Lecture 6: Morphology

LTAT.01.001 – Natural Language Processing
Kairit Sirts (kairit.sirts@ut.ee)
20.03.2019
Morphology

• Morphology studies the internal structure of words

availability NOUN +nom +pl

availabilities

avail +able +ity +es

avail_abil_itl_es
Morphemes

- Morphemes are the smallest units of language that carry a semantic meaning.

**availabilities**

- *Verbal root* = avail
- *Derivational suffix* that transforms an adjective into a noun = +able
- *Derivational suffix* that transforms a verb into an adjective = +ity
- *Nominative plural Inflectional suffix* = +es
Roots and stems

• Roots carry the basic indivisible meaning of a word
• A stem is the base of an inflected word

availables
avail +able +ity +es

• Avail is a root – it cannot be divided further
• Available is a stem
• Availability is a stem
• Availabilities is an inflected word
Exercise

• What could be the morphological analysis of the following words? Identify the roots and stems.

<table>
<thead>
<tr>
<th>Word</th>
<th>Root</th>
<th>Stems</th>
</tr>
</thead>
<tbody>
<tr>
<td>readers</td>
<td>read</td>
<td>read, reader</td>
</tr>
<tr>
<td>carefully</td>
<td>care</td>
<td>care, careful</td>
</tr>
<tr>
<td>ate</td>
<td>ate</td>
<td>ate</td>
</tr>
</tbody>
</table>
Lexical and grammatical morphemes

• Lexical morphemes carry themselves a semantic meaning. Most of them can stand on their own.
  • Boy, table, yellow, run, waste etc

• Grammatical morphemes cannot stand on their own. Their role is to modify the meaning of a lexical morpheme or specify the relationships between lexical morphemes
  • -s, -ing, -able, at, in, on
The role of grammatical morphemes

- Overlap with syntax and semantics

“I put the book on the table” vs “Panin raamatu lauale” (in Estonian)

“Giraffe bit the zebra” vs “Kaelkirjak hammustas sebrat” or “Sebrat hammustas kaelkirjak”
Derivational and inflectional morphemes

- Derivational morphemes change the semantic meaning of a word, they can also change the POS of the word
  
  avail VERB +able → available ADJ +ity → availability NOUN

- Inflectional morphemes change:
  - The number, gender, case etc of nouns, adjectives etc
  - The person, number, mood, tense etc of verbs

  dog +pl +gen → dog +s +’
  turn +past → turn +ed
Affixes

Grammatical morphemes are divided according to their attachment location:

- **Suffixes** attach to the end of the word:
  - Avail+able, table+s, go+ing
- **Prefixes** attach to the beginning of the word:
  - re+animate, mis+understand, non+compliant
- **Circumfixes** have two parts, one attaches in the beginning of the word and the other to the end:
  - In German: ge+komm+en, ge+spiel+t
- **Infixes** attach in the middle of the word:
  - In German: an+zu+fangen, ab+ge+fahren
Compounding

• Compounding is using two or more words (typically nouns) together to form a new meaning.

• In English, compounds can be written together:
  • notebook, bookstore, fireman

• ... or separately
  • Living room, dinner table, full moon

• In other languages, compounds are usually written together:
  • In Estonian: täis_kuu (full moon), elu_tuba (living room)

• In some languages compounding can be very productive
  • In German: Kraft_fahr_zeug-Haft_pflcht_versicherung (motor car liability insurance)
Tasks in computational morphology
Tasks in computational morphology

• Text normalization:
  • Stemming
  • Lemmatization

• Morphological analysis/parsing
• Morphological tagging/disambiguation
• Morphological generation
Stemming

• Used in information retrieval to reduce sparsity
• communication → communicat
• When searching for “communication” retrieve articles containing both “communication”, “communications”, “communicate”, “communicated”, “communicating”, “communicates” etc
  • Conflation set

• The most well-known stemming system is Snowball
  • Small string-based language that can be used to define stemming rules
  • A compiler compiles the rules into C or Java program
What are the stems of the following words

<table>
<thead>
<tr>
<th>Word</th>
<th>Stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>agreement</td>
<td>agreement</td>
</tr>
<tr>
<td>rotting</td>
<td>rot</td>
</tr>
<tr>
<td>rolling</td>
<td>roll</td>
</tr>
<tr>
<td>new</td>
<td>new</td>
</tr>
<tr>
<td>news</td>
<td>news</td>
</tr>
<tr>
<td>prove</td>
<td>prove</td>
</tr>
<tr>
<td>provable</td>
<td>prove</td>
</tr>
<tr>
<td>probe</td>
<td>probe</td>
</tr>
<tr>
<td>probable</td>
<td>probable</td>
</tr>
</tbody>
</table>
Lemmatization

• Linguistically motivated task
• The word swimming can have three lemmas:
  • swim (VERB)
  • swimming (ADJ)
  • swimming (NOUN)

Morphological parsing or stemming applies to many affixes other than plurals
Morphological analysis

• The task of finding all possible morphological tags for a word
• A morphological analysis consists of:
  • Lemma/stem
  • POS
  • Morphological attributes/features

• Question:
  Should morphological analysis be done in context or can you analyse each word in isolation?

• Answer: Can be done in isolation

(something) has lasted

<table>
<thead>
<tr>
<th>kestnud</th>
</tr>
</thead>
<tbody>
<tr>
<td>kest+nud //_V_ nud, //</td>
</tr>
<tr>
<td>kest=nu+d //_S_ pl n, //</td>
</tr>
<tr>
<td>kest=nud+0 //_A_ //</td>
</tr>
<tr>
<td>kest=nud+0 //_A_ sg n, //</td>
</tr>
<tr>
<td>kest=nud+d //_A_ pl n, //</td>
</tr>
</tbody>
</table>
Finite state morphological parsing

• The classical method for morphological parsing are finite state methods

• Two-level morphology (Koskenniemi, 1983)

• Resources necessary for developing a finite-state morphological parser:
  • Lexicon – a list of stems and affixes
  • Rules of morphotactics – which morphemes can occur with each POS and what is the ordering of the morphemes
  • Orthographic rules – describing the stem changes: city+s → cities
Morphotactics model as a finite state acceptor

How to accept words: caught, sing, walked, eating, speaks?

\[ \Sigma \text{ – alphabet} \]
\[ Q \text{ – set of states} \]
\[ S \text{ – set of start states} \]
\[ E \text{ – set of end states} \]
\[ \delta \text{ – state transition function} \]

Source: Jurafsky and Martin. Speech and Language Processing 2nd ed
Morphological word recognition

- Does a word belong to the language?
- Lexicon + morphotactics

Source: Jurafsky and Martin. Speech and Language Processing 2nd ed
Morphological parser as a finite state transducer

FST modeling English noun morphology

How can we analyse the words: boy, girls, woman’s, students’?

Σ – input alphabet
Δ – output alphabet
Q – set of states
S – set of start states
E – set of end states
δ – state transition function
ω - output function
Morphological segmentation

\texttt{avail\_abil\_iti\_es}

- Simplest morphological analysis
- Was very popular during the first decade of 2000s.
- Mostly unsupervised or weakly supervised methods
  - Minimum Description length principle
  - Bayesian models
  - Also sequence tagging with CRF
Morphological tagging/disambiguation

• Morphological tagging
  • Predict the morphological tag for each word in context choosing from all possible tags

• Morphological disambiguation
  • First perform morphological analysis
  • Then perform morphological disambiguation by choosing for each word the most appropriate analysis

• “Soft” morphological disambiguation
  • First perform morphological analyses and then use the analyses to influence the tagging decisions.
Morphological tagging

• Can be treated as a sequence tagging task
• Conceptually very similar to POS tagging – instead of POS tags there are now morphological tags
• The UD (universal dependencies) datasets also contain morphological analyses for many languages
Universal morphological features

<table>
<thead>
<tr>
<th>Lexical features</th>
<th>Inflectional features</th>
</tr>
</thead>
<tbody>
<tr>
<td>PronType</td>
<td>Gender</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
</tr>
<tr>
<td>Poss</td>
<td>NounClass</td>
</tr>
<tr>
<td>Reflex</td>
<td>Number</td>
</tr>
<tr>
<td>Foreign</td>
<td>Case</td>
</tr>
<tr>
<td>Abbr</td>
<td>Definite</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nominal*</th>
<th>Verbal*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>VerbForm</td>
</tr>
<tr>
<td>Animacy</td>
<td>Mood</td>
</tr>
<tr>
<td>NounClass</td>
<td>Tense</td>
</tr>
<tr>
<td>Number</td>
<td>Aspect</td>
</tr>
<tr>
<td>Case</td>
<td>Voice</td>
</tr>
<tr>
<td>Definite</td>
<td>Evident</td>
</tr>
<tr>
<td>Degree</td>
<td>Polarity</td>
</tr>
<tr>
<td></td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td>Polite</td>
</tr>
<tr>
<td></td>
<td>Clusivity</td>
</tr>
</tbody>
</table>
CONLL-U format

- A standard tabular format for certain type of annotated data
- Each word is in a separate line
- 10 tab-separated columns on each line:
  1. Word index
  2. The word itself
  3. Lemma
  4. Universal POS
  5. Language specific POS
  6. **Morphological features**
  7.-9. Information related to syntactic information
  10. Any other annotation
They buy and sell books.

I had no clue.
Morphological tagging

• Can be treated as a sequence tagging task
• Conceptually very similar to POS tagging – instead of POS tags there are now morphological tags
• The number of universal POS tags is 17?
• What is the number of morphological tags?
Morphological tagset sizes in UD corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>349</td>
</tr>
<tr>
<td>Chinese</td>
<td>31</td>
</tr>
<tr>
<td>Czech</td>
<td>2630</td>
</tr>
<tr>
<td>English</td>
<td>117</td>
</tr>
<tr>
<td>Estonian</td>
<td>662</td>
</tr>
<tr>
<td>Finnish</td>
<td>2052</td>
</tr>
<tr>
<td>French</td>
<td>228</td>
</tr>
<tr>
<td>German</td>
<td>684</td>
</tr>
<tr>
<td>Korean</td>
<td>11</td>
</tr>
<tr>
<td>Russian</td>
<td>693</td>
</tr>
</tbody>
</table>
Morphological tagging with CRF

- **MarMot** – Müller et al., 2013. *Efficient higher-order CRFs for Morphological Tagging*

- **Features:**
  - The current, preceding and next words as unigrams and bigrams
  - Word prefixes and suffixes up to 10 characters for rare words
  - The occurrence of capital letters, digits and special characters
  - Rare word has frequency $\leq 10$ in the training set
  - Combine all features with the POS tag, morphological tag and morphological features
Sequence tagging with CRF

\[ \hat{t}_1^n = \arg \max_{t_1^n} \exp \sum_{i=1}^n \frac{\rho(t_i, t_{i-1}^i, w_1^n)}{Z} \]

Tools

• CRFSuite
• CRF++
• CRFSharp
Construct word unigram features

- Старим +ADJ +Case=Ins|Degree=Pos|Gender=Masc
- Combine all features with the POS tag, morphological tag and all morphological features

- Старим+ADJ
- Старим+Case=Ins|Degree=Pos|Gender=Masc
- Старим+Case=Ins
- Старим+Degree=Pos
- Старим+Gender=Masc
Neural morphological tagging

- biLSTM encoder to construct contextual representations for words
- Very important! Character-level representations for words
  - biLSTM
  - CNN
- Predict the POS+features combination
  - But there can be valid combinations that were not observed in the training set
- Alternatively, predict POS and features separately or sequentially
Neural morphological tagging

Source: Tkachenko and Sirts, 2018. Modeling composite labels for neural morphological tagging
Morphological disambiguation

- First perform morphological analysis
- Then perform morphological disambiguation by choosing for each word the most appropriate analysis

<table>
<thead>
<tr>
<th>second</th>
<th>place</th>
<th>got</th>
<th>this time</th>
<th>rabbit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teise</td>
<td>koha</td>
<td>sai</td>
<td>seekord</td>
<td>Jänes</td>
</tr>
<tr>
<td>O+adt</td>
<td>S+sg_g</td>
<td>V+s</td>
<td>D</td>
<td>S+sg_n</td>
</tr>
<tr>
<td>O+sg_g</td>
<td>S+sg_n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P+adt</td>
<td>S+sg_p</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P+sg_g</td>
<td>V+o</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Neural morphological disambiguation

• Many possible ideas
• For example: Shen et al., 2016. The role of context in Neural Morphological Disambiguation

\[ r_i = \tanh(g_{stem_i} + g_{tag_i}) \]

\[ R = [r_1; r_2; \ldots; r_N] \]

\[ p(y_t = a | x) = \text{softmax}(R_{xt} \times h_t) \]
Morphological generation

• Lemmatization
  • Sheep --> sheep
  • Brought --> bring
  • Maksis --> maksma (in Estonian: payed --> to pay)
  • Tee --> tee (NOUN), tegema (VERB)

• Morphological synthesis
  • bring +Past --> brought
  • Sina +Plur +Part --> sind, sina(??)
  • Sina +PRON +Plur +Part --> sind (in Estonian: you)
  • Sina +NOUN +Plur +Part --> sina (in Estonian: blueness)
Morphological generation

• Conditioned generation: given some information generate some new information
• Essentially a translation task: translate input into output
  • Translate inflected word form into a lemma
  • Translate lemma + features into an inflected word form
• **Sequence-to-sequence** models or **Encoder-decoder** models
Sequence-to-sequence encoder-decoder architecture

Typically biLSTM

Encoder

Embed

bringen +Past

state

Decoder

Softmax

brought</w>
The UniMorph Project

• http://unimorph.org/
• Collect morphologically annotated lexicon data for many languages
• Currently data for 107 languages
• Many more are being annotated
• https://unimorph.github.io/
The role of computational morphology in NLP

• Still largely an unexplored area. Why?
• Intuitively, modeling morphology should help to reduce the vocabulary sparsity problems
• Morphological agreement:
  • For instance, verbs and nouns must agree in number
  • In German: *der Mann geht* vs *die Männer gehen* (the man goes vs men go)
• Potentially could be useful for many downstream tasks:
  • Machine translation
  • Natural language generation
  • Language modeling (for speech recognition)