Natural Language Processing Lab
Week 4 Spell Checker
Using Web Corpus

10620ISA 562100 自然語言處理實作

T7T8T9 張俊盛 Jason S. Chang 資電323

2018-0320

Example of Spelling Errors in Context

要先偵測錯字在句子的哪個地方，然後在從上下文找出最適合的候選字

- Non-word errors (from *Birkbeck Spelling Error Corpus*)
  - I felt very **strang** → I felt very **strange**
  - in the **weanter** when it was snowing → in the **winter** when it was snowing

- Real word errors 只能靠上下文來改字
  - at **brake** time → at **break** time
  - when the **brack** was finished → when the **break** was finished

Detecting Spelling Errors in Context

- Non–word errors 只要檢查每一個字有沒有在字典就好
  - 1–gram counts (for finding the most likely error)
    - big.txt
- Real word errors 很容易互相混淆的字的辭典
  - Use a set of confusable words (lab4.confusables.txt)
  - Use Linggle API or NetSpeak API
    - The lower trigram count, the higher error probability
    - min(count('when the brake'),
      count('the brake was'),
      count('brake was finished'))

Correcting Spelling Errors in Context

- Use your lab3 (to generate several candidates)
- Use Linggle API or NetSpeak API (to select the best correction)
  - Replace words in test phrase generating several candidates
    - count(‘when the **break** was finished’)

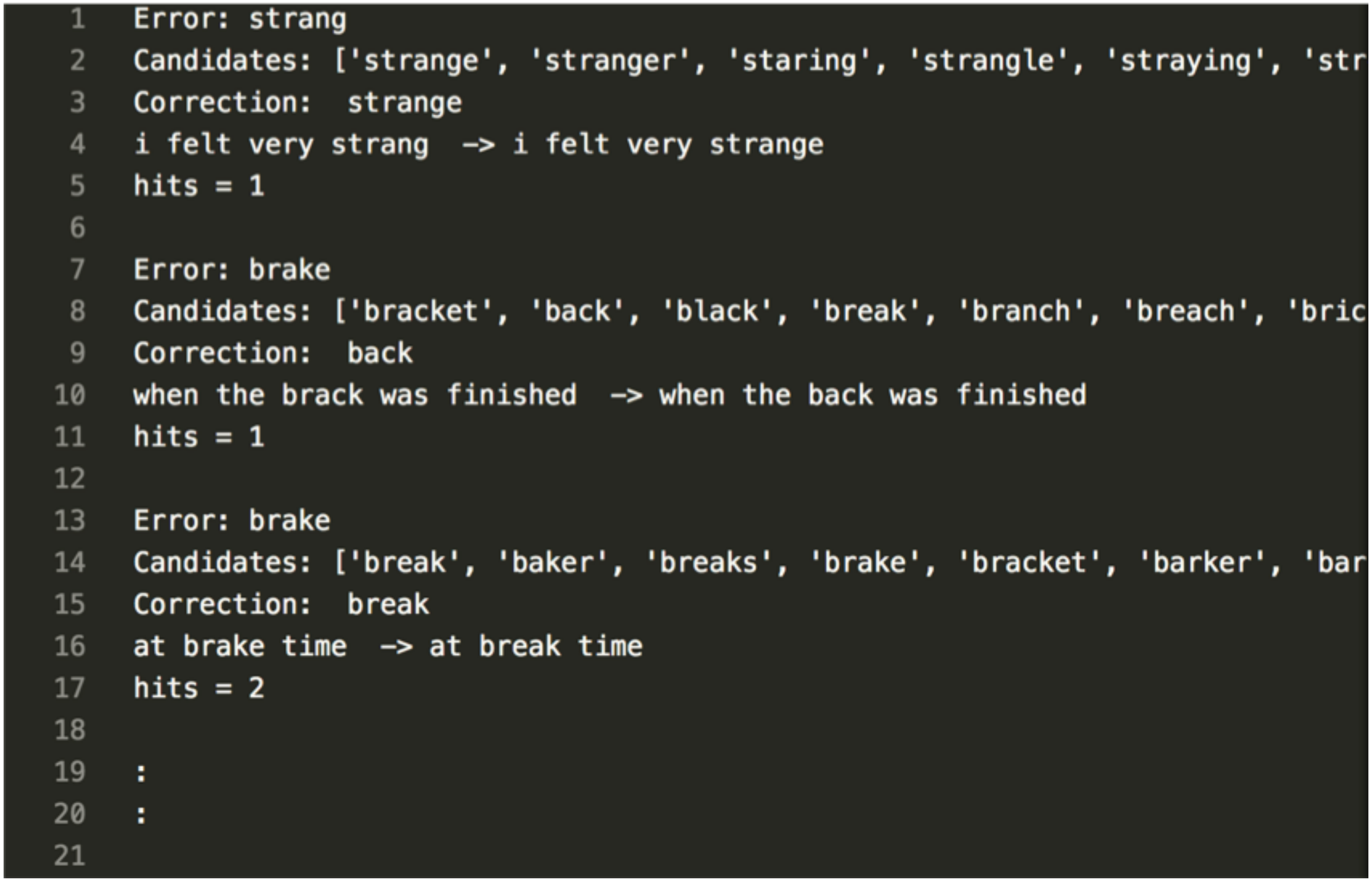
      = count(‘when the break’) * count(‘the break was’)

      * count(‘break was finished’)

Lab for week 4

- Training and Testing data
  - lab4.confusables.txt (www.alphadictionary.com/articles/confused_words.html)
  - lab4.test.1.txt (198 errors)
  - lab4.test.2.txt (0 errors)
    - Source: *Birkbeck Spelling Error Corpus* (ota.ox.ac.uk/headers/0643.xml)
- Reusable code
  - lab3.py
  - LinggleAPI.py
  - NetSpeakAPI.py
- Evaluation
  - Precision = #hits / #corrections
  - FalseAlarm = (#corrections – #hits) / #corrections

Output



```
Error: strang
Candidates: ['strange', 'stranger', 'staring', 'strangle', 'straying', 'str
Correction:  strange
i felt very strang  -> i felt very strange
hits = 1

Error: brake
Candidates: ['bracket', 'back', 'black', 'break', 'branch', 'breach', 'bric
Correction:  back
when the brack was finished  -> when the back was finished
hits = 1

Error: brake
Candidates: ['break', 'baker', 'breaks', 'brake', 'bracket', 'barker', 'bar
Correction:  break
at brake time  -> at break time
hits = 2

:
:

```

Discussion

- Where to find confusable words: dictionaries
  - Laurence Urdang's The Dictionary of Confusable Words
  - Andian Room's Dictionary of Confusable Words
  - Dave Dowling's The Wrong Word Dictionary
- Automatic generation of confusable words
  - Use spell.py to generate confusable words (take all edit-1 and edit-2 words not just the most frequent word)
  - Hovermale, Dennis & Mehay (2009). Real-word Spelling Correction for CALL (use CMUDict to generate confusable words)
    citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.6124&reprep1&type=pdf
- Edit Logs
  - WikEd Error Corpus (WikEd)
    romang.home.amu.edu.pl/wiked/wiked.html
  - Language Editing Dataset of Academic Texts (LEDAT)
    www.vtex.lt/en/ledat.html

References

- Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991.Context based spelling correction. Information Processingand Management, 23(5), 517—522.
- Islam, Aminul, and Diana Inkpen. ”Real–word spelling correction using Google Web IT 3–grams.” Proceedings of the2009 Conference on Empirical Methods in Natural LanguageProcessing: Volume 3–Volume 3. Association for Computa–tional Linguistics, 2009.
- Bergsma, Shane, Dekang Lin, and Randy Goebel. ”Web–Scale N–gram Models for Lexical Disambiguation.” IJCAI.Vol. 9. 2009.