LING 120: Language and Computers

Semester: Fall 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

6 September 2017

Class outline

1. Review of last week
2. Isolated word spelling error correction
    - Soundex
    - Edit distance

Reminder: Assignment 1 is due this week!!

Recap of last week

- Different kinds of errors, causes of spelling errors etc.
- The idea of N-grams and where they can be useful (e.g., contextual error correction, speech recognition, human-machine conversation etc).
- Non-word error detection (e.g., using large dictionaries covering different word forms + a set of rules to ignore certain words such as words beginning in upper case etc)

Recap of last week

- Different kinds of errors, causes of spelling errors etc.
- The idea of N-grams and where they can be useful (e.g., contextual error correction, speech recognition, human-machine conversation etc).
- Non-word error detection (e.g., using large dictionaries covering different word forms + a set of rules to ignore certain words such as words beginning in upper case etc)
- Review questions:
  - What is a non-word spelling error?

Recap of last week

- Different kinds of errors, causes of spelling errors etc.
- The idea of N-grams and where they can be useful (e.g., contextual error correction, speech recognition, human-machine conversation etc).
- Non-word error detection (e.g., using large dictionaries covering different word forms + a set of rules to ignore certain words such as words beginning in upper case etc)
- Review questions:
  - What is a non-word spelling error?
  - What is a real-word spelling error?

Recap of last week

- Different kinds of errors, causes of spelling errors etc.
- The idea of N-grams and where they can be useful (e.g., contextual error correction, speech recognition, human-machine conversation etc).
- Non-word error detection (e.g., using large dictionaries covering different word forms + a set of rules to ignore certain words such as words beginning in upper case etc)
- Review questions:
  - What is a non-word spelling error?
  - What is a real-word spelling error?
  - How do we categorize grammar errors? non-word or real-word?

Isolated word spelling correction

- Aim: suggest correction candidates for a mis-spelled word, irrespective of the surrounding words.
- 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.
- What are the three steps to do isolated word spelling correction?
  - Detect a error (let us say: dictionary based, or looking for improbable character n-grams)
  - Identify possible candidates for right spelling (from where?? - topic for today)
  - Rank them in terms of the most probable word (how? - topic for today)

Generating possible candidates for right spelling

- This is typically based on some notion of similarity between words.
- i.e., let us say I see a non-word error. The most intuitive way to look for correct words is to look for the closest words for this word in the dictionary.

Generating possible candidates for right spelling

- This is typically based on some notion of similarity between words.
- i.e., let us say I see a non-word error. The most intuitive way to look for correct words is to look for the closest words for this word in the dictionary.
- Let us say I typed "assingment" and a dictionary look-up told me it is a non-word (or may be the character trigram "ngm" never comes in English). What is the right word? How can you say?

Generating possible candidates for right spelling

- This is typically based on some notion of similarity between words.
- i.e., let us say I see a non-word error. The most intuitive way to look for correct words is to look for the closest words for this word in the dictionary.
- Let us say I typed "assingment" and a dictionary look-up told me it is a non-word (or may be the character trigram "ngm" never comes in English). What is the right word? How can you say?
- How do we quantify the notion of "closest word" and show possible candidates??

Grouping similar words: SOUNDEX algorithm

- Idea: Represent words in such a way that similar sounding words get the same representation
- Background: Was originally used to group different spelling variations of names in U.S. Census Data.
- Input: a Name
- Output: a code in which a alphabetic character is followed by 3 numbers.
- Using it: since similar sounding words will have same Soundex code, whenever you see a non-word error, look in the dictionary for words with same Soundex code.
- These will be the candidate words to suggest as correct spellings.

Soundex Procedure

1. Retain the first letter of the name and drop all occurrences of a, e, i, o, u, y, h, w.
2. Replace consonants with digits in the following manner:
   b, f, p, v : 1;
   c, g, j, k, q, s, x, z : 2
   d,t : 3
   l: 4
   m,n : 5
   r : 6
3. If there are two adjacent letters with same number (e.g., jj or cz, replace only with a single number)
4. If there are too few letters in the name, append zeros until there are three numbers.
5. If the name is too long, cut the SOUNDEX after 3 numbers.

(Note: There are several variants for this)
refer: `https://www.archives.gov/research/census/soundex.html`)

Using this idea in practice

One simple way:

- Convert all entries in a dictionary into soundex form, and build a new form of mapping where the entry is the soundex code and its values are all words with that soundex code in the original dictionary.

Using this idea in practice

One simple way:

- Convert all entries in a dictionary into soundex form, and build a new form of mapping where the entry is the soundex code and its values are all words with that soundex code in the original dictionary.
- Each time you see a spelling error in what the user typed:
    1. Calculate the soundex for that word and look for the entry with this soundex in the newly created mapping.
    2. All these words will be the possible candidates for the mis-spelt word!

Ranking candidate words: The idea of Edit Distance

- Idea: define a "distance" measure to find the closest word to a mis-spelling.
- How to calculate the distance?: in terms of the number of transformations such as: insertions, deletions, substitutions, transposition of characters etc.

Ranking candidate words: The idea of Edit Distance

- Idea: define a "distance" measure to find the closest word to a mis-spelling.
- How to calculate the distance?: in terms of the number of transformations such as: insertions, deletions, substitutions, transposition of characters etc.
- Example 1: if assignment was written as assingment, there is a case of one transposition (gn and ng).
- Example 2: if assignment was written as asignment, there is a case of one deletion.

Ranking candidate words: The idea of Edit Distance

- Idea: define a "distance" measure to find the closest word to a mis-spelling.
- How to calculate the distance?: in terms of the number of transformations such as: insertions, deletions, substitutions, transposition of characters etc.
- Example 1: if assignment was written as assingment, there is a case of one transposition (gn and ng).
- Example 2: if assignment was written as asignment, there is a case of one deletion.
- If each such transformation gets a score, the distance between a mis-spelt word and any valid word is the total transformation score.

Solving the edit distance problem

- For a human, edit distance problem seems like something related to your intuition of language.
- For a computer, it is a problem of choosing the "minimum edit distance" out of 1000s of possible options (i.e., to get from the mis-spelt "assingment" to correct candidate "assignment", it is also possible to first insert "ooooooooo" after "a" and deleting "ooooooooo" after a.

Solving the edit distance problem

- For a human, edit distance problem seems like something related to your intuition of language.
- For a computer, it is a problem of choosing the "minimum edit distance" out of 1000s of possible options (i.e., to get from the mis-spelt "assingment" to correct candidate "assignment", it is also possible to first insert "ooooooooo" after "a" and deleting "ooooooooo" after a.
- Okay, that is an extreme example, but a computer has to examine multiple options before finding the minimum edit distance option.
- This is done by something called "Dynamic Programming" in Computer Science. The basic intuition is that you solve the problem step by step (letter by letter), and finally solve the full problem by combining solutions to all these small problems.

Using this idea in practice

One simple way:

- Each time you see a spelling error in what the user typed:
  1. Get all words within an edit-distance of say one or two points.
  2. All these words can be the possible candidates for the mis-spelt word!
  3. You can rank them based on distance.

Ranking candidate words: Another solution

Using the notions of

- Transition probabilities (what is the probability that the next letter is "a" if the current letter is "b")
- Confusion probabilities (What is the probability that "k" gets confused as "l" in typing a word)

Isolated Word Spelling Correction: Conclusion

- What we discussed are a few simple ways of looking for possible corrections for a spelling error.
- Real-life solutions build on these approaches, and add more sophistication to that (More on that if you study further and take a course on natural language processing in future).

Next Class

1. Context sensitive real-word error correction
2. Readings for next class: Chapter 2
3. Additional References for today's topic:
    - Explaining how a computer calculates edit distance with dynamic programming: `https://www.youtube.com/watch?v=We3YDTzNXEk`
    - Also read the "Under the Hood 3" box in Chapter 2.

Attendance exercise for today

Visit `www.eogn.com/soundex/` and list a few examples of where Soundex will fail if you use it for spelling correction. Use some non-word errors and their correct versions and check which pairs have the same soundex and which pairs don't. Try to come up with 2 examples for each case and analyse why are the soundex codes same or different.
Example: Assignment can have two mis-spellings (among others). Asignment, Assingment - First one will have the same soundex as the original word. Second one won't have. Why? (sounds are different)