CS 4650/7650: Natural Language Processing

Jacob Eisenstein

Lecture 7: Morphology

September 9, 2014

On the inside of words: lumping

- For sentiment analysis, we may want to merge hate, hates, hated, hating into the **lemma** HATE.
- For topic classification, we want to merge computer and computers
- Sometimes this isn't so easy:
  - children→child
  - geese→goose
  - went→go
  - meeting→?

Lemmatization and stemming

Lemmatization is finding the right **lemma** for a **surface form**.

- cats → cat
- ponies → pony
- faster → fast
- better → good

Stemming is a simplification of this task.

- "Knowledge free," character-level transduction rules.
- cats → cat
- ponies → poni
- faster → fast
- better → bett

Stemming in NLP

- Code online:
  `http://tartarus.org/martin/PorterStemmer/python.txt`
- Should you stem (or lemmatize) your text?

Stemming in NLP

- Code online:
  `http://tartarus.org/martin/PorterStemmer/python.txt`
- Should you stem (or lemmatize) your text?
  - In IR, maybe (recall / precision tradeoff)
  - In (supervised) NLP, stems and lemmas can be **features**

On the inside of words: splitting

Morphology often indicates **when** events happen. For example, in French:

J'achete un vélo	I buy a bicycle (now)
J'acheterai un vélo	I will buy a bicycle
J'achetais un vélo	I was buying a bicycle
J'ai acheté un vélo	I bought a bicycle
J'acheterais un vélo	I would buy a bicycle

On the inside of words: splitting

Morphology often indicates **when** events happen. For example, in French:

J'achet**e** un vélo	I buy a bicycle (now)
J'achet**erai** un vélo	I will buy a bicycle
J'achet**ais** un vélo	I was buying a bicycle
J'ai achet**é** un vélo	I bought a bicycle
J'achet**erais** un vélo	I would buy a bicycle

On the inside of words: splitting

Morphology often indicates **when** events happen. For example, in French:

J'achete un vélo	I buy a bicycle (now)
J'acheterai un vélo	I will buy a bicycle
J'achetais un vélo	I was buy**ing** a bicycle
J'ai acheté un vélo	I b**ought** a bicycle
J'acheterais un vélo	I would buy a bicycle

On the inside of words: splitting

Morphology often indicates **when** events happen. For example, in French:

J'achete un vélo	I buy a bicycle (now)
J'acheterai un vélo	I will buy a bicycle
J'achetais un vélo	I was buying a bicycle
J'ai acheté un vélo	I bought a bicycle
J'acheterais un vélo	I would buy a bicycle

Morphology can also indicate **why** events happen.

Si tu vas a GT, tu **seras** rica	If you go to GT, you will be rich
Si tu vas a GT, tu **eres** rica.	If you go to GT, you are rich

On the inside of words: splitting

Morphology often indicates **when** events happen. For example, in French:

J'achete un vélo	I buy a bicycle (now)
J'acheterai un vélo	I will buy a bicycle
J'achetais un vélo	I was buying a bicycle
J'ai acheté un vélo	I bought a bicycle
J'acheterais un vélo	I would buy a bicycle

Morphology can also indicate **why** events happen.

Si tu vas a GT, tu **seras** rica	If you go to GT, you will be rich
Si tu vas a GT, tu **eres** rica.	If you go to GT, you are rich

And much more...

Morphology, orthography, and phonology

- Morphology

$$\text{berry} + \textsc{Plural} \rightarrow \text{berry+s}$$
$$\text{goose} + \textsc{Plural} \rightarrow \text{geese}$$

- Orthography: berry+s $\rightarrow$ berries
- Phonology

$$\text{cat+s} \rightarrow \textsc{cats}$$
$$\text{dog+s} \rightarrow \textsc{dogz}$$

Morphology, orthography, and phonology

- Morphology

$$\text{berry} + \textsc{Plural} \rightarrow \text{berry+s}$$
$$\text{goose} + \textsc{Plural} \rightarrow \text{geese}$$

- Orthography: berry+s → berries
- Phonology

$$\text{cat+s} \rightarrow \textsc{cats}$$
$$\text{dog+s} \rightarrow \textsc{dogz}$$

- Interactions
  - Homographs, e.g. read+$\textsc{Present}$ vs read+$\textsc{Past}$
  - Heterographs, e.g., Champion's vs champions

Morphemes

- **Stem**: "main" part of meaning.
  Usually can appear alone (free).
- **Affix**: modifier.
  Usually cannot appear alone (bound).
  - Prefixes, e.g. un+learn
  - Suffixes, e.g. learn+ed, apples, Mark's
  - Infixes, e.g. Tagalog: hingi (a request) + -um- (act) → h+um+ingi
  - Circumfixes, e.g. German: sagen (say) → ge+sag+t (said)

Circumfixes in Hebrew

(7)

Root	Pattern	Part of Speech	Phonological Form	Orthographic Form	Gloss
ktb	CaCaC	(v)	katav	כתב	'wrote'
ktb	hiCCiC	(v)	hixtiv	הכתיב	'dictated'
ktb	miCCaC	(n)	mixtav	מכתב	'a letter'
ktb	CCaC	(n)	ktav	כתב	'writing, alphabet'

[heb]

Types of morphology

Main types of morphology:

- **Inflection** creates different forms of a single word
- **Derivation** creates new words

Types of morphology

Main types of morphology:

- **Inflection** creates different forms of a single word
- **Derivation** creates new words

Other phenomena:

- **Cliticization** attaching phonologically-dependent affixes, e.g. Georgia's, j'accuse [fr]
- **Compounding** combines two words into a new word, e.g. cream → ice cream → ice cream cone → ice cream cone bakery
- **Portmanteaus** combine words, truncating one or both. smoke + fog → smog
  glass + asshole → glasshole

Types of morphology

Main types of morphology:

- **Inflection** creates different forms of a single word
- **Derivation** creates new words

Other phenomena:

- **Cliticization** attaching phonologically-dependent affixes, e.g. Georgia's, j'accuse [fr]
- **Compounding** combines two words into a new word, e.g. cream → ice cream → ice cream cone → ice cream cone bakery
- **Portmanteaus** combine words, truncating one or both. smoke + fog → smog
  glass + asshole → glasshole

Word formation is **productive**: new words are subject to all of these morphological processes.

Inflectional morphology

- Inflections add information about words.
- English inflectional morphology is very simple:

Affix	Syntactic/semantic effect	Examples
-s	NUMBER: plural	*cats*
-'s	possessive	*cat's*
-s	TENSE: present, SUBJ: 3sg	*jumps*
-ed	TENSE: past	*jumped*
-ed/-en	ASPECT: perfective	*eaten*
-ing	ASPECT: progressive	*jumping*
-er	comparative	*smaller*
-est	superlative	*smallest*

Case marking

- **Case** marking distinguishes the syntactic role of a **noun** in a sentence.
- In English, we distinguish the case of some pronouns:
  - He (NOMINATIVE) gave her (OBLIQUE) his (GENITIVE) guitar.
  - She gave him her guitar.
  - I gave you our guitar.
  - You gave me your guitar.

Case marking

- **Case** marking distinguishes the syntactic role of a **noun** in a sentence.
- In English, we distinguish the case of some pronouns:
  - He (NOMINATIVE) gave her (OBLIQUE) his (GENITIVE) guitar.
  - She gave him her guitar.
  - I gave you our guitar.
  - You gave me your guitar.
- Other languages mark the case of **all** nouns (e.g., Latin, Russian, Sanskrit, Tamil), often for more cases.

Case marking

- **Case** marking distinguishes the syntactic role of a **noun** in a sentence.
- In English, we distinguish the case of some pronouns:
  - He (NOMINATIVE) gave her (OBLIQUE) his (GENITIVE) guitar.
  - She gave him her guitar.
  - I gave you our guitar.
  - You gave me your guitar.
- Other languages mark the case of **all** nouns (e.g., Latin, Russian, Sanskrit, Tamil), often for more cases.
- In German, articles and adjectives are inflected for case:
  - Der alte Mann gab dem kleinen Affen die grosse Banane
  - The old man (NOM) gave the little monkey (DATIVE) the big banana (ACCUSATIVE)

Gender and number

Many languages inflect the article and adjective for gender and number, e.g. Spanish:

- El coche rojo pasó la luz roja: the red car ran the red light
- Los coches rojos pasó las luces rojas: the red cars ran the red lights

Gender and number

Many languages inflect the article and adjective for gender and number, e.g. Spanish:

- El coche rojo pasó la luz roja: the red car ran the red light
- Los coches rojos pasó las luces rojas: the red cars ran the red lights
- Article and adjective must **agree** for the sentence to be grammatical.
- In English, demonstrative determiners mark number, this book vs these books, and the determiner and noun must agree.

Gender and number

Gender is not necessarily binary.

- English pronouns include neuter it; German, Sanskrit, and Latin have the possibility of neuter gender for all nouns.
- Danish and Dutch distinguish **neuter** from **common** gender
- Other languages distinguish **animate** and **inanimate**

Gender and number

Gender is not necessarily binary.

- English pronouns include neuter it; German, Sanskrit, and Latin have the possibility of neuter gender for all nouns.
- Danish and Dutch distinguish **neuter** from **common** gender
- Other languages distinguish **animate** and **inanimate**

Number if not necessarily binary.

- Many languages, such as Arabic and Sanskrit, include a special **dual** number for two. English has residual traces of the dual number, with both vs all and either vs any.
- Some Austronesian languages have a **trial** number, for groups of 3.
- Some languages, including Arabic, have a **paucal** number, for small groups.

Tense and aspect in English

- English verbs are inflected for **tense** and number distinguishing
  - past (I ate)
  - present (I eat)
  - 3rd-person singular (She eats).
- They are also inflected for **aspect**, distinguishing perfective (I had eaten) and progressive (I am eating).
- Note that the perfective and the past tense are identical for regular verbs, e.g. we had talked, we talked.
- African-American English has a more complex system of tense and aspect, distinguishing completed and habitual actions (through auxiliary verbs, not morphology).

Tense and aspect in other languages

Many languages do not mark tense with morphology.
Indonesian uses time signals, e.g. Indonesian:

Saya makan apel	I eat an apple
Saya sedang makan apel	I am eating an apple
Saya telah makan apel	I already ate an apple
Saya akan makan apel	I will eat an apple

Tense and aspect in other languages

Romance languages distinguish many more tenses with morphology.

- Spanish has multiple past tenses: **preterite** and **imperfect**.
  - I ate onions yesterday vs I ate onions every day.
  - comí cebollas ayer vs comía cebollas cada día
- Spanish and French have endings for conditional and future,
  - comería cebollas vs comeré cebollas
- All of these are marked with time signals in English; future can also be marked this way in French and Spanish, e.g. voy a comer cebollas.

Person and number agreement in verbs

Parler

The verb *parler* "to speak", in French orthography and IPA transcription

	Indicative				Subjunctive		Conditional	Imperative
	Present	**Simple past**	**Imperfect**	**Simple future**	**Present**	**Imperfect**	**Present**	**Present**
je	parl-e /paʁl/	parl-ai /paʁle/	parl-ais /paʁlɛ/	parl-erai /paʁləʁe/	parl-e /paʁl/	parl-asse /paʁlas/	parl-erais /paʁləʁɛ/	
tu	parl-es /paʁl/	parl-as /paʁla/	parl-ais /paʁlɛ/	parl-eras /paʁləʁa/	parl-es /paʁl/	parl-asses /paʁlas/	parl-erais /paʁləʁɛ/	parl-e /paʁl/
il	parl-e /paʁl/	parl-a /paʁla/	parl-ait /paʁlɛ/	parl-era /paʁləʁa/	parl-e /paʁl/	parl-ât /paʁlɑ/	parl-erait /paʁləʁɛ/	
nous	parl-ons /paʁlɔ̃/	parl-âmes /paʁlɑm/	parl-ions /paʁljɔ̃/	parl-erons /paʁləʁɔ̃/	parl-ions /paʁljɔ̃/	parl-assions /paʁlasjɔ̃/	parl-erions /paʁləʁjɔ̃/	parl-ons /paʁlɔ̃/
vous	parl-ez /paʁle/	parl-âtes /paʁlɑt/	parl-iez /paʁlje/	parl-erez /paʁləʁe/	parl-iez /paʁlje/	parl-assiez /paʁlasje/	parl-eriez /paʁləʁje/	parl-ez /paʁle/
ils	parl-ent /paʁl/	parl-èrent /paʁlɛːʁ/	parl-aient /paʁlɛ/	parl-eront /paʁləʁɔ̃/	parl-ent /paʁl/	parl-assent /paʁlas/	parl-eraient /paʁləʁɛ/	

Other morphological inflections

- Comparative and superlative adjectives (e.g., taller, tallest)
- **Evidentiality**, e.g. Eastern Pomo verb suffixes

-ink'e	nonvisual sensory
-ine	inferential
-le	hearsay
-ya	direct knowledge

Other morphological inflections

- Comparative and superlative adjectives (e.g., taller, tallest)
- **Evidentiality**, e.g. Eastern Pomo verb suffixes

-ink'e	nonvisual sensory
-ine	inferential
-le	hearsay
-ya	direct knowledge

Morphology [edit]

Quileute features an interesting prefix system that changes depending on the physical characteristics of the person being spoken to. When speaking to a cross-eyed person, /ta/ is prefixed to each word. When speaking to a hunchback, the prefix /t͡ʃʼ/ is used. Additional prefixes are also used for short men (/s/), "funny people" (/t͡ɬ/), and people that have difficulty walking (/t͡ʃχ/).[7]

Index of synthesis

The **index of synthesis** measures the ratio of the number of morphemes in a given text to the number of words.

Language	Index of synthesis
Vietnamese	1.06
Yoruba	1.09
English	1.68
Old English	2.12
Swahili	2.55
Turkish	2.86
Russian	3.33
Inuit (Eskimo)	3.72

Number of unique surface forms in 10K parallel sentences from Europarl:

- English: 16k word types
- French: 22k
- German: 32k
- Finnish: 55k

Derivational morphology

- **nominalization**
  - V + -ation: computerization
  - V + -er: walker
  - Adj + -ness: fussiness
  - Adj + -ity: obesity
- **negation**: undo, unseen, misnomer
- **adjectivization**: V + -able : doable, thinkable, N + -al : tonal, national, N + -ous: famous, glamorous
- **abverbization**: ADJ + -ily: clumsily
- **lots more**: rewrite, phallocentrism,

A Turkish word

uygarlaştıramadıklarımızdanmışsınızcasına

uygar_laş_tır_ama_dık_lar_ımız_dan_mış_sınız_casına

"as if you are among those whom we were not able to civilize (=cause to become civilized)"

uygar: *civilized*
_laş: *become*
_tır: *cause somebody to do something*
_ama: *not able*
_dık: past participle
_lar: plural
_ımız: 1st person plural possessive (our)
_dan: *among* (ablative case)
_mış: past
_sınız: 2nd person plural (you)

K. Oflazer pc to J&M