Unsupervised learning of natural language morphology
John Goldsmith
March 1, 2010

http://linguistica.uchicago.edu

Word discovery

A good deal of work beginning in the late 1960s. Two widely-cited MIT dissertations in the mid 1990s on this, by Michael Brent and Carl de Marcken.

3749 sentences, 400,000 characters:
The Fulton County Grand Jury said Friday an investigation of Atlanta’s recent primary election produced no evidence that any irregularities took place. The jury further said in the end presentations that the City Executive Committee, which had overall charge of the election, deserves the praise and thanks of the City of Atlanta for the manner in which the election was conducted...

Select the lexicon \( L \) which minimizes the description length of the corpus \( C \). A lexicon \( L \) is a distribution \( \text{pr}_L \) over a subset of \( \Sigma^* \). \( L \)'s length is the length in bits in some specified format (the format matters!) and encoding. Any such distribution assigns a minimal encoding (up to trivial variants) to the corpus, and this encoding requires precisely \(-\log \text{pr}_L(C)\) bits. The description length of a corpus given lexicon \( L \) is defined as \(|L| - \log \text{pr}_L(C)\): select the lexicon that minimizes this quantity (as best you can). \(|L| \) comes into the picture because if we assume \( L \) is expressed in a binary-encoded format in which no morphology is a prefix of another, this encoding induces a natural probability distribution, with \( \text{pr}(l) \) proportional to \(2^{|l|} \).
**Big Picture question**

Can we build a picture of linguistics in which the goal is to specify a function mapping from the spaces of corpora $\times$ space of grammars such that for a fixed corpus, the optimal value of the function identifies the grammar that is in some *linguistic* sense correct? $g^* = \arg\max_g F(C, g)$, where $C$ is a given set of observations ("corpus"), and $g \in G$: how much is gained by restricting the set $G$? Such restrictions amount to an assumption about innate knowledge/Universal Grammar. An alternative strategy is (following Rissanen) to choose a Universal Turing Machine (UTM), and assign a probability to a grammar equal to $2^{-|I(g)|}$, where $|I(g)|$ is the length of the shortest implementation of grammar $g$ on this particular UTM. Does it matter that (1) this statement does not offer any hope that we can recognize the shortest implementation when we see it, or (2) we have no way to choose among UTMs: how do we determine whether UTM-choice matters, in a world of finite data and in which limits may not be taken?

If we want to tackle the problem of discovering linguistic structure, both phonology and syntax have the problem that their structure is heavily influenced by the nature of sound and perception (in the case of phonology) and of meaning and logical structure, in the case of syntax. Morphology is less influenced by such matters, and it is possible to emphasize both cross-linguistic variation and formal simplicity. *It is a good test case for language-learning from a computational point of view.*

The design of an appropriate objective function—explicating what the description length of a morphology is—is half the project; the other half is designing appropriate and workable discovery heuristics.

The goal is not to provide a morphology of English: it is to develop a language-independent morphology learner. Standard orthography (when it departs from phonemic representations) has rules that are similar to (and of the same type, in general) as the rules we find in phonology.

**Morph discovery: breaking words into pieces**

*What is the question?*

We identify morphemes due to frequency of occurrence: yes, but all of their sub-strings have at least as high a frequency, so frequency is only a small part of the matter; and due to the non-informativeness of their end with respect to what follows.

But those are *heuristics*: the real answer lies in formulating an FSA (with post-editing) that is simple, and generates the data.
List of stems:
\[ \sum_{t \in \text{Stems}} \sum_{i=1}^{|t|+1} -\log \text{pr}(t_i | t_{i-1}) \]

List of affixes:
\[ \sum_{f \in \text{Affixes}} \sum_{i=1}^{|f|+1} -\log \text{pr}(f_i | f_{i-1}) \]

Signatures:
\[ \sum_{\sigma \in \text{Signatures}} \left( \sum_{\text{stem } t \in \sigma} -\log \text{pr}(t) + \sum_{\text{suffix } f \in \sigma} -\log \text{pr}(f) \right) \]

\text{pr}(\text{word}) = \text{pr}(\sigma_W) * \text{pr}(t|\sigma_W) * \text{pr}(f|\sigma),
\]
where \( w \) is stem \( t \) + suffix \( f \); each stem belongs to a single signature.

PFSA \((V, E, L)\), with 4 distributions:
(a) \( pr_1() \) over \( E \) s.t. \( \sum_i pr_1(e_{ij}) = 1 \);
(b) \( pr_2() \) over \( V \);
(c) \( pr_3() \) over \( L \) (labels, i.e., morphemes), and
(d) \( pr_4() \) over \( \Sigma \), i.e., the alphabet used for \( L \).

Then \( \text{pr}(w) = \text{pr}(\text{path}_w) = \prod_{e \in \text{path}_w} pr_1(e) \).

\[
\text{FSA} = |V| + |E| + |L|, \text{ where } |v| = -\log \text{pr}_2(v) \text{.}
\]

\[
|E| = \sum_{e \in E} |e|, \text{ where } |e_{ij}| = |v_i| + |v_j| + |\text{ptr(label}_e)|, \text{ and}
\]

\[
|\text{ptr(label}_e)| = -\log \text{pr}_3(\text{label}_e). \text{, and}
\]

\[
|L| = \sum_{l \in L} |l|, \text{ where } |l| = -\sum_i \log \text{pr}_4(l_i). \text{.}
\]

Immediate issues: getting the morphology right

1. Real versus accidental subcases: When should sub-signatures be subsumed by the “mother” signature? When are two signatures

---

**Figure 2:** Bit cost of signature-based morphology

**Figure 3:** Word probability model: \( w \) is \( word, t \) stem, \( f \) suffix

**Figure 4:** More generally, an acyclic FSA. Natural identity between words and paths through the FSA: \( w \approx \text{path}_w \). There are various natural, and not so natural, ways to assign these distributions.

---

**English:** NULL - s - ed - ing - es - er - s - e - ly - y - al - ers - ism - man - a - (American, formula-al) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - ed - ate - ent - ment - t (burnt) - ism - man - est - ant - one - ated - ical - ance - tive - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - ment - t (burnt) - ism - man - est - ant - one - ated - ical - ance - tive - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ive - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less - d - (agreed) - ted - men - a - (American, formula-al) - n (blown/blown) - ful - or - ived - on - ian - ation - en - ies - ous - ate - ent - men - able - ness - ating - less -
two samples from the same multinomial distribution? In some cases, this seems like a question with a clear meaning, as in case (a). Case (b) is less clear. Case (e) is interestingly different.

(a) NULL-s vs NULL.ed.ing.s;
(b) NULL-s vs NULL-s’s
(c) NULL-ed.ing-s vs NULL-ed-ing-ment-s
(d) NULL-ed-er-ers-ing-s: how do we treat this?
(e) NULL-ed-ing-s (vs) NULL-ing-s (e.g., pull-pulling-pulls); similar question arises for all so-called strong English verbs (this is a linguistically common situation).

2. The role of “post-editing”: phonology and morphophonology.
   (a) final e-deletion in English
   (b) C-doubling (cut/cutting, hit/hitting; bite/bitten)
   (c) i/y alternation: beauty-beautiful; fly/flies;

   A calculation regarding a conjectured “phonological process” that falls half-way between heuristic and application of our DL-based objective function: Consider a process described as mapping \( X \rightarrow Y/\text{context} \). Rewrite the data as if that expressed an equivalence: we “divide” the data by that relation (for simplicity’s sake, we ignore the context). In this case, the result is a corpus from which all e’s have been deleted. What is the impact on the morphology that is induced from this new data? The lexical items are (of course) simpler (shorter). But the new morphology is much simpler than before, because signatures now collapse. NULL.ed.ing.s and e.ed.es.ing both map to NULL.d.ing.s. Each was of roughly the same order of magnitude; hence the bit cost of a pointer to the new signature is 1 bit less than that of the previous pointers, and that is a single bit of savings multiplied by thousands of times in the description length of the new corpus (quite independent of the missing es).

3. Succession of affixes: Stems of the signature NULL-s end in ship, ist, ment, ing. We can apply the analysis iteratively, re-analyzing all stems (and unanalyzed words), but this is not an adequate solution.

4. NULL-ed-ing-s vs. t-ted-ts-ting (Faulty MDL assumption?)

5. Clustering when no stem samples all its possible suffixes, but a family of them does: verbs in Romance languages.

Swahili
proud, loud

lord, hard, friend

buddh, special, capital

ship

dog, boy, girl

cultive, calcul

at, er

jump, walk, love, move

enjoy

inhibit

represent

boy

friend

thing

buddha

hard

ed

ing

s

ation

's

ion

ment

ship

ist

able

ly

er

est

Figure 5: FSA with morphemes labeling edges

Figure 6: FSA with morphemes as states
Figure 7: Simplified Swahili verbal morphology