LING 120: Language and Computers
Semester: FALL 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

30 Aug 2017

Class outline

- Question from last class
- Writing aids: Introduction
- Spelling correction: overview
- Grammar correction: overview
- Style checkers: overview

Speech Recognition and Back to Speech

If I record you speak, use ASR and convert it to text, and then use TTS to convert it again to speech - do you think I will get your speech sample back? Write a short explanation for your answer (few sentences). You can either write it now and give it, or post on Canvas forum for today and get attendance.

Your Answers

- Answer 1: No, because the TTS will have a built-in voice. Also, ASR could make mistake of disfluencies in speech.
- Answer 2: No, there will be errors in software when processing so much of data.
- Answer 3: Depends on the length of the sample. If it is one word, it may work.
- Answer 4: No. You can make something simple from complex, but not complex from simple.
- Answer 5: Yes they would be the same because ASR is speech to text, TTS is text to speech.
- Answer 6: No, because speech sample is deleted after speech recognition.
- Answer 7: No, even if everything is accurate, my specific tone/accent etc are gone.

Why it is a "no"

- There will be errors in ASR, and in TTS
- Specifics of a voice will be gone even if the TTS has my own customized voice.
- Depends on the length of the sample - is a great point! Haven't thought of that before.

Today's joke from Iowa State Daily
Something you will appreciate now

Q: What do you call a deer with no eyes?

Today's joke from Iowa State Daily

Something you will appreciate now

Q: What do you call a deer with no eyes?
A: No eye deer.

Writers' Aids: An overview

Topic 2: Writers' Aids

- Spelling correction
- Grammar correction
- Style checkers

Look at this text

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

- Keeping theories aside, clearly, we are able to make sense out of a mis-spelt text.
- So why bother about learning spelling, about winning spelling bee etc?

- Keeping theories aside, clearly, we are able to make sense out of a mis-spelt text.
- So why bother about learning spelling, about winning spelling bee etc?

Look at another text

My name is Susan. I'm forteen and I life in Germany. My hobbys are go to discos, sometimes I hear music in the radio. In the summer I go bathing in a lake. I haven't any brothers or sisters. We take busses to scool. I visit year 9 at my school. My birthday is on Friday. I hope I will become a new guitar. I'm looking forward to get a e-mail from you.
source: http://www.englisch-hilfen.de/en/exercises/
structures/error_text_letter.htm

- ▶ The text is written in a non-standard fashion, but it is not incomprehensible.
- ▶ Why bother about grammar then?? Are we just being too picky??

Why bother about spelling

- Mis-spellings can sometimes makes it difficult to understand what is being written
- Standard spelling makes it possible to communicate clearly across multiple cultures, dialects etc.
- Difficult to find the meaning of a mis-spelt word from a dictionary lookup (both for humans and for computers)
- If a name is mis-spelled, it is even more difficult to look up.
- For a machine, it is even more difficult than humans.
- It makes Text to speech conversion difficult
- In formal settings, mis-spellings will not give a favorable impression about the author

Why bother about grammar

- That example was an easy one. Bad grammar can cause misunderstandings due to miscomprehension
- It also poses problems for machines that interact with human language
- good writing includes proper grammar, and again, in formal settings, bad grammar does not give a good impression on the author.

Where/How are spell checkers used?

- Some are interactive spelling checkers - as you type, they highlight your mistakes, and offers you choices.
  $\Rightarrow$ computer has to quickly respond to your speed.

- Some give suggestions after you finish the entire thing

- Sometimes, they even automatically correct words whenever the spell checker thinks you are using wrong spelling (on phones, especially)
  $\Rightarrow$ harder than just detecting errors, but also potentially risky if the writer does not proof read.

Spelling Error Detection and Correction

- What is more difficult? detection or correction?

Spelling Error Detection and Correction

- What is more difficult? detection or correction?
- Let us take this sentence: "I will come back ater break".
  "ater" is the error - perhaps easy to detect. But what is the correction? later? water? after?

Spelling Error Detection and Correction

- What is more difficult? detection or correction?
- Let us take this sentence: "I will come back ater break". "ater" is the error - perhaps easy to detect. But what is the correction? later? water? after?
- Even detection can be difficult! Let us say there is this sentence: "Toll and sort people"
  - there is no spelling error if you look word by word. But clearly there is something wrong.

Spelling Error Detection and Correction

- What is more difficult? detection or correction?
- Let us take this sentence: "I will come back ater break". "ater" is the error - perhaps easy to detect. But what is the correction? later? water? after?
- Even detection can be difficult! Let us say there is this sentence: "Toll and sort people"
  - there is no spelling error if you look word by word. But clearly there is something wrong.
- Note: You need to detect to be able to correct.

Detecting and Correcting Grammar Errors

Again, two possible ways.

- ▶ Create (and program) large set of language rules (e.g., patterns such as "I have" but not "I has" is correct etc)

Detecting and Correcting Grammar Errors

Again, two possible ways.

- ▶ Create (and program) large set of language rules (e.g., patterns such as "I have" but not "I has" is correct etc)
- ▶ Or use large collections of **n-gram** frequencies to identify abnormal word combinations and tag them as grammar errors.

    - ▶ n-grams are sequences of n-words.
    - ▶ "I" is a unigram/1-gram. "I have" is a bi-gram/2-gram "I have a" is a trigram/3-gram. "I have a book" is a 4-gram. and so on.
    - ▶ Assuming I have large collections of n-gram frequencies from English language, any new n-gram I see (e.g., "I has a book" may have a frequency of 0 and hence it is a grammar error).

Detecting and Correcting Grammar Errors

Again, two possible ways.

- ▶ Create (and program) large set of language rules (e.g., patterns such as "I have" but not "I has" is correct etc)
- ▶ Or use large collections of **n-gram** frequencies to identify abnormal word combinations and tag them as grammar errors.

    - ▶ n-grams are sequences of n-words.
    - ▶ "I" is a unigram/1-gram. "I have" is a bi-gram/2-gram "I have a" is a trigram/3-gram. "I have a book" is a 4-gram. and so on.
    - ▶ Assuming I have large collections of n-gram frequencies from English language, any new n-gram I see (e.g., "I has a book" may have a frequency of 0 and hence it is a grammar error).

- ▶ Use this frequency information to then suggest alternatives to the writer.

Style Checkers

- These are related to spelling and grammar checkers, but are more "prescriptive" in nature
- If an organization has a certain style of writing (e.g., don't use passive voice, use data as a plural word etc.), the job of a style checker is to make sure authors stick to that style.

Style Checkers

- These are related to spelling and grammar checkers, but are more "prescriptive" in nature
- If an organization has a certain style of writing (e.g., don't use passive voice, use data as a plural word etc.), the job of a style checker is to make sure authors stick to that style.
- They would still require all the analysis that a spelling and grammar checker would demand.

More on Spelling Check and Spelling Correction

Causes of spelling errors on computers
Keyboard usage

- ▶ missing a space while typing can create a typo (the man becoming theman)
- ▶ or accidentally inserting a space.
- ▶ Note: both these can also result in real words.

Causes of spelling errors on computers
Keyboard usage

- missing a space while typing can create a typo (the man becoming theman)
- or accidentally inserting a space.
- Note: both these can also result in real words.
- proximity on keyboard: ran and tan (r and t are neighbors), okay and play, might and mihgt etc.

Causes of spelling errors on computers
Based on how they sound

- Two words sounding the same (read-red, cite-sight, their-there etc)

Causes of spelling errors on computers

Based on how they sound

- ▶ Two words sounding the same (read-red, cite-sight, their-there etc)
- ▶ Replacing a letter with similar sound (krack-crack, hole-whole etc)

Causes of spelling errors on computers
Based on how they sound

- Two words sounding the same (read-red, cite-sight, their-there etc)
- Replacing a letter with similar sound (krack-crack, hole-whole etc)
- Guessing spelling by hearing the sound (sientist)

... there are also other kinds of errors.

How does a spell checker work with these errors?

Broadly, we can classify all spelling and grammar error issues into three kinds of problems:

- ▶ non-word error detection and correction
- ▶ isolated word error detection and correction
- ▶ context dependent word error detection and correction (i.e., grammar check and correction)

Non-word errors

Isn't this easy? Why not just have a dictionary?

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- ▶ All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)
- ▶ What about plurals, past-tense and other "inflections" (should I have separate entries for eat, ate, eaten; car, cars etc?)

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)
- What about plurals, past-tense and other "inflections" (should I have separate entries for eat, ate, eaten; car, cars etc?)
- What should I do about words like don't?

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)
- What about plurals, past-tense and other "inflections" (should I have separate entries for eat, ate, eaten; car, cars etc?)
- What should I do about words like don't?
- Or those such as "at least" which are multi-word units?

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- ▶ All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)
- ▶ What about plurals, past-tense and other "inflections" (should I have separate entries for eat, ate, eaten; car, cars etc?)
- ▶ What should I do about words like don't?
- ▶ Or those such as "at least" which are multi-word units?
- ▶ Hyphenated and unhyphenated forms of the same word?

Non-word errors

Isn't this easy? Why not just have a dictionary?
What should the dictionary contain?

- All words? What about capitalization, punctuation? (are Yes, yes, yes: the same word?)
- What about plurals, past-tense and other "inflections" (should I have separate entries for eat, ate, eaten; car, cars etc?)
- What should I do about words like don't?
- Or those such as "at least" which are multi-word units?
- Hyphenated and unhyphenated forms of the same word?
- Abbreviations
- Names (how many names can I add to dictionary?)
- Foreign words?

Okay, then?

Assuming we take some decision about these, two tasks remain related to dictionary:

- ► Constructing a dictionary
- ► Figuring out some way to quickly retrieve information from the dictionary (how can I pick 10000th word in a dictionary of 6 million words in 0.5 seconds?)

Okay, then?

Assuming we take some decision about these, two tasks remain related to dictionary:

- ▶ Constructing a dictionary
- ▶ Figuring out some way to quickly retrieve information from the dictionary (how can I pick 10000th word in a dictionary of 6 million words in 0.5 seconds?)

... additionally, we will need some way to accomodate other aspects such as inflections, names etc.

(continued in next class)

Next Class ..

- Topic: Isolated word spelling correction - details
- Readings: Read up to Section 2.3 in Chapter 2 in the textbook.

Attendance exercise

Open MS Word on the lab computer and give examples for
different kinds of spelling errors (nonword, isolated word,
contextual) and the suggestions given by the software. You can
post in the forum for today on Canvas.