Structure Learning in Hidden Conditional Random Fields for Grapheme-to-Phoneme Conversion

Patrick Lehnen\textsuperscript{1}, Alexandre Allauzen\textsuperscript{2}, Thomas Lavergne\textsuperscript{2}, François Yvon\textsuperscript{2}, Stefan Hahn\textsuperscript{1}, Hermann Ney\textsuperscript{1,2}

\textsuperscript{1}Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Aachen, Germany
\textsuperscript{2}Univ. Paris-Sud, France and LIMSI/CNRS - Spoken Language Processing group
{lehnen,hahn,ney}@cs.rwth-aachen.de \{allauzen,lavergne,yvon\}@limsi.fr

Abstract

Accurate grapheme-to-phoneme (g2p) conversion is needed for several speech processing applications, such as automatic speech synthesis and recognition. For some languages, notably English, improvements of g2p systems are very slow, due to the intricacy of the associations between letter and sounds. In recent years, several improvements have been obtained either by using variable-length associations in generative models (joint-n-grams), or by recasting the problem as a conventional sequence labeling task, enabling to integrate rich dependencies in discriminative models. In this paper, we consider several ways to reconcile these two approaches. Introducing hidden variable-length alignments through latent variables, our Hidden Conditional Random Field (HCRF) models are able to produce comparative performance compared to strong generative and discriminative models on the CELEX database.

\textbf{Index Terms}: grapheme-to-phoneme conversion, G2P, HCRF, discriminative models, hidden conditional random fields

1. Introduction

In recent years, Conditional Random Fields [1] (CRFs) have been successfully applied to several language processing applications that can be formulated as sequence labelling tasks, such as part-of-speech (POS) tagging [1], chunking [2], speech recognition [3] and language modeling [4], to name a few. Grapheme to phoneme conversion (g2p) does not so easily lend itself to CRFs, since the training data not always contain the alignment information between individual graphemic and phonemic symbols. The usual approach is to automatically compute these alignments prior to training for instance using the BIO [5] labeling scheme. However, this solution requires the alignments to be provided or computed by an external knowledge source; furthermore, the choice of a specific labeling scheme introduces an undesirable bias in the training data. As a result, g2p is artificially expressed as a sequential letter classification task and fails to capture the variable-length nature of the linguistic grapheme. This issue is less of a problem for generative probabilistic models, which can model hidden alignment through latent variables [6, 7], at the price of a much less restrictive set of predictive features. A natural way to get the best of both worlds is to consider Hidden CRFs [8, 9, 10, 11], which can take hidden variables into account and which was shown, in [12], to deliver state-of-the-art performances at the expense of a high computational cost. This approach remains unsatisfactory, since the g2p mappings it captures are limited, and the features it uses only consider single letters or phonemes as unit.

We introduce a novel approach inspired by the phrase-based framework used in machine translation. This new HCRF model handles arbitrary mapping between graphemic and phonemic substrings since they are observed on the training data. The variable-length units is directly integrated in the model, enabling us to capture more complex dependencies. Moreover, by considering a pre-computed conversion table, the computational cost is drastically reduced.

In this paper, we compare performances achieved on the CELEX database [13] by these different kind of HCRF to a strong discriminative baseline [14] and the generative joint multigram model [6, 7]. We also describe how these approaches differ in terms of speed and the search space they explore. This paper is organised as follows: section 2 provides a short description of the HCRF approach, while section 3 proposes three implementations of this kind of models. Then features that are used by the different models are introduced in section 4 followed by the presentation of the experimental set-up and results in section 5.

2. Hidden Conditional Random Fields

Hidden Conditional Random Fields (HCRFs) [8, 9, 10, 11] estimates the conditional probability of a phoneme sequence \( y \in \mathbb{Y}^n \) given the observed grapheme sequence \( x \in \mathbb{X}^n \) by considering a set of latent variable \( s \) that represents in our case the segmentation of \( x \) and \( y \):

\[
p(y|x) = \sum_{s \in \mathbb{S}} p(y,s|x) = \sum_{s \in \mathbb{S}} \sum_{k \in \mathbb{K}} \exp H(y,s,x) \tag{1}
\]

The summation over the segmentations \( s \) is restricted by a set \( \mathbb{S} \) defined by the actual implementation of HCRFs. This publication will include three implementations of \( \mathbb{S} \) which are described in Sec. 3. The hypotheses are ranked with the help of a general feature description \( H(y,s,x) = \lambda^t h(y,s,x) \) composed of binary features \( h(y,s,x) \in \{0,1\} \) with their respective weights \( \lambda \). Estimation of the parameters \( \lambda \) is by maximization of the conditional log-likelihood \( L \) over the training corpus \( \{y_k,x_k\}_{k=1}^K \) taking into account Elastic-Net parameter priors:

\[
L(\lambda) = \sum_{k=1}^K \log p(y_k|x_k) - c_1 ||\lambda||_1 - \frac{1}{2} c_2 ||\lambda||_2^2.
\]
In this paper, we introduce a new model (System B) described in section 3.2 and inspired by the phrase-based approach in machine translation that allows arbitrary segmentation of both the grapheme and phoneme sequences, as long as they are observed in the training set.

### 3. Three ways to cope with hidden structure

As introduced in equation 1, the hidden variables $S$ defines the search space explored by the model, i.e. the possible segmentations of the grapheme and phoneme sequences and their associations. Without any restriction, this problem is untractable. However, word internal structure and its associated pronunciation suggests that this search space can be safely restricted in some way, hence allowing exact computation of the gradient and inference. Thus, in this paper we consider three different ways to restrict the search space.

As a baseline approach (see [15, 12] for further details), an external tool provides the alignment of the training examples that are used to recast the problem as a sequence labelling task using the so-called BIO scheme [5]. With this assumption, the model consists of a standard linear chain CRF (LC-CRF). Its main drawback is that it cannot handle the case where the phoneme sequence is longer than the grapheme sequence. The second approach (System A), described in [12] and in section 3.1, proposes an extension of the BIO scheme where, $S$ is by construction restricted such that any segment of graphemes can be associated to a segment of at most two phonemes. Without using an external tool to provide alignments, this restriction yields state of the art performances.

In this paper, we introduce a new model (System B) described in section 3.2 and inspired by the phrase-based approach in machine translation that allows arbitrary segmentation of both the grapheme and phoneme sequences, as long as they are observed in the training set.

#### 3.1. Description of HCRF (System A)

System A (introduced in [12]) uses phonemes annotated with special segmentation labels inspired by the BIO scheme [5] organized as in Hidden Markov Models (HMMs). Leaving the modeling of the phrase information to phrase-like features (Sec. 4 for more details). First, the grapheme sequence is represented as a chain of symbols in a finite state transducer (Fig. 1(a)). Each arc is duplicated for each possible grapheme symbol and weighted with the HCRF features (Sec. 4).

As shown in equation 1, the hidden variables $S$ defines the search space explored by the model, i.e. the possible segmentations of the grapheme and phoneme sequences and their associations. Without any restriction, this problem is untractable. However, word internal structure and its associated pronunciation suggests that this search space can be safely restricted in some way, hence allowing exact computation of the gradient and inference. Thus, in this paper we consider three different ways to restrict the search space.

As a baseline approach (see [15, 12] for further details), an external tool provides the alignment of the training examples that are used to recast the problem as a sequence labelling task using the so-called BIO scheme [5]. With this assumption, the model consists of a standard linear chain CRF (LC-CRF). Its main drawbacks is that it cannot handle the case where the phoneme sequence is longer than the grapheme sequence. The second approach (System A), described in [12] and in section 3.1, proposes an extension of the BIO scheme where, $S$ is by construction restricted such that any segment of graphemes can be associated to a segment of at most two phonemes. Without using an external tool to provide alignments, this restriction yields state of the art performances.

In this paper, we introduce a new model (System B) described in section 3.2 and inspired by the phrase-based approach in machine translation that allows arbitrary segmentation of both the grapheme and phoneme sequences, as long as they are observed in the training set.
all remaining features are applied. Known from HMMs the arcs indication continuation/beginning/doubling are weighed with a penalty $\delta_0/\delta_1/\delta_2$ (designated as HCRF features).

3.2. Description of HCRF (System B)

This system is inspired by the phrase based machine translation framework. In a preliminary step, the training corpus is aligned, e.g. with GIZA++ [17], to build the conversion table\(^3\) that consists in a set of segments of graphemes and their associated conversion in phoneme segments. It is worth noticing that the external alignment is only used to derive this conversion table and never used by the following steps.

For inference and decoding purpose, the System B is implemented as a finite-state transducer cascade including the following steps: the grapheme sequence is represented as an acceptor, that is composed with a segmentation transducer to consider all the segmentations; the conversion table is then applied by composition to generate the full search space that contains all the possible conversions of the input grapheme sequence; finally, the HCRF model is used to score this search space. For inference, the Viterbi hypothesis can be computed by finding the shortest path in this search space. Moreover, two additional steps can be carried on: the determinization that computes the summation of equation 2; and a composition with a n-gram model of phoneme sequence. This architecture is closely related to the proposal of [18].

4. Features

The features supporting the conditional probability $b(y, s, x) \in \{0, 1\}$ are a critical choice for CRF systems. Up to the authors knowledge, in [14] the currently best result on Celex was published. [14] use only surface features taking only the used letters on source side and the phonemes on target side into account. The authors decided to construct their systems based on the same kind of features. Basically three sets of features are used:

- **source-n-gram** Features depending only on one target symbol $y_n$ and a combination of source symbols $x_{s_n}^{s_n + \gamma_1}$ relative to the currently aligned source word (with $\gamma_1 \leq \gamma_2$).

- **target-n-gram** Features only describing the relation of a consecutive set of target symbols $y_{s_n}^{s_n + \gamma_1}$ including the current target symbol $y_n$.

- **joint-n-grams** Combinations of source-n-grams and target-n-grams.

System A models the search space directly via the source and target symbols. The choice of the parameters was $\gamma_1, \gamma_2 = -5, \ldots, 5, \gamma_1 + \gamma_2 + 1 \leq 6$, and $\delta \leq 3$. System A does not use joint-n-grams.

System B models the search space by tuples of source and target symbols, which is commonly known as phrases. Thus the smallest unit for a features is the source part of a phrase and the target part of a phrase. Using the source and target part of the phrase in the description of source-n-grams, target-n-grams and joint-n-grams as $x_m$ and $y_n$, the parameters are $\gamma_1, \gamma_2 = -1, \ldots, 1, \delta \leq 1$ and the joint-n-grams are all combinations of the source- and target-n-grams. In average the size of a source phrase with System B is 1.84 source symbols, letting the source features span over 5.53 source symbols in average. Which is about the same size as in System A.

5. Experimental Results

To evaluate the different approaches described in this paper, experimental results are reported using the English Celex corpus [19]. This data set is divided in three parts according to the previously published results of [20, 14] and it contains 39 995 training samples, 15 000 words for test and 5 000 word that are used as development set. The output vocabulary is made of 53 symbols (phonemes). The results are reported in terms of both the Levenshtein based phoneme error rate (PER) and the word error rate (WER).

5.1. Build-Up of System A

For System A the regularization parameters $c_1$ and $c_2$ were optimized with respect to the error rate on the development set to $c_1 = c_2 = \frac{1}{50}$, and a save number of iterations was estimated as 75. With 75 iterations all metrics (PER, WER, number of features, conditional log-likelihood) were converged. To avoid local optima the features with respect to the currently aligned source word $(y_n, x_{s_n})$ were initialized with IBM-1 probabilities $\lambda = -\log(p(e|f))$. It turned out that long contexts on source and target are needed to gain good performance. The result in line 4 of Tab. 1 is poor and improves greatly with source-n-grams (n up to 6) in line 5 of Tab. 1 and are further improved with target-n-grams in line 6 and 7 of Tab. 1. In earlier experiments the need for $\delta_0/\delta_1/\delta_2$ penalties was verified, as without the penalties, e.g. the performance of line 6 of Tab. 1 is reduced to a ten times larger PER ($\approx 25\%$).

Eq. 2 suggest the use of the summation over all possible segmentations in search to be symmetric to the training conditions. Actually the full summation is computationally infeasible as it involves the determination over a complex automaton, permitting to generate from one target symbol to two times as many target symbols as in the source sentence. We generate a $k$-best list with $k = 400$ and performed determination within the $k$-best through summing up all equal hypotheses. As is shown in line 8 of Tab. 1 the determination does not change the system performance significantly. As most of the time $p(y, s|x) \geq 50\%$ for the first best in the $k$-best list, only a small amount of the probability mass could change the recognition result.

5.2. Build-Up of System B

To build the System B, the first step aims to create the conversion table that consists in a set of segment of graphemes and their associated conversion in phoneme segments. This step is carried on by running GIZA++ with the standard setup to estimate the alignments between grapheme and phoneme sequences in both direction. Then the grow-diag-final-and heuristic [21] is applied to extract the conversion table. For all the results published in this article, the size of the segments on both sides (grapheme and phoneme) are limited to 3. This value was tuned experimentally on the development set.

For the decoder used during the inference and training steps, most operations directly use functionalities of the OpenFst library [22], except for the Forward-Backward algorithm and the interactions with the HCRF model, which rely on a in-house implementation\(^3\). The inference defined by Eq. 2 can be implemented with the determinization operation. However,
given the size of the search space, this operation can be very
time consuming. For efficiency purposes, we use an approxi-
mate determination that first extracts the k-best hypotheses,
with k = 400. During the inference step, a n-gram model of
phoneme sequence can be applied by composition after the
determination. The n-gram models are estimated with Kneser-
Ney discounting [23] using all the training set.

To restrict the size of the conversion table, the segment pairs
extracted from the alignment are limited to a size of 3. With
this restriction the references cannot be always produced given
the conversion table. A practical solution, used in many stud-
ies, is to resort to oracle references, corresponding to the best
reachable solution w.r.t to a given metric, the BLEU [24] in
the following experiments.

5.3. Influence of the latent variable
A popular way of modeling segmentations is to separate
the estimation of the segmentation and estimation of the target se-
quency. As a baseline the approach based on generative joint-
n-grams [20] estimated a segmentation on the training corpus
respecting the target sequence reference \( \text{ref} \). Based on the ref-
ence and the segmentation a LCCRF \( p(y, s | x) \) is estimated.
During search the generative joint-n-grams propose a hypothe-
sis including a best segmentation and best target sequence. The
best target sequence is omitted and only the best segmentation
is used to duplicate source symbols to provide slots for the es-
timation of target symbols in which the LCCRF is used to find
an optimal target labeling.

The resulting LCCRF (line 3 of Tab. 1) uses exactly the
same features as the HCRF in line 7 of Tab. 1 except the \( \delta_0, \delta_1, \delta_2, \delta_3 \) penalties, with a significant degradation in performance with
respect to PER and WER. Actually it is advantageous to model
the segmentation within the CRF. Moreover, we can observe that
the performances of System B are significantly worst than the
others. One difference with the others is the feature set that
uses segment as units instead of a single grapheme or phoneme.
The other difference is the way the search space is built that
makes System B drastically faster.

5.4. Comparison of the search spaces
For the sake of comparison, we provide statistics that character-
ize the search spaces explored by the System A and System B
two examples. For the word *aback*, System A considers a
search space of 13274 nodes and 207356 arcs that corresponds to
1.74 \( \times 10^{17} \) pathes. For the same word, System B only
explores 36686 pathes with 93 nodes and 1042 arcs. For the word
*bent*, System A consider a search space of 10336 nodes
and 155936 arcs that corresponds to 6.22 \( \times 10^{13} \) pathes,
while the corresponding search space explored by System B contains
7914 pathes made of 92 nodes and 1016 arcs. The huge reduc-
tion of the search space explains why System B is more than a
thousand times faster than System A for training and inference.

It is worth noticing that, while System A consider all the
possible segmentations of the grapheme sequence, System B
drastically reduces the number of segmentation by allowing a
limited segment length and by considering only the segments
observed in the training data. For instance, in these experi-
ments, System B uses an inventory of 5184 grapheme segments and
in average 408 different segmentations on the test set for
an average word length of 8.3 graphemes. Moreover, while the
System A tends to select for the test set 1-to-1 alignments (in
80% of the case) and 2-to-1 alignments (in 17% of the case), the
repartition for the System B differs: 36% of 1-to-1, 29% of

Table 1: Results on the Celex corpus, Line 1 and 2 provide baseline
results where [20] is the best found generative approach, and [14] the best found discriminative approach on this task.
Line 3 provides a result for a system leaving the modulation of
the segmentation to a the model [20] and using the same fea-
tures as in 7 except segmentation specific features (Sec. 5.3).
The next two blocks describe the build-up of the System A and
System B including their best result (Sec. 5.4).

<table>
<thead>
<tr>
<th></th>
<th>PER[%]</th>
<th>WER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Eva</td>
</tr>
<tr>
<td>1</td>
<td>[14]</td>
<td>10.8</td>
</tr>
<tr>
<td>2</td>
<td>joint n-grams [20]</td>
<td>11.4</td>
</tr>
<tr>
<td>3</td>
<td>LCCRF</td>
<td>2.8</td>
</tr>
<tr>
<td>4</td>
<td>((c_{a_j}, f_j) + (\delta_j))</td>
<td>52.5</td>
</tr>
<tr>
<td>5</td>
<td>(\alpha) + source n-grams</td>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
<td>(\beta) + target-2-grams</td>
<td>2.6</td>
</tr>
<tr>
<td>7</td>
<td>(\gamma) + target-3-grams</td>
<td>2.6</td>
</tr>
<tr>
<td>8</td>
<td>+ determination</td>
<td>2.6</td>
</tr>
<tr>
<td>9</td>
<td>(\delta_0)</td>
<td>3.2</td>
</tr>
<tr>
<td>10</td>
<td>+ determination</td>
<td>3.1</td>
</tr>
<tr>
<td>11</td>
<td>(\delta_2) + 4-gram LM</td>
<td>3.1</td>
</tr>
</tbody>
</table>

2-to-2, 17% of 2-to-1, 10% of 3-to-2, 5% of 3-to-3 and 1% of
2-to-3 (the others can be neglected).

To assess, whether this restriction of the search space may
explain the decrease in performance, the oracle hypothesis
is estimated for the test set as explained in section 5.2. The oracle
hypothesis exhibits a PER of 0.3% and a WER of 1.3%. These
results show that the restriction of search space is efficient since
the search space contains in average very competitive hypothe-
sis. Unfortunately, the feature set defined at the segment level
seems to be insufficient since the model is not able to select
among the search space such relevant hypothesis.

6. Conclusion
In this paper we compare three ways to define Hidden Condi-
tional Random Fields that can express a wide range of mapping
between grapheme and phoneme sequences. At the expense of a
degradation in performances, we introduce a model inspired
by the phrase-based machine translation framework that drasti-
cally reduces the computational cost by using an efficient way
to prune the search space. Experimental evidence tends to show
that the poor performances of this system is due to the feature
design and not to the strategy used to prune the search space.
In future work, we plan to overcome this issue by a tailored feature
engineering.

7. Acknowledgements
This work was partly realized as part of the Quaero Programme,
funded by OSEO, French State agency for innovation. H. Ney
was partially supported by a senior chair award from DIGITEO,
A French research cluster in Ile-de-France.
8. References


