CS 4650/7650, Lecture 7
Morphology

Jacob Eisenstein

September 10, 2013

Morphology and finite-state machines are two of my favorite topics. The formalisms are elegant, and it's an area where computer science and linguistics complement each nicely.

1 Morphology: words on the inside

So far we have been focusing on NLP at the word level. Today we go **inside of words**.

We've already hinted at a morphological problem by introducing the idea of **lemmas**, where *serve/served/serving* all have the lemma *serve*.

From the perspective of document classification, these multiple forms may just seem like an annoyance, which we can get rid of by lemmatization or stemming (more on this later).

But morphology conveys information which might be crucial for some applications:

- Information retrieval
  - With a query like *bagel*, we want to get hits for *bagels*.
  - Same for *corpus/corpora*, *goose/geese*.
  - But we don't always want all the inflected forms. For example, a query for *Apple* may not want hits for *apples*
- Time. Morphology often indicates when events happen. For example, in French:

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

- Causality. Consider the difference between the Spanish examples:

*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

- Lexical semantics: suppose *antichrist* is not in your sentiment dictionary. Do you think it is a positive or negative word?

In addition to recognizing morphology, there are applications in which we need to produce it.

- Translation: *you (pl) are smart* → *Ustedes son inteligentes* vs *Tu eres inteligente*
- Text generation: `(has-property you-pl smart)` → *ustedes son inteligentes*

1.1 Morphology, Orthography, and Phonology

- **Morphology** describes how meaning is constructed from combining affixes. For example, it is a morphological fact of English that adding the affix $+S$ to many nouns creates a plural.

$$berry + \textsc{Plural} \rightarrow berry{+}s$$

- **Orthography** specifically relates to writing.
  For example,

$$berry{+}s \rightarrow berries$$

  is an orthographic rule. We have lots of these in English, which is one reason English spelling is difficult.

- Morphological rules also include stem changes, such as *goose*+PLURAL → *geese*.

- **Phonology** describes how sounds combine. For example, the different pronunciations of the final *s* in *cats* (s) and *dogs* (z) follow from a phonological rule (example (25) in the Bender text, page 30).
- In English, morphologically distinct words may be pronounced differently even when they are spelled the same, and this can reflect morphological differences. *read*+PRESENT vs. *read*+PAST.
- Conversely, morphological variants may be spelled differently even when they sound the same, like *The Champions' league* vs *The Champion's league* vs *The Champions league*.

1.2 Productivity

One idea for dealing with morphology is to build a morphologically aware dictionary:

- Map each **surface form** to its underlying **lemma**
- Include meaning of morphology: tense, number, animacy, posession, etc.
- Then when you encounter a surface form, just look it up.

duck	*duck*/N+SG
ducks	*duck*/N+PL
duck	*duck*/VB+PRESENT
ducks	*duck*/VBZ+PRESENT

Will this work? Besides the problem of ambiguity, still another problem is that morphology is **productive**, meaning that it applies to new words. If you only know the words *Google* or *iPad*, you can immediately understand their inflected forms.

- Have you Googled that yet?
- I have owned three iPads.

Derivational morphology (more on this later) is productive in another way: you can produce new words by applying morphological changes to existing words. hyper+un+desire+able+ity

In some languages, derivational morphology can create extremely complicated words. The J&M textbook has a fun example from Turkish:

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

In the homework, you'll see examples from Swahili, which also has complex morphology. A dictionary of all possible surface forms in such languages would be gargantuan. So instead of building a static dictionary, we will try to model the underlying morphological and orthographic rules.

2 Morphemes

Two broad classes: **stems** and **affixes**.

- Intuitively, stems are the "main" part of meaning, affixes are the modifiers
- Typically, **stems** can appear on their own (they are **free**) and affixes cannot (they are **bound**).
- Types of affixes:
  - **Prefixes**: *un+learn*, *pre+view*.

    * These examples are derivational. English has few inflectional prefixes, but other languages have many.
    * For example, in Swahili: *u-na-kata* versus *u-me-kata* distinguishes *you are cutting* versus *you have cut*. *na* and *me* are prefixes, *kata* is the root.

– **Suffixes**: *I learn+ed*, *She learn+s*, *three apple+s*, *four fox+es*.

– **Circumfixes** go around the stem.

    * None in English.
    * German has a circumfix for the past participle: *sagen* (say) → *ge+sag+t* (said)
    * French negation can be seen as a circumfix: *Je mange+NEG→Je ne mange pas.* (I do not eat).
    * More generally, morphemes can be non-continuous, as in the Hebrew example (7) in the Bender reading (page 12).

(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]

    In this example, the root *ktb* (related to writing) is combined with patterns that indicate where to insert vowels to produce different parts-of-speech and meanings.

– **Infixes** go inside the stem.

    * Tagalog: *hingi+AGENT→h+um+ingi*
    * Lakhota: *m'ani* (he walks), *ma-wá-ni* (I walk). The *wá* marks agreement with a first-person singular subject; it is an infix for this word, although it is a prefix in other words.
    * English: *absolutely+fucking→absofuckinglutely*, but **absfuckingsolutely* arguably doesn't work.

3 Types of morphology

- **Inflection** creates different forms of a single word:
  - tense: *to be, being, I am, you are, he is, I was*
  - number: *book, books*
  - case: *he, his, her, they, them, their*
- **Derivation** creates new words:
  grace → *disgrace* → *disgraceful* → *disgracefully*
- **Cliticization** combines *Georgia*+*'s* into *Georgia's*; the possessive clitic *'s* is syntactically independent but phonologically dependent.
  - Pronouns appear as clitics in French, e.g., *j'accuse* (I accuse), as does negation *Je n'accuse personne* (I don't accuse anyone).
  - Another example is from Hebrew: *l'shana tova* (literally for year good, meaning happy new year); the preposition *for* appears as a clitic.
- **Compounding** combines two words in a new word:
  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 processes

3.1 Inflectional morphology

Inflectional morphology adds information about words. English has a very simple system of inflectional morphology, compared to many languages.

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*

- English nouns are marked for number and possession; many language also mark nouns for **case**, which is the syntactic role that the noun plays in the sentence.
  - In English, we do 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.*

    Specifically, we distinguish the **nominative** case of personal pronouns (except for 2nd-person), and the **genitive** case; all other uses are the **oblique** case.
  - Other languages – such as Latin, Russian, Sanskrit, and Tamil, mark the case of all nouns. These languages have additional cases, such as dative (indirect object), accusative (direct object), and vocative (address).
  - In German, noun is not inflected for case, but the articles and adjectives are:
    - *Der alte Mann gab dem kleinen Affen die grosse Banane* (based on example 49 from Bender)
    - The old man (NOM) gave the little monkey (DATIVE) the big banana (ACCUSATIVE)

    - Notice how *der*, *dem*, and *die* all mean the, but carry the case marking.
    - The adjectives are also marked for case.
  - Many languages – such as Romance languages – mark the gender and number of nouns by inflecting the article and adjective. 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
    - Note that the article and adjective must **agree** for the sentence to be grammatical.
    - In English, demonstrative determiners mark number, *this book* vs *these books*. In English, the determiner and noun must agree, e.g. **this books*.
  - Gender is not necessarily binary.
    - English pronouns include neuter *it*; German, Sanskrit, and Latin do this 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 three.
    - Some languages, including Arabic, have a **paucal** number, for small groups.

- English verbs are inflected for tense and number distinguishing past (*I ate*), present (*I eat*), and 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*.
  - Many languages (e.g., Chinese and Indonesian), do not mark tense with morphology. Indonesian uses time signals.

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

  - Romance languages distinguish many more tenses than English with morphology.
    - Spanish has multiple past tenses: **preterite** and **imperfect**, distinguishing, e.g. *I ate onions yesterday* from *I ate onions every day*. These are distinguished by morphology: *comí cebollas ayer*, *comía cebollas cada día*.
    - Spanish and French have endings for conditional (*comería cebollas*) and future (*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*.
  - Romance also have separate forms for every combination of number and person.
    - *(yo hablo / tu hablas / ella habla / nosotros hablamos / vosotros hableis / ellas hablan)*
    - *(je parle / tu parles / elle parle / nous parlons / vous parlez / ils parlent)*
    - In Spanish, they they eliminate pronouns (pro-drop) in cases where the morphology makes it clear (unless they want to add emphasis). Chinese is also a pro-drop language?
    - This doesn't happen in French, maybe because many different spellings (*parle/parles/parlent*) sound the same.
    - In English, we only distiniguish 3rd-person singular.

- Adjectives in English mark comparative and superlative (*taller*, *tallest*). As we have seen, they can mark gender and number in other languages.

- Other things can be marked with affixes, such as **evidentiality** – how the speaker came to know the information. In Eastern Pomo (a California language), there are verb suffixes for four evidential categories:

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

Example (41) in the text shows evidentiary marking in Turkish, *Ahmet geldi* (Ahmet came, witnessed by the speaker) vs *Ahmet gelmi s* (not witnessed by the speaker)

The **index of synthesis** measures the ratio of the number of morphemes in a given text to the number of words. Languages with complex morphology are called **synthetic**; languages with simple morphology are called **isolating** or **analytic**. English is relatively, but not extremely, analytic.

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

An approximation of the index of synthesis is the type-token ratio. Can you see why? If you count the number of unique surface forms in 10K *parallel* sentences from Europarl, you get:

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

3.2 Derivational Morphology

Derivational morphology is a way to create new words and change part-of-speech.

- **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*, ...

You can create totally new words this way.
word → *wordify* → *wordification* → *wordificationism* → *antiwordificiationism* → *hyperantiwordificationism*

3.3 Irregularities

English morphology contains a lot of irregularities: *know/knew/known*, *foot/feet*, *go/went*.. if you're not a native speaker, learning these was probably a pain in the neck.

- the good news is, there are fewer of these all the time! for example, the past tense of *show* used to be *shew*, like *know/knew* (the past participle is still *shown*).
- the bad news is, the most common words will be the last to change (if ever).

Attaching affixes can cause orthographic and phonological changes:

- walk + ed = walked, but frame+ed = framed, emit+ed = emitted, easy + ier = easier
- this is usually due to phonetic or orthographic *constraints*
- *optimality theory* is an approach to systematizing such interacting constraints. There's a lot of research on finite state models of optimality theory, but you'll have to take a linguistics course for that [KB05].

References

[KB05] Lauri Karttunen and Kenneth R Beesley. Twenty-five years of finite-state morphology. *Inquiries Into Words, a Festschrift for Kimmo Koskenniemi on his 60th Birthday*, pages 71–83, 2005.