LING 120: Language and Computers

Semester: FALL 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

1 Sep 2017

Class outline

- Clarification about terms used yesterday
- A group exercise preparing you for next week

Clarification about terms used on Wednesday

- Non-word error detection
- Isolated word error correction
- Contexual spelling correction and grammatical error correction (for real-word errors)

Non-word error detection - Dictionaries

- Dictionary based - but it does not cover all cases, as we saw in the last class.
- Usual solution: maintain a dictionary with different word forms etc (add some exceptions - like ignore words with caps etc.) ..
- Most spell-checking tools support personalized spell-checking. (What does that mean??)

Non-word error detection - N-gram methods

- To some extent, we can detect misspellings even without a dictionary.
- Before talking about "how", what is a character n-gram?

Non-word error detection - N-gram methods

- To some extent, we can detect misspellings even without a dictionary.
- Before talking about "how", what is a character n-gram?
- If we collect a large list of valid English words, we can have an estimate of character sequence frequencies in English.
- We can then use it to check for abnormal character sequences in a word, to identify spelling errors. (Makes sense??)

Non-word error detection - N-gram methods

- To some extent, we can detect misspellings even without a dictionary.
- Before talking about "how", what is a character n-gram?
- If we collect a large list of valid English words, we can have an estimate of character sequence frequencies in English.
- We can then use it to check for abnormal character sequences in a word, to identify spelling errors. (Makes sense??)
- example: "thi" is perhaps a valid character trigram in English, but "qki" is not.

Isolated word spelling correction

- Aim: suggest correction candidates for a mis-spelled word, irrespective of the surrounding words.
- Question to test if you did the reading: What are the three steps to do isolated word spelling correction?

Isolated word spelling correction

- Aim: suggest correction candidates for a mis-spelled word, irrespective of the surrounding words.
- Question to test if you did the reading: What are the three steps to do isolated word spelling correction?
  - Detect a error (dictionary or n-gram method)
  - Identify possible candidates for right spelling (from where??)
  - Rank them in terms of the most probable word (how?)

Group Exercise

- Form into groups of 2–3 people, and do the exercise described in the given sheet.
- Work together, perhaps spend about 20-25 minutes on the problem, and we will use the remaining time for discussion of what you did.
- This counts as your attendance for today.

Background Information

- n-grams are sequences of n-words.
- In the sentence: "I have a book", "I" is a word unigram/1-gram. "I have" is a word bi-gram/2-gram "I have a" is a word trigram/3-gram. "I have a book" is a word 4-gram. and so on.
- "I hav" is a character 5 gram with 5 characters - I, space, h, a, v
- If I have a large collection of text files (called **corpus**), I can compile large dictionaries of such word and character n-grams of any arbitrary n, and how frequently are they seen in this corpus.
- An example of such a large text corpus is: Wikipedia

Next Week

- Topic 1: Isolated spelling error correction
- Topic 2: Contextual spelling error correction
- Deadlines: Assignment 1 deadline on September 8th.
- Reminder: No class on Monday.
- Readings: Finish reading Chapter 2