自然語言處理實作 (2003-)

雜訊通道模型 Noisy Channel Model
＝通道模型 Channel Model
＋ 語言模型 Ngram Model

張俊盛 jason@nlplab.cc

助教：許瑋芩 <winnie@nlplab.cc>

郭俊豪 <isaackch@nlplab.cc>

雜訊通道模型 Noisy Channel Model

- 兩部分 Two parts
  - 通道模型 Channel Model
  - 語言模型 Ngram Model
- 例子
  - 拼字、文法改錯 spelling and grammatical error correct
    - The trains are very crowdy in morning.
    - The trains are very **crowded** in **the** morning.
  - Real word errors
    - Input: He sold everything **expect** the house.
    - Output: He sold everything **except** the house.
  - Verb tense errors
    - Input: He **sell** everything except the house.
    - Out: He **sold** everything except the house.

語言模型 Language Model (通用 general)

- 一億詞英國國家語料庫
  British National Corpus (100 million or 10^8 words)

- 用工具製作機率模型
  Language model built using a tool

  - https://github.com/kpu/kenlm

通道語言 (Probabilistic) Channel Models

- 拼字錯誤 spelling errors
  - 用 spell.py 產生候選詞 Use **spell.py** to generate candidates
    - crowdy ==> crowded

- 虛詞錯誤 Commonly confusable function words
  - 用清單產生候選詞 Use a list to generate candidates
    - a, an, any, the, empty
      (e.g., a => an OR any => a)

    - at, from, on, by, in, to, for, of, with, empty
      (e.g., for => on as in "feedback for")

本次作業 Lab Work

- 任務 Tasks

  - 拼字改錯 spelling error

  - 文法改錯 grammatical errors

    - 誤用、多餘的冠詞、介詞 wrong/useless det./prep.

- 加分題 bonus

  - verb tense error (use the file provided: **inflect.txt**)

  - 缺少冠詞、介詞 missing articles and preposition

參考資料

- Bergsma, Shane, Dekang Lin, and Randy Goebel. "Web-scale N-gram models for lexical disambiguation." (2009). storage.googleapis.com/pub-tools-public-publication-data/pdf/35387.pdf
- Bryant, Christopher, and Ted Briscoe. "Language model based grammatical error correction without annotated training data. BEA 2018. https://www.aclweb.org/anthology/W18-0529.pdf https://github.com/chrisjbryant/lmgec-lite
- Bryant, Christopher, et al. "The BEA-2019 shared task on grammatical error correction." BEA 2019. https://www.aclweb.org/anthology/W19-4406.pdf
- Bryant, Christopher, et al. "The BEA-2019 shared task on grammatical error correction." BEA 2019
- Stahlberg, Felix, Christopher Bryant, and Bill Byrne. "Neural grammatical error correction with finite state transducers." arXiv preprint arXiv:1903.10625 (2019).