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.

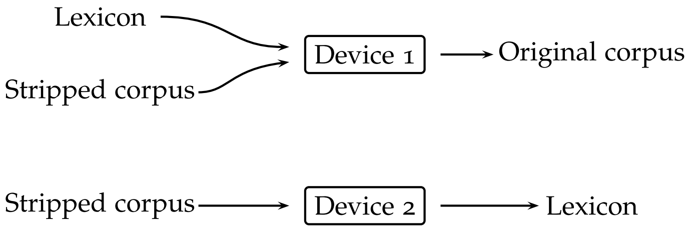

Figure 1: The two problems of word segmentation

3749 sentences, 400,000 characters:

TheFultonCountyGrandJurysaidFridayaninvestigationofAtl anta'srecentprimaryelectionproducednoevidencethatan yirregularitiestookplace.f Thejuryfurthersaidinterm-endpresentmentsthattheCityE xecutiveCommittee,whichhadover-allchargeoftheelecti on,deservesthepraiseandthanksoftheCityofAtlantaforthem annerinwhichtheelectionwasconducted ...

The Fulton County Grand Ju ry s aid Friday an investi gation of At l anta 's recent prim ary e lection produc ed no e videnc e that any ir regul ar it i es took place . Thejury further s aid in term - end present ment s thatthe City Ex ecutive Commit t e e ,which had over - all charg e ofthe e lection , d e serv e s the pra is e and than k softhe City of At l anta forthe man ner in whichthe e lection was conduc ted.

Select the lexicon $\mathcal{L}$ which minimizes the description length of the corpus $\mathcal{C}$. A lexicon $\mathcal{L}$ is a distribution $pr_{\mathcal{L}}$ over a subset of Σ^*. $\mathcal{L}$'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 $-logpr(\mathcal{C})$ bits. The description length of a corpus given lexicon $\mathcal{L}$ is defined as $|\mathcal{L}| - logpr_{\mathcal{L}}\mathcal{C}$: select the lexicon that minimizes this quantity (as best you can). $|\mathcal{L}|$ comes into the picture because if we assume $\mathcal{L}$ is expressed in a binary-encoded format in which no morphology is a prefix of another, this encoding induces a natural probability distribution, with $pr(l)$ 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 $g \in \mathcal{G}$: how much is gained by restricting the set $\mathcal{G}$? Such restrictions amount to an assumption about innate knowledge/Univeral Grammar. An alternative strategy is (following Rissanen) to choose a Universal Turing Machine (UTM), and assign a probability to a grammar equal to $2^{-|l(g)|}$, where $|l(g)|$ 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?

$g^* = \arg\max_g F(C, g)$, where C is a given set of observations ("corpus"). Classical MDL offers the joint probability of the data and model as its candidate for F.

Why **morphology**?

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.*

2 goals: objective function and learning heuristics

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.

Why conventional orthography? Why not phonemes?

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:

$$\sum_{t \in Stems} \sum_{i=1}^{|t|+1} -log\, pr(t_i | t_{i-1})$$

List of affixes:

$$\sum_{f \in Affixes} \sum_{i=1}^{|f|+1} -log\, pr(f_i | f_{i-1})$$

Signatures:

$$\sum_{\sigma \in Signatures} \left(\sum_{stem\, t \in \sigma} -log\, pr(t) + \sum_{suffix\, f \in \sigma} -log\, pr(f) \right)$$

Figure 2: Bit cost of signature-based morphology

$pr(word) = pr(\sigma_W) * pr(t|\sigma_w) * pr(f|\sigma)$,
where word w = stem t + suffix f; each stem belongs to a single signature. .

Figure 3: Word probability model: w is $word$, t $stem$, f $suffix$

PFSA $(\mathcal{V}, \mathcal{E}, \mathcal{L})$, with 4 distributions:
(a) $pr_1(\,)$over $\mathcal{E}$ s.t. $\sum_j pr_1(e_{i,j}) = 1$; (b) $pr_2()$ over $\mathcal{V}$;
(c) $pr_3()$ over $\mathcal{L}$ (labels, i.e., morphemes), and
(d) $pr_4()$ over Σ, i.e., the alphabet used for $\mathcal{L}$.
Then $pr(w) = pr(path_w) = \prod_{e \in path_w} pr_1(e)$.;
$|FSA| = |\mathcal{V}| + |\mathcal{E}| + |\mathcal{L}|$.
$|\mathcal{V}| = \sum_{v \in \mathcal{V}} |v|$, where $|v| = -log pr_2(v)$.
$|\mathcal{E}| = \sum_{e \in \mathcal{E}} |e|$, where $|e_{ij}| = |v_i| + |v_j| + |ptr(label_e)|$, and $|ptr(label_e)| = -log pr_3(label_e)$.
$|\mathcal{L}| = \sum_{l \in \mathcal{L}} |l|; |l| = -\sum_i log pr_4(l_i)$.

Figure 4: More generally, an acyclic FSA. Natural identity between words and paths through the FSA: $w \approx path_w$. There are various natural, and not so natural, ways to assign these distributions.

Signatures	Exemplar	Descr. Length (model)	Corpus Count	Stem Count	Source
NULL-s	accommodation	12996.7	13787	978	SF1
's-NULL	a*a*u	4237.23	8263	324	SF1
NULL-ly	according	3436.6	3391	259	SF1
NULL-ed-ing-s	account	886.936	2852	76	SF1
ed.ing	allott	1036.02	272	71	SF1
NULL.ed	abolish	1308.03	392	91	SF1
NULL.ed.s	accent	646.789	859	51	SF1
NULL.ing.s	boat	592.372	1060	46	SF1
NULL.ing	abound	1078.03	528	76	SF1
NULL.ed.ing	absorb	503.885	364	37	SF1
ing.s	awaken	172.814	29	11	SF1
ed.ing.s	fad	56.9268	13	3	SF1
's-NULL-s	afternoon	967.65	4258	83	SF1
e-ed-es-ing	accus	480.75	1345	40	Known stems to
e.ed.es	advanc	497.055	702	38	Check sigs
e.ed	acquiesc	825.969	311	58	Check sigs
e.ed.ing	anticipat	337.05	189	24	Known stems to
e.es.ing	battl	208.905	478	16	Known stems to
e.ing	abid	395.385	128	27	SF1
ed.es	aggravat	330.992	146	23	Check sigs
es.ing	celebrat	254.894	72	17	SF1
ed.es.ing	experienc	55.0602	35	3	From known sten
ies-y	abilit	899.932	642	66	SF1
NULL-al-s	addition	310.116	485	24	SF1
NULL.al	dramatic	87.2327	65	6	Check sigs
NULL-ly-s	absolute	320.709	468	25	SF1

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

English: NULL - s - ed - ing - es- er - 's - e - ly - y - al - ers - in - ic - tion - ation - en - ies - ion - able - ity - ness - ous - ate - ent - ment - t (*burnt*) - ism - man - est - ant - ence - ated - ical - ance - tive - ating - less - d (*agreed*) - ted - men - a (*Americana, formul-a/-ate*) - n (*blow/blown*) - ful - or - ive - on - ian - age - ial - o (*command-o, concert-o*) ...

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-beatiful; 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 $X \rightarrow Y/context$. 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 *e*s).

French: s - es - e- er - ent - ant - a - ée - é - és - ie - re - ement - tion - ique - ait - èrent - on - ées - te - ation - is - aient - al - ité - eur - aire - it - isme - en - age - ion - aux - ier - ale - iste - ien - t - eux - ance - ence - elle - iens - euse - ants - ienne - sion ...

$e \rightarrow \emptyset / -ed, -ing$

$corpus \Rightarrow corpus/e \approx \emptyset.$

creeps is now spelled *crps*, and *creeping* is *crping*.

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

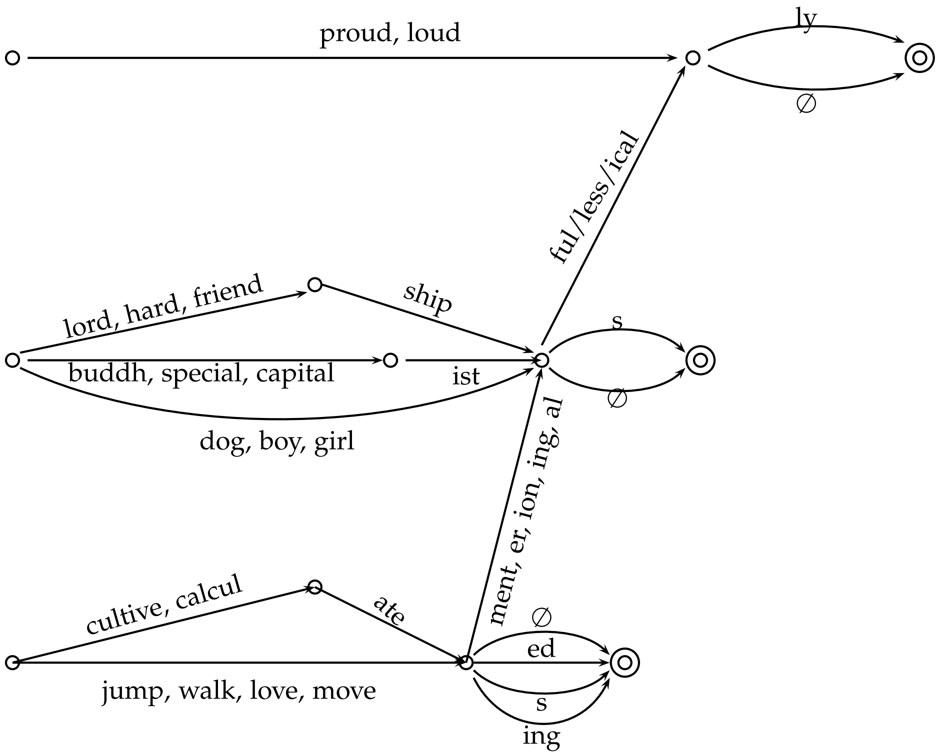

Figure 5: FSA with morphemes labeling edges

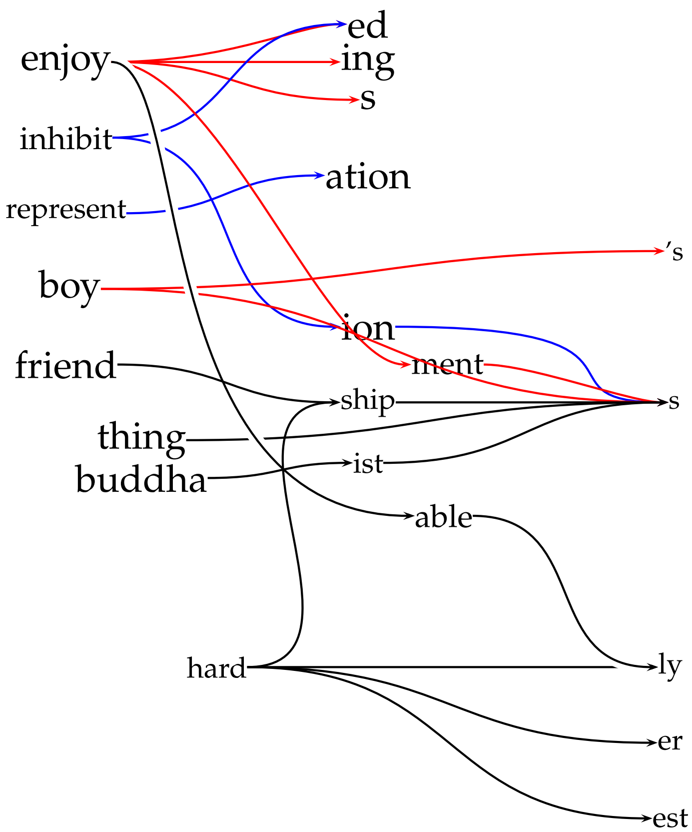

Figure 6: FSA with morphemes as states

Figure 7: Simplified Swahili verbal morphology