Week 2

2016年7月19日 星期二　　上午9:30

Welcome to **Week Two** of the **Introduction to Natural Language Processing** class.

In Week One, we covered some introductory concepts such as language ambiguity and the fundamentals of linguistics. You also got access to the first weekly quiz. You have three attempts to do this quiz before the due date.

We are now in Week Two. You may want to start working on the **first programming assignment** this week. We have **not covered** all the material needed for this first assignment yet but you can read it now in order to (1) understand what it is about and (2) to make sure that you have all the necessary **Python** libraries (e.g., the right versions of numpy and scikit-learn) installed.

If you have signed up for this class only to watch the lectures, you don't need to worry about the programming assignments. If you have signed up for a certificate, you will need to get at least a **70**% on **each** assignment in the course, including the ten quizzes, the three programming assignments, and the final exam.

This first assignment is on the topic of **Dependency Parsing** (DP). DP is one way in which computers can learn to understand the relationship between the words in a sentence. For example the sentence "John likes Mary" is converted into a dependency parse tree, with "likes" as the root of the tree and both "John" and "Mary" as the children of this root node. The root node indicates the main meaning of the sentence, i.e., that the sentence is about 'liking'. The root node has two children, "John" and "Mary". The first one of them indicates the role of the "liker" and the second one indicates the role of the "liked". More information about Dependency Parsing will be given in a future lecture.

In Week Two, I will cover Parts of Speech, Morphology, Text Similarity, and Text Preprocessing. I will also introduce NACLO, the North American Computational Linguistics Olympiad (www.nacloweb.org), a competition for high school students interested in NLP and Linguistics.

2.1

NLP

02.01 Parts of Speech
NATURAL LANGUAGE PROCESSING

Introduction to NLP

Parts of speech

Syntactic Categories

- Substitution test:

Nathalie likes { *black* / *Persian* / *tabby* / *small* } *cats.*

- Open (lexical) and closed (functional) categories:

No-fly-zone	*the*
twerk	*in*

02.01 Parts of Speech NATURAL LANGUAGE PROCESSING

Example

The dog chased the yellow bird.

- Parts of speech
  - eight (or so) general types
  - nouns, verbs, adjectives…

Nouns

- Examples
  - *dog, tree, computer, idea*
- Nouns vary in
  - number (singular, plural)
  - gender (masculine, feminine, neuter)
  - case (nominative, genitive, accusative, dative)
- Case example in Latin
  - *Singular: puer (nominative), puerum (accusative), puerī (genitive)*
  - *Plural: puerī (nominative), puerōs (accusative), puerōrum (genitive)*
- Gender example in German
  - *Mädchen* (neuter gender)

Jabberwocky (Lewis Carroll)

`Twas **brillig**, and the **slithy toves**
Did gyre and gimble in the **wabe**:
All **mimsy** were the borogoves,
And the **mome raths outgrabe**.

- What are the parts of speech for the words in bold?

Answers

`Twas **brillig**, and the **slithy toves**
Did gyre and gimble in the **wabe**:
All **mimsy** were the borogoves,
And the **mome raths outgrabe**.

- Wabe, borogoves
  - Nouns (after "the")
- brillig
  - adjective?
  - noun? ("noon")
- mimsy
  - adjective
- slighty toves
  - adjective+noun?
  - noun+verb? ("the bell tolls")
- mome raths outgrabe
  - Adjective+noun+verb?
  - Noun+verb+adverb? ("birds fly outside")

02.01 Parts of Speech
NATURAL LANGUAGE PROCESSING

Why is this an Important Example?

- Computers see text that they don't really understand.
- They have to use some prior knowledge.
- They reason probabilistically.
- They use context.
- The can be wrong.

Pronouns

- Examples
  - *she, ourselves, mine*
- Pronouns vary in
  - person
  - gender
  - number
  - case (in English: nominative, accusative, possessive, 2nd possessive)
- Reflexive and anaphoric forms
  - *herself, each other*

Samantha gave her a haircut.
Samantha gave herself a haircut.

reflexive

02.01 Parts of Speech NATURAL LANGUAGE PROCESSING

Determiners and Adjectives

- Determiners
  - Articles
    - *the, a*
  - Demonstratives
    - *this, that*
- Adjectives
  - describe properties
  - attributive and predicative adjectives
  - agreement
    - in gender, number
  - comparative and superlative forms
    - derivative and periphrastic
  - positive form

shorter shortest. | more most difficult
short.

Verbs

- Describe
  - actions, activities, and states (*throw, walk, have*)
- English
  - four verb forms
- tense
  - present, past, future
- other inflection
  - number, person
- gerunds and infinitive
- aspect
  - progressive, perfective
- Voice
  - active, passive

02.01 Parts of Speech NATURAL LANGUAGE PROCESSING

Verbs

- Participles, auxiliaries
- Arguments:
  - The dog sleeps (intransitive)
  - The dog chased the cat (transitive)
  - Mary gave the dog a bone (ditransitive)
- Irregular verbs
- Richer inflections
  - E.g., French, Latin, and Finnish

不及物动词
及物动词.
A verb that takes two objects.

Verb Conjugation in French

Present	Past	Imperfect
je vais *I go*	je suis allé(e) *I went*	j'allais *I used to go*
tu vas *you go*	tu es allé(e) *you went*	tu allais *you used to go*
il va *he goes*	il est allé(e) *he went*	il allait *he used to go*
nous allons *we go*	nous sommes allé(e)s *we went*	nous allions *we used to go*
vous allez *you go*	vous êtes allé(e)s *you went*	vous alliez *you used to go*
ils vont *they go*	ils sont allé(e)s *they went*	ils allaient *they used to go*

Conditional	Future	*Subjunctive*
j'irais *I would go*	j'irai *I will go*	que j'aille *that I go*
tu irais *you would go*	tu iras *you will go*	que tu ailles *that you go*
il irait *he would go*	il ira *he will go*	qu'il aille *that he go*
nous irions *we would go*	nous irons *we will go*	que nous allions *that we go*
vous iriez *you would go*	vous irez *you will go*	que vous alliez *that you go*
ils iraient *they would go*	ils iront *they will go*	qu'ils aillent *that they go*

Other Parts of Speech

- Adverbs
  - happily, here, never
- Prepositions
  - of, through, in
- Particles
  - Phrasal verbs
  - the plane took off / take it off
- Particles vs. prepositions
  - *She ran up a bill/hill*

Other Parts of Speech

- Coordinating conjunctions
  - *and, or, but*
- Subordinating conjunctions

– and, or, but

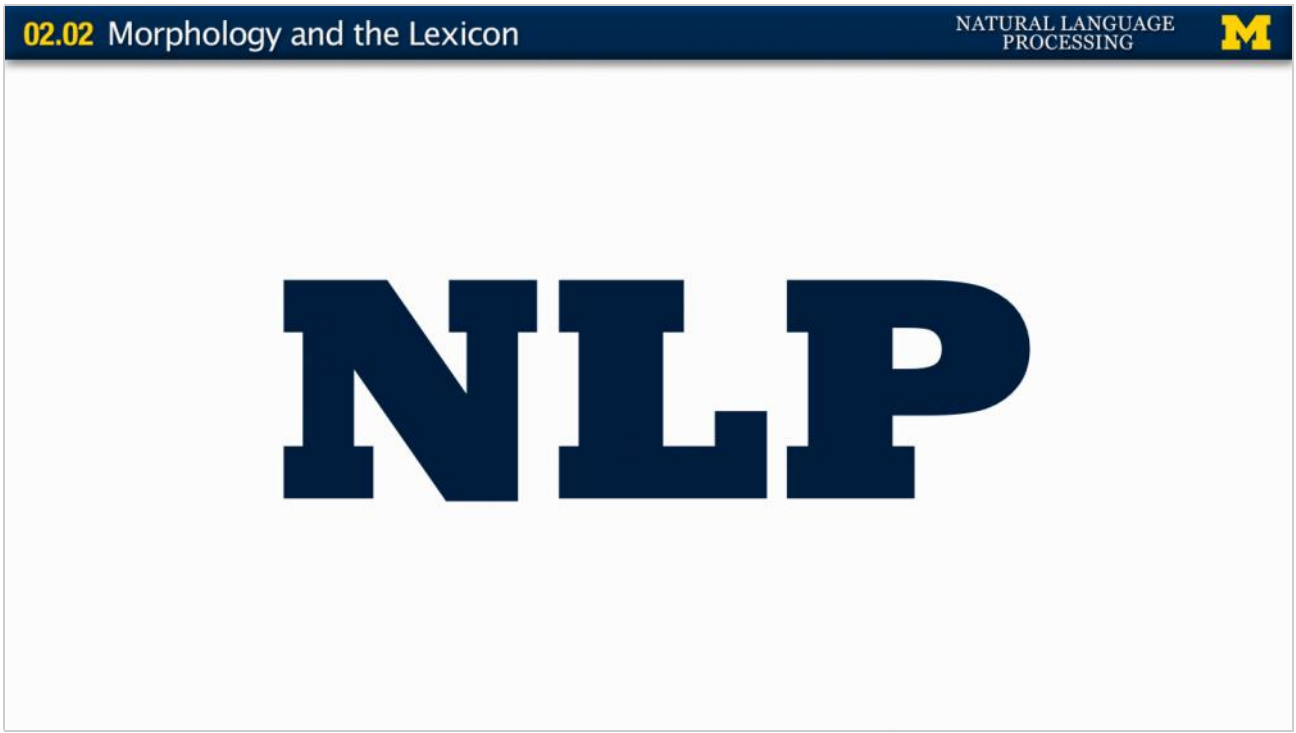
- Subordinating conjunctions
  – if, because, that, although
- Interjections
  – Ouch!

Part of Speech Tags

```
NN    /* singular noun */
IN    /* preposition */
AT    /* article */
NP    /* proper noun */
JJ    /* adjective */
,     /* comma */
NNS   /* plural noun */
CC    /* conjunction */
RB    /* adverb */
VB    /* un-inflected verb */
VBN   /* verb +en (taken, looked (passive,perfect)) */
VBD   /* verb +ed (took, looked (past tense)) */
CS    /* subordinating conjunction */
```


02.01 Parts of Speech
NATURAL LANGUAGE PROCESSING
NLP

2.2


02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING
NLP

Introduction to NLP

Morphology and the Lexicon

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

Mental Lexicon

- What is the meaning of cat? Its pronunciation? Part of speech?
- What is the meaning of wug?
- What is the meaning of cluvious?
- Compare traftful and traftless?
- Morphology of these words
- Intuition and productivity
- "Runs"
- Allomorphs – "cats/oxen", "played/swung"
- Affixes

start / ending

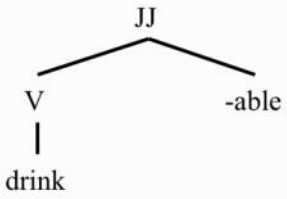
不规则的变化

ox – oxen

swing – swung

Derivational Morphology

- Er (many examples)
- What do these morphemes mean?
  - Ness, able, ing, re, un, er (adj)
  - JJ → V + "-able"
- Recursion:
  - unconcernednesses
- Ambiguous – undoable
- Not ambiguous – unbelievable – why?



turn adj. to n.

negation.

two interpretations.

undo-able
un-doable

Answer to the Quiz

- Undoable
  - unable to be done
  - able to be undone
- Unbelievable
  - unable to be believed
  - ? able to be unbelieved

unnatural

Morphological Examples

- Reduplication
  - amigo = friend, amimígo = friends (in Pangasinan) [Rubino 2001]
  - savali = he travels, savavali = they travel (in Samoan)
- Circumfixes
  - spielen – gespielt (in German)
- Pig Latin
  - appyhay
- Verlan
  - "céfran", "ripou" (from "l'envers", "Français", "pourri")
- Massa-*freakin'*-chusetts
  - where can you insert "freakin'" in "education"?

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

Answer to the Quiz

- The "freakin'" infix is inserted
- … to the left of the syllable that bears the main stress

  edu-*freakin'*-cation

  * educa-*freakin'*-tion

  * e-*freakin'*-ducation
- though there can be exceptions

Morphemes

- Stems, affixes
- Concatenative morphology
- Templatic morphology (e.g., Semitic languages):
  - lmd (learn), lamad (he studied), limed (he taught), lumad (he was taught)

Inflectional Morphology

- Tense, number, person, mood, aspect
- Five verb forms in English (am are was were been) ex.
- 40+ forms in French
- Six cases in Russian:
  http://www.departments.bucknell.edu/russian/language/case.html
- Up to 40,000 forms in Turkish (you cause X to cause Y to … do Z) recurse set of rules to make long words.

Morphological Analysis

- sleeps = sleep + V + 3P + SG
- done = do + V + PP

Turkish Vowel Harmony

	Front		Back	
	Unrounded	Rounded	Unrounded	Rounded
High	i	ü	ı	u
Low	e	ö	a	o

- Back vowels
  - in the room → odada
  - at the door → kapıdā
- Front vowels
  - at home → evde
  - at the lake → gölde
  - on the bridge → köprüde

NACLO Problem

- Turkish
  - www.naclo.cs.cmu.edu/problems2010/F.pdf

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

NACLO Solution

- Turkish
  - www.naclo.cs.cmu.edu/problems2010/FS.pdf

Agglutinative Languages

How to convert.

- How does English become Turkish?

if	we	will	be	able	to	make	...	become	strong
if	we	will	be	able	to	make	...	become	strong
...	strong	become	to	make	be	able	will	if	we
...	sağlam	+laş		+tır		+abil	+ecek	+se	+k

⇓

... sağlamlaştırabileceksek

reorder labels.

Slide from Kemal Oflazer

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

アメフト	amefuto	Ame(rican) Foot(ball)
アイスクリーム	aisu kurīmu	ice cream
アイドル	aidoru	idol
アパート	apāto	apartment
バイク	baiku	bike
バリアフリー	bariafurī	barrier free
コンピューター	konpyūtā	computer
デスク	desuku	desk (at a news agency)
ラマ	dorama	drama (on TV)
エレベーター	erebētā	elevator
エスカレーター	esukarētā	escalator
フライドポテト	furaidopoteto	fried potato (French fries)
グラス	gurasu	glass (for drinking)
ハッピーエンド	happīendo	happy end(ing)
ホットケーキ	hottokēki	hotcake (pancake)
カシューナッツ	kashū nattsu	cashew nut
コーヒー	kōhī	coffee
クラブ	kurabu	club
キーボード	kībōdo	keyboard
キャンペーン	kyanpēn	campaign
キャップ	kyappu	cap
パソコン	pāsokon	perso(nal) com(puter)
パーソナルコンピューター	pāsonaru konpyūtā	personal computer
レジュメ	rejume	resume
レストラン	resutoran	restaurant
リモコン	rimokon	remo(te) con(trol)
サラダ	sarada	salad
タバコ	tabako	tobacco
テレビゲーム	terebigēmu	television game
ゼミナール	zemināru	seminar

Japanese.

音译的词.

Introduction to NLP

Other Levels of Linguistic Analysis

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

Semantics

- Semantics
  - Lexical semantics and compositional semantics
- Lexical Semantics
  - Hypernyms, hyponyms, antonyms, meronyms and holonyms (part-whole relationship, tire is a meronym of car), synonyms, homonyms
  - Senses of words, polysemous words
  - Collocations
    - *white hair, white wine*
  - Idioms
    - *to kick the bucket*
- *Compositional Semantics*
  - How to understand the meaning of a sentence based on the meaning of its components.

meaning of individual words; and relationship of words.

stock market.

Pragmatics

beyond literal meaning of the sentence

- The study of how knowledge about the world and language conventions interact with literal meaning.
- Speech acts
- Resolution of anaphoric relations
- Modeling of speech acts in dialogue

02.02 Morphology and the Lexicon
NATURAL LANGUAGE PROCESSING

Other Areas

- Sociolinguistics
  - interactions of social organization and language.
- Historical linguistics
  - change over time.
- Linguistic typology
- Language acquisition
  - L1 and L2
- Psycholinguistics
  - real-time production and perception of language

how people produce and perceive language in real time

2.3

Text similarity

Introduction

02.03 Introduction
NATURAL LANGUAGE PROCESSING

Text Similarity

What is it importance.

- People can express the same concept (or related concepts) in many different ways. For example, "the plane leaves at 12pm" vs "the flight departs at noon"
- Text similarity is a key component of Natural Language Processing
- If the user is looking for information about cats, we may want the NLP system to return documents that mention kittens even if the word "cat" is not in them.
- If the user is looking for information about "fruit dessert", we want the NLP system to return documents about "peach tart" or "apple cobbler".
- A speech recognition system should be able to tell the difference between similar sounding words like the "Dulles" and "Dallas" airports.
- This set of lectures will teach you how text similarity can be modeled computationally.

speech.

Human Judgments of Similarity

tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
computer	keyboard	7.62
computer	internet	7.58
plane	car	5.77
train	car	6.31
telephone	communication	7.50
television	radio	6.77
media	radio	7.42
drug	abuse	6.85
bread	butter	6.19
cucumber	potato	5.92

→ full mark.

10 max.

[Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, "Placing Search in Context: The Concept Revisited", ACM Transactions on Information Systems, 20(1):116-131, January 2002]

02.03 Introduction
NATURAL LANGUAGE PROCESSING

Human Judgments of Similarity

delightful	wonderful	A	8.65
modest	flexible	A	0.98
clarify	explain	V	8.33
remind	forget	V	0.87
get	remain	V	1.6
realize	discover	V	7.47
argue	persuade	V	6.23
pursue	persuade	V	3.17
plane	airport	N	3.65
uncle	aunt	N	5.5
horse	mare	N	8.33

- used to train or evaluate systems.

[SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. 2014. Felix Hill, Roi Reichart and Anna Korhonen. Preprint pubslished on arXiv. arXiv:1408.3456]

Automatic Similarity Computation

spain	0.679
belgium	0.666
netherlands	0.652
italy	0.633
switzerland	0.622
luxembourg	0.610
portugal	0.577
russia	0.572
germany	0.563
catalonia	0.534

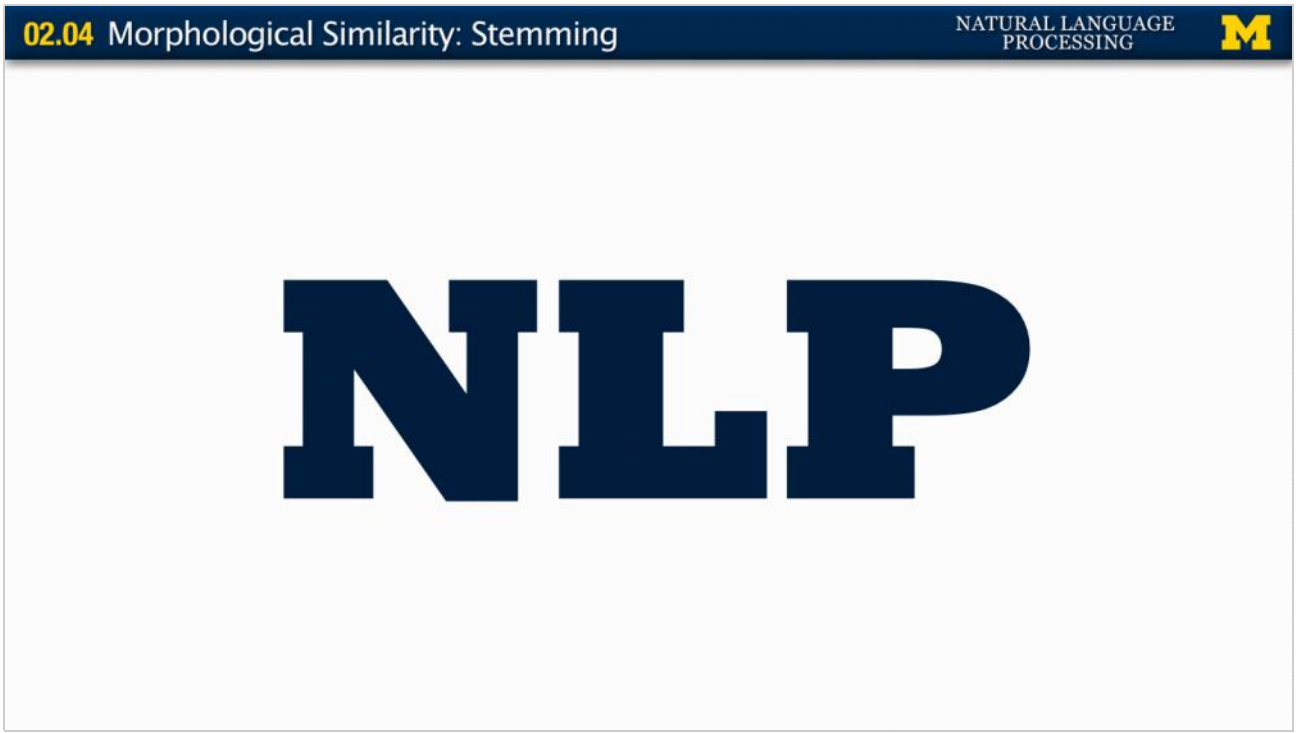
- Words most similar to "France"
- Computed using "word2vec"
- [Mikolov et al. 2013]

Types Of Text Similarity

- Many types of text similarity exist:
  - Morphological similarity (e.g., respect-respectful)
  - Spelling similarity (e.g., theater-theatre)
  - Synonymy (e.g., talkative-chatty)
  - Homophony (e.g., raise-raze-rays)
  - Semantic similarity (e.g., cat-tabby)
  - Sentence similarity (e.g., paraphrases)
  - Document similarity (e.g., two news stories on the same event)
  - Cross-lingual similarity (e.g., Japan-Nihon)


02.03 Introduction
NATURAL LANGUAGE PROCESSING
NLP

2.4


02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING
NLP

Text Similarity

Morphological Similarity: Stemming

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Morphological Similarity

- Words with the same root:
  - scan (base form)
  - scans, scanned, scanning (inflected forms)
  - scanner (derived forms, suffixes)
  - rescan (derived forms, prefixes)
  - rescanned (combinations)

Stemming

- To stem a word is to reduce it to a base form, called the *stem*, after removing various suffixes and endings and, sometimes, performing some additional transformations
- Examples
  - *scanned* → *scan*
  - *indication* → *indicate*
- In practice, prefixes are sometimes preserved, so *rescan* will not be stemmed to *scan*

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Porter's Stemming Method

- Porter's stemming method is a rule-based algorithm introduced by Martin Porter in 1980
- The paper ("An algorithm for suffix stripping") has been cited more than 7,000 times according to Google Scholar
- The input is an individual word. The word is then transformed in a series of steps to its stem
- The method is not always accurate

Porter's Algorithm

- Example 1:
  - Input = *computational*
  - Output = *comput*
- Example 2:
  - Input = *computer*
  - Output = *comput*
- The two input words end up stemmed the same way

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Porter's Algorithm

measure of a word.

- The *measure* of a word is an indication of the number of syllables in it
  - Each sequence of consonants is denoted by C
  - Each sequence of vowels is denoted as V
  - The initial C and the final V are optional
  - So, each word is represented as [C]VCVC ... [V], or [C](VC){k}[V], where *k* is its *measure*

repeated k times

how many sequences of VC we have

Examples of Measures

- k=0: I, AAA, CNN, TO, GLEE
- k=1: OR, EAST, BRICK, STREET, DOGMA
- k=2: OPAL, EASTERN, DOGMAS
- k=3: EASTERNMOST, DOGMATIC

02.04 Morphological Similarity: Stemming NATURAL LANGUAGE PROCESSING

Porter's Algorithm

- The initial word is then checked against a sequence of transformation patterns, in order.
- An example pattern is:
  - `(m>0) ATION -> ATE` `medication -> medicate`
- Note that this pattern matches *medication* and *dedication*, but not *nation*.
- Whenever a pattern matches, the word is transformed and the algorithm restarts from the beginning of the list of patterns with the transformed word.
- If no pattern matches, the algorithm stops and outputs the most recently transformed version of the word.

Example Rules

- Step 1a

SSES	-> SS	presses	->	press
IES	-> I	lies	->	li
SS	-> SS	press	->	press
S	-> ø	lots	->	lot

y -> i

- Step 1b

(m>0) EED -> EE refereed -> referee

(doesn't apply to bleed since m('BL')=0)

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Example Rules

- Step 2

(m>0) ATIONAL	-> ATE	inflational	->	inflate
(m>0) TIONAL	-> TION	notional	->	notion
(m>0) IZER	-> IZE	nebulizer	->	nebulize
(m>0) ENTLI	-> ENT	intelligentli	->	intelligent
(m>0) OUSLI	-> OUS	analogousli	->	analogous
(m>0) IZATION	-> IZE	realization	->	realize
(m>0) ATION	-> ATE	predication	->	predicate
(m>0) ATOR	-> ATE	indicator	->	indicate
(m>0) IVENESS	-> IVE	attentiveness	->	attentive
(m>0) ALITI	-> AL	realiti	->	real
(m>0) BILITI	-> BLE	abiliti	->	able

Example Rules

- Step 3

(m>0) ICATE -> IC	replicate -> replic
(m>0) ATIVE -> ø	informative -> inform
(m>0) ALIZE -> AL	realize -> real
(m>0) ICAL -> IC	electrical -> electric
(m>0) FUL -> ø	blissful -> bliss
(m>0) NESS ->	tightness -> tight

- Step 4

(m>1) AL -> ø	appraisal -> apprais
(m>1) ANCE -> ø	conductance -> conduct
(m>1) ER -> ø	container -> contain
(m>1) IC -> ø	electric -> electr
(m>1) ABLE -> ø	countable -> count
(m>1) IBLE -> ø	irresistible -> irresist
(m>1) EMENT -> ø	displacement -> displac
(m>1) MENT -> ø	investment -> invest
(m>1) ENT -> ø	respondent -> respond

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Examples

- Example 1:
  - Input = *computational*
  - Step 2: replace *ational* with *ate*: *computate*
  - Step 4: replace *ate* with *ø*: *comput*
  - Output = *comput*
- Example 2:
  - Input = *computer*
  - Step 4: replace *er* with *ø*: *comput*
  - Output = *comput*
- The two input words end up stemmed the same way

External Pointers

- Online demo
  - http://text-processing.com/demo/stem/
- Martin Porter's official site
  - http://tartarus.org/martin/PorterStemmer/

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

Quiz

- How will the Porter stemmer stem these words?

`construction`	?	Construct ✓
`increasing`	?	increas ✓
`unexplained`	?	unexplain ✓
`differentiable`	?	diff. ✗

- Check the Porter paper (or the code for the stemmer) in order to answer these questions.
- Is the output what you expected? If not, explain why.

Answers to the Quiz

construction	?
increasing	?
unexplained	?
differentiable	?

construction	**construct**
increasing	**increas**
unexplained	**unexplain**
differentiable	**differenti**

→ not taking into account prefix.

☆ where Porter's algorithm fall short

02.04 Morphological Similarity: Stemming
NATURAL LANGUAGE PROCESSING

NACLO Problem

☆ Exercise. Can solve it myself

- Thorny Stems, NACLO 2008 problem by Eric Breck
  - http://www.naclo.cs.cmu.edu/assets/problems/NACLO08h.pdf

Solution to the NACLO Problem

Solution to the prev. slide

- Thorny Stems
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-HS.pdf

02.04 Morphological Similarity: Stemming NATURAL LANGUAGE PROCESSING

NLP

2.5

NLP

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Text Similarity

Spelling Similarity: Edit Distance

Spelling Similarity

- Typos:
  - Brittany Spears -> Britney Spears
  - Catherine Hepburn -> Katharine Hepburn
  - Reciept -> receipt
- Variants in spelling:
  - Theater -> theatre

typo

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Who Is This?

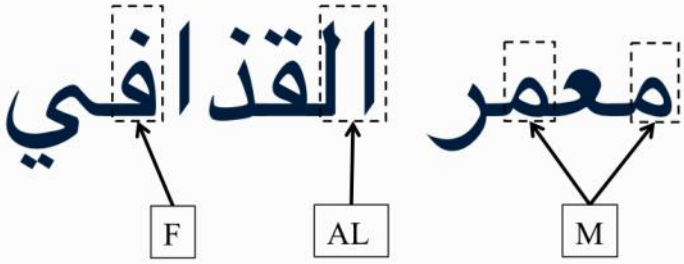
معمر القذافي

Hints

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

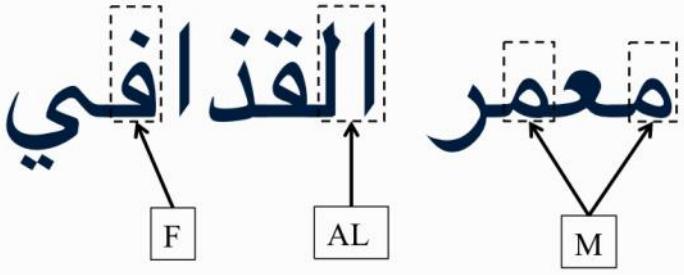
Hints

Hints



02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Hints



Muammar (al-)Gaddafi, or **Moamar Khadafi,** or …

Quiz

How many different transliterations can there be?

m u o a m mm a e r	el al El Al ø	Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

A Lot!

m u o a m mm a e r		el al El Al ø		Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y		
8	x	5	x	360	=	14,400

lot of possibilities

Edit Operations

behavior ← insertion

- behaviour – behavior (insertion/deletion) (“al”)
- string – spring (substitution) (“k”–”q”)
- sleep – slept (multiple edits)

spring
string

sleep → slept
- Replace. replace.
- swap, replace

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Levenshtein Method

- Based on dynamic programming
- Insertions, deletions, and substitutions usually all have a cost of 1.

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	5								

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Recurrence Relation

- Definitions
  - $s_1(i)$ – i^{th} character in string s_1
  - $s_2(j)$ – j^{th} character in string s_2
  - $D(i,j)$ – edit distance between a prefix of s_1 of length i and a prefix of s_2 of length j
  - $t(i,j)$ – cost of aligning the i^{th} character in string s_1 with the j^{th} character in string s_2

- Recursive dependencies

$$D(i,0)=i$$
$$D(0,j)=j$$
$$D(i,j)=\min[$$
$$D(i-1,j)+1$$
$$D(1,j-1)+1$$
$$D(i-1,j-1)+t(i,j)$$
$$]$$

- Simple edit distance:

$$t(i,j)=0 \ \textit{iff} \ s_1(i)=s_2(j)$$
$$t(i,j)=1, \ \textit{otherwise}$$

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2	2							
e	3								
n	4								
d	5								

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2					
r	2	2							
e	3								
n	4								
d	5								

$D(i-1, j-1) + t(i, j)$

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

02.05 Spelling Similarity: Edit Distance NATURAL LANGUAGE PROCESSING

Example

keep track of the cell

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Edit Distance Matrix

Edit Transcript

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Other Costs

- Damerau modification
  - Swaps of two adjacent characters also have a cost of 1
  - E.g., Lev("cats","cast") = 2, Dam("cats","cast") = 1

Damerau swap cost of 1

02.05 Spelling Similarity: Edit Distance NATURAL LANGUAGE PROCESSING

Quiz

- Some distance functions can be more specialized.
- Why do you think that the edit distances for these pairs are as follows?
  - Dist ("sit clown","sit down") = 1
  - Dist ("qeather","weather") = 1, *but* Dist ("leather","weather") = 2

handwriting

keyboard adjacent

Quiz Answers

- Dist("sit down","sit clown") is lower in this example because we want to model the type of errors common with optical character recognition (OCR)
- Dist("qeather","weather") < Dist("leather","weather") because we want to model spelling errors introduced by "fat fingers" (clicking on an adjacent key on the keyboard)

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Quiz: Guess the Language

DNA sequence.

AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC

Quiz Answer

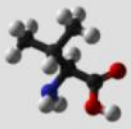
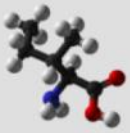
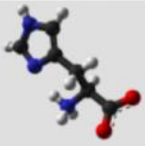
- This is a genetic sequence (nucleotides AGCT)

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC
```

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

Other Uses of Edit Distance

- In biology, similar methods are used for aligning non-textual sequences
  - Nucleotide sequences, e.g., GTTCGTGATGGAGCG, where A=adenine, C=cytosine, G=guanine, T=thymine, U=uracil, "-"=gap of any length, N=either one of ACGTU, etc.
  - Amino acid sequences, e.g., FMELSEDGIEMAGSTGVI, where A=alanine, C=cystine, D=aspartate, E=glutamate, F=phenylalanine, Q=glutamine, Z=either glutamate or glutamine, X="any", etc. The costs of alignment are determined empirically and reflect evolutionary divergence between protein sequences. For example, aligning V (valine) and I (isoleucine) is lower-cost than aligning V and H (histidine).

Valine	Isoleucine	Histidine
		

External URLs

- Levenshtein demo
  - http://www.let.rug.nl/~kleiweg/lev/
- Biological sequence alignment
  - http://www.bioinformatics.org/sms2/pairwise_align_dna.html
  - http://www.sequence-alignment.com/sequence-alignment-software.html
  - http://www.ebi.ac.uk/Tools/msa/clustalw2/
  - http://www.animalgenome.org/bioinfo/resources/manuals/seqformats

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

NACLO Problem

- "Nok-Nok", NACLO 2009 problem by Eugene Fink:
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-B.pdf

Solution to the NACLO Problem

- "Nok-Nok"
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-BS.pdf

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

NACLO Problem

sequence alignment.

- "The Lost Tram", NACLO 2007 problem by Boris Iomdin:
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-F.pdf

Solution to the NACLO problem

- "The Lost Tram"
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-FS.pdf

02.05 Spelling Similarity: Edit Distance
NATURAL LANGUAGE PROCESSING

NLP

2.6

个语言学竞赛.

NLP

02.06 NACLO
NATURAL LANGUAGE PROCESSING

Introduction to NLP

NACLO

NACLO

- Competition in Linguistics (including Computational Linguistics)
  - Since 2007
  - http://www.naclo.cs.cmu.edu
- Best individual US performers so far:
  - Adam Hesterberg (2007)
  - Hanzhi Zhu (2008)
  - Rebecca Jacobs (2007–2009) – 3 team golds + 2 individual medals
  - Ben Sklaroff (2010)
  - Morris Alper (2011)
  - Alex Wade (2012, 2013) – 2 team golds + 2 individual golds + 1 individual silver
  - Darryl Wu (2012, 2014)
- Other strong countries:
  - Russia, UK, Netherlands, Poland, Bulgaria, South Korea, Canada, China
- IOL – the International contest
  - Since 2003
  - IOL 2013 in Manchester, IOL 2014 in Beijing, IOL 2015 in Bulgaria
  - http://www.ioling.org
- Other high school competitions, e.g., IMO, IOI, IPhO, IChO, IBO, IOAA, etc.

02.06 NACLO
NATURAL LANGUAGE PROCESSING

Consider these phrases in Ancient Greek (in a Roman-based transcription) and their unordered English translations:

(A) *ho tōn hyiōn dulos*	(1) the donkey of the master
(B) *hoi tōn dulōn cyrioi*	(2) the brothers of the merchant
(C) *hoi tu emporu adelphoi*	(3) the merchants of the donkeys
(D) *hoi tōn onōn emporoi*	(4) the sons of the masters
(E) *ho tu cyriu onos*	(5) the slave of the sons
(F) *ho tu oicu cyrios*	(6) the masters of the slaves
(G) *ho tōn adelphōn oicos*	(7) the house of the brothers
(H) *hoi tōn cyriōn hyioi*	(8) the master of the house

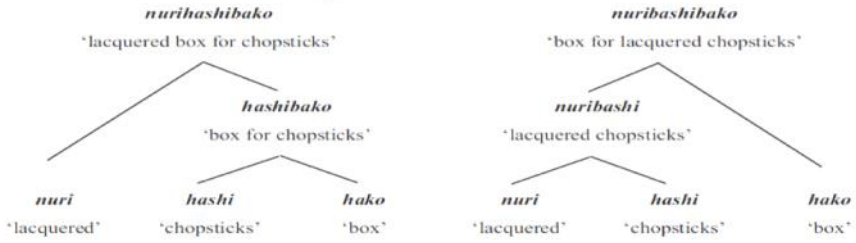
C1. Place the number of the correct English translation in the space following each Greek sentence. Explain your answers!

C2. Translate into Ancient Greek:
the houses of the merchants;
the donkeys of the slave
Explain your answers!

A Donkey in Every House, by Ivan Derzhanski

In English, we can combine two nouns to get a compound noun, such as in 'mailbox' or 'sandcastle'. We can do this in Japanese as well, but just sticking the two words together isn't enough. Instead, the words themselves undergo predictable changes:

Compound words can then be compounded again, creating compounds with three or more members. Study the diagrams below carefully. You'll notice that the order in which the compound is built affects both the meaning and the final form of the word.



Fakepapershelfmaker, by Willie Costello

02.06 NACLO
NATURAL LANGUAGE PROCESSING

An excerpt from a well known text is shown below. It is in two languages (X and Y) that are closely linguistically related to each other and also to English. However the two versions are not perfect translations of one another.

Text in language X

X1. Rödluvan: Men mormor, varför har du så stora ögon?
X2. "Mormor": Det är bara för att jag skall se dig bättre, mitt barn.
X3. Rödluvan: Men mormor, varför har du så stora öron?
X4. "Mormor": Det är bara för att jag skall höra dig bättre, mitt barn.
X5. Rödluvan: Men mormor, varför har du så stora tänder?
X6. "Mormor": Det är bara för att jag skall kunna äta upp dig!

(almost) the same text in language Y

Y1. - Så store ører du har, bestemor, sa Rødhette.
Y2. - Det er fordi jeg skal kunne høre deg bedre, svarte ulven.
Y3. - Så store øyne du har, bestemor, sa Rødhette.
Y4. - Det er fordi jeg skal kunne se deg bedre, svarte ulven.
Y5. - Så store hender du har, bestemor, sa Rødhette.
Y6. - Det er fordi jeg skal kunne klemme deg bedre, svarte ulven.
Y7. - Så stor munn du har, bestemor, sa Rødhette.
Y8. - Det er fordi jeg skal kunne ete deg bedre, svarte ulven.

Rødhette, by Dragomir Radev

To the right is a Japanese word written in the ***tenji*** ("dot characters") writing system. The large dots represent the raised bumps; the tiny dots represent empty positions.

karaoke 

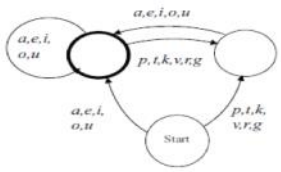
A1. The following *tenji* words represent ***atari***, ***haiku***, ***katana***, ***kimono***, ***koi***, and ***sake***. Which is which? You don't need to know either Japanese or Braille to figure it out; you'll find that the system is highly logical.

a. _____ 	b. _____ 
c. _____ 	d. _____ 
e. _____ 	f. _____ 

Tenji Karaoke, by Patrick Littell

Finite-state automata (FSA) are a type of abstract "machine" with many possible uses. One possible use is to guess what language a document (such as a webpage) is in. If we make an automaton that can distinguish between possible English words and impossible ones, and then give it a webpage with a bunch of words that are impossible in English (like "*aioaepa*" or "*ragaiiare*"), we can be pretty sure that the webpage isn't written in English. (Or, at least, isn't *entirely* written in English.)

Here is a finite state automaton that can distinguish between possible and impossible words in Rotokas, a language spoken on the island of Bougainville off the coast of New Guinea. Rotokas has a very simple system of sounds and allows us to create a very small FSA.



finite-state automata

An FSA works like a board game. Choose a word, and place your pencil on the space marked "Start". Going through the letters of the word one at a time, move your pencil along the path marked with that letter. If the word ends and you're at a space marked with a thicker circle, the word succeeds: it's a possible Rotokas word! If the word ends and you're not at a thicker circle, or you're midway through the word and there's no path corresponding to the next letter, the word fails: it's *not* a possible Rotokas word!

Try it out with these possible and impossible words; the automaton should accept all the possible words and reject the impossible ones.

Possible Rotokas words		Impossible Rotokas words	
tauo	*kareveiepa*	*grio*	*ouag*
puraveva	*ovokirovuia*	*ovgi*	*vonoka*
avaopa	*ouragaveva*	*gataap*	*oappa*

11. Now, using the automaton above, put a check mark next to each possible Rotokas word:

___ *iu*	___ *uente*	___ *voav*
___ *idau*	___ *urioo*	___ *uaia*
___ *oire*	___ *raorao*	___ *oratreopaveiepa*

Rotokas aw-TOM-uh-tuh, by Patrick Littell

writing system.

On her visit to Armenia, Millie has gotten lost in Yerevan, the nation's capital. She is now at the Metropoliten (subway) station named Shengavit, but her friends are waiting for her at the station named Barekamutyun. Can you help Millie meet up with her friends?

1. Assuming Millie takes a train in the right direction, which will be the first stop after Shengavit?

Note that all names of stations listed below appear on the map.

a. Gortsaranayin
b. Zoravar Andranik
c. Charbakh
d. Garegin Njdehi Hraparak
e. none of the above

2. After boarding at Shengavit, how many stops will it take Millie to get to Barekamutyun (don't include Shengavit itself in the number of stops)?

Lost in Yerevan, by Dragomir Radev

NACLO: Computational Problems

computational challenges

- http://clair.si.umich.edu/naclo/resources/resources.html

- List of computational problems:
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-O.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-J.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-F.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-H.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-N.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-Q.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-K.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-O.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-R.pdf
  - http://www.naclo.cs.cmu.edu/problems2011/F.pdf
  - http://www.naclo.cs.cmu.edu/problems2011/M.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/D.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/E.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/I.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/K.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-E.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-G.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-J.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-M.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-F.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-H.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-I.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-A.pdf
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-H.pdf

Convert text into usable format.

2.7

Introduction to NLP

Preprocessing

Normalize
stem
morpholog
capitalize.

02.07 Preprocessing NATURAL LANGUAGE PROCESSING

Text Preprocessing

- Removing non-text (e.g., ads, javascript)
- Dealing with text encoding (e.g., Unicode)
- Sentence segmentation ☆ Boundary of sentences.
- Normalization same word spelled differently → convert to the same word.
  - labeled/labelled, extra-terrestrial/extraterrestrial, extra terrestrial
- Stemming
  - computer/computation
- Morphological analysis
  - car/cars inflexive analysis.
- Capitalization
  - Now/NOW, led/LED organization name vs words with original meaning
- Named entity extraction ☆.
  - USA/usa

Text Preprocessing

- Types vs. Tokens
  - To be or not to be
- Tokenization:
  - ALS vs. A.L.S.
  - Paul's, Willow Dr., Dr. Willow, New York, ad hoc, can't
  - "The New York-Los Angeles flight" vs. "Minneapolis-St.Paul"
  - Numbers, e.g., (888) 555-1313, 1-888-555-1313
  - Dates, e.g., Jan-13-2012, 20120113, 13 January 2012, 01/13/12
  - URLs

type: any sequence of character that rep. specific words

4 types.

New York - Los Angeles

single #

02.07 Preprocessing NATURAL LANGUAGE PROCESSING

Word Segmentation

Chinese alike. no boundary in between words

- 金属製品製造の日立金属は19日、世界最大手の鉄鋳物メーカー「ワウパカ　ファウンドリー　ホールディングス」(米国・デラウェア州)を米投資ファンドから買収し、完全子会社にすると発表した。買収額は13億ドル(約1330億円)で、10月中にも手続きを終える。

Word Segmentation

- Arabic:

  كتاب → no words boundary. 先词 词隔不是书写的问题，并非分词

- Japanese:

  この本は重い。

  (kono hon ha omoi)

- German: long words.

  Finanzdienstleistung = financial services

  3 words merged together.
  problem: find articles about financial issue.

- Chinese:

  电视 (television)
  电 (diàn = electric) 视 (shì = to look at)

Text Preprocessing

ニューヨーク (New York) は、アメリカ合衆国ニューヨーク州にある都市

- Kanji, Katakana, Hiragana, Rōmaji, (numbers)
- Nyūyōku wa, Amerikagasshūkoku nyūyōku-shū ni aru toshi

Sentence Boundary Recognition

- Decision trees
- Features
  - punctuation
  - formatting
  - fonts
  - spacing
  - capitalization
  - case
  - use of abbreviations, e.g., Dr., a.m.
- Example
  - If there is no space after a period, don't assume that there is a sentence boundary

02.07 Preprocessing NATURAL LANGUAGE PROCESSING

For the rest of this course. text used are properly pre-processed

NLP