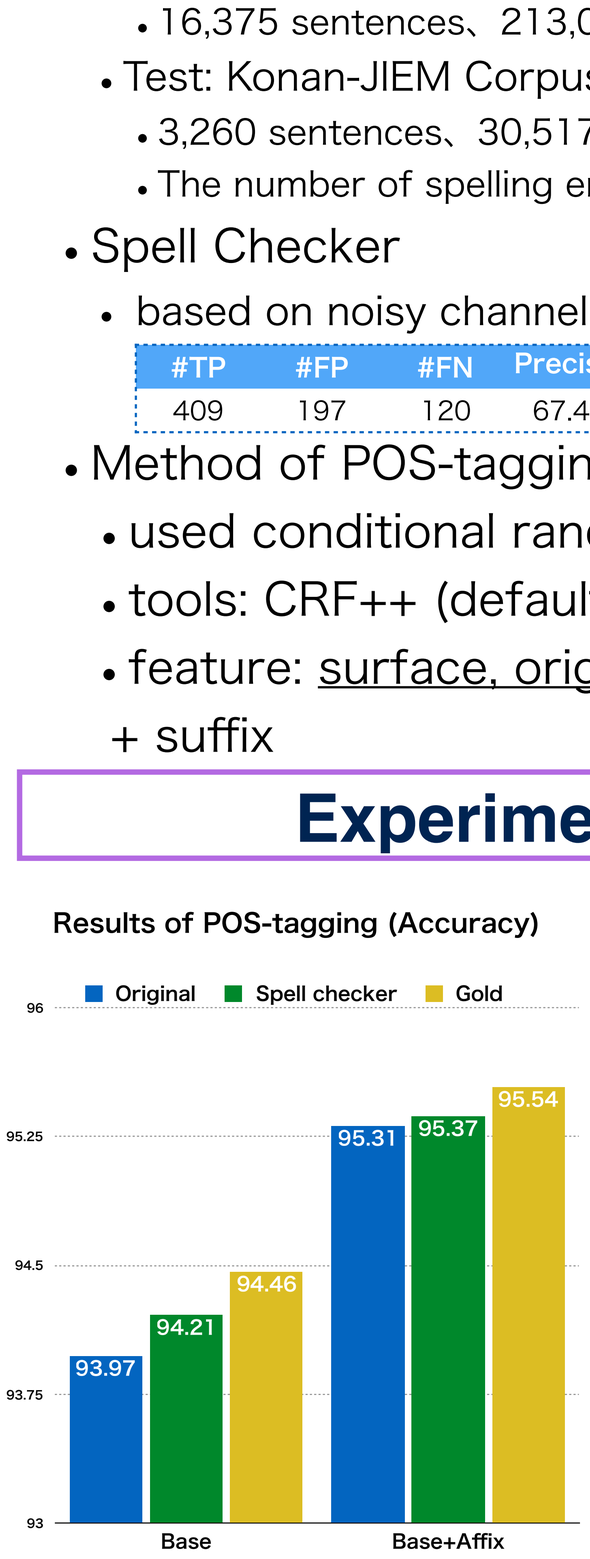
- Data
  - Train: in-house data
    - 16,375 sentences、213,017 tokens
  - Test: Konan-JIEM Corpus
    - 3,260 sentences、30,517 tokens
    - The number of spelling errors: 654 (Unknown errors: 487, Known errors: 167)
- Spell Checker
  - based on noisy channel model

#TP	#FP	#FN	Precisio	Recall	F-score
409	197	120	67.49	77.32	72.07

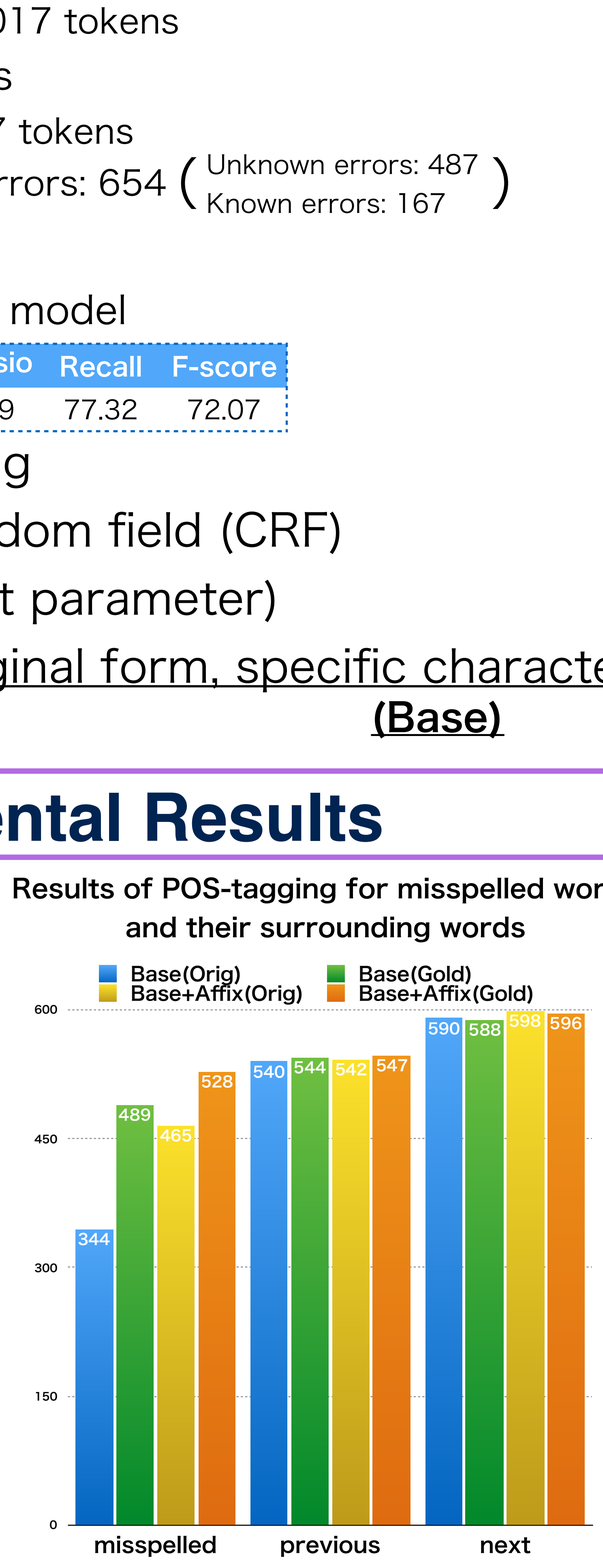
- Method of POS-tagging
  - used conditional random field (CRF)
  - tools: CRF++ (default parameter)
  - feature: surface, original form, specific character (Base) + suffix

Experimental Results

Results of POS-tagging (Accuracy)



Results of POS-tagging for misspelled words and their surrounding words



Extent of performance degradation due to spelling errors

- Comparing the results of POS-tagging
  - Base+Affix (Orig) v.s. Base+Affix (Gold)
    - 95.31% -> 95.54%、0.23↑

POS-tagging performance dropped 0.23% due to spelling errors

- Comparing the number of correct POS
  - the number of correct POS for misspelled words increased
    - i.e. 344 -> 465, 489 -> 528
  - for the number of correct POS for surrounding words, there was nearly no difference

Types of spelling errors

- The effect of affix information for spelling errors (344 -> 465)
  - by using affix information, POS-tagger could identify the correct POS for approximately 120 misspelled words
- Unknown word error v.s. known word error
  - Analyze the words that Base+Affix (Original) can not identify
    - unknown: 143/487 (29%), known: 46/167 (27.5%)

the ratio are not difference between unknown and known

Effects of a spell checker

- by using spell checker, the accuracy improves 0.06%

Spell checker does not have positive effect for POS-tagging

It is sufficient to assign POS tags using affix information

- The number of spelling errors that were correctly assigned to POSs with spell checker (74)

Base+Affix (Original)	Base (Spell checker)
pepole/Noun, singular	people/Noun, plural
tow/Noun, singular	two/Numeral

- The number of spelling errors that were incorrectly assigned to POSs with spell checker (49)

Base+Affix (Original)	Base (Spell checker)
tero/Noun, singular (corr: terrorist)	to/Noun, plural
tittle/Noun, singular (corr: title)	little/Adjective