SO... COMPUTERS HAVE MASTERED PLAYING CHESS AND DRIVING CARS ACROSS THE DESERT, BUT CAN'T HOLD FIVE MINUTES OF NORMAL CONVERSATION?

PRETTY MUCH.

Source: xkcd

NATURAL LANGUAGE PROCESSING

amitabha mukerjee
iit kanpur
“मोहल्ले का एक लड़का”

“A monkey came in through the window and ate up my lunch.”
The magic of language

- Language is about conveying meaning
- Language is one-dimensional – Meaning is multi-dimensional

- Challenges
  Sounds along one-dimension express multi-dimensional aspects of reality
  - Same sounds map to different meanings [Polysemy]
  - Same meanings map to different sounds [Synonymy]
Myths about language

• grammar is about whether language is correct or incorrect
  
  *It’s me.*
  
  *Ganesh is at home?*
  
  *There are many small-small holes in this dress.*

• Modern view: grammar is about usage
  
  • descriptive, not prescriptive
Myths about language

• **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → Is Ganesh at home?

  It’s me (accusative) → “It’s I”

  There are many small-small holes in this dress.

• words are separated by spaces.

• how many sounds are there in English? 26
Myths about language

• **grammar** is about the correct and incorrectness of language.

  *Ganesh is at home? → Is Ganesh at home?*
  *It's me (accusative) → “It's I”*
  *There are many small-small holes in this dress.*

• words are separated by spaces.
  *words = meaningful bits of sound*

• alphabets are **not** the sounds of language
Levels of Linguistic Analysis

Phonology

Morphology

Syntax

Semantics

Boolean Logic:

\[ \exists x \exists y \text{ boy}(x) \land \text{ loc}(y) \land \text{lives-at}(x,y) \]
Semantics vs Pragmatics:

Find $x$. 

Diagram: Right triangle with sides 3 cm, 4 cm, and unknown side $x$. The hypotenuse is the side opposite the right angle.
Pragmatics: Direct vs Indirect meaning

Traditional thinking:
- Semantics
  - Direct meaning
- Pragmatics
  - Indirect meaning
Pragmatics

- You can’t hold two watermelons in one hand
- Iranian proverb
Pragmatics: Meaning in Context

Traditional levels of analysis:

- **Semantics**: composition from lexical meaning of words [direct meaning]

- **Pragmatics**: social / contextual meaning ; [indirect meaning]

Psycholinguists:

  Retrieval of pragmatic meaning is often faster
LEVELS OF STRUCTURE IN LANGUAGE
Language Structure: Levels

boys like girls
Language Structure: Levels

- Phonology
- Lexicon
- Syntax [Morphology]
- Discourse
- Semantics / Compositionality
- Pragmatics / Discourse
Language Structure: Levels

- **Phonology**: sounds of speech
  phoneme /b/ /oy/ /z/

- **Lexicon**: set of meaning-bearing units, *lexemes*

- **Syntax**: composing lexemes *composition*
  - **Word** = base + affixes / suffixes
  - **Phrase** = [ [ [boys ] like] girls]

- **Discourse**: Boy likes girl. They meet.
NLP: Goals

NL Understanding
Language $\rightarrow$ NLP $\rightarrow$ Decision

NL Translation
Language 1 $\rightarrow$ Machine $\rightarrow$ Language 2
Translation

NL description (Generation)
Situation $\rightarrow$ NLP $\rightarrow$ Language
Phonemes

• Which sounds change a meaning?
  
  *pin, tin, kin, fin, thin, sin, shin*
  
  *dim, din, ding, did, dig, dish*
  
  *pin, pen, pan, pun, pain, pine, pawn*

• **Phoneme** = minimum distinction in sound that changes meaning

• Phonemes at middle of syllable: **vowel**
  
  start or end: **consonant**
Vocal organs

tube model of vocal tract (for most neutral vowel)
Vowels: Formants

Formant frequencies: peaks in the harmonic spectrum of vowel sounds

first three: F1, F2, F3

http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html
Partitioning the speech sound space

[petitot 1989], [gardenfors 00]
## stop consonants

<table>
<thead>
<tr>
<th>Voiceless</th>
<th>Voiced</th>
<th>Nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td>inaspirate</td>
<td>aspirated</td>
<td>in-</td>
</tr>
<tr>
<td>क ख ग घ ङ</td>
<td>क्ष ज झ ञ</td>
<td>त थ द ध न</td>
</tr>
<tr>
<td>[velar]</td>
<td>[palatal]</td>
<td>[retroflex]</td>
</tr>
<tr>
<td>प फ ब भ म</td>
<td></td>
<td>[dental]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[labial]</td>
</tr>
</tbody>
</table>
## Consonants

<table>
<thead>
<tr>
<th>stop consonants</th>
<th>voiceless</th>
<th>voiced</th>
<th>nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>inaspirate</td>
<td>aspirated</td>
<td>in-</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
<td>g</td>
<td>gh</td>
</tr>
<tr>
<td>kh</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>chh</td>
<td>j[dz]</td>
<td>jh[dzh]</td>
</tr>
<tr>
<td>T</td>
<td>Th</td>
<td>D</td>
<td>Dh</td>
</tr>
<tr>
<td>t</td>
<td>th</td>
<td>d</td>
<td>dh</td>
</tr>
<tr>
<td>p</td>
<td>ph</td>
<td>b</td>
<td>bh</td>
</tr>
</tbody>
</table>
“cats” → “cat” + /s/

“boys” → “boy” + /z/
WHY ISN’T THE PLURAL OF SMURF SMURVES?

ONE HOUSE, TWO HICE.

For some reason, nobody wants to talk to us!

Source: urbanblah

Morphology and syntax
Syntax (morphosyntax)

- Regularity in how larger structures are assembled from units or smaller structures
  - **morphology**
    - cook-er / read-er / *-ercook
  - **phrase syntax**
    - smart woman / *woman smart
  - **sentence syntax**
    - boys like girls / girls like boys / *like boys girls
Lexicon vs Grammar

- Assumption: larger structures are assembled from smaller ones

- *Q. Is this assumption valid?*

- Smallest meaning-bearing structures = unit

- **morpheme**: less likely to appear independently
  - -er, -s, -ly, -able

- **lexeme**
  - cat, boy, smart, undergraduate student, cook, cooker
Lexicon vs Grammar

- lexicon = mental inventory of units = set of all lexemes

- Is “cats” a lexeme?

  cook $\rightarrow$ cooks : grammatical (rule-driven, inflection)
  $\rightarrow$ cooker : cook + er (not fully a rule; derivation)

Older thinking: lexicon is separate from grammar
at present: lexicon - grammar is a continuum
• **Syntax:** how words can be assembled into phrases / sentences:
  - *I found an unopened bottle of wine*
  - *I found a bottle unopened of wine*

• **Morphology:** internal form of words
  - *unopened* – not *openuned* or any other order

• But this distinction is not crisp (since notion of “morpheme” or “word” is graded) → **Morphosyntax**
Morpheme examples

- निरीक्षक = नि- [रीक्ष] -क
- prefix suffix

- bound / free morphemes: 
  -क vs -कर्ता (e.g. अपहरणकर्ता)

- Morphemes often cause changes to the stem
  - bAngla: kin- , buy
    Ami kinAm uni kenen kenAkATA
    I buy+PAST he (honorific) buy+PRES
    buying (noun)
Stemming (baby lemmatization)

- Assumption: surface form = root . affix

- Reduce a word to the main morpheme

  `automate` → `automat`
  `automates` → `automat`
  `automatic` → `automat`
  `automation` → `automat`
  `run` → `run`
  `runs` → `run`
  `running` → `run`

- Widely used in Information Retrieval
Porter Stemmer (1980)

- Most common algorithm for stemming English
  - Results suggest it’s at least as good as other stemming options
- Multiple sequential phases of reductions using rules, e.g.
  - sses → ss
  - ies → i
  - ational → ate
  - tional → tion

- [http://tartarus.org/~martin/PorterStemmer/](http://tartarus.org/~martin/PorterStemmer/)
Stemming example

Candidate = candid + ate

This is a poorly constructed example using the Porter stemmer.

This is a poorli construct example us the Porter stemmer.

http://maya.cs.depaul.edu/~classes/ds575/porter.html

Code:
http://snowball.tartarus.org/algorithms/english/stemmer.html
Inflection vs Derivation
Inflections and Derivations

- **Inflection**: e.g. $\text{sing} \rightarrow \text{sang}$; $\text{cat} \rightarrow \text{cats}$
  - variation in form due to tense, person, etc.
  - does not change primary meaning,
  - same part-of-speech
  - applies to nearly entire class of lexemes

- **Derivation**: e.g. $\text{sing} \rightarrow \text{singer}$
  - changes meaning, changes part-of-speech
  - Like much in grammar, not very crisp distinction
    - e.g. $\text{cyclic} \rightarrow \text{cyclical} = \text{derivation}$
    - treat as new word
Productive Morphemes

- A morpheme is productive if it applies to all words of a given type.
- Inflections – almost fully productive
- Derivations – very limited
Semantics of morphemes

- **inflections:**
  - e.g. “-ed” : past tense = events in the past
    - *The course started last week.*
  - *But:* often does not refer to past, e.g.:
    - *I thought the course started next week.*
    - *If the course started, everyone would be pleased.*

- past time = **primary** or most common characteristic
- many other interpretations possible (in many languages)
  - → *past tense* = grammatical form, varied semantics
Derivations

- e.g. **ungentlemanly**: un + gentle + man + ly

- not all lexemes of a class will take all these particles, nor will they have the same meaning.

- how to break up (**parse**) the lexeme?
  - [ [un+gentle] + man ] + ly
  - [un + [gentle + man] + ly

many interpretations are possible
Derivations : Parsing

- Differing parses $\rightarrow$ different semantics:
- e.g. unlockable
  - “can’t be locked” or “can be unlocked”?

Huddleston & Pullum 05
Derivations : Ambiguity

- Semantics : not fully systematic – e.g. anomalous usage of *un*-
  - *loosen* same as *unloosen*
Semantics of composition

- **derivations:**
  
e.g. “-er”: usually agentive – builder, writer, teacher
  But may be instrumental – e.g. cooker
  - However, meaning is constrained (not arbitrary)

- **compounds:** composed from multiple lexemes
  - doghouse, darkroom (endocentric, tatpurusha):
    ‘house’, ‘room’ is the head
  - redcoat, Hindi: nllakanTha (exocentric, bahuvrihi):
    refers to neither red nor coat
Models of Syntax
पांच फिरंगी अफसरों ___ फांसी पर ___ दिया

what can go in the blanks?

what can NOT go there?
Syntax

Sentences are built from “words”.

*boys* *like* *girls*
*germans* *drink* *beer*

sentence = noun  verb  noun
Syntactic Composition

• Constituency: *like girls* = verb phrase VP
  head: *like* V
  constituent: *girls* N-plural

• Grammatical Function (maps to semantics?):
  subject: boys
  predicate: like
  arguments: boys, girls

• Hierarchy and Control
One Version of Constituent Structure

- Lexicon:
  
  the, a, small, nice, big, very, boy, girl, sees, likes

- Grammatical sentences:
  1. (the) boy (likes a girl)
  2. (the small) girl (likes the big girl)
  3. (a very small nice) boy (sees a very nice boy)

- Ungrammatical sentences:
  1. *(the) boy (the girl)
  2. *(small) boy (likes the nice girl)
N-gram language models
Word Segmentation
NLP Tasks

Word segmentation:

- **Chinese:** 浮法像蝴蝶.
  
  (“float like a butterfly”)

- **Hindi**
  
  पांचफिंगीअफसरोंकोफ्रांसीपरलटकादिया

  - Q. Letter-or Syllable-based?
  - Which positions have low “sequence” probability?
NLP tasks : Probabilistic Models

- Other problems
  - Machine Translation:
    - \( P(\textbf{high} \text{ winds \ tonite}) > P(\textbf{large} \text{ winds \ tonite}) \)
  - Spell Correction
    - The office is about fifteen \textit{minuets} from my house
    - \( P(\text{about fifteen} \ \textit{minutes} \text{ from}) > P(\text{about fifteen} \ \textit{minuets} \text{ from}) \)
  - Speech Recognition
    - \( P(\text{I saw a van}) >> P(\text{eyes awe of an}) \)
  - Verb argument structure discovery
    - Via factorization of syntactic parses to discover
    - Argument structure (syntax ?)
    - Selection preference (semantics)
  - + Summarization, question-answering, etc.,
Models of Syntax

• Linguistic Rules and Hierarchies:
  o Phonology:
  o Morphology:
  o Lexical:

\{ \text{categories + rules} \} = \text{syntax (e.g. CFG)}

• Probabilistic models
  o Bayesian models – PCFG
  o N-grams
Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words:
  \[ P(W) = P(w_1, w_2, w_3, w_4, w_5 \ldots w_n) \]

- Related task: probability of an upcoming word:
  \[ P(w_5 | w_1, w_2, w_3, w_4) \]

- A model that computes either of these:
  \[ P(W) \quad \text{or} \quad P(w_n | w_1, w_2, \ldots w_{n-1}) \]

  is called a language model.

- Better: the grammar  
  But language model or LM is standard
Shannon Entropy

- Predict the next word/letter, given \((n-1)\) previous items
  \[ F_n = \text{entropy} = \sum p_i \log p_i \]

- Probabilities \(p_i\) (of \(n\)-grams) from corpus:
  - \(F_0\) (only alphabet) = \(\log_2 27\) = 4.76 bits per letter
  - \(F_1\) (1-gram frequencies \(p_i\)) = 4.03 bits
  - \(F_2\) (bigram frequencies) = 3.32 bits
  - \(F_3\) (trigrams) = 3.1 bits
  - \(F_{\text{word}}\) = 2.62 bits

  (avg word entropy = 11.8 bits per 4.5 letter word)

Shannon Entropy: Human

- Ask human to guess the next letter:

  THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
  ----ROO-------NOT-V------I-------SM-----OBL---

  READING LAMP ON THE DESK SHED GLOW ON
  REA--------O-------D----SHED-OLD--O-

  POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
  P-L-S-----O---BU--L-S--O--------SH------RE--C-------

- 69% guessed on 1st attempt [“-” = 1st attempt]

Shannon Entropy: Human

- Count number of attempts:

  THERE IS NO REVERSE ON A MOTORCYCLE A
  1 1 1 5 1 1 2 1 1 2 1 1 5 1 1 7 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
  FRIEND OF MINE FOUND THIS OUT
  8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
  RATHER DRAMATICALLY THE OTHER DAY
  4 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

- Entropy: $F_1 = 3.2, 4.0$  $F_{10} = 1.0, 2.1$  $F_{100} = $

LANGUAGE MODELING

NL Corpora
Creating a Corpus

1961: W. Nelson Francis and Henry Kucera of Brown Univ
500 samples of 2,000 words each from various text genres
→ American English

1970s: Lancaster-Oslo-Bergen corpus: British English
also 500 x 2000 = 1mn words – genres similar to Brown Corpus
Geoffrey Leech of Lancaster U.

1994: British National Corpus – 100mn words
Oxford U, Lancaster, Longman / Chambers dictionaries
10% : transcripts of spoken English

2000s: Google corpora: American english 155 bn words; British : 34bn

[Lindquist 2009]: Corpus linguistics and the description of English
## The Brown Corpus

<table>
<thead>
<tr>
<th>Category</th>
<th># Texts</th>
<th>% Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Press: reportage (newspapers)</td>
<td>44</td>
<td>8.8%</td>
</tr>
<tr>
<td>B Press: editorial (including letters to the editor)</td>
<td>27</td>
<td>5.4%</td>
</tr>
<tr>
<td>C Press: reviews (theatre, books, music, dance)</td>
<td>17</td>
<td>3.4%</td>
</tr>
<tr>
<td>D Religion</td>
<td>17</td>
<td>3.4%</td>
</tr>
<tr>
<td>E Skills and hobbies</td>
<td>36</td>
<td>7.2%</td>
</tr>
<tr>
<td>F Popular lore</td>
<td>48</td>
<td>9.6%</td>
</tr>
<tr>
<td>G Belles letters, biography, memoirs etc.</td>
<td>75</td>
<td>15.0%</td>
</tr>
<tr>
<td>H Miscellaneous (mainly government documents)</td>
<td>30</td>
<td>6.0%</td>
</tr>
<tr>
<td>J Learned (academic texts)</td>
<td>80</td>
<td>16.0%</td>
</tr>
<tr>
<td>K General fiction (novels and short stories)</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>L Mystery and detective fiction</td>
<td>24</td>
<td>4.8%</td>
</tr>
<tr>
<td>M Science fiction</td>
<td>6</td>
<td>1.2%</td>
</tr>
<tr>
<td>N Adventure and Western fiction</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>P Romance and love story</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>R Humour</td>
<td>9</td>
<td>1.8%</td>
</tr>
<tr>
<td>Non-fiction subtotal</td>
<td>374</td>
<td>75%</td>
</tr>
<tr>
<td>Fiction subtotal</td>
<td>126</td>
<td>25%</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>100%</td>
</tr>
</tbody>
</table>

News: political, sports, society "spot news", financial, cultural)

[Lindquist 2009]: Corpus linguistics and the description of English
Parallel Corpora
Congress MP from Haryana Birender Singh said at a programme that "once someone had told me that Rs 100 crore was required to get a Rajya Sabha berth. But he said he got it for Rs 80 crore and saved Rs 20 crore. Now will people who are willing to invest Rs 100 crore, ever think of the poor country."
Matching on parallel Corpus

电脑坏了。
  The computer is broken.
电脑死机了。
  My computer has frozen.
我想玩电脑。
  I want to play on the computer.
我家没有电脑。
  I don't have a computer at home.
我有一台电脑。
  I have a computer.
你有两台电脑吗？
  Do you have two computers?
电脑坏了。
The computer is broken.
电脑死机了。
My computer has frozen.
我想玩电脑。
I want to play on the computer.
我家没有电脑。
I don't have a computer at home.
我有一台电脑。
I have a computer.
你有两台电脑吗？
Do you have two computers?

电脑：diànnǎo, computer
[ 电：diàn lightning, electricity  脑：nǎo brain ]
电脑坏了。
    The computer is broken.
电脑死机了。
    My computer has frozen.
我想玩电脑。
    I want to play on the computer.
我家没有电脑。
    I don't have a computer at home.
我有一台电脑。
    I have a computer.
你有两台电脑吗？
    Do you have two computers?

有："in possession of"
    [ 又 ("hand") + 肉 ("meat") = a hand holding meat ]
LANGUAGE MODELING

Generalization and zeros
The perils of overfitting

- N-grams only work well for word prediction if the test corpus looks like the training corpus
  - In real life, it often doesn’t
  - We need to train robust models that generalize!
- One kind of generalization: Zeros!
  - Things that don’t ever occur in the training set
    - But occur in the test set
Training set:
... denied the allegations
... denied the reports
... denied the claims
... denied the request

\[ P(“offer” \mid \text{denied the}) = 0 \]

Test set
... denied the offer
... denied the loan
Actual Probability Distribution:
Actual Probability Distribution:
“Smoothing”

- Develop a model which decreases probability of seen events and allows the occurrence of previously unseen n-grams
- a.k.a. “Discounting methods”
- “Validation” – Smoothing methods which utilize a second batch of test data.

based on Manning and Schütze
Smoothing

![Graph showing C(w) for different words such as the, her, cherries, Maria, word 5, word 6, word 7, and others. The graph indicates varying C(w) values with 'the' having the highest value at 2.](image)
Smoothing: +1
Smoothing: +1
Spelling correction w bigram language model

“a stellar and versatile across whose combination of sass and glamour…”

Counts from the Corpus of Contemporary American English with add-1 smoothing

- \( P(\text{actress}|\text{versatile}) = 0.000021 \)
  \[ P(\text{whose}|\text{actress}) = 0.0010 \]
- \( P(\text{across}|\text{versatile}) = 0.000021 \)
  \[ P(\text{whose}|\text{across}) = 0.000006 \]

\[
P(“\text{versatile actress whose”}) = 0.000021 \times 0.0010 = 210 \times 10^{-10}
\]
\[
P(“\text{versatile across whose”}) = 0.000021 \times 0.000006 = 1 \times 10^{-10}
\]
Estimating N-gram Probabilities
Probabilistic Language Modeling

- Goal: determine if a sentence or phrase has a high acceptability in the language
  → compute the probability of the sequence of words
    
    E.g. “its water is so transparent that”

- \( P(\text{its, water, is, so, transparent, that}) \)
Probabilistic Language Modeling

\[ P(W) = P(w_1, w_2, w_3, w_4, w_5 \ldots w_n) \]

- Related task: probability of an upcoming word:
  \[ P(w_5 | w_1, w_2, w_3, w_4) \]
Reliability vs. Discrimination

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)
The Chain Rule

- Chain Rule in General
  \[
P(x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | x_{1:i-1})
  \]

- Proof:
  - Holds for \( n=2 \) (Product rule)
  - Assume is true for \( X = x_1 \ldots x_{n-1} \).
    
    \[
P(X, x_n) = P(X) P(x_n | X) \quad \rightarrow \text{General chain rule}
    \]
Markov Assumption

- Simplifying assumption: Depends only on $k$-nearby text

- **First-order** Markov Process ($k=1$): 
  
  $$ P(\text{the} \mid \text{its water is so transparent that}) \quad P(\text{the} \mid \text{that}) $$

- or **Second-order** ($k=2$): 
  
  $$ P(\text{the} \mid \text{its water is so transparent that}) \quad P(\text{the} \mid \text{transparent that}) $$

Andrei Markov, 1856-1922, Russia
Estimating bigram probabilities

- The Maximum Likelihood Estimate

\[
P(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}
\]

\[
P(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}
\]
N-gram Text Generation
Sentence Generation

Unigram Model: No dependencies on previous words

\[ P(w_1 w_2 \ldots w_n) \approx \prod_{i} P(w_i) \]

Bigram Model: Depends on 1 previous word

\[ P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1}) \]
The Corpus matters

- What corpus was used to generate these:

  **Bigram**
  
  What means, sir. I confess she? then all sorts, he is trim, captain.
  Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
  What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

  **Trigram**
  
  Sweet prince, Falstaff shall die. Harry of Monmouth’s grave.
  This shall forbid it should be branded, if renown made it empty.
  Indeed the duke; and had a very good friend.
  Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done.

  **Quadrigram**
  
  King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in;
  Will you not tell me who I am?
  It cannot be but so.
  Indeed the short and the long. Marry, ’tis a noble Lepidus.
The Corpus matters

- What corpus was used to generate these:

**Bigram**

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

**Trigram**

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions
N-gram frequency falls rapidly with N.

- Shakespeare Corpus: N=884,647 tokens, V=29,066
- Shakespeare produced 300,000 bigram types out of $V^2=844$ million possible bigrams.
  - So 99.96% of the possible bigrams were never seen (have zero entries in the table)
- Quadrigrams worse: Shakespeare had very specific patterns of usage
Limitations of N-gram models

- **Advantages:**
  - Does not require expensive annotated corpora
  - Annotations are often disputed
  - Efficacy of intermediate representations are doubtful
- **We can extend to trigrams, 4-grams, 5-grams**
  - Corpus size must grow exponentially larger
- **Main Disadvantage:** *Long-distance dependencies:*
  
  “The computer which I had just put into the machine room on the fifth floor crashed.”
Practical Issues

- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

\[
\log(p_1 \cdot p_2 \cdot p_3 \cdot p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4
\]
Google N-Gram Release, August 2006

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

That’s why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.
serve as the incoming 92
serve as the incubator 99
serve as the independent 794
serve as the index 223
serve as the indication 72
serve as the indicator 120
serve as the indicators 45
serve as the indispensable 111
serve as the indispensible 40
serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
Computational Morphology
Computational Analysis

- **[Goldsmith 01]**

  Information-Theoretic ideas - Minimum Description Length

  Which “signature” (pattern) will result in the most compact description of the corpus?

<table>
<thead>
<tr>
<th>Signature</th>
<th>Example</th>
<th>Stem # (type)</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL.ed.ing</td>
<td>betray betrayed</td>
<td>69</td>
<td>864</td>
</tr>
<tr>
<td></td>
<td>betraying</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NULL.ed.ing.s</td>
<td>remain remained</td>
<td>14</td>
<td>516</td>
</tr>
<tr>
<td></td>
<td>remaining</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>remaining remains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NULL.s.</td>
<td>cow cows</td>
<td>253</td>
<td>3414</td>
</tr>
<tr>
<td>e.ed.es.ing</td>
<td>notice noticed</td>
<td>4 62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>noticing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computational Analysis

• [Dasgupta & V.Ng 07]
  • Simple concatenation not enough for more agglutinated languages.
  • Attempt to discover root word form. (denial → deny)
  • Assumption: if compound word is common, then root word will also: Word-Root Frequency Ratios (WRFR)

<table>
<thead>
<tr>
<th>Correct Parses</th>
<th>Incorrect Parses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Root</strong></td>
</tr>
<tr>
<td>bear-able</td>
<td>bear</td>
</tr>
<tr>
<td>attend-ance</td>
<td>attend</td>
</tr>
<tr>
<td>arrest-ing</td>
<td>arrest</td>
</tr>
<tr>
<td>sub-group</td>
<td>group</td>
</tr>
<tr>
<td>re-cycle</td>
<td>cycle</td>
</tr>
<tr>
<td>un-settle</td>
<td>settle</td>
</tr>
</tbody>
</table>
STATISTICAL NATURAL LANGUAGE PARSING

POS-Tagging
POS Tagging Approaches

- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge (e.g. ENGTWOL 95)

- **Stochastic**: Trained on human annotated corpora like the Penn Treebank
  - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), log-linear models, support vector machines
  - **Rule learning**: Transformation Based Learning (TBL)

- Many English POS-taggers are publicly available

- Hindi / Bangla POS tagger:
  - [http://nltr.org/snltr-software/](http://nltr.org/snltr-software/)
Deciding on a POS tagset

NOUN          The DOG barked.          WE saw YOU.
VERB          The dog BARKED.          It IS impossible.
ADJECTIVE     He's very OLD.          I've got a NEW car.
DETERMINATIVE THE dog barked.          I need SOME nails.
ADVERB        She spoke CLEARLY.       He's VERY old.
PREPOSITION   It's IN the car.         I gave it TO Sam.
COORDINATOR   I got up AND left.       It's cheap BUT strong.

SUBORDINATOR  It's odd THAT they were late. I wonder WHETHER it's still there.
INTERJECTOR   OH, HELLO, WOW, OUCH

from [huddleston-pullum 05] Student's intro to English Grammar

Coordinator / subordinator: markers for coordinate / subordinate clauses
POS distinctions based on analysis of syntax and semantics
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordin. Conjunction</td>
<td><em>and, but, or</em></td>
<td>SYM</td>
<td>Symbol</td>
<td><em>+, %, &amp;</em></td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td><em>one, two, three</em></td>
<td>TO</td>
<td>&quot;to&quot;</td>
<td><em>to</em></td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td><em>a, the</em></td>
<td>UH</td>
<td>Interjection</td>
<td><em>ah, oops</em></td>
</tr>
<tr>
<td>EX</td>
<td>Existential ‘there’</td>
<td><em>there</em></td>
<td>VB</td>
<td>Verb, base form</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td><em>mea culpa</em></td>
<td>VBD</td>
<td>Verb, past tense</td>
<td><em>ate</em></td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/sub-conj</td>
<td><em>of, in, by</em></td>
<td>VBG</td>
<td>Verb, gerund</td>
<td><em>eating</em></td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td><em>yellow</em></td>
<td>VBN</td>
<td>Verb, past participle</td>
<td><em>eaten</em></td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td><em>bigger</em></td>
<td>VBP</td>
<td>Verb, non-3sg pres</td>
<td><em>eat</em></td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td><em>wildest</em></td>
<td>VBZ</td>
<td>Verb, 3sg pres</td>
<td><em>eats</em></td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td><em>1, 2, One</em></td>
<td>WDT</td>
<td>Wh-determiner</td>
<td><em>which, that</em></td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td><em>can, should</em></td>
<td>WP</td>
<td>Wh-pronoun</td>
<td><em>what, who</em></td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td><em>llama</em></td>
<td>WP$</td>
<td>Possessive wh-</td>
<td><em>whose</em></td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td><em>llamas</em></td>
<td>WRB</td>
<td>Wh-adverb</td>
<td><em>how, where</em></td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td><em>IBM</em></td>
<td>$</td>
<td>Dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td><em>Carolinas</em></td>
<td>#</td>
<td>Pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td><em>all, both</em></td>
<td>“</td>
<td>Left quote</td>
<td>(‘ or “)</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td><em>’s</em></td>
<td>”</td>
<td>Right quote</td>
<td>(’ or ”)</td>
</tr>
<tr>
<td>PP</td>
<td>Personal pronoun</td>
<td><em>I, you, he</em></td>
<td>(</td>
<td>Left parenthesis</td>
<td>(, (, {, &lt;)</td>
</tr>
<tr>
<td>PPS$</td>
<td>Possessive pronoun</td>
<td><em>your, one’s</em></td>
<td>)</td>
<td>Right parenthesis</td>
<td>(), }, &gt;)</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td><em>quickly, never</em></td>
<td>,</td>
<td>Comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td><em>faster</em></td>
<td>;</td>
<td>Sentence-final punc</td>
<td>( ! ?)</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td><em>fastest</em></td>
<td>:</td>
<td>Mid-sentence punc</td>
<td>( ; ; ... ; - )</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
<td><em>up, off</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 8.6** Penn Treebank Part-of-Speech Tags (Including Punctuation)

Figure: jurafsky-martin ch.8 (2000)
"I miss the good old days when all we had to worry about was nouns and verbs."
Stochastic POS-tagging

- Markovian assumption: tag depends on limited set of previous tags

- HMM:
  
  \[
  \text{maximize } P(\text{word} | \text{tag}) \times P(\text{tag} | \text{previous n tags})
  \]

- Maximize the probability for whole sentence, not single word

\[
S = \arg \max_{t_1 \ldots t_n} \prod_{i=1, n} P(w_i | t_i) P(t_i | t_{i-1})
\]
Stochastic POS-tagging

- Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN
- People/NNS continue/VBP to/TO inquire/VB the_DT reason/NN for/IN the_DT race/NN for/IN outer/JJ space/NN

- to race vs. the race
Stochastic POS-tagging

- *to/TO race*  
- *the/DT race*

- $P(VB|TO) \cdot P(race|VB) = 0.00001$
- $P(NN|TO) \cdot P(race|NN) = 0.000007$

- $P(VB|TO) = 0.34$  
- $P(race|VB) = 0.00003$

- $P(NN|TO) = 0.021$  
- $P(race|NN) = 0.00041$
GROUND LANGUAGE MODELS

Unsupervised POS and Syntax: Grounded Models
Language Acquisition: Domains

- Perceptual input

- Discovery Targets:
  - semantics: objects, 2-agent actions, relations
  - lexicon: nominal, transitive verbs, preposition
  - lexical categories: N VT P Adj
  - constructions: PP VP S
  - sense extension (metaphor) [nayak/mukerjee (AAAI-12)]

[heid/simmel 1944] [hard/tversky 2003]
Linguistic input

- input = description commentaries transcribed into text
  - 48 descriptions in English / 10 : Hindi

- Unconstrained description by different subjects:
  - the little square hit the big square
  - they're hitting each other
  - the big square hit the little square
  - circle and square in [unitelligible stammer]
  - the two squares stopped fighting

- छोटा बक्सा बडा बक्सा मे कुछ बातचीत होती है
  - little box big box between some talk happens
POS categories - Unsupervised

[mukerjee nayak 12] based on ADIOS
[solan rupin edelman 05]
Language Structures : Verbs

1. \[
\left( \begin{array}{c}
\text{the} \\
\text{big} \\
\text{large}
\end{array} \right) \rightarrow \text{square} \rightarrow \left( \begin{array}{c}
\text{scares} \\
\text{approaches} \\
\text{chases}
\end{array} \right) \rightarrow \left( \begin{array}{c}
\text{the} \\
\text{small} \\
\text{little}
\end{array} \right)
\]

2. \[
\left( \begin{array}{c}
\text{the} \\
\text{ball} \\
\text{box} \\
\text{door} \\
\text{square}
\end{array} \right) \rightarrow \left( \begin{array}{c}
\text{circle} \\
\text{it}
\end{array} \right) \rightarrow \left( \begin{array}{c}
\text{moved} \\
\text{moves} \\
\text{runs}
\end{array} \right)
\]
Hindi Acquisition: Word learning

<table>
<thead>
<tr>
<th>[BS]</th>
<th>[SS]</th>
<th>[C]</th>
<th>[IN]</th>
</tr>
</thead>
<tbody>
<tr>
<td>word(s)</td>
<td>word(s)</td>
<td>word(s)</td>
<td>word(s)</td>
</tr>
<tr>
<td>बक्सा</td>
<td>बक्सा</td>
<td>गोला</td>
<td>अन्दर</td>
</tr>
<tr>
<td>baksA/box</td>
<td>baksA/box</td>
<td>golA/ball</td>
<td>andar/in</td>
</tr>
<tr>
<td>बडा</td>
<td>छोटा</td>
<td>बक्सा</td>
<td>बाहर</td>
</tr>
<tr>
<td>bada/big</td>
<td>chota/small</td>
<td>ke/-</td>
<td>bA-har/out</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.73</td>
</tr>
<tr>
<td>.77</td>
<td>.37</td>
<td>.62</td>
<td>.83</td>
</tr>
<tr>
<td>.44</td>
<td></td>
<td>.54</td>
<td>.80</td>
</tr>
<tr>
<td>.18</td>
<td></td>
<td>.25</td>
<td>1.30</td>
</tr>
<tr>
<td>.90</td>
<td></td>
<td>.63</td>
<td>.78</td>
</tr>
<tr>
<td>.27</td>
<td></td>
<td>.73</td>
<td>.73</td>
</tr>
</tbody>
</table>
Incipient Syntax

\[
\begin{array}{c}
\text{डब्बे (dabbA/box)} \\
\text{बक्से (bakse/box)}
\end{array}
\rightarrow
\begin{array}{c}
\text{के (ke/-)}
\end{array}
\rightarrow
\begin{array}{c}
\text{वाहर (bAhAr/out)} \\
\text{आ (aa/come)} \\
\text{भाग (bhAg/run)}
\end{array}
\rightarrow
\begin{array}{c}
\text{जाता (jAtA/goes)}
\end{array}
\]