Lexicalized phonotactic word segmentation

Margaret M. Fleck
Department of Computer Science
University of Illinois
Urbana, IL 61801, USA
mfleck@cs.uiuc.edu

Abstract

This paper presents a new unsupervised algorithm (WordEnds) for inferring word boundaries from transcribed adult conversations. Phone ngrams before and after observed pauses are used to bootstrap a simple discriminative model of boundary marking. This fast algorithm delivers high performance even on morphologically complex words in English and Arabic, and promising results on accurate phonetic transcriptions with extensive pronunciation variation. Expanding training data beyond the traditional miniature datasets pushes performance numbers well above those previously reported. This suggests that WordEnds is a viable model of child language acquisition and might be useful in speech understanding.

1 Introduction

Words are essential to most models of language and speech understanding. Word boundaries define the places at which speakers can fluently pause, and limit the application of most phonological rules. Words are a key constituent in structural analyses: the output of morphological rules and the constituents in syntactic parsing. Most speech recognizers are word-based. And, words are entrenched in the writing systems of many languages.

Therefore, it is generally accepted that children learning their first language must learn how to segment speech into a sequence of words. Similar, but more limited, learning occurs when adults hear speech containing unfamiliar words. These words must be accurately delimited, so that they can be added to the lexicon and nearby familiar words recognized correctly. Current speech recognizers typically misinterpret such speech.

This paper will consider algorithms which segment phonetically transcribed speech into words. For example, Figure 1 shows a transcribed phrase from the Buckeye corpus (Pitt et al., 2005; Pitt et al., 2007) and the automatically segmented output. Like almost all previous researchers, I use human-transcribed input to work around the limitations of current speech recognizers.

In most available datasets, words are transcribed using standard dictionary pronunciations (henceforth "dictionary transcriptions"). These transcriptions are approximately phonemic and, more importantly, assign a constant form to each word. I will also use one dataset with accurate phonetic transcriptions, including natural variation in the pronunciation of words. Handling this variation is an important step towards eventually using phone lattices or features produced by real speech recognizers.

This paper will focus on segmentation of speech between adults. This is the primary input for speech recognizers. Moreover, understanding such speech is the end goal of child language acquisition. Models tested only on simplified child-directed speech are incomplete without an algorithm for upgrading the understander to handle normal adult speech.

2 The task in more detail

This paper uses a simple model of the segmentation task, which matches prior work and the available datasets. Possible enhancements to the model are discussed at the end.

```
   "all the kids in there       # are people that have kids # or that are having kids"
IN  REAL: ohlThikidsinner       # ahrpiyp@lThA?HAvkids      # ohrThADurHAviynqkids
    DICT: ahlThiykidzinTher     # ahrpiyp@lThAtHAvkidz      # owrThAtahrHAvinqkidz
OUT REAL: ohl Thi kids inner    # ahr piyp@l ThA? HAv kids  # ohr ThADur HAviynq kids
    DICT: ahl Thiy kidz in Ther # ahr piyp@l ThAt HAv kidz  # owr ThAt ahr HAvinq kidz
```

Figure 1: Part of Buckeye corpus dialog 2101a, in accurate phonetic transcription (REAL) and dictionary pronunciations (DICT). Both use modified arpabet, with # marking pauses. Notice the two distinct pronunciations of "that" in the accurate transcription. Automatically inserted word boundaries are shown at bottom.

2.1 The input data

This paper considers only languages with an established tradition of words, e.g. not Chinese. I assume that the authors of each corpus have given us reasonable phonetic transcriptions and word boundaries. The datasets are informal conversations in which debatable word segmentations are rare.

The transcribed data is represented as a sequence of phones, with neither prosodic/stress information nor feature representations for the phones. These phone sequences are presented to segmentation algorithms as strings of ASCII characters. Large phonesets may be represented using capital letters and punctuation or, more readably, using multi-character phone symbols. Well-designed (e.g. easily decodable) multi-character codes do not affect the algorithms or evaluation metrics in this paper. Testing often also uses orthographic datasets.

Finally, the transcriptions are divided into "phrases" at pauses in the speech signal (silences, breaths, etc). These pause phrases are **not** necessarily syntactic or prosodic constituents. Disfluencies in conversational speech create pauses where you might not expect them, e.g. immediately following the definite article (Clark and Wasow, 1998; Fox Tree and Clark, 1997). Therefore, I have chosen corpora in which pauses have been marked carefully.

2.2 Affixes and syllables

A theory of word segmentation must explain how affixes differ from free-standing function words. For example, we must explain why English speakers consider "the" to be a word, but "-ing" to be an affix, although neither occurs by itself in fluent prepared English. We must also explain why the Arabic determiner "Al-" is not a word, though its syntactic and semantic role seems similar to English "the".

Viewed another way, we must show how to estimate the average word length. Conversational English has short words (about 3 phones), because most grammatical morphemes are free-standing. Languages with many affixes have longer words, e.g. my Arabic data averages 5.6 phones per word.

Pauses are vital for deciding what is an affix. Attempts to segment transcriptions without pauses, e.g. (Christiansen et al., 1998), have worked poorly. Claims that humans can extract words without pauses seem to be based on psychological experiments such as (Saffran, 2001; Jusczyk and Aslin, 1995) which conflate words and morphemes. Even then, explicit boundaries seem to improve performance (Seidl and Johnson, 2006).

Another significant part of this task is finding syllable boundaries. For English, many phone strings have multiple possible syllabifications. Because words average only 1.26 syllables, segmenting pre-syllabified input has a very high baseline: 100% precision and 80% recall of boundary positions.

2.3 Algorithm testing

Unsupervised algorithms are presented with the transcription, divided only at phrase boundaries. Their task is to infer the phrase-internal word boundaries. The primary worry in testing is that development may have biased the algorithm towards a particular language, speaking style, and/or corpus size. Addressing this requires showing that different corpora can be handled with a common set of parameter settings. Therefore a test/training split within one corpus serves little purpose and is not standard.

Supervised algorithms are given training data with all word boundaries marked, and must infer word boundaries in a separate test set. Simple supervised algorithms perform extremely well (Cairns et al., 1997; Teahan et al., 2000), but don't address our main goal: **learning** how to segment.

Notice that phrase boundaries are not randomly

selected word boundaries. Syntactic and communicative constraints make pauses more likely at certain positions than others. Therefore, the "supervised" algorithms for this task train on a representative set of word boundaries whereas "unsupervised" algorithms train on a biased set of word boundaries. Moreover, supplying **all** the word boundaries for even a small amount of data effectively tells the supervised algorithms the average word length, a parameter which is otherwise not easy to estimate.

Standard evaluation metrics include the precision, recall and F-score [1] of the phrase-internal boundaries (BP, BR, BF), of the extracted word tokens (WP, WR, WF), and of the resulting lexicon of word types (LP, LR, LF). Outputs don't look good until BF is at least 90%.

3 Previous work

Learning to segment words is an old problem, with extensive prior work surveyed in (Batchelder, 2002; Brent and Cartwright, 1996; Cairns et al., 1997; Goldwater, 2006; Hockema, 2006; Rytting, 2007). There are two major approaches. *Phonotactic* methods model which phone sequences are likely within words and which occur primarily across or adjacent to word boundaries. *Language modelling* methods build word ngram models, like those used in speech recognition. Statistical criteria define the "best" model fitting the input data. In both cases, details are complex and variable.

3.1 Phonotactic Methods

Supervised phonotactic methods date back at least to (Lamel and Zue, 1984), see also (Harrington et al., 1989). Statistics of phone trigrams provide sufficient information to segment adult conversational speech (dictionary transcriptions with simulated phonology) with about 90% precision and 93% recall (Cairns et al., 1997), see also (Hockema, 2006). Teahan et al.'s compression-based model (2000) achieves BF over 99% on orthographic English. Segmentation by adults is sensitive to phonotactic constraints (McQueen, 1998; Weber, 2000).

To build unsupervised algorithms, Brent and Cartwright suggested (1996) inferring phonotactic constraints from phone sequences observed at phrase boundaries. However, experimental results are poor. Early results using neural nets by Cairns et al. (1997) and Christiansen et al (1998) are discouraging. Rytting (2007) seems to have the best result: 61.0% boundary recall with 60.3% precision [2] on 26K words of modern Greek data, average word length 4.4 phones. This algorithm used mutual information plus phrase-final 2-phone sequences. He obtained similar results (Rytting, 2004) using phrase-final 3-phone sequences.

Word segmentation experiments by Christiansen and Allen (1997) and Harrington et al. (1989). simulated the effects of pronunciation variation and/or recognizer error. Rytting (2007) uses actual speech recognizer output. These experiments broke useful new ground, but poor algorithm performance (BF $\leq 50\%$ even on dictionary transcriptions) makes it hard to draw conclusions from their results.

3.2 Language modelling methods

So far, language modelling methods have been more effective. Brent (1999) and Venkataraman (2001) present incremental splitting algorithms with BF about 82% [3] on the Bernstein-Ratner (BR87) corpus of infant-directed English with disfluencies and interjections removed (Bernstein Ratner, 1987; Brent, 1999). Batchelder (2002) achieved almost identical results using a clustering algorithm. The most recent algorithm (Goldwater, 2006) achieves a BF of 85.8% using a Dirichlet Process bigram model, estimated using a Gibbs sampling algorithm.[4]

Language modelling methods incorporate a bias towards re-using hypothesized words. This suggests they should systematically segment morphologically complex words, so as to exploit the structure they share with other words. Goldwater, the only author to address this issue explicitly, reports that her algorithm breaks off common affixes (e.g. "ing", "s"). Batchelder reports a noticable drop in performance on Japanese data, which might relate to its more complex words (average 4.1 phones).

[1] $F = \frac{2PR}{P+R}$ where P is the precision and R is the recall.

[2] These numbers have been adjusted so as not to include boundaries between phrases.

[3] Numbers are from Goldwater's (2006) replication.

[4] Goldwater numbers are from the December 2007 version of her code, with its suggested parameter values: $\alpha_0 = 3000$, $\alpha_1 = 300$, $p\# = 0.2$.

4 The new approach

Previous algorithms have modelled either whole words or very short (e.g. 2-3) phone sequences. The new approach proposed in this paper, "lexicalized phonotactics," models extended sequences of phones at the starts and ends of word sequences. This allows a new algorithm, called WordEnds, to successfully mark word boundaries with a simple local classifier.

4.1 The idea

This method models sequences of phones that start or end at a word boundary. When words are long, such a sequence may cover only part of the word e.g. a group of suffixes or a suffix plus the end of the stem. A sequence may also include parts of multiple short words, capturing some simple bits of syntax.

These longer sequences capture not only purely phonotactic constraints, but also information about the inventory of lexical items. This improves handling of complex, messy inputs. (Cf. Ando and Lee's (2000) kanji segmenter.)

On the other hand, modelling only partial words helps the segmenter handle long, infrequent words. Long words are typically created by productive morphology and, thus, often start and end just like other words. Only 32% of words in Switchboard occur both before and after pauses, but many of the other 68% have similar-looking beginnings or endings.

Given an inter-character position in a phrase, its *right and left contexts* are the character sequences to its right and left. By convention, phrases input to WordEnds are padded with a single blank at each end. So the middle position of the phrase "afunjoke" has right context "joke␣" and left context "␣afun." Since this is a word boundary, the right context looks like the start of a real word sequence, and the left context looks like the end of one. This is not true for the immediately previous position, which has right context "njoke␣" and left context "␣afu."

Boundaries will be marked where the right and left contexts look like what we have observed at the starts and ends of phrases.

4.2 Statistical model

To formalize this, consider a fixed inter-character position in a phrase. It may be a word boundary (b) or not ($\neg b$). Let r and l be its right and left contexts. The input data will (see Section 4.3) give us $P(b|r)$ and $P(b|l)$. Deciding whether to mark a boundary at this position requires estimating $P(b|r,l)$.

To express $P(b|r,l)$ in terms of $P(b|l)$ and $P(b|r)$, I will assume that r and l are conditionally independent given b. This corresponds roughly to a unigram language model. Let $P(b)$ be the probability of a boundary at a random inter-character position. I will assume that the average word length, and therefore $P(b)$, is not absurdly small or large.

$P(b|r,l)$ is $\frac{P(r,l|b)P(b)}{P(r,l)}$. Conditional independence implies that this is $\frac{P(r|b)P(l|b)P(b)}{P(r,l)}$, which is $\frac{P(r)P(b|r)P(l)P(b|l)}{P(b)P(r,l)}$. This is $\frac{P(b|r)P(b|l)}{QP(b)}$ where $Q = \frac{P(r,l)}{P(r)P(l)}$. Q is typically not 1, because a right and left context often co-occur simply because they both tend to occur at boundaries.

To estimate Q, write $P(r,l)$ as $P(r,l,b) + P(r,l,\neg b)$. Then $P(r,l,b)$ is $\frac{P(r)P(b|r)P(l)P(b|l)}{P(b)}$. If we assume that r and l are also conditionally independent given $\neg b$, then a similar equation holds for $P(r,l,\neg b)$. So $Q = \frac{P(b|r)P(b|l)}{P(b)} + \frac{P(\neg b|r)P(\neg b|l)}{P(\neg b)}$

Contexts that occur primarily inside words (e.g. not at a syllable boundary) often restrict the adjacent context, violating conditional independence given $\neg b$. However, in these cases, $P(b|r)$ and/or $P(b|l)$ will be very low, so $P(b|r,l)$ will be very low. So (correctly) no boundary will be marked.

Thus, we can compute $P(b|r,l)$ from $P(b|r)$, $P(b|l)$, and $P(b)$. A boundary is marked if $P(b|r,l) \geq 0.5$.

4.3 Estimating context probabilities

Estimation of $P(b|r)$ and $P(b|l)$ uses a simple ngram backoff algorithm. The details will be shown for $P(b|l)$. $P(b|r)$ is similar.

Suppose for the moment that word boundaries are marked. The left context l might be very long and unusual. So we will estimate its statistics using a shorter lefthand neighborhood l'. $P(b|l)$ is then estimated as the number of times l' occurs before a boundary, divided by the total number of times l' occurs in the corpus.

The suffix l' is chosen to be the longest suffix of l which occurs at least 10 times in the corpus, i.e. often enough for a reliable estimate in the presence

corpus	language	transcription	sm size	med size	lg size	pho/wd	wd/phr	hapax
BR87	English	dictionary	33K	–	–	2.9	3.4	31.7
Switchboard	English	dictionary	34K	409K	3086K	3.1	5.9	33.8
Switchboard	English	orthographic	34K	409K	3086K	[3.8]	5.9	34.2
Buckeye	English	dictionary	32K	290K	–	3.1	5.9	41.9
Buckeye	English	phonetic	32K	290K	–	2.9	5.9	66.0
Arabic	Arabic	dictionary	30K	405K	–	5.6	5.9	60.3
Spanish	Spanish	dictionary	37K	200K	–	3.7	8.4	49.1

Table 1: Key parameters for each test dataset include the language, transcription method, number of words (small, medium, large subsets), average phones per word, average words per phrase, and percent of word types that occur only once (hapax). Phones/word is replaced by characters/word for the orthographic corpus.

of noise.[5] l' may cross word boundaries and, if our position is near a pause, may contain the blank at the lefthand end of the phrase. The length of l' is limited to N_{max} characters to reduce overfitting.

Unfortunately, our input data has boundaries only at pauses (#). So applying this method to the raw input data produces estimates of $P(\#|r)$ and $P(\#|l)$. Because phrase boundaries are not a representative selection of word boundaries, $P(\#|r)$ and $P(\#|l)$ are not good estimates of $P(b|r)$ and $P(b|l)$. Moreover, initially, we don't know $P(b)$.

Therefore, WordEnds bootstraps the estimation using a binary model of the relationship between word and phrase boundaries. To a first approximation, an ngram occurs at the end of a phrase if and only if it can occur at the end of a word. Since the magnitude of $P(\#, l)$ isn't helpful, we simply check whether it is zero and, accordingly, set $P(b|l)$ to either zero or a constant, very high value.

In fact, real data contains phrase endings corrupted by disfluencies, foreign words, etc. So WordEnds actually sets $P(b|l)$ high only if $P(\#|l)$ is above a threshold (currently 0.003) chosen to reflect the expected amount of corruption.

In the equations from Section 4.2, if either $P(b|r)$ or $P(b|l)$ is zero, then $P(b|r, l)$ is zero. If both values are very high, then Q is $\frac{P(b|r)P(b|l)}{P(b)} + \epsilon$, with ϵ very small. So $P(b|r, l)$ is close to 1. So, in the bootstrapping phase, the test for marking a boundary is independent of $P(b)$ and reduces to testing whether $P(\#|r)$ and $P(\#|l)$ are both over threshold.

So, WordEnds estimates $P(\#|r)$ and $P(\#|l)$ from the input data, then uses this bootstrapping method ($N_{max} = 5$) [6] to infer preliminary word boundaries. The preliminary boundaries are used to estimate $P(b)$ and to re-estimate $P(b|r)$ and $P(b|l)$, using $N_{max} = 4$. Final boundaries are then marked.

5 Mini-morph

In a full understanding system, output of the word segmenter would be passed to morphological and local syntactic processing. Because the segmenter is myopic, certain errors in its output would be easier to fix with the wider perspective available to this later processing. Because standard models of morphological learning don't address the interaction with word segmentation, WordEnds does a simple version of this repair process using a placeholder algorithm called Mini-morph.

Mini-morph fixes two types of defects in the segmentation. Short fragments are created when two nearby boundaries represent alternative reasonable segmentations rather than parts of a common segmentation. For example, "treestake" has potential boundaries both before and after the s. This issue was noted by Harrington et al. (1988) who used a list of known very short words to detect these cases. See also (Cairns et al., 1997). Also, surrounding words sometimes mislead WordEnds into undersegmenting a phone sequence which has an "obvious" analysis using well-established component words.

Mini-morph classifies each word in the segmentation as a fragment, a word that is reliable enough to use in subdividing other words, or unknown status.

[5] A single character is used if no suffix occurs 10 times.

[6] Values for N_{max} were chosen empirically. They could be adjusted for differences in entropy rate, but this is very similar across the datasets in this paper.

Because it has only a feeble model of morphology, Mini-morph has been designed to be cautious: most words are classified as unknown.

To classify a word, we compare its frequency w as a word in the segmentation to the frequencies p and s with which it occurs as a prefix and suffix of words in the segmentation (including itself). The word's fragment ratio f is $\frac{2w}{p+s}$.

Values of f are typically over 0.8 for freely occurring words, under 0.1 for fragments and strongly-attached affixes, and intermediate for clitics, some affixes, and words with restricted usage. However, most words haven't been seen enough times for f to be reliable. So a word is classified as a fragment if $p + s \geq 1000$ and $f \leq 0.2$. It is classified as a reliable word if $p + s \geq 50$ and $f \geq 0.5$.

To revise the input segmentation of the corpus, Mini-morph merges each fragment with an adjacent word if the newly-created merged word occurred at least 10 times in the input segmentation. When mergers with both adjacent words are possible, the algorithm alternates which to prefer. Each word is then sudivided into a sequence of reliable words, when possible. Because words are typically short and reliable words rare, a simple recursive algorithm is used, biased towards using shorter words. [7]

WordEnds calls Mini-morph twice, once to revise the preliminary segmentation produced by the bootstrapping phase and a second time to revise the final segmentation.

6 Test corpora

WordEnds was tested on a diverse set of seven corpora, summarized in Table 1. Notice that the Arabic dataset has much longer words than those used by previous authors. Subsets were extracted from the larger corpora, to control for training set size. Goldwater's algorithm, the best performing of previous methods, was also tested on the small versions. [8]

The first three corpora all use dictionary transcriptions with 1-character phone symbols. The Bernstein-Ratner (BR87) corpus was described above (Section 3.2). The Arabic corpus was created by removing punctuation and word boundaries from the Buckwalter version of the LDC's transcripts of Gulf Arabic Conversational Telephone Speech (Appen, 2006). Filled pauses and foreign words were kept as is. Word fragments were kept, but the telltale hyphens were removed. The Spanish corpus was produced in a similar way from the Callhome Spanish dataset (Wheatley, 1996), removing all accents. Orthographic forms were used for words without pronunciations (e.g. foreign, fragments)

The other two English dictionary transcriptions were produced in a similar way from the Buckeye corpus (Pitt et al., 2005; Pitt et al., 2007) and Mississippi State's corrected version of the LDC's Switchboard transcripts (Godfrey and Holliman, 1994; Deshmukh et al., 1998). These use a "readable phonetic" version of arpabet. Each phone is represented with a 1–2 character code, chosen to look like English orthography and to ensure that character sequences decode uniquely into phone sequences. Buckeye does not provide dictionary pronunciations for word fragments, so these were transcribed as "X". Switchboard was also transcribed using standard English orthography.

The Buckeye corpus also provides an accurate phonetic transcription of its data, showing allophonic variation (e.g. glottal stop, dental/nasal flaps), segment deletions, quality shifts/uncertainty, and nasalization. Some words are "massively" reduced (Johnson, 2003), going well beyond standard phonological rules. We represented its 64 phones using codes with 1–3 characters.

7 Test results

Table 2 presents test results for the small corpora. The numbers for the four English dictionary and orthographic transcriptions are very similar. This confirms the finding of Batchelder (2002) that variations in transcription method have only minor impacts on segmenter performance. Performance seems to be largely determined by structural and lexical properties (e.g. word length, pause frequency).

For the English dictionary datasets, the primary overall evaluation numbers (BF and WF) for the two algorithms differ less than the variation created by tweaking parameters or re-running Goldwater's (randomized) algorithm. Both degrade similarly on the phonetic version of Buckeye. The most visible overall difference is speed. WordEnds processes

[7] Subdivision is done only once for each word type.

[8] It is too slow to run on the larger ones.

		WordEnds					Goldwater				
corpus	transcription	BP	BR	BF	WF	LF	BP	BR	BF	WF	LF
BR87	dictionary	**94.6**	73.7	82.9	70.7	36.6	89.2	**82.7**	**85.8**	**72.5**	**56.2**
Switchboard	dictionary	**91.3**	80.5	**85.5**	**72.0**	**37.4**	73.9	**93.5**	82.6	65.8	27.8
Switchboard	orthographic	**90.0**	75.5	**82.1**	**66.3**	**33.7**	73.1	**92.4**	81.6	63.6	28.4
Buckeye	dictionary	**89.7**	82.2	**85.8**	**72.3**	**37.4**	74.6	**94.8**	83.5	68.1	26.7
Buckeye	phonetic	**71.0**	64.1	**67.4**	**44.1**	**28.6**	49.6	**95.0**	65.1	35.4	12.8
Arab	dictionary	**88.1**	68.5	**77.1**	**56.6**	**40.4**	47.5	**97.4**	63.8	32.6	9.5
Spanish	dictionary	**89.3**	48.5	62.9	38.7	16.6	69.2	**92.8**	**79.3**	**57.9**	**17.0**

Table 2: Results for WordEnds and Goldwater on the small test corpora. See Section 2.3 for definitions of metrics.

		medium w/out morph			medium			large		
corpus	transcription	BF	WF	LF	BF	WF	LF	BF	WF	LF
Switchboard	dictionary	90.4	78.8	39.4	93.0	84.8	44.2	94.7	88.1	44.3
Switchboard	orthographic	89.6	77.4	37.3	91.6	81.8	41.1	94.1	87.0	41.1
Buckeye	dictionary	91.2	80.3	41.5	93.7	86.1	47.8	–	–	–
Buckeye	phonetic	72.1	48.4	27.1	75.0	54.2	28.2	–	–	–
Arab	dictionary	85.7	69.1	49.5	86.4	70.6	50.0	–	–	–
Spanish	dictionary	75.1	52.2	19.7	76.3	55.0	20.2	–	–	–

Table 3: Results for WordEnds on the medium and large datasets, also on the medium dataset without Mini-morph. See Table 1 for dataset sizes.

each small dataset in around 30-40 seconds. Goldwater requires around 2000 times as long: 14.5-32 hours, depending on the dataset.

However, WordEnds keeps affixes on words whereas Goldwater's algorithm removes them. This creates a systematic difference in the balance between boundary recall and precision. It also causes Goldwater's LF values to drop dramatically between the child-directed BR87 corpus and the adult-directed speech. For the same reason, WordEnds maintains good performance on the Arabic dataset, but Goldwater's performance (especially LF) is much worse. It is quite likely that Goldwater's algorithm is finding morphemes rather than words.

Datasets around 30K words are traditional for this task. However, a child learner has access to much more data, e.g. Weijer (1999) measured 1890 words per hour spoken near an infant. WordEnds performs much better when more data is available (Table 3). Numbers for even the harder datasets (Buckeye phonetic, Spanish) are starting to look promising. The Spanish results show that data with infrequent pauses can be handled in two very different ways: aggressive model-based segmentation (Goldwater) or feeding more data to a more cautious segmenter (WordEnds).

The two calls to Mini-morph sometimes make almost no difference, e.g. on the Arabic data. But it can make large improvements, e.g. BF +6.9%, WF +10.5%, LF +5.8% on the BR corpus. Table 3 shows details for the medium datasets. Its contribution seems to diminish as the datasets get bigger, e.g. improvements of BF +4.7%, WF +9.3%, LF +3.7% on the small dictionary Switchboard corpus but only BF +1.3%, WF +3.3%, LF +3.4% on the large one.

8 Some specifics of performance

Examining specific mistakes confirms that WordEnds does not systematically remove affixes on English dictionary data. On the large Switchboard corpus, "-ed" is never removed from its stem and "-ing" is removed only 16 times. The Mini-morph post-processor misclassifies, and thus segments off, some affixes that are homophonous with free-standing words, such as "-en"/"in" and "-es"/"is". A smarter model of morphology and local syntax could probably avoid this.

There is a visible difference between English "the" and the Arabic determiner "Al-". The English determiner is almost always segmented off. From the medium-sized Switchboard corpus, only 434 lexical items are posited with "the" attached to a following word. Arabic "Al" is sometimes attached and sometimes segmented off. In the medium Arabic dataset, the correct and computed lexicons contain similar numbers of words starting with Al (4873 and 4608), but there is only partial overlap (2797 words). Some of this disagreement involves foreign language nouns, which the markup in the original corpus separates from the determiner.[9]

Mistakes on twenty specific items account for 24% of the errors on the large Switchboard corpus. The first two items, accounting for over 11% of the mistakes, involve splitting "uhhuh" and "umhum". Most of the rest involve merging common collocations (e.g. "a lot") or splitting common compounds that have a transparent analysis (e.g. "something").

9 Discussion and conclusions

Performance of WordEnds is much stronger than previous reported results, including good results on Arabic and promising results on accurate phonetic transcriptions. This is partly due to good algorithm design and partly due to using more training data. This sets a much higher standard for models of child language acquisition and also suggests that it is not crazy to speculate about inserting such an algorithm into the speech recognition pipeline.

Performance would probably be improved by better models of morphology and/or phonology. An ngram model of morpheme sequences (e.g. like Goldwater uses) might avoid some of the mistakes mentioned in Section 8. Feature-based or gestural phonology (Browman and Goldstein, 1992) might help model segmental variation. Finite-state models (Belz, 2000) might be more compact. Prosody, stress, and other sub-phonemic cues might disambiguate some problem situations (Hockema, 2006; Rytting, 2007; Salverda et al., 2003).

However, it is not obvious which of these approaches will actually improve performance. Additional phonetic features may not be easy to detect reliably, e.g. marking lexical stress in the presence of contrastive stress and utterance-final lengthening. The actual phonology of fast speech may not be quite what we expect, e.g. performance on the phonetic version of Buckeye was slightly **improved** by merging nasal flap with n, and dental flap with d and glottal stop. The sets of word initial and final segments may not form natural phonological classes, because they are partly determined by morphological and lexical constraints (Rytting, 2007).

Moreover, the strong performance from the basic segmental model makes it hard to rule out the possibility that high performance could be achieved, even on data with phonetic variation, by throwing enough training data at a simple segmental algorithm.

Finally, the role of child-directed speech needs to be examined more carefully. Child-directed speech displays helpful features such as shorter phrases and fewer reductions (Bernstein Ratner, 1996; van de Weijer, 1999). These features may make segmentation easier to learn, but the strong results presented here for adult-directed speech make it trickier to argue that this help is necessary for learning.

Moreover, it is not clear how learning to segment child-directed speech might make it easier to learn to segment speech directed at adults or older children. It's possible that learning child-directed speech makes it easier to learn the basic principles of phonology, semantics, or higher-level linguistic structure. This might somehow feed back into learning segmentation. However, it's also possible that its only raison d'être is social: enabling earlier communication between children and adults.

[9] The author does not read Arabic and, thus, is not in a position to explain why the annotaters did this.

Acknowledgments

Many thanks to the UIUC prosody group, Mitch Marcus, Cindy Fisher, and Sharon Goldwater.

References

Rie Kubota Ando and Lillian Lee. 2000. Mostly-Unsupervised Statistical Segmentation of Japanese. *Proc ANLP-NAACL* 2000:241–248.

Appen Pty Ltd. 2006. Gulf Arabic Conversational Telephone Speech, Transcripts Linguistic Data Consortium, Philadelphia

Eleanor Olds Batchelder 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. Cognition 83, pp. 167–206.

Anja Belz 2000. Multi-Syllable Phonotactic Modelling. 5th ACL SIGPHON, pp. 46–56.

Nan Bernstein Ratner. 1987. The phonology of parent child speech. In K. Nelson and A. Van Kleeck (Eds.), Children's Language: Vol 6, Lawrence Erlbaum.

Nan Bernstein Ratner 1996. From "Signal to Syntax": But what is the Nature of the Signal? In James Morgan and Katherine Demuth (eds) Signal to Syntax, Lawrence Erlbaum, Mahwah, NJ.

Michael R. Brent. 1999. An Efficient, Probabalistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning* 1999:71–105.

Michael R. Brent and Timothy A. Cartwright. 1996. Distributional Regularity and Phonotactic Constraints are Useful for Segmentation *Cognition* 1996:93–125.

C. P. Browman and L. Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49:155–180.

Paul Cairns, Richard Shillcock, Nick Chater, and Joe Levy. 1997. Bootstrapping Word Boundaries: A Bottom-up Corpus-based Approach to Speech Segmentation. *Cognitive Psychology*, 33:111–153.

Morten Christiansen and Joseph Allen 1997. Coping with Variation in Speech Segmentation GALA 1997.

Morten Christiansen, Joseph Allen, Mark Seidenberg. 1998. Learning to Segment Speech Using Multiple Cues: A Connectionist Model. Language and Cognitive Processes 12/2–3, pp. 221-268.

Herbert H. Clark and Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. Cognitive Psychology 37:201–242.

N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker and J. Picone. 1998. Resegmentation of Switchboard. Proc. Intern. Conf. on Spoken Language Processing:1543-1546.

Jean E. Fox Tree and Herbert H. Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. Cognition 62(2):151–167.

John J. Godfrey and Ed Holliman. 1993. Switchboard-1 Transcripts. Linguistic Data Consortium, Philadelphia, PA.

Sharon Goldwater. 2006. Nonparametric Bayesian Models of Lexical Acquisition. Ph.D. thesis, Brown Univ.

Jonathan Harrington, Gordon Watson, and Maggie Cooper. 1989. Word boundary detection in broad class and phoneme strings. Computer Speech and Language 3:367–382.

Jonathan Harrington, Gordon Watson, and Maggie Cooper. 1988. Word Boundary Identification from Phoneme Sequence Constraints in Automatic Continuous Speech Recognition. Coling 1988, pp. 225–230.

Stephen A. Hockema. 2006. Finding Words in Speech: An Investigation of American English. *Language Learning and Development*, 2(2):119-146.

Keith Johnson 2003. Massive reduction in conversational American English. Proc. of the Workshop on Spontaneous Speech: Data and Analysis.

Peter W. Jusczyk and Richard N. Aslin. 1995. Infants' Detection of the Sound Patterns of Words in Fluent Speech. Cognitive Psychology 29(1)1–23.

Lori F. Lamel and Victor W. Zue. 1984. Properties of Consonant Sequences within Words and Across Word Boundaries. *Proc. ICASSP* 1984:42.3.1–42.3.4.

James M. McQueen. 1998. Segmentation of Continuous Speech Using Phonotactics. Journal of Memory and Language 39:21–46.

Mark Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. Speech Communication, 45, 90-95.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond., E. Hume, and E. Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release) Department of Psychology, Ohio State University, Columbus, OH

C. Anton Rytting 2004. Greek Word Segmentation using Minimal Information. HLT-NAACL 2004, pp. 78–85.

C. Anton Rytting 2007. Preserving Subsegmental Variation in Modelling Word Segmentation. Ph.D. thesis, Ohio State, Columbus OH.

J. R. Saffran. 2001 Words in a sea of sounds: The output of statistical learning. Cognition 81:149-169.

Anne Pier Salverda, Delphine Dahan, and James M. McQueen. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90:51–89.

Amanda Seidl and Elizabeth K. Johnson. 2006. Infant Word Segmentation Revisited: Edge Alignment Facilitates Target Extraction. Developmental Science 9(6):565–573.

W. J. Teahan, Y. Wen, R. McNab, I. H. Witten 2000 A compression-based algorithm for Chinese word segmentation. Computational Linguistics 26/3, pp. 375–393.

Anand Venkataraman. 2001. A Statistical Model for Word Discovery in Transcribed Speech. *Computational Linguistics*, 27(3):351–372.

A. Weber. 2000 Phonotactic and acoustic cues for word segmentation. Proc. 6th Intern. Conf. on Spoken Language Processing, Vol. 3: 782-785. pp

Joost van de Weijer 1999. Language Input for Word Discovery. Ph.D. thesis, Katholieke Universiteit Nijmegen.

Barbara Wheatley. 1996. CALLHOME Spanish Transcripts. Linguistic Data Consortium, Philadelphia.