Computational Morphology: Machine learning of morphology

Yulia Zinova

09 April 2014 – 16 July 2014

Introduction: History

- Disconnect between computational work on syntax and computational work on morphology.
- Work on computational syntax traditionally involved work on parsing based on hand-constructed rule sets.
- In the early 1990s, the paradigm shifted to statistical parsing methods.
- Rule formalisms (context-free rules, Tree-Adjoining grammars, unification-based formalisms, and dependency grammars) remained much the same, statistical information was added in the form of probabilities associated with rules or weights associated with features.

Introduction: History

- Rules and their probabilities were learned from treebanked corpora (+ some more recent work on inducing probabilistic grammars from unannotated text)
- No equivalent statistical work on morphological analysis (one exception being Heemskerk, 1993).
- Nobody started with a corpus of morphologically annotated words and attempted to induce a morphological analyzer of the complexity of a system such as Koskenniemis (1983)
- such corpora of fully morphologically decomposed words did not exist, at least not on the same scale as the Penn Treebank.
- Work on morphological induction that did exist was mostly limited to uncovering simple relations between words, such as the singular versus plural forms of nouns, or present and past tense forms of verbs.
- Part of the reason for this: handconstructed morphological analyzers actually work fairly well.

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?
- The likelihood that a particular prepositional phrase is associated with a head verb versus the head of the nearest NP

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?
- The likelihood that a particular prepositional phrase is associated with a head verb versus the head of the nearest NP
- There is ambiguity in morphology too
- Example?

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?
- The likelihood that a particular prepositional phrase is associated with a head verb versus the head of the nearest NP
- There is ambiguity in morphology too
- Example?
- It is common for complex inflectional systems to display massive syncretism so that a given form can have many functions

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?
- The likelihood that a particular prepositional phrase is associated with a head verb versus the head of the nearest NP
- There is ambiguity in morphology too
- Example?
- It is common for complex inflectional systems to display massive syncretism so that a given form can have many functions
- What's the difference?

Ambiguities...

- Syntax abounds in structural ambiguity, which can often only be resolved by appealing to probabilistic information
- Example?
- The likelihood that a particular prepositional phrase is associated with a head verb versus the head of the nearest NP
- There is ambiguity in morphology too
- Example?
- It is common for complex inflectional systems to display massive syncretism so that a given form can have many functions
- What's the difference?
- Often this ambiguity is only resolvable by looking at the wider context in which the word form finds itself, and in such cases importing probabilities into the morphology to resolve the ambiguity would be pointless

Statistical morphology

- Increased interest in statistical modeling morphology and the unsupervised or lightly supervised induction of morphology from raw text corpora.
- One recent piece of work on statistical modeling of morphology is Hakkani-Tur et al. (2002)
- What: n-gram statistical morphological disambiguator for Turkish.
- How: break up morphologically complex words and treat each component as a separate tagged item, on a par with a word in a language like English.

Korean morphology

- A related approach to tagging Korean morpheme sequences is presented in Lee et al. (2002).
- Formalism: syllable trigrams used to calculate the probable tags for unknown morphemes within a Korean *eojeol*, a space-delimited orthographic word.
- For eojeol-internal tag sequences involving known morphemes, the model uses a standard statistical language-modeling approach.
- With unknown morphemes, the system backs off to a syllable-based model, where the objective is to pick the tag that maximizes the tag-specific syllable n-gram model.
- The model presumes that syllable sequences are indicative of part-of-speech tags, which is statistically true in Korea
- For example, the syllable conventionally transcribed as park is highly associated with personal names, since Park is one of the the most common Korean family names.

Agglutinative languages

- Agglutinative languages such as Korean and Turkish are natural candidates for such approach
- In such languages, words can consist of often quite long morpheme sequences
- The sequences obey word-syntactic constraints, and each morpheme corresponds fairly robustly to a particular morphosyntactic feature bundle, or tag.
- Such approaches are harder to use in more “inflectional” languages where multiple features tend to be bundled into single morphs.
- As a result, statistical n-gram language-modeling approaches to morphology have been mostly restricted to agglutinative languages.

Transition to unsupervised methods

- Last couple of decades: automatic methods for the discovery of morphological alternations
- particular attention to unsupervised methods

Morphological learning

- First sense: the discovery, from a corpus of data, that the word *eat* has alternative forms *eats, ate, eaten* and *eating*.
- Goal: find a set of morphologically related forms as evidenced in a particular corpus
- Second sense: learn that the past tense of regular verbs in English involves the suffixation of *-ed*, and from that infer that a new verb, such as *google*, would be *googled* in the past tense.
- Goal: to infer a set of rules from which one could derive new morphological forms for words for which we have not previously seen those forms
- Which sense is stronger?

Morphological learning

- First sense: the discovery, from a corpus of data, that the word *eat* has alternative forms *eats, ate, eaten* and *eating*.
- Goal: find a set of morphologically related forms as evidenced in a particular corpus
- Second sense: learn that the past tense of regular verbs in English involves the suffixation of *-ed*, and from that infer that a new verb, such as *google*, would be *googled* in the past tense.
- Goal: to infer a set of rules from which one could derive new morphological forms for words for which we have not previously seen those forms
- Which sense is stronger?
- The second sense is the stronger sense and more closely relates to what human language learners do.

Stronger sense

- Earlier supervised approaches to morphology: stronger sense
- System by Rumelhart and McClelland (1986) proposed a connectionist framework which, when presented with a set of paired present- and past-tense English verb forms, would generalize from those verb forms to verb forms that it had not seen before.
- “generalize” does not mean “generalize correctly” (a lot of criticism of the Rumelhart and McClelland work)
- Other approaches to supervised learning of morphological generalizations include van den Bosch and Daelemans (1999) and Gaussier (1999).

Weaker sense

- Supervised approaches assume that the learner is presented with a set of alternations that are known to be related to one another by some predefined set of morphological alternations.
- How the teacher comes by that set?
- This is the question that the work on unsupervised learning of morphology addresses itself (find the set of alternate forms as evidenced in a particular corpus).
- Once one has a list of alternation exemplars, one could apply a supervised technique to learn the appropriate generalizations; (Yarowsky and Wicentowski 2001).
- Most of the work in the past ten years has been on unsupervised approaches, so we will focus in this discussion on these.

Goldsmith, 2001

- *Linguistica, minimum description length* (MDL) approach
- System for learning of affixation alternations.
- Available on the Internet
- Goldsmiths system starts with an unannotated corpus of text of a language
- The original paper demonstrated the application to English, French, Spanish, Italian, and Latin and derives a set of signatures along with words that belong to those signatures.
- Signatures are sets of affixes that are used with a given set of stems.
- Example: one signature in English is (using Goldsmiths notation) NULL.er.ing.s, which includes the stems *blow, bomb, broadcast, drink, dwell, farm, feel*, all of which take the suffixes *null*, *-er, -ing* and *-s* in the corpus that Goldsmith examines.

Signatures vs. paradigms

- Signatures are not equivalent to paradigms.
- First reason: *NULL.er.ing.s* contains not only the clearly inflectional affixes *-ing* and *-s*, but the (apparently) derivational affix *-er*.
- Whether or not one believes in a strict separation of derivational from inflectional morphology, most morphologists would consider endings such as *-s* and *-ing* as constituting part of the paradigm of regular (and most irregular) verbs in English, whereas *-er* would typically not be so considered.
- Second reason: the set is not complete, the past tense affix is absent, but it does show up in other signatures, such as NULL.ed.er.ing.s (for *attack, back, demand, and flow*).
- Summary: it is not trivial to go from signatures to paradigms.
- Note: Goldsmith is only concerned with affixation, so no alternations like *blow/blew* are in the system.

Goldsmith: System

- Two steps.
- The first step derives candidate signatures and signature-class membership
- The second step evaluates the candidates.

Candidate Generation

- The generation of candidates requires a method for splitting words into potential morphemes.
- Based on weighted mutual information, the method first starts by generating a list of potential affixes.
- Starting at the right edge of each word in the corpus, which has been padded with an end-of-word marker “#”, collect the set of possible suffixes up to length six (the maximum length of any suffixes in the set of languages that Goldsmith was considering), and then for each of these suffixes, compute the following metric, where N_k here is the total number of k-grams:

$$\frac{freq(n_1, n_2 \ldots n_k)}{N_k} \log \frac{freq(n_1, n_2 \ldots n_k)}{\prod_1^k freq(n_i)}$$

Getting the signatures

- The first 100 top ranking candidates are chosen
- Words in the corpus are segmented according to these candidates
- Where that is possible, the best parse for each word was chosen according to the *take-all-splits* approach.
- Suffixes that are not optimal for at least one word are discarded.
- Result: a set of stems and associated suffixes including the null suffix
- The alphabetized list of suffixes associated with each stem constitutes the signature for that stem.
- Remove all signatures associated with but one stem and all signatures involving one suffix.
- Remaining signatures are called *regular signatures*, and these constitute the set of suffixes associated with at least two stems.

Candidate Evaluation

- The set of signatures and associated stems constitutes a proposal for the morphology of the language: it provides suggestions on how to decompose words into a stem plus suffix(es).
- It needs to be evaluated: not all the suggested morphological decompositions are useful, and a metric is needed to evaluate the utility of each proposed analysis.
- Goldsmith proposes an evaluation metric based on minimum description length.
- The best proposal will be the one that allows for the most compact description of the corpus (in terms of the morphological decomposition) and the morphology itself.
- This is a standard measure in text compression: a good compression algorithm is one that minimizes the size of the compressed text plus the size of the model that is used to encode and decode that text.

Goldsmith: testing

- Goldsmith tested his method on corpora from English, French, Italian, Spanish, and Latin.
- For each of these languages, he lists the top ten signatures.
- Goldsmith also evaluated the results for English and French.
- Having no gold standard against which to compare, he evaluated the results subjectively, classifying the analyses into the categories good, wrong analysis, failed to analyze, spurious analysis.
- Results for English, for 1,000 words, were 82.9% in the good category, with 5.2% wrong, 3.6% failure, and 8.3% spurious.
- Results for French were roughly comparable.

Schone and Jurafsky, 2001

- Uses semantic, orthographic, and syntactic information derived from unannotated corpora to arrive at an analysis of inflectional morphology.
- The system is evaluated for English, Dutch, and German using the CELEX corpus (Baayen et al., 1996).
- Goldsmith relies solely on orthographic (or phonological) features.

Semantics?

- Without semantic information, it would be hard to tell that *ally* should not be analyzed as *all+y*, and since Goldsmiths approach does not attempt to induce spelling changes, it would be hard to tell that *hated* is not *hat+ed*.
- On the other hand, semantics by itself is not enough.
- Morphological derivatives may be semantically distant from their bases consider reusability versus use so that it can be hard to use contextual information as evidence for a morphological relationship.
- Furthermore, contextual information can be weak for common function words,
- There is effectively no information that would lead one to prevent *as* being derived from *a+s*.

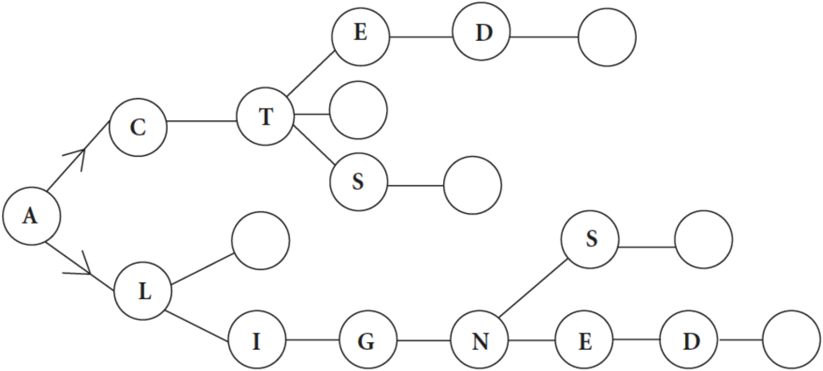
Schone and Jurafsky, 2001

- considers circumfixes
- automatically identifies capitalizations by treating them similarly to prefixes
- incorporates frequency information
- uses distributional information to help identify syntactic properties,
- uses transitive closure to help find variants that may not have been found to be semantically related but which are related to mutual variants
- Schone and Jurafsky use the term *circumfix* somewhat loosely to denote apparently true circumfixes such as the German past participle circumfix get, as well as combinations of prefixes and suffixes more generally.

Procedure: finding affixes

- Strip off prefixes that are more common than some predetermined threshold.
- Take the original lexicon, plus the potential stems generated in the previous step, and build a trie out of them.
- Posit potential suffixes wherever there is a branch in the trie, where a branch is a subtrie of a node where splitting occurs.
- Armed with a set of potential suffixes, one can obtain potential prefixes by starting with the original lexicon, stripping the potential suffixes, reversing the words, building a trie out of the reversed words, and finding potential suffixes of these reversed strings, which will be a set of potential prefixes in reverse.
- Identify candidate circumfixes, defined as prefix-suffix combinations that are attached to some minimum number of stems that are also shared by other potential circumfixes. The stems here are actually called pseudostems since, of course, they may not actually correspond to morphological stems.

Trie?



Output

- The output of these steps for English, German, and Dutch, with particular settings for the minima mentioned above, produces a large set of rules (about 30,000 for English) of which some are reasonable (e.g. -s → *null*, -ed → -ing) but many of which are not (e.g. s- → *null*, as induced from such seeming alternations as *stick/tick* or *spark/park*.)

Yarowsky and Wicentowski (2001)

- A lightly supervised method for inducing analyzers of inflectional morphology.
- The method consists of two basic steps.
- First a table of alignments between root and inflected forms is estimated from data.
- The table might contain the pair *take/took*, indicating that *took* is a particular inflected form of *take*.
- Second, a supervised morphological analysis learner is trained on a weighted subset of the table.
- In considering possible alignment pairs, Yarowsky and Wicentowski concentrate on morphology that is expressed by suffixes or by changes in the root (as in *take/took*).