Psych 229: Language Acquisition

Lecture 6

Words & Models

Gambell & Yang 2006: Computational model of word segmentation

modeling statistical learning (TPs)

The modeling of statistical learning is straightforward, though it may be useful to make the details of our implementation clear. The model consists of two stages: training and testing. During the training stage, the learner gathers transitional probabilities over adjacent syllables in the learning data. The testing stage does not start until the entire learning data has been processed, and statistical learning is applied to the same data used in the training stage.

Another technical detail also needs to be spelled out: the TPs are gathered without stress information. That is, when counting syllable frequencies, the learner does not distinguish, say, a stressed syllable /ba/ from among the unstressed one.

That is, there is a word boundary AB and CD if if TP(A→B) >TP(B→C) < TP(C→D). The conjectured word boundaries are then compared against the target segmentation. Scoring is done for each utterance, using the definition of precision and recall in (1)

results

Modeling shows that the statistical learning (Saffran et al., 1996) does not reliably segment words such as those in child-directed English. Specifically, precision is 41.6%, recall is 23.3%. In other words, about 60% of words postulated by the statistical learner are not English words, and almost 80% of actual English words are not extracted. This is so even under favorable learning conditions:

- the child has syllabified the speech perfectly,
- the child has neutralized the effect of stress among the variants of syllables, which reduces the sparse data problem,
- and the data for segmentation is the same as the data used in training, which eliminates the sparse data problem

Gambell & Yang 2006: Computational model of word segmentation

What happened?

We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason. A necessary condition on the use of TP local minima to extract words is that words must consist of multiple syllables. If the target sequence of segmentation contains only monosyllabic words, it is clear that statistical learning will fail. A sequence of monosyllabic words require a word boundary after each syllable; a statistical learner, on the other hand, will only place a word boundary between two sequences of syllables for which the TPs within are higher than that in the middle. Indeed, in the artificial language learning experiment of Saffran et al. (1996) and much subsequent work, the pseudowords are uniformly three syllables long. However, the case of child-directed English is quite different. The fact that the learning data consists of 226,178 words but only 263,660 syllables suggests that the overwhelming majority of word tokens are monosyllabic. More specifically, a monosyllabic word is followed by another monosyllabic word 85% of time. As long as this is the case, statistical learning cannot work.

Gambell & Yang 2006: Computational model of word segmentation

Would more data help?…probably not

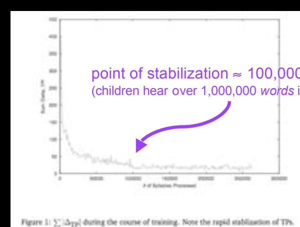

Figure 1: $\sum |\delta_{TP}|$ during the course of training. Note the rapid stabilization of TPs.

Gambell & Yang 2006: Computational model of word segmentation

What about other models(Swingley (2005)) that have success on data like this?

Swingley's corpus study makes use of multiple source of statistical information. Specifically, it maintains three kinds of information units: single syllables, adjacent syllable pairs (bigrams), and adjacent syllable triples (trigrams). Four types of statistical information are accumulated: the frequencies of these three units, in addition to he mutual information between adjacent syllable pairs (I_{AB}).[7] These numbers are then ranked along a percentile scale, much like standardized tests.

How plausible is it that infants track all of this and know the special threshold for "yes, it's a word"?

Finally, issues remain in the interpretation of Swingley's results. It is true that overall precision may be quite high for certain values of θ but it is worth noting that most of the three-syllable words determined by Swingley's criteria are wrong: the precision is consistently under 25-30% (Swingley, ibid; Figure 1) regardless the value of θ. Moreover, the statistical criteria in (3) produce very low recalls. Swingley does not provide raw data but the performance plots in his paper show that the maximum number of correctly extracted words does not appear to exceed 400-500. Given that Swingley's corpus contains about 1,800 distinct word types (ibid; p96), the recall is at best 22-27%.

Results not so good on precision either…

In sum, the corpus study of Swingley (2005) considers a number of statistical regularities that could be extracted in the linguistic data. The extraction of these regularities, and the criteria postulated for finding word boundaries, are not always supported by independent evidence. Even if these assumptions were motivated, the segmentation results remain poor, particularly for recall and longer words.

Gambell & Yang 2006: Computational model of word segmentation

Here's an idea…(and a language-independent one at that)

(4) The Unique Stress Constraint (USC): A word can bear at most one primary stress (a strong syllable).

Is this the only language-independent constraint?

How easy is it to identify "main stress" in a word (especially cross-linguistically)?

Why do they think this might work?

Its pervasiveness can be appreciated if we–at least those of us that are old enough–warped ourselves back to the 1977 premiere of *Star Wars*. Upon hearing "chewbacca" and "darthvader" for the very first time, it must have been immediately clear that the former utterance is one word, the latter is two (though whatever they meant was altogether a different matter). Both sequences are three syllables so length is not a useful guide. Yet "chewbacca" contains only one primary stress, which falls on /ba/, whereas "darthvader" contains two primary stresses, which fall on /darth/ and /va/ respectively: USC immediately segments the utterances correctly. Likewise, we believe that USC provides important clues for word boundaries for an infant learner whose situation is not much different from first time *Star Wars* viewers.

First, and most directly, USC may give the learner many isolated words for free. This, so far as we know, constitutes the only known mechanism that takes advantage of the abundance of single word utterances (Brent & Siskind, 2001).

Second, and somewhat indirectly, USC can constrain the use of statistical learning. For example, the syllable consequence $S_1W_1W_2W_3S_2$ cannot be segmented by USC alone, but it may still provide highly informative cues that facilitate the application of other segmentation strategies. For instance, the learner knows that the sequence consists of two words, as indicated by two strong syllables.[8] Moreover, it also knows that in the window between S_1 and S_2 must lie a word boundary (or boundaries)–and that may be what statistical learning using local minima may be able to locate.

Gambell & Yang 2006: Computational model of word segmentation

Here's one model…

In the first model, we apply statistical learning when USC does not automatically identify word boundaries. In the training stage, TPs are gathered as before. In the testing stage, the learner scans a sequence of input syllables from left to right:

a. If two strong syllables are adjacent (i.e., "… S_1S_2 …"), a word boundary is postulated in between.
b. If there are more than one (weak) syllables between two strong ones (i.e, $S_1W...WS_2$), then a word boundary is postulated where the pairwise TP is at the local minimum.

Hey, not bad!

The improvement in segmentation results is remarkable: when constrained by USC, statistical learning with local minimum achieves precision of 73.5% and recall of 71.2%.

In fact, these figures are comparable to the highest performance reported in the literature (Brent, 1999a), which nevertheless uses uses a computationally prohibitive algorithm that iteratively optimizes over the entire lexicon. By contrast, the computational complexity of the present model is exactly that of computation of transitional probabilities, which appears to be less costly but still leaves much to be desired.

Gambell & Yang 2006: Computational model of word segmentation

What about algebraic learning?

Therefore, if the child has learned the word "big", she might be able to recognize "big" in the utterance "bigsnake" and extract "snake" as a result. For concreteness, call this bootstrapping process *subtraction* (Gambell & Yang, 2003). Furthermore, the subtraction strategy is evidenced by familiar observations of young children's speech. The irresistible segmentation errors (e.g., "I was have" from be-have, "hiccing up" from hicc-up, "two dults" from a-adult) suggest that subtraction does take place (cf. Peters, ibid). Recent work (Bortfeld, Morgan, et al., 2005) demonstrates that infants as young as 6 months old may use this bootstrapping strategy. For word sequences such as XY, where Y is a novel word, infants prefer those that are paired with a familiar X, such as "Mommy", the child's name, and others that may be developmentally appropriate for this stage.

Under algebraic learning, the learner has a lexicon which stores previously segmented words. No statistical training of the TPs is used. As before, the learner scans the input from left to right. If it recognizes a word that has been stored in the lexicon, it puts the word aside and proceeds to the remainder of the string. Again, the learner will use USC to segment words in the manner of (5a): in our modeling, this constraint handles most cases of segmentation. However, USC may not resolve word boundaries conclusively. This happens when the learner encounters $S_1W_1^nS_2$: the two S's stand for strong syllables, and there are n syllables in between, where W_i^j stands for the substring that spans from the ith to the jth weak syllable. In the window of W_1^n, two possibilities may arise.

Gambell & Yang 2006: Computational model of word segmentation

Strong Weak1 Weak2…Weakn Strong

(6) a. If both $S_1W_1^{i-1}$ and $W_{j+1}^n S_2$ ($i < j$) are, or are part of, known words on both sides of $S_1W_1^nS_2$, then W_i^j must be a word,[9] and the learner adds W_i^j as a new word into the lexicon. This is straightforward.
b. Otherwise, a word boundary lies somewhere in W_1^n, and USC does not provide reliable information. This is somewhat more complicated.

S W1 = known word
W2…Wn S = known word

(7) a. **Agnostic**: the learner ignores the strings $S_1W_1^nS_2$ altogether and proceeds to segment the rest of the utterance. No word is added to the lexicon.
b. **Random**: the learner picks a random position r ($1 \leq r \leq n$ and splits W_1^n into two substrings W_1^r and W_{r+1}^n, as parts of the two words containing S_1 and S_2 respectively.[10] Again, no word is added to the lexicon.

Agnostic: ignore this string

The logic behind the agnostic learner is that the learner is non-committal if the learning data contains uncertainty unresolvable by "hard" linguistic constraints such as USC.[11]

Random: pick a division point at random
S W1 W2 [word boundary] W3…Wn S

While the agnostic learner does not make a decision when such situations arise, it can be expected that the words in the sequence $S_1W_1^nS_2$ will mostly like appear in combinations with other words in future utterances, where USC may directly segment them out. The random learner is implemented as a baseline comparison, though we suspect that in actual language acquisition, the learner may invoke the language-specific Metrical Segmentation Strategy, rather than choosing word boundaries randomly, in ambiguous contexts such as $S_1W_1^nS_2$.

Gambell & Yang 2006: Computational model of word segmentation

Model	Precision	Recall	F-measure ($\alpha = 0.5$)
SL	41.6%	23.3%	0.298
SL + USC (5)	73.5%	71.2%	0.723
Algebraic agnostic (7a)	85.9%	89.9%	0.879
Algebraic random (7b)	95.9%	93.4%	0.946

Table 1: Performance of four models of segmentation. SL stands for the statistical learning model of Saffran et al. (1996), while the other three models are described in the text.

It may seem a bit surprising that the random algebraic learner yields the best segmentation results but this is not unexpected. The performance of the agnostic learner suffers from deliberately avoiding segmentation in a substring where word boundaries lie. The random learner, by contrast, always picks out some word boundary, which is very often correct. And this is purely due to the fact that words in child-directed English are generally short.

Gambell & Yang 2006: Computational model of word segmentation

Conclusions

- The segmentation process can get off the ground only through the use of language-independent means: experience-independent linguistic constraints such as USC and experience-dependent statistical learning are the only candidates among the proposed strategies for word segmentation.
- Statistical learning does not scale up to realistic settings of language acquisition.
- Simple principles on phonological structures such as USC can constrain the applicability of statistical learning and improve its performance, though the computational cost of statistical learning may still be prohibitive.
- Algebraic learning under USC, which has trivial computational cost and is in principle universally applicable, outperforms all other segmentation models.

Statistical learning (Saffran et al., 1996) surely ranks among the most important discoveries of our cognitive abilities. Yet it remains to be seen, contrary to a number of claims (Bates & Elman, 1996; Seidenberg, 1997, etc.), whether statistical learning serves as an alternative to innate and domain-specific knowledge of language (Universal Grammar, broadly speaking). In addition, as the present study shows, it remains an open question whether statistical learning using local minima is used in actual word segmentation in the first place.

Gambell & Yang 2006: Computational model of word segmentation

Conclusions

First, does the ability to learn diminish the need for Universal Grammar? Here we concur with Saffran et al. (1997), who are cautious about the interpretation of their results.

The same logic applies to the success of statistical learning in segmenting artificial language: it presupposes the learner knowing what kind of statistical information to keep track of. After all, an infinite range of statistical correlations exists: e.g., What is the probability of a syllable rhyming with the next? What is the probability of two adjacent vowels being both nasal? The fact that infants can use statistical learning in the first place entails that, at the minimum, they know the relevant unit of information over which correlative statistics is gathered: in this case, it is the syllables, rather than segments, or front vowels, or labial consonants.

Constraints on learning (innate biases)

It is worth reiterating that our critical stance on statistical learning refers only to a specific kind of statistical learning that exploits local minima over adjacent linguistic units (Saffran et al., 1996). Rather, we simply wish to reiterate the conclusion from decades of machine learning research that no learning, statistical or otherwise, is possible without appropriate prior assumptions on the representation of the learning data and a constrained hypothesis space. Recent work on the statistical learning over non-adjacent phonological units has turned out some interesting limitations on the kind of learnable statistical correlations (Newport & Aslin, 2004, Aslin, Newport, & Hauser, 2004; Toro, Sinnett, & Soto-Faraco, in press; Peña, Bonatti, Nespor, & Mehler, 2002; for visual learning tasks, see Tucker-Brown, Junge, & Scholl, submitted; Catena & Scholl, submitted). The present work, then, can be viewed as an attempt to articulate the specific linguistic constraints that might be built in for successful word segmentation to take place.

Gambell & Yang 2006: Computational model of word segmentation

Discussion Questions

What about other languages besides English? (Turkish, Mohawk - polysynthetic languages)

An example from Chukchi, a polysynthetic, incorporating, and agglutinating language:

Təmeyŋəlevtpəγtərkən.

t-ə-meyŋ-ə-levt-pəγt-ə-rkən

1.SG.SUBJ-great-head-hurt-PRES.1

'I have a fierce headache.' (Skorik 1961: 102)

Təmeyŋəlevtpəγtərkən has a 5:1 morpheme-to-word ratio with 3 incorporated lexical morphemes (meyŋ 'great', levt 'head', pəγt 'ache').

What does it mean that the USC+Algebraic learner actually identifies words much quicker than real children seem to? [~Bruno]

Gómez & Lakusta 2004: Categorization

Nouns, Verbs, Adjectives…

Given the important role of category information in linguistic productivity, a critical question is how children might achieve such generalization.

One Idea: Semantic Bootstrapping

A widely held view, referred to as the semantic bootstrapping hypothesis, is that young children discover lexical categories by first noting semantic or referential information.[1] By this view, learners are equipped with knowledge of innate categories, such as noun and verb, as well as knowledge of grammatical functions, such as subject and object (Grimshaw, 1981; Pinker, 1984). Children identify semantic referents in the world by means of perceptual processing and then link these to innate knowledge of syntactic categories and functions.

Another Idea: Distributional Learning

A very different view assumes that distributional relationships among form-based cues are central to category-based abstraction (Braine, 1987; Gerken, Landau & Remez, 1990; Gleitman & Wanner, 1982; Morgan & Demuth, 1996; Morgan & Newport, 1981; Redington, Chater & Finch, 1998). Examples of such cues are relative location of words in strings, phonological regularities within words of a class and co-occurrence relations between classes. With regard to phonological regularities within a class, functor categories tend to have shorter vowel durations, weaker amplitudes and simplified syllable structure compared to lexical categories such as noun and verb (Morgan, Shi & Allopena, 1996; Shi, Morgan & Allopena, 1998).

Gómez & Lakusta 2004: Categorization

What babies can do…

& Allopena, 1998). Newborn infants are sensitive to such differences (Shi, Werker & Morgan, 1999) and by 7 months of age, infants can recognize and track specific functor elements in running speech (Höhle & Weissenborn, 2003). Nouns and verbs are also distinguishable by means of phonological cues.

What babies might do…

If infants are able to identify categories in the speech stream by means of their phonological properties, they might then use this information to learn the predictive relationships between categories. In English, for example, children must learn that 'the' and 'a' precede nouns and not verbs, whereas 'will' and 'can' precede verbs but not nouns. An infant who has learned that particular functors predict particular lexical forms (i.e. one who has identified categories in speech and the relationships between them) will have a considerable advantage with respect to the later task of mapping between meaning and form, compared to the toddler who only begins this process once semantic knowledge is more fully in place (Gómez & Gerken, 2000; Naigles, 2002).

Gómez & Lakusta 2004: Categorization

Category abstraction task

Table 1 *A paradigm for investigating category abstraction. Learners are exposed to the pairings shown below except for those denoted by empty cells. Learners are then tested to see if they will generalize correctly to the withheld strings (denoted by empty cells)*

	X_1	X_2	X_3	X_4	X_5
a_1 = the	boy	girl	ball	dog	cat
a_2 = a	boy	girl	ball	dog	
	Y_1	Y_2	Y_3	Y_4	Y_5
b_1 = will	jump	run	play	sleep	eat
b_2 = can	jump	run	play	sleep	

Previous work (aX, bY paradigm)

Interestingly, although learners readily acquire the legal positions of words with respect to which occur first versus second (Smith, 1969), categories and their relationships (i.e. that words belong to particular a, b, X, and Y classes, and that a-words go with Xs and not Ys) are virtually impossible to acquire unless some subset of the X- and Y-category members are marked with salient conceptual or perceptual cues