LING 120: Language and Computers

Semester: Fall 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

8 September 2017

Class Outline

1. Last class' question
2. Context sensitive spelling correction
3. Assignment 2 description

Question from last class

Visit `www.eogn.com/soundex/` and list a few examples of where Soundex will fail. 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)

Your Answers

1. Cases where Soundex will work and show a correct alternative:
    - Calculate, calculade (C-424); Discussion-Discusion (D-225)

Your Answers

1. Cases where Soundex will work and show a correct alternative:
   - Calculate, calculade (C-424); Discussion-Discusion (D-225)
   - But, there are also these false positives that can come up: Discussion-Decagon (D225!)
     (for more such false positive examples, check out: `https://goo.gl/2sz3yW`)

Your Answers

1. Cases where Soundex will work and show a correct alternative:
   - Calculate, calculade (C-424); Discussion-Discusion (D-225)
   - But, there are also these false positives that can come up: Discussion-Decagon (D225!)
     (for more such false positive examples, check out: `https://goo.gl/2sz3yW`)
2. Cases where Soundex will not work:
   - Discussion-Discssion (D-225, D-250)
   - laf, laugh - L100 and L200
   - night, nite - N-230, N-300
3. Another major problem: For long words, you may never get right suggestions (Why?)

Your Answers

1. Cases where Soundex will work and show a correct alternative:
   - Calculate, calculade (C-424); Discussion-Discusion (D-225)
   - But, there are also these false positives that can come up: Discussion-Decagon (D225!)
     (for more such false positive examples, check out: `https://goo.gl/2sz3yW`)
2. Cases where Soundex will not work:
   - Discussion-Discssion (D-225, D-250)
   - laf, laugh - L100 and L200
   - night, nite - N-230, N-300
3. Another major problem: For long words, you may never get right suggestions (Why?) Revolution, Revolutionize, Revolutionary, Revolutionist, Revolutionization- all get the same Soundex!

Today's topic: Context Sensitive Spelling Correction

What is the problem to solve?

- ... detecting and correcting real word spelling errors.
- i.e., words are not spelt wrong - they are spelt wrong in that context.
- Grammar checking is considered a context sensitive spelling correction process.
- Since everything is dependent on context, can we say every word is a potential error? (What? How?)

Some examples

- Let us take this sentence: "There house is nice". There are two possible "correct" options.
  - The house is nice.
  - Their house is nice. (more likely)
  - The house there is nice.
- Or another: "The teams was successful". There are again two possible "correct" options.
  - The team was successful.
  - The teams were successful. (can we say for certain what is more likely?)
- How does a computer go about detecting such errors?

What are the causes of contexual word errors?

- "False Friends": *bekommen* in German means *get*. So, a German native speaker, when writing English may confuse between become and get.

What are the causes of contexual word errors?

- "False Friends": *bekommen* in German means *get*. So, a German native speaker, when writing English may confuse between become and get.
- Words sound the same: their vs there.

What are the causes of contexual word errors?

- "False Friends": *bekommen* in German means *get*. So, a German native speaker, when writing English may confuse between become and get.
- Words sound the same: their vs there.
- Influence of the sentence structure in the writer's native language (many Indian English speakers make errors with articles because several languages do not have them).

What are the causes of contexual word errors?

- "False Friends": *bekommen* in German means *get*. So, a German native speaker, when writing English may confuse between become and get.
- Words sound the same: their vs there.
- Influence of the sentence structure in the writer's native language (many Indian English speakers make errors with articles because several languages do not have them).
- Not knowing the rules of the language (e.g., subject-verb agreement. *He has* but not *He have*

What are the causes of contexual word errors?

- "False Friends": *bekommen* in German means *get*. So, a German native speaker, when writing English may confuse between become and get.
- Words sound the same: their vs there.
- Influence of the sentence structure in the writer's native language (many Indian English speakers make errors with articles because several languages do not have them).
- Not knowing the rules of the language (e.g., subject-verb agreement. *He has* but not *He have*
- Gender errors due to native language background (e.g, one language has the Gender for Sun as male. The other one has female.)

...

Correcting such errors

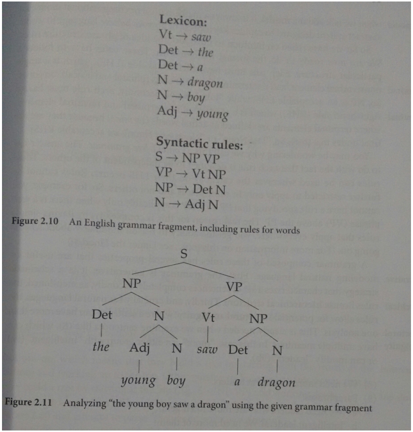
- Grammar based word correction
- Error pattern based word correction
- Probability based word correction
- Meaning based word correction

Grammar based word correction

- Idea: encode the rules of language into a computer program. As the program tries to build a syntactic structure of the language, if there is no matching grammar rule, then it breaks, which is an indication that the sentence has an error.

Grammar based word correction

- Idea: encode the rules of language into a computer program. As the program tries to build a syntactic structure of the language, if there is no matching grammar rule, then it breaks, which is an indication that the sentence has an error.
- This is what I mean when I said "syntactic structure".



Lexicon:
Vt → *saw*
Det → *the*
Det → *a*
N → *dragon*
N → *boy*
Adj → *young*

Syntactic rules:
S → NP VP
VP → Vt NP
NP → Det N
N → Adj N

Figure 2.10 An English grammar fragment, including rules for words

S
NP
VP
Det
N
Vt
NP
the
Adj
N
saw
Det
N
young
boy
a
dragon

Figure 2.11 Analyzing "the young boy saw a dragon" using the given grammar fragment

Pros and Cons

- Pros: Very effective - specific error identification and feedback is possible.
- Cons: Such a grammar has to be painstakingly prepared first, and suiting to the needs of a computational approach (which is time consuming and expensive)
- Language is continuously changing - so new expressions, new syntactic structures keep coming. This approach will not work if we don’t keep updating.

Error pattern based word correction

- Idea: Prepare a large set of rules of error patterns in language. Whenever there is a match to a rule, flag an error and suggest a solution (e.g. rule: if a plural word is followed by *is*, flag an error, and suggest using *are* instead).

Error pattern based word correction

- Idea: Prepare a large set of rules of error patterns in language. Whenever there is a match to a rule, flag an error and suggest a solution (e.g. rule: if a plural word is followed by *is*, flag an error, and suggest using *are* instead).
- Pros: As long as we know what kind of errors the users make, this is a straight forward and effective process
- Cons: Rule making takes time and expertise and money. Again, may be it is difficult to exhaustively cover every single error pattern.
- However, all spell checkers you use today have some kind of a rule-engine inside it, along with something based on an n-gram approach.

Probability based word correction

- Idea: use frequencies of word n-grams in a large collection of texts to estimate what are the likely and unlikely n-grams in the language.

Probability based word correction

- Idea: use frequencies of word n-grams in a large collection of texts to estimate what are the likely and unlikely n-grams in the language.
- E.g, Let us take this sentence: "John came form the house".
- Since we don't know what word is an error in this context, let us start with the assumption that each word is a potential error.
- Candidate words: let us leave John (poor guy!). Came - come, lame, tame, tame, cane etc; form: from, dorm, norm etc.; house - hos

getting from word level candidates to sentence level suggestions

- Try to make sentences with all these possible candidates replacing one candidate word at a time.
- From a large corpus of English texts, estimate the likelihood of seeing each sentences (probability)
- The sentence with highest probability gets the top-rank in suggestion list.

Pros and Cons

- Pros: Works around the problem of writing a lot of language specific rules which requires time and effort.
- Cons: Lot of calculations with large corpus-computationally intensive! (thankfully, computer programs have efficient way of organizing and retrieving data)
- Cons: No direct way to handle unknown words or phrases

Note: Real life spelling and grammar checkers use a combination of error pattern based and probability based methods.

Meaning based word correction

- Idea: Find words that do not fit into the meaning of the rest of the sentence. Replace them with words that are "semantically appropriate"
- e.g., "It is my sincere *hole* that you will recover swiftly" - hole seems semantically inappropriate.
- "hope" suits better here.
- Problem: How do you choose the related word given a context? - this is studied in natural language processing under something called "distributional representation of language".

Assignment 2 Description

- Two questions - one each on isolated and contextual spelling correction
- Requires you to analyze what the word processor tools show you and interpret the causes of the errors and suggestions.
- Carries 10 marks.
- Guidelines are on Canvas.
- Deadline: 23 September 2017

Next Week

- Topic 3: Language Tutoring Systems
- Readings: Chapter 3 from the Textbook
- Reminder: Submit Assignment 1!

Attendance Question for today

Write answers to any two of these scenarios on a sheet of paper and return it to me with your name on it.

- 2 examples of grammar errors caused due to a change in word order
- 2 examples of grammar errors caused due to usage of wrong tense
- 2 examples of grammar errors caused because of gender-differences between the author's native language and English
- 2 examples of using similar sounding words instead of each other (e.g., their-there, hole-whole -now, don't use these examples!)