A joint model of word segmentation and phonological variation for English word-final /t/-deletion

Benjamin Börschinger[1,3] and **Mark Johnson**[1] and **Katherine Demuth**[2]

(1) Department of Computing, Macquarie University

(2) Department of Linguistics, Macquarie University

(3) Department of Computational Linguistics, Heidelberg University

{benjamin.borschinger, mark.johnson, katherine.demuth}@mq.edu.au

Abstract

Word-final /t/-deletion refers to a common phenomenon in spoken English where words such as /*wɛst*/ "west" are pronounced as [*wɛs*] "wes" in certain contexts. Phonological variation like this is common in naturally occurring speech. Current computational models of unsupervised word segmentation usually assume idealized input that is devoid of these kinds of variation. We extend a non-parametric model of word segmentation by adding phonological rules that map from underlying forms to surface forms to produce a mathematically well-defined joint model as a first step towards handling variation and segmentation in a single model. We analyse how our model handles /t/-deletion on a large corpus of transcribed speech, and show that the joint model can perform word segmentation and recover underlying /t/s. We find that Bigram dependencies are important for performing well on real data and for learning appropriate deletion probabilities for different contexts.[1]

1 Introduction

Computational models of word segmentation try to solve one of the first problems language learners have to face: breaking an unsegmented stream of sound segments into individual words. Currently, most such models assume that the input consists of sequences of phonemes with no pronunciation variation across different occurrences of the same word type. In this paper we describe an extension of the Bayesian models of Goldwater et al. (2009) that incorporates phonological rules to "explain away" surface variation. As a concrete example, we focus on word-final /t/-deletion in English, although our approach is not limited to this case. We choose /t/-deletion because it is a very common and well-studied phenomenon (see Coetzee (2004, Chapter 5) for a review) and segmental deletion is an interesting test-case for our architecture. Recent work has found that /t/-deletion (among other things) is indeed common in child-directed speech (CDS) and, importantly, that its distribution is similar to that in adult-directed speech (ADS) (Dilley et al., to appear). This justifies our using ADS to evaluate our model, as discussed below.

Our experiments are consistent with longstanding and recent findings in linguistics, in particular that /t/-deletion heavily depends on the immediate context and that models ignoring context work poorly on real data. We also examine how well our models identify the probability of /t/-deletion in different contexts. We find that models that capture bigram dependencies between underlying forms provide considerably more accurate estimates of those probabilities than corresponding unigram or "bag of words" models of underlying forms.

In section 2 we discuss related work on handling variation in computational models and on /t/-deletion. Section 3 describes our computational model and section 4 discusses its performance for recovering deleted /t/s. We look at both a situation where word boundaries are pre-specified and only inference for underlying forms has to be performed; and the problem of jointly finding the word boundaries and recovering deleted underlying /t/s. Section 5 discusses our findings, and section 6 concludes with directions for further research.

[1] The implementation of our model as well as scripts to prepare the data will be made available at http://web.science.mq.edu.au/~bborschi. We can't release our version of the Buckeye Corpus (Pitt et al., 2007) directly because of licensing issues.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1508–1516,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

2 Background and related work

The work of Elsner et al. (2012) is most closely related to our goal of building a model that handles variation. They propose a pipe-line architecture involving two separate generative models, one for word-segmentation and one for phonological variation. They model the mapping to surface forms using a probabilistic finite-state transducer. This allows their architecture to handle virtually arbitrary pronunciation variation. However, as they point out, combining the segmentation and the variation model into one joint model is not straight-forward and usual inference procedures are infeasible, which requires the use of several heuristics. We pursue an alternative research strategy here, starting with a single well-studied example of phonological variation. This permits us to develop a joint generative model for both word segmentation and variation which we plan to extend to handle more phenomena in future work.

An earlier work that is close to the spirit of our approach is Naradowsky and Goldwater (2009), who learn spelling rules jointly with a simple stem-suffix model of English verb morphology. Their model, however, doesn't naturally extend to the segmentation of entire utterances.

/t/-deletion has received a lot of attention within linguistics, and we point the interested reader to Coetzee (2004, Chapter 5) for a thorough review. Briefly, the phenomenon is as follows: word-final instances of /t/ may undergo deletion in natural speech, such that */wɛst/* "west" is actually pronounced as *[wɛs]* "wes".[2] While the frequency of this phenomenon varies across social and dialectal groups, within groups it has been found to be robust, and the probability of deletion depends on its phonological context: a /t/ is more likely to be dropped when followed by a consonant than a vowel or a pause, and it is more likely to be dropped when following a consonant than a vowel as well. We point out two recent publications that are of direct relevance to our research. Dilley et al. (to appear) study word-final variation in stop consonants in CDS, the kind of input we ideally would like to evaluate our models on. They find that "infants largely experience statistical distributions of non-canonical consonantal pronunciation variants [including deletion] that mirror those experienced by adults." This both directly establishes the need for computational models to handle this dimension of variation, and justifies our choice of using ADS for evaluation, as mentioned above.

Coetzee and Kawahara (2013) provide a computational study of (among other things) /t/-deletion within the framework of Harmonic Grammar. They do not aim for a joint model that also handles word segmentation, however, and rather than training their model on an actual corpus, they evaluate on constructed lists of examples, mimicking frequencies of real data. Overall, our findings agree with theirs, in particular that capturing the probability of deletion in different contexts does not automatically result in good performance for recovering individual deleted /t/s. We will come back to this point in our discussion at the end of the paper.

3 The computational model

Our models build on the Unigram and the Bigram model introduced in Goldwater et al. (2009). Figure 1 shows the graphical model for our joint Bigram model (the Unigram case is trivially recovered by generating the $U_{i,j}$s directly from L rather than from $L_{U_{i,j-1}}$). Figure 2 gives the mathematical description of the graphical model and Table 1 provides a key to the variables of our model.

The model generates a latent sequence of *underlying word-tokens* $U_1, \ldots, U_n$. Each word token is itself a non-empty sequence of segments or phonemes, and each U_j corresponds to an underlying word form, prior to the application of any phonological rule. This generative process is repeated for each utterance i, leading to multiple utterances of the form $U_{i,1}, \ldots, U_{i,n_i}$ where n_i is the number of words in the i^{th} utterance, and $U_{i,j}$ is the j^{th} word in the i^{th} utterance. Each utterance is padded by an observed utterance boundary symbol $\$$ to the left and to the right, hence $U_{i,0} = U_{i,n_i+1} = \$$.[3] Each $U_{i,j+1}$ is generated conditionally on its predecessor $U_{i,j}$ from $L_{U_{i,j}}$, as shown in the first row of the lower plate in Figure 1. Each L_w is a distribution over the possible words that can follow a token of w and L is a global distribution over possible words, used as back-off for all L_w. Just as in Goldwater et al. (2009), L is drawn from a Dirichlet Process (DP) with base distribution B and concentration

[2] Following the convention in phonology, we give underlying forms within "/.../" and surface forms within "[...]".

[3] Each utterance terminates as soon as a $\$$ is generated, thus determining the number of words n_i in the i^{th} utterance. See Goldwater et al. (2009) for discussion.

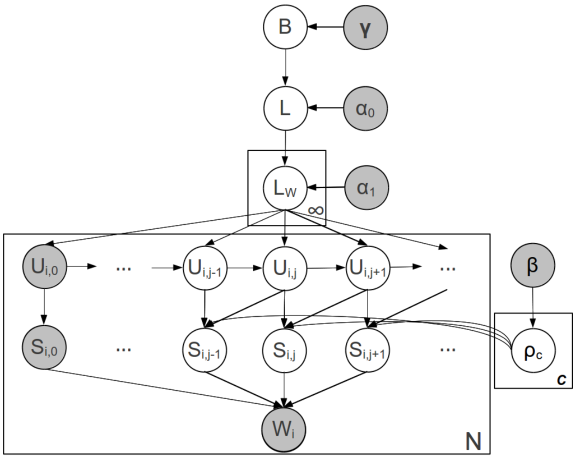

Figure 1: The graphical model for our joint model of word-final /t/-deletion and Bigram word segmentation. The corresponding mathematical description is given in Figure 2. The generative process mimics the intuitively plausible idea of generating underlying forms from some kind of syntactic model (here, a Bigram language model) and then mapping the underlying form to an observed surface-form through the application of a phonological rule component, here represented by the collection of rule probabilities ρ_c.

$$
\begin{aligned}
L \mid \boldsymbol{\gamma}, \alpha_0 &\sim DP(\alpha_0, B(\cdot \mid \boldsymbol{\gamma})) \\
L_w \mid L, \alpha_1 &\sim DP(\alpha_1, L) \\
\rho_c \mid \beta &\sim Beta(1,1) \\
U_{i,0} &= \$ \\
S_{i,0} &= \$ \\
U_{i,j+1} \mid U_{i,j}, L_{U_{i,j}} &\sim L_{U_{i,j}} \\
S_{i,j} \mid U_{i,j}, U_{i,j+1}, \boldsymbol{\rho} &= P_R(\cdot \mid U_{i,j}, U_{i,j+1}) \\
W_i \mid S_{i,1}, \ldots, S_{i,n_i} &= \textsc{Cat}(S_{i,0}, \ldots, S_{i,n_i})
\end{aligned}
$$

Figure 2: Mathematical description of our joint Bigram model. The lexical generator $B(\cdot \mid \boldsymbol{\gamma})$ is specified in Figure 3 and P_R is explained in the text below. CAT stands for concatenation without word-boundaries, n_i refers to the number of words in utterance i.

Variable	Explanation
B	base distribution over possible words
L	back-off distribution over words
L_w	distribution over words following w
$U_{i,j}$	underlying form, a word
$S_{i,j}$	surface realization of $U_{i,j}$, a word
ρ_c	/t/-deletion probability in context c
W_i	observed segments for i^{th} utterance

Table 1: Key for the variables in Figure 1 and Figure 2. See Figure 3 for the definition of B.

parameter α_0, and the word type specific distributions L_w are drawn from a $DP(L, \alpha_1)$, resulting in a hierarchical DP model (Teh et al., 2006). The base distribution B functions as a lexical generator, defining a prior distribution over possible words. In principle, B can incorporate arbitrary prior knowledge about possible words, for example syllable structure (cf. Johnson (2008)). Inspired by Norris et al. (1997), we use a simpler possible word constraint that only rules out sequences that lack a vowel (see Figure 3). While this is clearly a simplification it is a plausible assumption for English data.

Instead of generating the observed sequence of segments W directly by concatenating the underlying forms as in Goldwater et al. (2009), we map each $U_{i,j}$ to a corresponding surface-form $S_{i,j}$ by a probabilistic rule component P_R. The values over which the $S_{i,j}$ range are determined by the available phonological processes. In the model we study here, the phonological processes only include a rule for deleting word-final /t/s but in principle, P_R can be used to encode a wide variety of phonological rules. Here, $S_{i,j} \in \{U_{i,j}, \textsc{Delf}(U_{i,j})\}$ if $U_{i,j}$ ends in a /t/, and $S_{i,j} = U_{i,j}$ otherwise, where $\textsc{Delf}(u)$ refers to the same word as u except that it lacks u's final segment.

We look at three kinds of contexts on which a rule's probability of applying depends:

1. a *uniform* context that applies to every word-final position

2. a *right* context that also considers the following segment

3. a *left-right* context that additionally takes the preceeding segment into account

For each possible context c there is a probability ρ_c which stands for the probability of the rule applying in this context. Writing

$$\gamma \sim Dir(\langle 0.01, \ldots, 0.01 \rangle)$$

$$B(w = x_{1:n} \mid \gamma) = \begin{cases} \frac{[\prod_{i=1}^{n} \gamma_{x_i}]\gamma_{\#}}{Z} & \text{if } \mathrm{V}(w) \\ 0.0 & \text{if } \neg\mathrm{V}(w) \end{cases}$$

Figure 3: Lexical generator with possible word-constraint for words in Σ^+, Σ being the alphabet of available phonemes. $x_{1:n}$ is a sequence of elements of Σ of length n. γ is a probability vector of length $|\Sigma| + 1$ drawn from a sparse Dirichlet prior, giving the probability for each phoneme and the special word-boundary symbol $\#$. The predicate V holds of all sequences containing at least one vowel. Z is a normalization constant that adjusts for the mass assigned to the empty and non-possible words.

contexts in the notation familiar from generative phonology (Chomsky and Halle, 1968), our model can be seen as implementing the following rules under the different assumptions:[4]

uniform /t/ → ∅ / $\underline{\quad}]_{\text{word}}$

right /t/ → ∅ / $\underline{\quad}]_{\text{word}}\ \beta$

left-right /t/ → ∅ / $\alpha\ \underline{\quad}]_{\text{word}}\ \beta$

We let β range over V(owel), C(onsonant) and \$ (utterance-boundary), and α over V and C. We define a function CONT that maps a pair of adjacent underlying forms $U_{i,j}, U_{i,j+1}$ to the context of the final segment of $U_{i,j}$. For example, CONT(*/wɛst/*,*/əv/*) returns "C $\underline{\quad}]_{\text{word}}$ V" in the *left-right* setting, or simply "$\underline{\quad}]_{\text{word}}$" in the *uniform* setting. CONT returns a special NOT context if $U_{i,j}$ doesn't end in a /t/. We stipulate that $\rho_{\text{NOT}} = 0.0$. Then we can define P_R as follows:

$$P_R(\text{DELFINAL}(u) \mid u, r)) = \rho_{\text{CONT}(u,r)}$$
$$P_R(u \mid u, r) = 1 - \rho_{\text{CONT}(u,r)}$$

Depending on the context setting used, our model includes one (*uniform*), three (*right*) or six (*left-right*) /t/-deletion probabilities ρ_c. We place a uniform Beta prior on each of those so as to learn their values in the LEARN-ρ experiments below.

Finally, the observed unsegmented utterances W_i are generated by concatenating all $S_{i,j}$ using the function CAT.

We briefly comment on the central intuition of this model, i.e. why it can infer underlying from surface forms. Bayesian word segmentation models try to compactly represent the observed data in terms of a small set of units (word types) and a short analysis (a small number of word tokens). Phonological rules such as /t/-deletion can "explain away" an observed surface type such as *[wɛs]*] in terms of the underlying type */wɛst/* which is independently needed for surface tokens of *[wɛst]*. Thus, the */t/*→ ∅ rule makes possible a smaller lexicon for a given number of surface tokens. Obviously, human learners have access to additional cues, such as the meaning of words, knowledge of phonological similarity between segments and so forth. One of the advantages of an explicitly defined generative model such as ours is that it is straight-forward to gradually extend it by adding more cues, as we point out in the discussion.

3.1 Inference

Just as for the Goldwater et al. (2009) segmentation models, exact inference is infeasible for our joint model. We extend the collapsed Gibbs breakpoint-sampler described in Goldwater et al. (2009) to perform inference for our extended models. We refer the reader to their paper for additional details such as how to calculate the Bigram probabilities in Figure 4. Here we focus on the required changes to the sampler so as to perform inference under our richer model. We consider the case of a single surface string W, so we drop the i-index in the following discussion.

Knowing W, the problem is to recover the underlying forms $U_1, \ldots, U_n$ and the surface forms $S_1, \ldots, S_n$ for unknown n. A major insight in Goldwater's work is that rather than sampling over the latent variables in the model directly (the number of which we don't even know), we can instead perform Gibbs sampling over a set of boundary variables $b_1, \ldots, b_{|W|-1}$ that jointly determine the values for our variables of interest where $|W|$ is the length of the surface string W. For our model, each $b_j \in \{0, 1, t\}$, where $b_j = 0$ indicates absence of a word boundary, $b_j = 1$ indicates presence of a boundary and $b_j = t$ indicates presence of a boundary with a preceeding underlying */t/*. The relation between the b_j and the $S_1, \ldots, S_n$ and $U_1, \ldots, U_n$ is illustrated in Figure 5. The required sampling equations are given in Figure 4.

[4] For *right* there are three and for *left-right* six different rules, one for every instantiation of the context-template.

$$
\begin{aligned}
P(b_j = 0 \mid \boldsymbol{b^{-j}}) &\propto P(w_{12,u} \mid w_{l,u}, \boldsymbol{b^{-j}}) \times P_r(w_{12,s} \mid w_{12,u}, w_{r,u}) \times P(w_{r,u} \mid w_{12,u}, \boldsymbol{b^{-j}} \oplus \langle w_{l,u}, w_{12,u} \rangle) && (1) \\
P(b_j = t \mid \boldsymbol{b^{-j}}) &\propto P(w_{1,t} \mid w_{l,u}, \boldsymbol{b^{-j}}) \times P_r(w_{1,s} \mid w_{1,t}, w_{2,u}) \times P(w_{2,u} \mid w_{1,t}, \boldsymbol{b^{-j}} \oplus \langle w_{l,u}, w_{1,t} \rangle) \\
&\quad \times P_r(w_{2,s} \mid w_{2,u}, w_{r,u}) \times P(w_{r,u} \mid w_{2,u}, \boldsymbol{b^{-j}} \oplus \langle w_{l,u}, w_{1,t} \rangle \oplus \langle w_{1,t}, w_{2,u} \rangle) && (2) \\
P(b_j = 1 \mid \boldsymbol{b^{-j}}) &\propto P(w_{1,s} \mid w_{l,u}, \boldsymbol{b^{-j}}) \times P_r(w_{1,s} \mid w_{1,s}, w_{2,u}) \times P(w_{2,u} \mid w_{1,s}, \boldsymbol{b^{-j}} \oplus \langle w_{l,u}, w_{1,s} \rangle) \\
&\quad \times P_r(w_{2,s} \mid w_{2,u}, w_{r,u}) \times P(w_{r,u} \mid w_{2,u}, \boldsymbol{b^{-j}} \oplus \langle w_{l,u}, w_{1,s} \rangle \oplus \langle w_{1,s}, w_{2,u} \rangle) && (3)
\end{aligned}
$$

Figure 4: Sampling equations for our Gibbs sampler, see figure 5 for illustration. $b_j = 0$ corresponds to no boundary at this position, $b_j = t$ to a boundary with a preceeding underlying /t/ and $b_j = 1$ to a boundary with no additional underlying /t/. We use $\boldsymbol{b^{-j}}$ for the statistics determined by all but the j^{th} position and $\boldsymbol{b^{-j}} \oplus \langle r, l \rangle$ for these statistics plus an additional count of the bigram $\langle r, l \rangle$. $P(w \mid l, \boldsymbol{b})$ refers to the bigram probability of $\langle l, w \rangle$ given the the statistics $\boldsymbol{b}$; we refer the reader to Goldwater et al. (2009) for the details of calculating these bigram probabilities and details about the required statistics for the collapsed sampler. P_R is defined in the text.

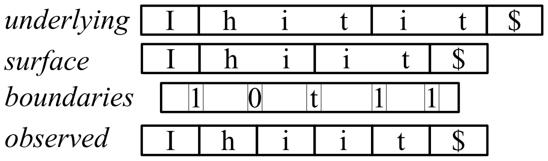

Figure 5: The relation between the observed sequence of segments (bottom), the boundary variables $b_1, \ldots, b_{|W|-1}$ the Gibbs sampler operates over (in squares), the latent sequence of surface forms and the latent sequence of underlying forms. When sampling a new value for $b_3 = t$, the different word-variables in figure 4 are: $w_{12,u}$=$w_{12,s}$=*hiit*, $w_{1,t}$=*hit* and $w_{1,s}$=*hi*, $w_{2,u}$=$w_{2,s}$=*it*, $w_{l,u}$=*I*, $w_{r,u}$=*$*. Note that we need a boundary variable at the end of the utterance as there might be an *underlying* /t/ at this position as well. The final boundary variable is set to 1, not *t*, because the /t/ in *it* is observed.

orthographic	I don’t intend to
transcript	/*aɪ ɾ oʊ n ɪ n t ɛ n d ə*/
idealized	/*aɪ d oʊ n t ɪ n t ɛ n d t ʊ*/
t-drop	/*aɪ d oʊ n ɪ n t ɛ n d t ʊ*/

Figure 6: An example fragment from the Buckeye-corpus in orthographic form, the fine transcript available in the Buckeye corpus, a fully idealized pronunciation with canonical dictionary pronunciations and our version of the data with dropped /t/s.

4 Experiments

4.1 The data

We are interested in how well our model handles /t/-deletion in real data. Ideally, we'd evaluate it on CDS but as of now, we know of no available large enough corpus of accurately hand-transcribed CDS. Instead, we used the Buckeye Corpus (Pitt et al., 2007) for our experiments, a large ADS corpus of interviews with English speakers that have been transcribed with relatively fine phonetic detail, with /t/-deletion among the things manually annotated. Pointing to the recent work by Dilley et al. (to appear) we want to emphasize that the statistical distribution of /t/-deletion has been found to be similar for ADS and CDS, at least for read speech.

We automatically derived a corpus of 285,792 word tokens across 48,795 utterances from the Buckeye Corpus by collecting utterances across all interviews and heuristically splitting utterances at speaker-turn changes and indicated silences. The Buckeye corpus lists for each word token a manually transcribed pronunciation in context as well as its canonical pronunciation as given in a pronouncing dictionary. As input to our model, we use the canonical pronunciation unless the pronunciation in context indicates that the final /t/ has been deleted in which case we also delete the final /t/ of the canonical pronunciation Figure 6 shows an example from the Buckeye Corpus, indicating how the original data, a fully idealized version and our derived input that takes into account /t/-deletions looks like.

Overall, /t/-deletion is a quite frequent phenomenon with roughly 29% of all underlying /t/s being dropped. The probabilities become more peaked when looking at finer context; see Table 3 for the empirical distribution of /t/-dropping for the six different contexts of the *left-right* setting.

4.2 Recovering deleted /t/s, given word boundaries

In this set of experiments we are interested in how well our model recovers /t/s when it is provided with the gold word boundaries. This allows us to investigate the strength of the statistical signal for the deletion rule without confounding it with the word segmentation performance, and to see how the different contextual settings *uniform*, *right* and *left-right* handle the data. Concretely, for the example in Figure 6 this means that we tell the model that there are boundaries between /aɪ/, /doʊn/, /ɪntɛnd/, /tu/ and /liv/ but we don't tell it whether or not these words end in an underlying /t/. Even in this simple example, there are 5 possible positions for the model to posit an underlying /t/. We evaluate the model in terms of F-score, the harmonic mean of recall (the fraction of underlying /t/s the model correctly recovered) and precision (the fraction of underlying /t/s the model predicted that were correct).

In these experiments, we ran a total of 2500 iterations with a burnin of 2000. We collect samples with a lag of 10 for the last 500 iterations and perform *maximum marginal decoding* over these samples (Johnson and Goldwater, 2009), as well as running two chains so as to get an idea of the variance.[5]

We are also interested in how well the model can infer the rule probabilities from the data, that is, whether it can learn values for the different ρ_c parameters. We compare two settings, one where we perform inference for these parameters assuming a uniform Beta prior on each ρ_c (LEARN-ρ) and one where we provide the model with the empirical probabilities for each ρ_c as estimated off the gold-data (GOLD-ρ), e.g., for the *uniform* condition 0.29. The results are shown in Table 2.

Best performance for both the Unigram and the Bigram model in the GOLD-ρ condition is achieved under the *left-right* setting, in line with the standard analyses of /t/-deletion as primarily being determined by the preceding and the following context. For the LEARN-ρ condition, the Bigram model still performs best in the *left-right* setting but the Unigram model's performance drops in all settings and is now worst in the *left-right* and best in the *uniform* setting.

		uniform	*right*	*left-right*
Unigram	LEARN-ρ	56.52	39.28	23.59
	GOLD-ρ	62.08	60.80	66.15
Bigram	LEARN-ρ	60.85	62.98	77.76
	GOLD-ρ	69.06	69.98	73.45

Table 2: F-score of recovered /t/s with known word boundaries on real data for the three different context settings, averaged over two runs (all standard errors below 2%). Note how the Unigram model always suffers in the LEARN-ρ condition whereas the Bigram model's performance is actually best for LEARN-ρ in the *left-right* setting.

	C_C	C_V	C_$	V_C	V_V	V_$
empirical	0.62	0.42	0.36	0.23	0.15	0.07
Unigram	0.41	0.33	0.17	0.07	0.05	0.00
Bigram	0.70	0.58	0.43	0.17	0.13	0.06

Table 3: Inferred rule-probabilities for different contexts in the left-right setting from one of the runs. "C_C" stands for the context where the deleted /t/ is preceded and followed by a consonant, "V_$" stands for the context where it is preceded by a vowel and followed by the utterance boundary. Note how the Unigram model severely under-estimates and the Bigram model slightly over-estimates the probabilities.

In fact, comparing the inferred probabilities to the "ground truth" indicates that the Bigram model estimates the true probabilities more accurately than the Unigram model, as illustrated in Table 3 for the *left-right* setting. The Bigram model somewhat overestimates the probability for all post-consonantal contexts but the Unigram model severely underestimates the probability of /t/-deletion across all contexts.

4.3 Artificial data experiments

To test our Gibbs sampling inference procedure, we ran it on artificial data generated according to the model itself. If our inference procedure fails to recover the underlying /t/s accurately in this setting, we should not expect it to work well on actual data. We generated our artificial data as follows. We transformed the sequence of canonical pronunciations in the Buckeye corpus (which we take to be underlying forms here) by randomly deleting final /t/s using empirical probabilities as shown in Table 3 to generate a sequence of artificial surface forms that serve as input to our models. We

[5] As manually setting the hyper-parameters for the DPs in our model proved to be complicated and may be objected to on principled grounds, we perform inference for them under a vague gamma prior, as suggested by Teh et al. (2006) and Johnson and Goldwater (2009), using our own implementation of a slice-sampler (Neal, 2003).

		uniform	right	left-right
Unigram	LEARN-ρ	94.35	23.55 (+)	63.06
	GOLD-ρ	94.45	94.20	91.83
Bigram	LEARN-ρ	92.72	91.64	88.48
	GOLD-ρ	92.88	92.33	89.32

Table 4: F-score of /t/-recovery with known word boundaries on artificial data, each condition tested on data that corresponds to the assumption, averaged over two runs (standard errors less than 2% except (+) = 3.68%)).

	Unigram	Bigram
LEARN-ρ	33.58	55.64
GOLD-ρ	55.92	57.62

Table 5: /t/-recovery F-scores when performing joint word segmention in the *left-right* setting, averaged over two runs (standard errors less than 2%). See Table 6 for the corresponding segmentation F-scores.

did this for all three context settings, always estimating the deletion probability for each context from the gold-standard. The results of these experiments are given in table 4. Interestingly, performance on these artificial data is considerably better than on the real data. In particular the Bigram model is able to get consistently high F-scores for both the LEARN-ρ and the GOLD-ρ setting. For the Unigram model, we again observe the severe drop in the LEARN-ρ setting for the *right* and *left-right* settings although it does remarkably well in the *uniform* setting, and performs well across all settings in the GOLD-ρ condition. We take this to show that our inference algorithm is in fact working as expected.

4.4 Segmentation experiments

Finally, we are also interested to learn how well we can do word segmentation and underlying /t/-recovery jointly. Again, we look at both the LEARN-ρ and GOLD-ρ conditions but focus on the *left-right* setting as this worked best in the experiments above. For these experiments, we perform simulated annealing throughout the initial 2000 iterations, gradually cooling the temperature from 5 to 1, following the observation by Goldwater et al. (2009) that without annealing, the Bigram model gets stuck in sub-optimal parts of the solution space early on. During the annealing stage, we prevent the model from performing inference for underlying /t/s so that the annealing stage can be seen as an elaborate initialisation scheme, and we perform joint inference for the remaining 500 iterations, evaluating on the last sample and averaging over two runs. As neither the Unigram nor the Bigram model performs "perfect" word segmentation, we expect to see a degradation in /t/-recovery performance and this is what we find indeed. To give an impression of the impact of /t/-deletion, we also report numbers for running only the segmentation model on the Buckeye data with no deleted /t/s and on the data with deleted /t/s. The /t/-recovery scores are given in Table 5 and segmentation scores in Table 6. Again the Unigram model's /t/-recovery score degrades dramatically in the LEARN-ρ condition. Looking at the segmentation performance this isn't too surprising: the Unigram model's poorer token F-score, the standard measure of segmentation performance on a word token level, suggests that it misses many more boundaries than the Bigram model to begin with and, consequently, can't recover any potential underlying /t/s at these boundaries. Also note that in the GOLD-ρ condition, our joint Bigram model performs almost as well on data with /t/-deletions as the word segmentation model on data that includes no variation at all.

The generally worse performance of handling variation as measured by /t/-recovery F-score when performing joint segmentation is consistent with the finding of Elsner et al. (2012) who report considerable performance drops for their phonological learner when working with induced boundaries (note, however, that their model does not perform joint inference, rather the induced boundaries are given to their phonological learner as ground-truth).

5 Discussion

There are two interesting findings from our experiments. First of all, we find a much larger difference between the Unigram and the Bigram model in the LEARN-ρ condition than in the GOLD-ρ condition. We suggest that this is due to the Unigram model's lack of dependencies between underlying forms, depriving it of an important source of evidence. Bigram dependencies provide additional evidence for underlying /t/ that are deleted on the surface, and because the Bigram model identifies these underlying /t/ more accurately, it can also estimate the /t/ deletion probability more accurately.

	Unigram	Bigram
LEARN-ρ	54.53	72.55 (2.3%)
GOLD-ρ	54.51	73.18
NO-ρ	54.61	70.12
NO-VAR	54.12	73.99

Table 6: Word segmentation F-scores for the /t/-recovery F-scores in Table 5 averaged over two runs (standard errors less than 2% unless given). NO-ρ are scores for running just the word segmentation model with no /t/-deletion rule on the data that includes /t/-deletion, NO-VAR for running just the word segmentation model on the data with no /t/-deletions.

For example, /t/ dropping in "don't you" yields surface forms "don you". Because the word bigram probability $P(\text{you} \mid \text{don't})$ is high, the bigram model prefers to analyse surface "don" as underlying "don't". The Unigram model does not have access to word bigram information so the underlying forms it posits are less accurate (as shown in Table 2), and hence the estimate of the /t/-deletion probability is also less accurate. When the probabilities of deletion are pre-specified the Unigram model performs better but still considerably worse than the Bigram model when the word boundaries are known, suggesting the importance of non-phonological contextual effects that the Bigram model but not the Unigram model can capture. This suggests that for example word predictability in context might be an important factor contributing to /t/-deletion.

The other striking finding is the considerable drop in performance between running on naturalistic and artificially created data. This suggests that the natural distribution of /t/-deletion is much more complex than can be captured by statistics over the phonological contexts we examined. Following Guy (1991), a finer-grained distinction for the preceeding segments might address this problem.

Yet another suggestion comes from the recent work in Coetzee and Kawahara (2013) who claim that "[a] model that accounts perfectly for the overall rate of application of some variable process therefore does not necessarily account very well for the actual application of the process to individual words." They argue that in particular the extremely high deletion rates typical of high frequency items aren't accurately captured when the deletion probability is estimated across all types. A look at the error patterns of our model on a sample from the Bigram model in the LEARN-ρ setting on the naturalistic data suggests that this is in fact a problem. For example, the word "just" has an extremely high rate of deletion with $\frac{1746}{2442} = 0.71\%$. While many tokens of "jus" are "explained away" through predicting underlying /t/s, the (literally) extra-ordinary frequency of "jus"-tokens lets our model still posit it as an underlying form, although with a much dampened frequency (of the 1746 surface tokens, 1081 are analysed as being realizations of an underlying "just").

The /t/-recovery performance drop when performing joint word segmentation isn't surprising as even the Bigram model doesn't deliver a very high-quality segmentation to begin with, leading to both sparsity (through missed word-boundaries) and potential noise (through misplaced word-boundaries). Using a more realistic generative process for the underlying forms, for example an Adaptor Grammar (Johnson et al., 2007), could address this shortcoming in future work without changing the overall architecture of the model although novel inference algorithms might be required.

6 Conclusion and outlook

We presented a joint model for word segmentation and the learning of phonological rule probabilities from a corpus of transcribed speech. We find that our Bigram model reaches 77% /t/-recovery F-score when run with knowledge of true word-boundaries and when it can make use of both the preceeding and the following phonological context, and that unlike the Unigram model it is able to learn the probability of /t/-deletion in different contexts. When performing joint word segmentation on the Buckeye corpus, our Bigram model reaches around above 55% F-score for recovering deleted /t/s with a word segmentation F-score of around 72% which is 2% better than running a Bigram model that does not model /t/-deletion.

We identified additional factors that might help handling /t/-deletion and similar phenomena. A major advantage of our generative model is the ease and transparency with which its assumptions can be modified and extended. For future work we plan to incorporate into our model richer phonological contexts, item- and frequency-specific probabilities and more direct use of word predictability. We also plan to extend our model to handle additional phenomena, an obvious candidate being /d/-deletion.

Also, the two-level architecture we present is not limited to the mapping being defined in terms of rules rather than constraints in the spirit of Optimality Theory (Prince and Smolensky, 2004); we plan to explore this alternative path as well in future work.

To conclude, we presented a model that provides a clean framework to test the usefulness of different factors for word segmentation and handling phonological variation in a controlled manner.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This research was supported under Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593).

References

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Haper & Row, New York.

Andries W. Coetzee and Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguisic Theory*, 31:47–89.

Andries W. Coetzee. 2004. *What it Means to be a Loser: Non-Optimal Candidates in Optimality Theory*. Ph.D. thesis, University of Massachusetts , Amherst.

Laura Dilley, Amanda Millett, J. Devin McAuley, and Tonya R. Bergeson. to appear. Phonetic variation in consonants in infant-directed and adult-directed speech: The case of regressive place assimilation in word-final alveolar stops. *Journal of Child Language*.

Micha Elsner, Sharon Goldwater, and Jacob Eisenstein. 2012. Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Jeju Island, Korea. Association for Computational Linguistics.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Gregory R. Guy. 1991. Contextual conditioning in variable lexical phonology. *Language Variation and Change*, 3(2):223–39.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, June. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Mark Johnson. 2008. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Jason Naradowsky and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1531–1536, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31:705–767.

Dennis Norris, James M. Mcqueen, Anne Cutler, and Sally Butterfield. 1997. The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34(3):191 – 243.

Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye corpus of conversational speech.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.