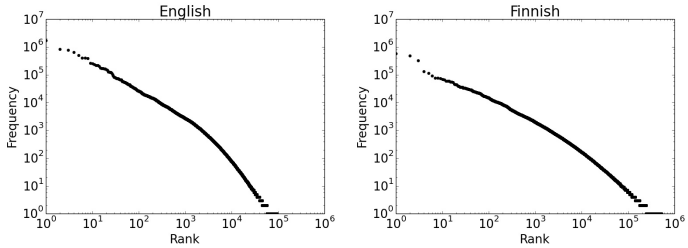
Accelerated Natural Language Processing
Lecture 2
Morphology

Sharon Goldwater
(based on slides by Philipp Koehn)

17 September 2019

School of informatics

Two plots from last time



How Many Different Words?

10,000 sentences from the Europarl corpus

Language	Different words
English	16k
French	22k
Dutch	24k
Italian	25k
Portuguese	26k
Spanish	26k
Danish	29k
Swedish	30k
German	32k
Greek	33k
Finnish	55k

Why the difference? Morphology.

Today's Lecture

- What is morphology, how does it differ across languages, and why does it matter for NLP?
- What's the difference between a stem, lemma, and affix?
- What are the characteristics of derivational and inflectional morphology?
- What is an FSM, and what is the relationship between FSMs and regular languages?

Interlude/reminder: types and tokens

The word *word* is ambiguous.

- Word **type**: "10k sentences from English Europarl have 16k different words" (unique strings, lexical items)
- Word **token**: "English Europarl has 54m words" (possibly repeated instances)

A CAT AND A BROWN DOG CHASED A BLACK DOG:
10 tokens, 7 types.

What is morphology?

The study of wordforms and word formation.

- Structured relationships between words:

play, played, replay, player
played, walked, jumped

- Units of meaning (**morphemes**) and their ordering (**morphotactics**):

de+salin+ate+ion but not ate+salin+ion+de

Why does morphology matter?

- Information retrieval: return pages with related forms.
- Language modelling: make predictions about unseen words
- Machine translation and language understanding: signals differences in meaning (might be expressed using word order in other languages).

Why does morphology matter?

Example (Russian):

zhenshina devochke dala knigu
woman+NOM girl+DAT gave book+ACC
'the woman gave the girl a book'

vs.

zhenshine devochka dala knigu
woman+DAT girl+NOM gave book+ACC
'the girl gave the woman a book'

A noun's **case marking** (a kind of morphology) indicates its role in the sentence, where English uses word order and prepositions.

Morphemes: Stems and Affixes

- Two types of morphemes
  - stems: small, cat, walk
  - affixes: +ed, un+
- Four types of affixes
  - suffix
  - prefix
  - infix
  - circumfix

Stems vs. Lemmas

- Lemma: the canonical form or dictionary form of a set of words
  - fly, flies, flew and flying all have the lemma fly.
  - walk, walks, walked and walking all have the lemma walk.
  - walker, walkers have the lemma walker.

Stems vs. Lemmas

- Lemma: the canonical form or dictionary form of a set of words
  - fly, flies, flew and flying all have the lemma fly.
  - walk, walks, walked and walking all have the lemma walk.
  - walker, walkers have the lemma walker.
- Stem: definitions can vary, but often: the part of the word that is common to all its variants
  - stem of produce, production is produc.
  - stem of walk, walks, walked, walking, walker, walkers is walk.
  - Do fly, flies, flew, flying have a common stem fl?
    Or maybe only fly and flying share a stem: fly.
    Decision may depend on application.

Suffix

- Plural of nouns

  cat+s
- Comparative and superlative of adjectives

  small+er
- Formation of adverbs

  great+ly
- Verb tenses

  walk+ed
- All inflectional morphology in English uses suffixes

Prefix

- In English: these typically change the meaning
- Adjectives

  un+friendly
  dis+interested

- Verbs

  re+consider

- Some language use prefixing much more widely

Other types of morphology

Mainly in non-English languages; check textbook or online.

- Infixes
- Circumfixes
- Reduplication
- Root and pattern

Not that easy...

- Affixes are not always simply attached
- In writing, some letters may be changed/added/removed
  - walk+ed
  - frame+d
  - emit+ted
  - carr(–y)+ied
- In speaking, some sounds may be changed/added/removed
  - Compare the final sound: cats [s] vs dogs [z] vs foxes [əz]

Irregular Forms

- Some words have irregular forms:
  - is, was, been
  - eat, ate, eaten
  - go, went, gone
- Irregular forms tend to be the most frequent (and vice versa)

Inflectional Morphology

- In English, we inflect
  - *nouns* for count (plural: +s) and for possessive case (+'s)
  - *verbs* for tense (+ed, +ing) and a special 3rd person singular present form (+s)
  - *adjectives* in comparative (+er) and superlative (+est) forms.
- In German, we inflect
  - *nouns* for count and case
  - *verbs* for tense, person, and count
  - *adjectives* for count, case, gender, and definiteness
  - *determiners* for count, case and gender

Forms of the German the

Case	Singular			Plural		
	male	fem.	n.	male	fem.	n.
nominative (**s**ubject)	der	die	das	die	die	die
genitive (**p**ossessive)	des	der	des	der	der	der
dative (**i**ndirect **o**bject)	dem	der	dem	den	den	den
accusative (direct **o**bject)	den	die	das	die	die	die

Phrase/role: [The A]/**s** put [the B]/**o** [of the C]/**p** [on the D]/**io**

Not only many different forms,
but each form is highly ambiguous.

Inflectional vs. Derivational Morphology

- Inflectional morphology typically
  - does not change basic meaning or part of speech
  - expresses grammatical features or relations between words
  - applies to all words of the same part of speech
- Derivational morphology
  - may change the part of speech or meaning of a word
  - is not driven by syntactic relations outside the word
  - may be "picky": drama+(t)ize but not traged(-y)+ize
  - applies closer to the stem; whereas inflection occurs at word edges: govern+ment+s, centr+al+ize+d

Derivational Morphology

- Changing the part of speech, e.g. noun to verb

  word → wordify

- Is it a real word?
- Consulting Google (a few years ago):
  - 8,840 hits: e.g., wordify mugs, tshirts and magnets
- Google now returns over 75k hits. (Why?)

Derivational Morphology

- Changing the verb back to a noun

  wordify → wordification (8k hits on Google)

- A person/thing who engages in wordification

  wordification → wordificator (was 8 hits, now 21k: another app!)

- A person/thing who wordifies

  wordify → wordifier (1500 hits on Google)

- What is the difference between a wordifier and a wordificator?

Derivational Morphology

- Turning wordification into a ideology:

  wordification → wordificationism (was just 1 hit:)

> I think you're confusing the term "Democracy" with "Capitalism"; I think you mean "Has Capitalism failed"?
>
> No. It hasn't.
>
> I agree, Hambone; I'm just trying to correct the wordificationism.
>
> Where in the world did you get the word "wordificationism"? Not in the Merriam-Webster dictionary, not in the Thesaurus...

Derivational Morphology

- An adherent of wordificationism

  wordificationism → wordificationist

- Used to have 0 hits on Google, now you get these slides!

- We created a new word!

Compounds

- Creating new words by merging multiple words

- (Somewhat) rare in English

  home work → homework
  web site → website

- More common in other languages (like German)

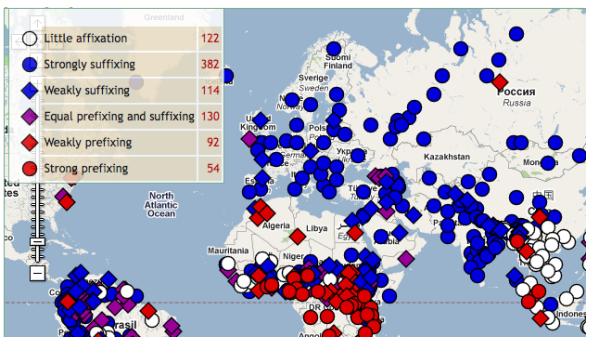
Acronyms/Initialisms

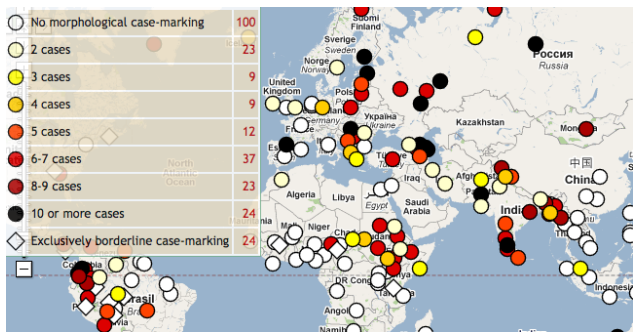
- Wikileaks / Guardian, document 2007-081-100110-0444:

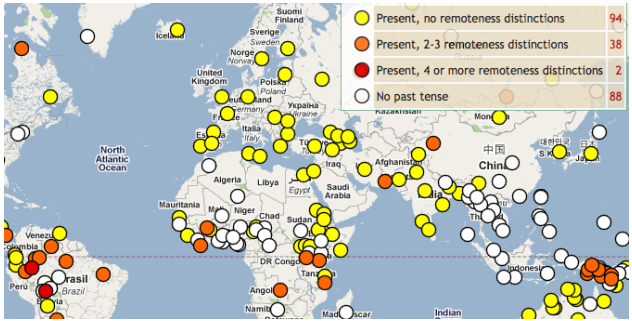
> OGA operating in TF Catamount sector moved into Malekshay for operation. LN Shum Khan ran at the sight of the approaching CFA's. CF utilized the escalation of force doctrine and shouted to stop, fired warning shots and then fired to wound. The LN was hit in the ankle and treated by Element medics on scene. It was determined through discussions with local Elders that the man was a deaf mute that was nervous of the CF operation. Solatia was made in the form of supplies and the Element mission progressed

Morphology differs across languages

- Usually a trade-off between morphology and syntax (word order)
  - Some languages have no verb tenses
    - → use explicit time references (yesterday)
  - Case inflection determines roles of noun phrase
    - → use fixed word order instead
    - → use prepositional phrases instead of cased noun phrases
- Examples from the World Atlas of Language Structures (wals.info)
  - prefixes vs. suffixes
  - cases (zero to more than ten)
  - past tense remoteness distinctions







So...

How to deal with all this computationally?

What do we even want to be able to do?

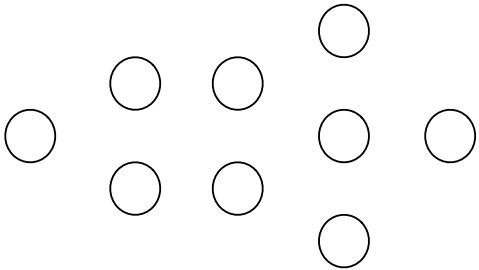
Tasks

- Recognition
  - given: wordform (string of characters)
  - wanted: yes/no decision if it is in the language
- Generation
  - given: lemma and morphological properties
  - wanted: correctly inflected wordform
- Analysis
  - given: wordform
  - wanted: lemma and morphological properties

Word Lists

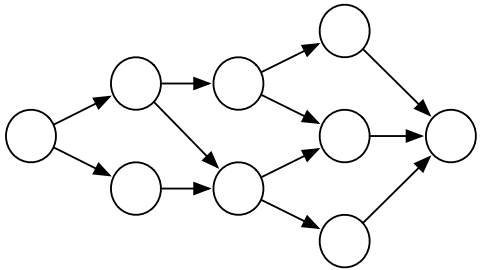
- Simple Solution
  - create a list of all wordforms and their morphological properties
  - solve tasks by checking against list
- But...
  - list can become very long
  - list fails to generalize for productive morphology
- Instead: use finite state machines
  (also called finite state automatons)

Finite State Machines: States



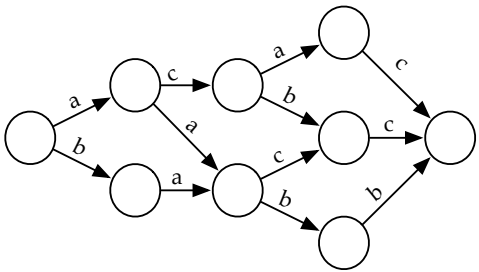
places we may find ourselves in

Finite State Machines: Transitions



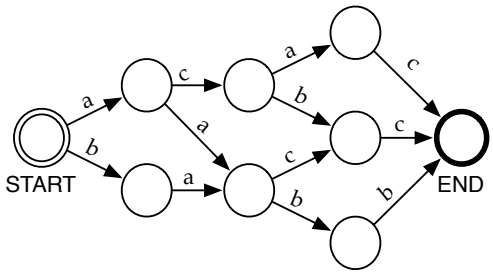
moving between the states

Finite State Machines: Emissions



emissions: letters produced at each transition

Finite State Machines: Start and End



begin at start state, finish at end state

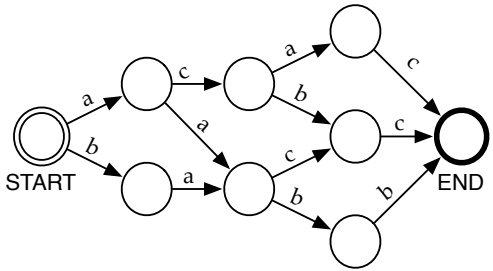
The language of an FSM

Every FSM defines a **formal language**:

- The set of strings that can be generated by moving from start to end states, emitting symbols on each transition.
- Equivalently, the set of strings that can be **recognized** by matching input characters to emission symbols.

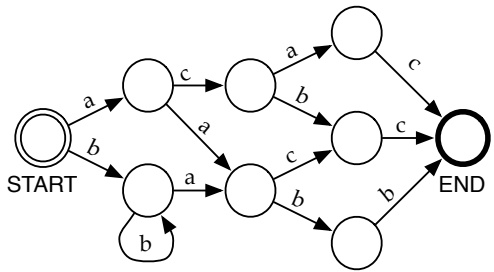
The language of an FSM may be finite or infinite.

FSM with Finite Language



generated language: { acac, acbc, aacc, aabb, bacc, babb }

FSM with Infinite Language



generated language: { acac, acbc, aacc, aabb, bacc, babb, bbacc, bbabb, bbbacc, bbbabb, bbbbacc, bbbbabb, ... }

Regular Languages

- Languages produced by FSMs are called **regular languages**
- Many convenient properties (e.g., straightforward to determine if a word is in the language)
- Not all languages are regular
  example: $a^n b^n$ = { ab, aabb, aaabbb, aaaabbbb, ... }
  (would require an infinite number of states)

Regular Expressions

- Reg. languages can also be described with **regular expressions**.
- Every RegEx is equivalent to some FSM (and vice versa).

  Example: `ac(ac|bc) | aa(cc|bb) | bb*a(bb|cc)`

  where '|' means "or" and 'x^*' means "zero or more x's".
- RegExs are common in programming to describe sets of strings.
  - `ls *.jpg`
  - `if ($word =~ /^[A-Z].*/) { $name = 1; }`
  - `if ($name =~ /[WB]ill/) { print "Will or Bill"; }`

Chomsky Hierarchy

- Chomsky discussed four major classes of formal languages

  3. **regular** (generated by finite state machines, usually assumed sufficient to describe phonology and morphology)
  2. **context-free** (will be covered in later lectures on syntax)
  1. **context-sensitive** (possibly needed for some natural language phenomena)
  0. **recursively enumerable** (anything a computer program can produce)
- (There are also many classes of "sub-regular" languages.)

Chomsky Hierarchy

- Language classes further down the list are increasingly complex
  - can describe more languages
  - but languages in the class are more difficult to compute

    for instance: for a type-0 language it is not generally possible to determine if a specified word can be generated by the language
- Linguists argue about which (if any) of these classes natural languages belong to, but most phenomena of interest can be described by context-free languages.

Questions for review

- What is morphology, how does it differ across languages, and why does it matter for NLP?
- What's the difference between a stem, lemma, and affix?
- What are the characteristics of derivational and inflectional morphology?
- What is an FSM, and what is the relationship between FSMs and regular languages?
- (To be answered next time: how do we use FSMs for morphology?)

Exercises

1. Look at the FSM on slide 38, where there is a state that has a self-loop labelled 'b'. Suppose we added another self-loop to the same state, labelled 'c'. Which of the following strings is NOT accepted by the new FSM?

   acac aacc bacc bbacc bcacc bcbabb bacbb bccabb

2. What is the lemma of each of the following words? How many affixes does each word have, and are they derivational or inflectional?

   located dreamy stole standardizes

Reminders

1. Labs this week:
   - Check Learn to see which lab to go to and for additional prep instructions.
   - When you arrive in lab, sit down and work with a partner! Discuss and help each other. Pass the keyboard back and forth.

2. Tutorials next week (starting Tuesday):
   - You'll be automatically enrolled in one when you register, so try to do that by the end of this week.
   - Also, start going through probability tutorial (linked from week 2 Reading).