Language and Linguistics

Winston Wu

9/17/2020
From Last Class

• How does n affect the generated sentences?
• Any interesting results?
  • **Bigrams:** "How 'Up in London?" Harry got someone's hand. "Unusually bright." Professor McGonagall's head in the next few words on to the more important day to Mrs. Figg.
  • **Trigrams:** "How is Harry Potter, just as Ron found him. "See?" said Hagrid. "An ' see here, Professor Snape." And it didn't even listen to her.
  • **4-grams:** "yeh, yer wrong!" said Harry, trying to knock in a nail with the piece of fruitcake Aunt Petunia had rushed them out of the ground. The cat's tail twitched and its eyes narrowed. Nothing like this man had ever been seen on Privet Drive.
  • **5-grams:** He was wearing long robes, a purple cloak that swept the ground, and high-heeled, buckled boots. His blue eyes were light, bright, and sparkling behind half-moon spectacles and his nose was still pink where his mother had rubbed it.
Discussion

“All GPT-3 really has is a tunnel-vision understanding of how words relate to one another; it does not, from all those words, ever infer anything about the blooming, buzzing world.”

Thoughts?
Today

• Writing Systems
• Words and Morphology
• Language in 5 assignment
How do we categorize languages?
Language

• There are around 7000 languages and dialects
• A system of communication
  • Exact definitions depend on your perspective
• For us, language is a **sequence**
  • Text is a sequence of sentences/words/characters
  • Speech is a sequence of sound waves
  • Sign language is a sequence of motions
Languages are related! Stay tuned for our lecture on Historical Linguistics.
Typology

• Group languages according to their **features**
• One example: Word Order
  • SOV (Japanese, Tamil, Turkish)
  • SVO (English, Chinese, Fula)
  • VSO (Arabic, Tongan, Welsh)
  • The others (VOS, OVS, OSV) are not common
  • 95% of languages, Subject comes before Object
    • Why?

SOV (Japanese)
Watashitachi wa Nihongo o hanasu.
we TOP Japanese OBJ speak
'We speak Japanese.'

SVO (English)
He ate the pudding.

VSO (Arabic)
Qatala l- malik-u l- malikat-a
kill DEF king NOM+DEF DEF queen ACC
'The king killed the queen.'
Vocabulary: Not Typology

• Fula has 82 words for cattle
• Chinese has many kinship terms
  • tangge 'elder male paternal cousin'
  • tangdi 'younger male paternal cousin'
  • tangjie 'elder female paternal cousin'
  • tangmei 'younger female paternal cousin'
  • biaoge 'elder male maternal cousin'
  • biaodi 'younger male maternal cousin'
  • biaojie 'elder female maternal cousin'
  • biaomei 'younger female maternal cousin'
Writing Systems

- A set of symbols (graphemes) and rules (orthography)
- **Orthography**: a set of rules for writing a language, spelling, hyphenization, capitalization, word breaks, emphasis, punctuation
Writing Systems: Logographic

• Each symbol (logogram) represents a word or morpheme
  • Early symbols were based on pictures of what they signified
  • Chinese, Mayan glyphs, Cuneiform, Egyptian hieroglyphs
  • Latin script also has some, e.g. ‘2’ or ‘&’
  • Most are logosyllabic (each character is a syllable)

https://www.zdic.net/hans/魚
Writing Systems: Syllabary

• Each symbol represents a syllable or mora
  • Japanese, Cherokee, Linear B
• Often derived from logograms
Writing Systems: Alphabet

- Each symbol (letter) represents a speech sound
- Latin, Greek, Cyrillic alphabets

Vowels
- a e i o u y æ ø
- [a] [e] [i] [o] [u] [y] [æ] [ø]

Diphthongs
- ai au ei eu ua uе ui uо
- [ai] [au] [ei] [eu] [ua] [ue] [ui] [uo]

Consonants
- b c ch d f g gn h j
- [b] [cʰ] [d] [f] [g] [ŋ] [h] [j]
- k l m n p ph q r s
- [k] [l] [m] [n] [p] [f] [k] [r] [s] [z]
- sc th ti v x z
- [skʃ] [t] [tʰ] [ts/tʃ] [v] [ks/gz] [dz]

<table>
<thead>
<tr>
<th>α</th>
<th>A</th>
<th>alpha</th>
<th>a</th>
<th>father</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>B</td>
<td>beta</td>
<td>b</td>
<td>big</td>
</tr>
<tr>
<td>γ</td>
<td>G</td>
<td>gamma</td>
<td>γ, n</td>
<td>God, ankle</td>
</tr>
<tr>
<td>δ</td>
<td>Δ</td>
<td>delta</td>
<td>d</td>
<td>door</td>
</tr>
<tr>
<td>ε</td>
<td>Ε</td>
<td>epsilon</td>
<td>e</td>
<td>mgt</td>
</tr>
<tr>
<td>ζ</td>
<td>Z</td>
<td>zeta</td>
<td>z, dz</td>
<td>zeal, kugzu</td>
</tr>
<tr>
<td>η</td>
<td>H</td>
<td>eta</td>
<td>η</td>
<td>obey</td>
</tr>
<tr>
<td>θ</td>
<td>Θ</td>
<td>theta</td>
<td>θ</td>
<td>thing</td>
</tr>
<tr>
<td>I</td>
<td>Iota</td>
<td>i</td>
<td>pit, police</td>
<td></td>
</tr>
<tr>
<td>Κ</td>
<td>K</td>
<td>kappa</td>
<td>k</td>
<td>keep</td>
</tr>
<tr>
<td>Λ</td>
<td>Lambda</td>
<td>l</td>
<td>law</td>
<td></td>
</tr>
<tr>
<td>Μ</td>
<td>m</td>
<td>mu</td>
<td>m</td>
<td>mother</td>
</tr>
<tr>
<td>Ν</td>
<td>n</td>
<td>nu</td>
<td>n</td>
<td>number</td>
</tr>
<tr>
<td>Ξ</td>
<td>xi (ksee)</td>
<td>x</td>
<td>fox</td>
<td></td>
</tr>
<tr>
<td>Ο</td>
<td>O</td>
<td>omicron</td>
<td>o</td>
<td>ngt</td>
</tr>
<tr>
<td>Π</td>
<td>pi</td>
<td>p</td>
<td>poor</td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>P</td>
<td>rho</td>
<td>r, rh</td>
<td>rod</td>
</tr>
<tr>
<td>σ</td>
<td>Sigma</td>
<td>s</td>
<td>save</td>
<td></td>
</tr>
<tr>
<td>τ</td>
<td>T</td>
<td>tau</td>
<td>τ</td>
<td>time</td>
</tr>
<tr>
<td>υ</td>
<td>upsilon</td>
<td>u, y</td>
<td>German ü</td>
<td></td>
</tr>
<tr>
<td>φ</td>
<td>Phi</td>
<td>phi</td>
<td>phone</td>
<td></td>
</tr>
<tr>
<td>χ</td>
<td>Chi</td>
<td>ch</td>
<td>German ich</td>
<td></td>
</tr>
<tr>
<td>ψ</td>
<td>Psi</td>
<td>ps</td>
<td>tipsy</td>
<td></td>
</tr>
<tr>
<td>Ω</td>
<td>Omega</td>
<td>o</td>
<td>vote</td>
<td></td>
</tr>
</tbody>
</table>

14
Writing Systems: Abjad

- Each symbol represents a consonant
- Phoenician, which gave rise to...
- “Impure abjad” has diacritics to mark vowels
- Aramaic, which gave rise to Arabic, Hebrew scripts

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Written Position</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>vowel written below consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>vowel written below consonant</td>
<td>[ŋ] with succeeding yod</td>
</tr>
<tr>
<td>9</td>
<td>vowel written below consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>vowel written below consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>vowel written below consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>vowel written above consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>vowel written above consonant</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>isolated vowel written on its own</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>9</td>
<td>isolated vowel written on its own</td>
<td>[ŋ]</td>
</tr>
</tbody>
</table>
Writing Systems: Abugida

- Represents consonants and vowels, but the vowels are secondary
  - Indic scripts, Ethiopic (Ge’ez)

<table>
<thead>
<tr>
<th>position</th>
<th>syllable</th>
<th>pronunciation</th>
<th>base form</th>
<th>script</th>
</tr>
</thead>
<tbody>
<tr>
<td>above</td>
<td>कृ</td>
<td>/ke/</td>
<td>क /k(a)/</td>
<td>Devanagari</td>
</tr>
<tr>
<td>below</td>
<td>कृ</td>
<td>/ku/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left</td>
<td>कि</td>
<td>/ki/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td>को</td>
<td>/ko/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>around</td>
<td>कृ</td>
<td>/kə/</td>
<td>क /ka/</td>
<td>Tamil</td>
</tr>
<tr>
<td>surround</td>
<td>कृ</td>
<td>/kie/</td>
<td>क /ka/</td>
<td>Khmer</td>
</tr>
<tr>
<td>within</td>
<td>कृ</td>
<td>/ki/</td>
<td>क /ka/</td>
<td>Kannada</td>
</tr>
<tr>
<td>within</td>
<td>कृ</td>
<td>/ki/</td>
<td>क /ka/</td>
<td>Telugu</td>
</tr>
<tr>
<td>below and extend to the right</td>
<td>कृज</td>
<td>/kja/</td>
<td>क /ka/</td>
<td>Javanese</td>
</tr>
<tr>
<td>below and extend to the left</td>
<td>कृज</td>
<td>/kru/</td>
<td>क /ka/</td>
<td>Javanese</td>
</tr>
</tbody>
</table>
Writing Systems: Other

- Abugida: each symbol is consonant+vowel
  - Brahmic scripts
  - Vowel diacritics in Devanagari

क का कि की कू कृ कू के के को को

- Featural writing system: each symbol represents a feature
  - Hangul

[Diagram of Hangul characters]
Challenges for NLP

• Most research has been done on Western languages
  • Latin script
  • Latin script (~50 characters) vs Chinese characters (~20,000)
• Transliteration, Transcription, Romanization
  قطر → qṭr or qaṭar

• Some languages can be written in multiple scripts
  • Japanese has hiragana, katakana, kanji
  • Turkic languages can be written in Arabic, Cyrillic, and Latin script

"I like coffee”

コヒーが好きです

coohiī  ga  su.ī  desu

katakana  hiragana
A Case Study: Latin

- Affixes encode the word’s function in the sentence
- The verb encodes the subject
  - ‘ego’ is optional
- Same as
  - tibi librum do
  - librum tibi do
  - do tibi librum
  - ...
- Different from
  - ego te libro do
  - mihi tu librum das
  - me tibi liber dat
  - ...

<table>
<thead>
<tr>
<th>ego</th>
<th>tibi</th>
<th>librum</th>
<th>do</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOM</td>
<td>DAT</td>
<td>ACC</td>
<td>1st sg pres</td>
</tr>
<tr>
<td>I</td>
<td>you</td>
<td>book</td>
<td>give</td>
</tr>
</tbody>
</table>
Morphology

• Morphology is the study of form and structure
• First part of [http://mt-class.org/jhu/slides/lecture-words.pdf](http://mt-class.org/jhu/slides/lecture-words.pdf)
WALS

• World Atlas of Language Structures
• https://wals.info
Other resources

• Linguistic Fundamentals for Natural Language Processing (Bender, 2013)
  • Available for free through the JHU library
Language in 5 minutes

- [http://mt-class.org/jhu/lin10.html](http://mt-class.org/jhu/lin10.html)
  - Make it 5 minutes instead of 10
  - Must pick a language you are not familiar with
- See example [here](http://mt-class.org/jhu/lin10.html)
- Signup sheet will be posted on Piazza
Survey

1. How might the challenges we talked about today (regarding scripts, words, morphology) affect any of the tasks we talked about last week?
2. Questions, comments, concerns, suggestions