Text preprocessing
What is text preprocessing?

- Cleaning up a text for further analysis
- A huge problem that is underestimated by almost everyone
- What kinds of text?
  - Newspaper articles
  - Emails
  - Tweets
  - Blog posts
  - Scans
  - Web pages
- A skill in high demand
Common tasks

1. Sentence boundary detection
2. Tokenization
3. Normalization
4. Lemmatization
Sentence boundary detection

- Find sentences. How are they defined?
- Find sentence punctuation (., ?, !)
  - How about “;”? Does it divide sentences?
  - “One more remains: the southern states.”
- Problematic when lots of abbreviations
  - “The I.R.S.” “5.23”
- Can’t always rely on input (typos, OCR errors, etc.)
  - “In fact. they indicated . . .”
  - “overall. So they . . .”
How do you determine sentence boundaries in Chinese or Japanese or Latin with no punctuation?

Can capital letter show sentence beginning?
• . . . on the bus. Later, they were . . .
• . . . that is when Bob came to the . . .

Quotes
• “You still do that?” John asked.
Crazy talk. George Bush hasn't quite gone to war yet, but he's already murdering the language. John Sutherland on how conflict throws up new phrases. Special report: George Bush's America. Special report: terrorism in the US.


"Words," Elaine Showalter declares, "do not fail us. Words are what will help us through this crisis." The American critic has it right. And, given the difficult times ahead, we should examine our linguistic options as carefully as the military do theirs. Conflict throws up new coinages. Desert Storm (that name oddly redolent of Mills & Boon) gave currency to: "mother of all battles", "line in the sand", "surgical strike", "collateral damage" and, as its sad linguistic legacy, "Gulf war syndrome".
Tokenization

- Splitting up words from an input document
- How hard can that be? What is a word? Issues:
  - Compounds
    - Well-known vs. well known
    - Auto body vs. autobody
    - Rail road vs. railroad
    - On-site vs. onsite
    - E-mail vs. email
    - Shut down (verb) vs. shutdown (noun)
    - Takeoff (noun) vs. take off (verb)
Tokenization

- Clitics (how many words?)
  - “Le voy a dar” vs. “Voy a darle”
  - “don't, won't, she'll”
  - “et cetera” “vice versa” “cannot” one or two words?
- Hyphenation at end of line
  - Rab-bit, en-tourage, enter-taining
- Capitalization
- Normalization sometimes refers to this cleanup
- It’s easy to underestimate this task!
- Related: sentence boundary detection
Tokenize this!

- Sample page
Normalization

- Make all tokens of a given type equivalent
  - Capitalization
    - “The cats” vs. “Cats are”
  - Hyphenation
    - Pre-war vs. prewar
    - E-mail vs. email
  - Expanding abbreviations
    - e.g. vs. for example
  - Spelling errors/variations
    - IBM vs. I.B.M.
    - Behavior vs. behaviour
POS tagging: introduction

- Part-of-speech assignment (tagging)
- Label each word with its part-of-speech
  - Noun, preposition, adjective, etc.

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

- State of art: 95%+ for English
- Often 1 wd/sent error
- Syntagmatic approach: consider close tags
- Frequency (‘dumb’) approach: over 90%
- Various standardized tagsets
Why are POS helpful?

- Pronunciation
  - I will lead the group into the lead smelter.
- Predicting what words can be expected next
  - Personal pronoun (e.g., I, she) ____________
- Stemming (web searches)
  - -s means singular for verbs, plural for nouns
- Translation
  - (E) content +N ⇒ (F) contenu +N
  - (E) content +Adj ⇒ (F) content +Adj or satisfait +Adj
Why are POS helpful?

- Having POS is prerequisite to syntactic parsing
  - Syntax trees
- POS helps distinguish meaning of words
  - “bark” dog or tree?
    - They stripped the bark. It shouldn't bark at night.
  - “read” past or present?
    - He read the book. He's going to read the book.
Why are POS helpful?

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.
  - people  organizations  places
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Extract pieces of information relevant to a specific application, e.g. used car ads:
  - make  model  year  mileage  price
  - For sale, 2002 Toyota Prius, 20,000 mi, $15K or best offer. Available starting July 30, 2006.
Why are POS helpful?

• For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.
  • agent  patient  source  destination  instrument
    • John drove Mary from Austin to Dallas in his Toyota Prius.
    • The hammer broke the window.

• Also referred to as “case role analysis,” “thematic analysis,” and “shallow semantic parsing”
Annotating POS

- Textbook tags: noun, adjective, verb, etc.
- Most English sets have about 40-75 tags
Annotating POS

Noun (person, place or thing)

» Singular (NN): dog, fork
» Plural (NNS): dogs, forks
» Proper (NNP, NNPS): John, Springfields
» Personal pronoun (PRP): I, you, he, she, it
» Wh-pronoun (WP): who, what

• Verb (actions and processes)
  » Base, infinitive (VB): eat
  » Past tense (VBD): ate
  » Gerund (VBG): eating
  » Past participle (VBN): eaten
  » Non 3rd person singular present tense (VBP): eat
Tagsets

- Brown corpus tagset (87 tags)
- Claws7 tagset (146 tags)
How hard is POS tagging?

- Easy: Closed classes
  - conjunctions: *and, or, but*
  - pronouns: *I, she, him*
  - prepositions: *with, on*
    - determiners: *the, a, an*
- Hard: open classes (verb, noun, adjective, adverb)
How hard is POS tagging?

- Harder:
  - *provided*, as in “I’ll go provided John does.”
  - *there*, as in “There aren’t any cookies.”
  - *might*, as in “I might go.” or “I might could go.”
  - *no*, as in “No, I won’t go.”
How hard is POS tagging?

• “Like” can be a verb or a preposition
  • I like/VBP candy.
  • Time flies like/IN an arrow.

• “Around” can be a preposition, particle, or adverb
  • I bought it at the shop around/IN the corner.
  • I never got around/RP to getting a car.
  • A new Prius costs around/RB $25K.
How hard is POS tagging?

- Degree of ambiguity in English (based on Brown corpus)
  - 11.5% of word types are ambiguous.
  - 40% of word tokens are ambiguous.
- Average POS tagging disagreement among expert human judges for the Penn treebank was 3.5%
  - Based on correcting the output of an initial automated tagger, which was deemed to be more accurate than tagging from scratch.
- Baseline: Picking the most frequent tag for each specific word type gives about 90% accuracy
  - 93.7% if use model for unknown words for Penn Treebank tagset.
How hard is it done?

• **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge.

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

• Generally, learning-based approaches have been found to be more effective overall, taking into account the total amount of human expertise and effort involved.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.

classifier

NNP
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

• Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

John saw the saw and decided to take it to the table.
Using Probabilities

- Is “can” a noun or a modal verb?
- We know nouns follow “the” 90% of the time
- Modals never do so “can” must be a noun.
- Nouns are followed by verbs 90% of the time
- So “can” is probably a modal verb in “cars can”
Sample Markov Model for POS

- **Det**: 0.95 to Noun, 0.05 from Noun
- **Noun**: 0.9 to Verb, 0.1 from Verb
- **Verb**: 0.5 to stop, 0.5 from stop
- **PropNoun**: 0.1 from start, 0.4 to Det
- **start**: 0.1 to Det, 0.25 to PropNoun
- **stop**: 0.1 from Verb, 0.25 to Verb

States: Det, Noun, Verb, PropNoun, start, stop
Lemmatization

- What is frequency of “to be”?  
  - Just # of “be”?
Lemmatization

- What is frequency of “to be”?  
  - Just # of “be”?  
  - No we want to include “be, are, is, am, etc.”  
  - The lemma of “to be” includes these.
Lemmatization

- What is frequency of “to be”?  
  – Just # of “be”?  
  – No we want to include “be, are, is, am, etc.”  
  – The lemma of “to be” includes these.  
  – What would the lemma of “chair” include?
Lemmatization

- What is frequency of “to be”?  
  - Just # of “be”?  
  - No we want to include “be, are, is, am, etc.”  
  - The lemma of “to be” includes these.  
  - What would the lemma of “chair” include?  
    • Chair, chairs
Computational morphology

- Developing/using computer applications that involve morphology

- Analysis: parse/break a word into its constituent morphemes

- Generation: create/generate a word from its constituent morpheme
Word classification

- Part-of-speech category
  - Noun, verb, adjective, adverb, etc.
- Simple word vs. complex word
  - One morpheme vs. more morphemes
  - Open-class/lexical word vs.
- closed-class/function(al)/stop word
  - Productive/inventive use vs. restricted use
Word-structure diagrams

- Each morpheme is labelled (root, affix type, POS)
- Each step is binary (2 branches)
- Each stage should span a real word
Portuguese morphology

- Verb conjugation
  - 63 possible forms
  - 3 major conjugation classes, many sub-classes
  - Over 1000 (semi)productive verb endings

- Noun pluralization
  - Almost as simple as English

- Adjective inflection
  - Number
  - Gender
Portuguese verb (falar)

falando
falado
falar falares falar falarmos falardes falarem
falo falas fala falamos falais falam
falava falavas falava falávamos faláveis falavam
falei falaste falou falamos falastes falaram
falara falaras falara faláramos faláreis falaram
falarei falarás falará falaremos falareis falarão
falaria falarias falaria falaríamos falaríeis falariam
fala falai
fale fales fale falem
falasse falasses falasse falássemos falásseis falassem
falar falares falar falarmos falardes falarem
Finnish complexity

- Nouns
  - Cases, number, possessive affixes
  - Potentially 840 forms for each noun

- Adjectives
  - As for nouns, but also comparative, superlative
  - Potentially 2,520 forms for each

- Verbs
  - Potentially over 10,000 forms for each
Complexity

- Varying degrees of morphological richness across languages
  qasuiirsarvingsarsingitluinarnarpuq
  “someone did not find a completely suitable resting place”
  Dampfschiffahrtsdirektorsstellvertretersgemahlin
## English complexity (WSJ)

<table>
<thead>
<tr>
<th>superconductivity</th>
<th>disproportionately</th>
<th>overspecialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>telecommunications</td>
<td>constitutionality</td>
<td>counterproductive</td>
</tr>
<tr>
<td>misrepresentations</td>
<td>superconductivity</td>
<td>administration's</td>
</tr>
<tr>
<td>biotechnological</td>
<td>deoxyribonucleic</td>
<td>enthusiastically</td>
</tr>
<tr>
<td>immunodeficiency</td>
<td>mischaracterizes</td>
<td>nonmanufacturing</td>
</tr>
<tr>
<td>nonparticipation</td>
<td>pharmaceuticals'</td>
<td>recapitalization</td>
</tr>
<tr>
<td>responsibilities</td>
<td>superspecialized</td>
<td>unapologetically</td>
</tr>
<tr>
<td>unconstitutional</td>
<td>administrations</td>
<td>anthropological</td>
</tr>
<tr>
<td>capitalizations</td>
<td>cerebrovascular</td>
<td>competitiveness</td>
</tr>
<tr>
<td>computerization</td>
<td>confidentiality</td>
<td>confrontational</td>
</tr>
<tr>
<td>congressionally</td>
<td>criminalization</td>
<td>discombobulated</td>
</tr>
<tr>
<td>discontinuation</td>
<td>dispassionately</td>
<td>dissatisfaction</td>
</tr>
<tr>
<td>diversification</td>
<td>entrepreneurial</td>
<td>experimentation</td>
</tr>
<tr>
<td>extraordinarily</td>
<td>inconsistencies</td>
<td>instrumentation</td>
</tr>
<tr>
<td>internationally</td>
<td>liberalizations</td>
<td>micromanagement</td>
</tr>
<tr>
<td>microprocessors</td>
<td>notwithstanding</td>
<td>pharmaceuticals</td>
</tr>
<tr>
<td>philosophically</td>
<td>professionalism</td>
<td>proportionately</td>
</tr>
</tbody>
</table>
Morphological constraints

- dog+s, walk+ed, big(g)+est, sight+ing+s, punish+ment+s
  - *s+dog, *ed+walk, *est+big, *sight+s+ing, *punish+s+ment
- big+er, hollow+est
  - *interesting+er, *ridiculous+est
Base (citation) form

- Dictionaries typically don’t contain all morphological variants of a word
- Citation form: base form, lemma
- Languages, dictionaries differ on citation form
  - Armenian: verbs listed with first person sg.
  - Semitic languages: triliteral roots
  - Chinese/Japanese: character stroke order
Derivational morphology

- Changes meaning and/or category (do+able, adjourn+ment, deposition, un+lock, teach+er)
- Allows leveraging words of other categories (import)
- Not very productive
- Derivational morphemes usually surround root
Variation: morphology

217 air conditioning system
24  air conditioner system
 ▪ 1  air condition system
 ▪ 4  air start motor
 ▪ 48  air starter motor
 ▪ 131 air starting motor
 ▪ 91 combustion gases
 ▪ 16 combustible gases
 ▪ 5 washer fluid
 ▪ 1 washing fluid

4 synchronization solenoid
 ▪ 19 synchronizing solenoid
 ▪ 85 vibration motor
 ▪ 16 vibrator motor
 ▪ 118 vibratory motor
 ▪ 1  blowby / airflow indicator
 ▪ 12 blowby / air flow indicator
 ▪ 18 electric system
 ▪ 24 electrical system
 ▪ 3 electronic system
 ▪ 1 electronics system

1 cooling system pressurization pump group
 ▪ 103 cooling system pressurizing pump group
Traditional analysis

d/ba7riyjuiuynnveiq

Prefix
Root
Suffix
Ending
The PC-Kimmo system

- System for doing morphology
- Distributed by SIL for fieldwork, text analysis
- Components
  - Lexicons: inventory of morphemes
  - Rules: specify patterns
  - Word grammar (optional): specify word-level constraints on order, structure of morpheme classes
Sample rule, table, automaton

;;; Optional syncope rule
;;; Note: free variation
;;; L: Lu+ad+s+pastEd
;;; S: L00ad0s0pastEd
RULE
"u:0 => [L|T'] _ +:@ VW" 4 6

<table>
<thead>
<tr>
<th>u</th>
<th>L</th>
<th>+</th>
<th>VW</th>
<th>@</th>
<th>T'</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>L</td>
<td>@</td>
<td>VW</td>
<td>@</td>
<td>T'</td>
</tr>
<tr>
<td>1: 0 2 1 1 1 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2: 3 2 1 1 1 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: 1 0 4 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: 1 0 0 1 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample parses

PC-KIMMO> recognize gWEdsutudZildubut
gWE+d+s+?u+^tudZil+du+b+ut
Dub+my+Nomz+Perf+bend_over+OOC+Midd+Rfx

PC-KIMMO> recognize adsukWaxWdubs
ad+s+?u+^kWaxW+du+b+s
Your+Nomz+Perf+help+OOC+Midd+his/hers
Sample constituency graph
Sample generation

PC-KIMMO>generate  ad+^pastEd=al?txW
adpastEdal?txW

PC-KIMMO>generate  ad+s+?u+^kWaxW+du+b+s
adsukWaxWdubs

PC-KIMMO>generate  Lu+ad+s+al?txW
Luadsal?txW
Ladsal?txW
PC-KIMMO>recognize ?acqW|a?stq1sCnCsa
?ac+qW|a?=stq=ls+Cn+Csa    stative+ache=fire=head+SubjITrx1s+again

ASPTENSE
?ac+
stative+
VFrame
Root3
Root2
LSUFF
=ls
=head
Root1
FSUFF
=stq
ROOT
qW|a?
ache
Armenian word graph

<table>
<thead>
<tr>
<th>Word</th>
<th>NDet</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDecl</td>
<td>ART</td>
</tr>
<tr>
<td>NBase</td>
<td>CASE</td>
</tr>
<tr>
<td>ROOT</td>
<td>PLURAL</td>
</tr>
<tr>
<td>tjpax'dowt'iwn</td>
<td>+ny'r</td>
</tr>
<tr>
<td>woe_tribulation</td>
<td>+plural</td>
</tr>
<tr>
<td></td>
<td>+1sPoss.</td>
</tr>
<tr>
<td></td>
<td>+ov</td>
</tr>
<tr>
<td></td>
<td>+Inst</td>
</tr>
<tr>
<td></td>
<td>+s</td>
</tr>
</tbody>
</table>