Phrasal Cohort Based Unsupervised Discriminative Language Modeling

Puyang Xu and Sanjeev Khudanpur

Center for Language and Speech Processing,
Johns Hopkins University, USA
{puyangxu,khudanpur}@jhu.edu

Brian Roark

Center for Spoken Language Understanding
Oregon Health and Sciences University, USA
roark@cslu.ogi.edu

Abstract

Simulated confusions enable the use of large text-only corpora for discriminative language modeling by hallucinating the likely recognition outputs that each (correct) sentence would be confused with. In [1], a novel approach was introduced to simulate confusions using phrasal cohorts derived directly from recognition output. However, the described approach relied on transcribed speech to derive cohorts. In this paper, we extend the phrasal cohort technique to the fully unsupervised scenario, where transcribed data are completely absent. Experimental results show that even if the cohorts are extracted from *untranscribed* speech, the unsupervised training can still achieve over 40% of the gains of the supervised approach. The results are presented on NIST data sets for a state-of-the-art LVCSR system.

Index Terms: unsupervised training, discriminative language modeling

1. Introduction

A language model (LM) assigns probabilities to word sequences, and is an important component of automatic speech recognition (ASR) systems. Standard LMs are usually parameterized as generative n-gram models. Training of such models requires a text corpus and estimating a large set of conditional n-gram distributions with the objective of maximizing the likelihood of the corpus. Because of the simple estimation procedure and the abundance of text resources, such n-gram models often can be trained at very large scales.

Despite the predominant use, generative LMs are not trained to optimize the system performance measured by word error rate (WER). Discriminative LMs (DLM), which directly target the errors that the system produces, have been shown to be able to achieve consistent gains over well-estimated generative n-gram models [2, 3, 4, 5]. Unlike generative LMs, training DLMs usually requires more than just clean text. In order to identify the set of acoustically confusable hypotheses that the recognizer outputs, we have to decode the speech corresponding to the text. Once we obtain the set of competitors that the reference text is confused with, a global linear model can be trained to maximize the likelihood ratio of the lowest-error hypothesis to its set of competitors.

Unfortunately, transcribed speech data are usually expensive to acquire. The amount of training data available for DLMs can sometimes be quite limited for a lot of tasks. In contrast, it is often much easier to obtain a large text corpus. In order to apply discriminative language modeling in the absence of sufficient transcribed data, a lot of work has focused on training DLMs from only text resources. Typically, such approaches require modeling the ASR errors and simulating the confusion set for discriminative training. The weighted finite state transducer (FST) provides a flexible framework for this purpose. It allows us to simulate different levels of ASR confusions successively. Previous methods usually approach this problem by first modeling the phonetic similarities [6, 7, 8, 9]. The transduction cost from one phone to another can be estimated from the acoustic model distance or phone recognition errors.

Recently in [1], the authors presented extensive experimental results comparing a number of confusion generation techniques, including a new method that completely bypasses modeling the phone-level confusions, which they called the phrasal cohort method. Compared with phone-based techniques, the cohort approach is more straightforward and has a simpler procedure of generating the confusion set. Empirically, they found that among the three techniques they investigated, the phrasal cohort method yields the largest WER gains, recovering over half of the gains of the fully supervised methods.

This paper is intended as a detailed study of the phrasal cohort method and as a direct extension to [1], addressing two important issues not covered in that paper: First, the authors of [1] assumed a set of transcribed speech data, from which phrasal cohorts can be extracted. In such a semi-supervised scenario where some transcribed data are available, the benefits of unsupervised DLM are best evaluated on top of the supervised DLM trained on the available transcribed resources. Instead of using the cross-validation setup presented in [1], this paper will describe a new experimental setup in which such results can be presented; Second, transcribed resources are not always attainable. As an example, if we move to a new domain, it is unlikely that we have sufficient data to train supervised DLMs or extract phrasal cohorts. In this paper, we will extend the phrasal cohort approach to the fully *unsupervised* setting, where the phrase-level errors are estimated from *untranscribed* speech. The proposed method is also a direct extension of the word cohort method described in [10]. We find that even in the absence of transcribed data, the phrasal cohort method can achieve about 40% of the WER reduction of the supervised method. Such a fully unsupervised approach is more promising than assuming a transcribed corpus for learning the ASR errors. Because the transcribed data may as well be used to build a supervised DLM on top of which additional unsupervised training often leads to no gains.

We want to point out that regardless of whether the cohorts are derived from transcribed or untranscribed data, we consider the DLMs to be unsupervised. Because they are estimated from hallucinated confusion sets as opposed to real confusion sets in the supervised case.

The rest of the paper is organized as follows: We describe in section 2 the extraction of phrasal cohorts and how to use them to train DLMs unsupervisedly. The baseline ASR system and

the data setup are described in section 3. Empirical results are presented in section 4, followed by future work and conclusions in section 5 and 6.

2. The Phrasal Cohort Approach

2.1. Extraction of phrasal cohorts

The cohort for a phrase p, as denoted by $C(p)$, is defined as the set of similar-sounding phrases that are potential erroneous outputs where the correct transcription is p. Table 1 contains a few examples of phrasal cohorts.

Table 1: *Example of cohort*

p	$C(p)$
that to you	that too you
<s> she	<s> and she, <s> oh she
to you	to be you

Such phrase-level transformation rules can be easily extracted from data. In cases where transcribed speech is available, we can align the reference string with each one of the ASR hypotheses according to the edit distance. The number of hypotheses to align can be a hyper-parameter to tune. Every time we find a position where the two strings differ, we search for the left and right pivots, namely the nearest position where the two strings agree. The phrases between the pivots give us a transformation rule from the reference to the corrupted ASR output. The extraction then continues from the next word to the right pivot. Note that in the examples we show in the table, pivots are also part of the phrases, thus they can be thought of as context dependent rules.

In the fully *unsupervised* setting, we do not have access to the reference string, phrasal cohorts can be extracted from each one of the pair-wise alignments between the ASR hypotheses. Below is an example of two hypotheses of the same utterance.

1-best: <s> *what* kind of a company is it </s>
2-best: <s> *well* kind of a company is it </s>

Following the same procedure described above, we can identify the competing phrases "<s> *what* kind" and "<s> *well* kind". Since we do not know which phrase is correct, we extract the transformation rules symmetrically. Specifically, we have the transformation rule "<s> *what* kind —> <s> *well* kind", as well as "<s> *well* kind —> <s> *what* kind".

Obviously, such symmetric modeling of the ASR errors is much coarser. But constrained by the pivots, the two phrases usually reflect some kind of confusions that exist in the ASR system. It is therefore not unreasonable to expect one phrase in the ASR output when the other is the correct transcript.

2.2. Hallucination of N-best List

Phrasal cohorts are essentially a set of transformation rules. Whether they are derived from transcribed or untranscribed speech, they can be used in the same way to hallucinate n-best lists from text. As shown in Figure 1, given the reference string, we first construct a confusion network using the phrasal cohorts in Table 1. An LM is often used to score each path in the network, the top n hypotheses are then extracted and used as the simulated confusion set for discriminative training. Note that as described in [1], the confusion network can also be weighted by

Figure 1: *Construction of confusion network*

the probability of each transformation rule, namely the probability of the correct phrase getting recognized to its erroneous version. However, we have found that the use of such probabilities does not have much impact. From untranscribed data, where correctness of the phrase can not be determined, it is also not straightforward to estimate such probabilities. Therefore, in our experiments, we only use an LM to compose with the confusion network, and the rule probabilities are not used for the n-best extraction.

2.3. Local N-best Training

Hallucinating n-best lists using phrasal cohorts requires constructing confusion networks and extracting best-scoring paths from them using an LM. This process usually involves FST compositions and can be very time-consuming. To allow for faster experiments and more detailed analysis of the approach, we propose a much more efficient training strategy that does not require the global n-best extraction.

Instead of building a confusion network, every time a phrase p with non-empty $C(p)$ appears in the training text, we construct a set of competitors by putting each phrase in $C(p)$ in place of p with the rest of the sentence unchanged. In the example of Figure 1, when we see the phrase "<s> she", we construct the below training instance, which can be thought of as a *local* n-best list.

Ref: <s> she didn't announce that to you </s>
1-best: <s> and she didn't announce that to you </s>
2-best: <s> oh she didn't announce that to you </s>

Note that in such local n-best lists, the cohort phrases are not weighted by LMs or any other scores. However, in practice, we have found that this approach generally gives more or less the same results compared with the standard training strategy, despite having a much simpler training pipeline. In this work, we rely on this strategy to efficiently tune the various hyper-parameters in our approach (e.g. length of phrases, depth of n-best, order of features, ...). The final results are however, still based on the global n-best extraction.

3. Data & Setup

Our experiments are conducted on the English conversational telephone speech (CTS) dataset. We use the same state-of-the-art baseline system as the one described in [1]. The acoustic models are trained on 2000 hours of transcribed data. Please refer to [1] for a detailed description of the acoustic models. The baseline LM is a 4-gram LM trained on the transcript of the 2000 hours and about 1 billion words of web text.

For training DLMs, the 2000 hours of transcribed speech are divided into 20 folds. The 100-best list for each fold is produced using an LM trained on all the LM text excluding the transcript of that fold. Such cross-decoding strategy is commonly used for discriminative language modeling [5]. The hope is to remove the LM bias and that the confusions on each fold would potentially resemble those on the test data.

For training unsupervised DLMs, we use the transcript of 8 folds, with a total of 10 million words as the training text from which we hallucinate confusion sets. Their corresponding speech and the real n-best hypotheses produced by the decoder are set aside and never used. We also reserve another 8 folds of data as the speech from which we extract phrasal cohorts. The 8 folds of speech amount to 690 hours and are used as either transcribed or untranscribed data. The goal of our experiments is to investigate the benefit of unsupervised discriminative language modeling under various conditions of data availability (We will investigate the conditions of 2, 4 and 8 folds for both text and speech). Our results are reported on the NIST RT04 Fall development and evaluation sets, consisting of 38K and 37K words respectively.

4. Experimental results

4.1. Cohorts from Transcribed Speech

In this setting, we have access to the reference transcript of the 690 hours of speech. The phrasal cohorts are obtained by aligning the reference string with the 1-best hypothesis. Aligning with more hypotheses in the n-best list does not seem to help.

We also tune the maximal length of the phrase on the *correct* side of the rule using the local training strategy. We find that with the pivots included, length of more than 3 usually does not bring further improvement. In other words, it is sufficient to model only the insertion errors (length 2) and the errors where one word gets falsely recognized as an arbitrary number of words(length 3).

Keeping the pivots in the phrases proves to be advantageous. Such context dependency not only results in larger WER reduction, but also makes the training faster. Without the pivots, we will have many transformation rules where the correct side is a single word, and the number of training instances with the local training strategy will become much larger.

After tuning the various properties of the phrasal cohorts, we hallucinate the 100-best lists on the 8 folds of text using the baseline LM. To train the DLMs, we use the standard averaged perceptron algorithm with unigram, bigram and trigram features. Going up to 4-gram features does not help much. After each epoch of perceptron training, we compute the WER on the development set, and the training stops when the WER does not improve for 5 consecutive epochs. Note that the baseline score (including the AM and the baseline LM score) is only combined with the perceptron score at test time, and the mixing factors are optimized on the development data.

Table 2 shows the WER results on the development set with various amounts of LM text and transcribed speech. For the unsupervised DLMs, the amount of LM text does not appear to have much impact, the best WER can be obtained using the smallest LM corpus (3M) in our experiments. On the other hand, increasing the amount of speech beyond a certain point does not seem to help either. Our best WER can be obtained with as little as 133 hours of speech.

Table 2: *WER results with cohorts derived from transcribed speech. 1-best WER:* **22.8**

Text\Speech	17h	35h	70h	133h	337h	693h
3M	22.5	22.4	22.3	**22.2**	22.4	22.3
5M	22.5	22.4	22.3	22.2	22.3	22.4
10M	22.5	22.4	22.4	22.3	22.3	22.2
Sup DLM	22.5	22.5	22.2	22.1	21.8	**21.7**

The bottom row of the Table 2 shows the WER of supervised DLMs trained on the transcribed data. As we can see, assuming such a set of transcribed speech, we can train a supervised DLM that gives us 1.1% absolute WER reduction (21.7%). If we train a DLM unsupervisedly on the hallucinated data using phrasal cohorts derived from the set of transcribed speech, we can achieve 0.6% absolute WER reduction, recovering more than half of the gains achieved by the supervised approach. Such improvement also holds on the evaluation set, the results on which are presented in Table 5 at the end of section 4.2, where we summarize all the results. It is also worth mentioning that in low-resource conditions with only 17h or 35h of transcribed speech, the unsupervised DLMs performs as well as the supervised DLMs.

As we discussed in the introduction, a more interesting question to ask in this semi-supervised scenario, is whether the improvement from unsupervised training can add to the gains of supervised training on the transcribed resources. Unfortunately the answer is no. As demonstrated in Table 3, unsupervised training brings no extra WER reduction. It also appears to have a tendency to degrade the model performance as we add more hallucinated data. In other words, if we have a set of transcribed data where we can train supervised DLMs, it is probably not necessary to extract phrasal cohorts from it and carry out additional unsupervised training.

Table 3: *Additional "No-benefit" of Unsupervised training. 1-best WER: 22.8*

Transcribed speech	133h	337h	693h
Sup DLM	22.1	21.8	21.7
Sup + 3M Unsup	22.1	21.9	21.7
Sup + 5M Unsup	22.2	22.0	21.8
Sup + 10M Unsup	22.2	22.1	21.8

4.2. Cohorts from Untranscribed Speech

With the negative results presented in Table 3, we want to emphasize that the real strength of the phrasal cohort approach is in the fully unsupervised scenario. In this section, we will present such results.

In this setting, we do not have access to the reference transcript of the 690 hours of speech. As we described in section 2, we can extract cohorts symmetrically, considering all of the pair-wise alignments between the top-k ASR hypotheses.

As in the previous set of experiments, the choices of various hyper-parameters (length of phrases, k, ...) are found using the fast local training strategy. The optimized set of phrasal cohorts is then used for hallucinating the 100-best list from text. The maximal length of phrases on both side of the rule is chosen to be 4, with pivot words also as part of the phrases. The cohorts are derived from the top-5 hypotheses. We use the same perceptron algorithm with up to trigram features to train the DLMs.

Table 4 demonstrates the WER results on the development set. In this fully unsupervised setting, we consistently get 0.4% absolute WER reduction across almost all data conditions. With as little as 17 hours of speech, we in fact achieve 0.5% WER improvement. This is only very slightly lower than the 0.6% improvement shown in Table 2, where the cohorts are extracted from transcribed data. It is also worth pointing out that with 17 hours of speech, the completely unsupervised approach actually

Table 4: *WER results with cohorts derived from untranscribed speech. 1-best WER:* **22.8**

Text\Speech	17h	35h	70h	133h	337h	693h
3M	22.4	22.4	22.4	22.4	22.4	22.4
5M	22.4	22.5	22.4	22.4	22.4	22.4
10M	**22.3**	22.4	22.4	22.4	22.4	22.4

outperforms the standard supervised DLM by 0.2% (22.3% vs. 22.5%). However, the benefit of the proposed method does not seem to scale, the WER improvement seems to quickly saturate with the minimal amount of data.

Another interesting observation we had for this set of experiments is that the global n-best extraction using an LM consistently produces 0.1%-0.2% better WERs than the local training strategy, of which the WER numbers are not shown here. This was not the case when the cohorts are derived from transcribed data. The reason for such distinction is very likely to be the symmetric way that we extract cohorts from untranscribed data. Considering all pair-wise alignments in the absence of reference strings can produce cohort sets with significant level of noise. The local n-best training considers all of them while the LM-derived global n-best list is a much smaller and cleaner set of hypotheses.

To further compare supervised and unsupervised DLMs, and also compare cohorts derived from transcribed and untranscribed data, we apply the tuned model in each category (supervised DLM on the transcribed 690h; unsupervised DLM with cohorts derived from the transcribed 690h; unsupervised DLM with cohorts derived from the untranscribed 690h) on the evaluation set. The WER results are shown in Table 5. On the

Table 5: *Summary*

	Dev WER	Eval WER
ASR 1best	22.8	25.7
Sup DLM	21.7	24.8
Uns DLM; Trans cohorts	22.2	25.2
Uns DLM; Untrans cohorts	**22.3**	**25.3**

evaluation set, we see similar numbers of WER reduction as on the development set. First, we are able to corroborate the results presented in [1]: The unsupervised DLM trained with cohorts derived from transcribed speech can achieve over half of the WER reduction of supervised DLM (third row). Second, as the main focus of this paper, we want to highlight that in the complete absence of transcribed data, as indicated by the bottom row, the phrasal cohort approach can still achieve over 40% of the gains of supervised DLM. The 0.4% absolute WER improvement over the 1-best on the evaluation set is also statistically significant ($p < 0.001$).

5. Future Work

The scalability of unsupervised approaches is always an interesting problem. For the purpose of extensive comparisons with supervised approaches, our experiments only use up to 690 hours of speech and up to 10 million words of text. However, there is no reason not to use more. The amounts of text and untranscribed speech are significantly larger than transcribed speech, it would be very interesting to see whether the proposed fully unsupervised training can achieve more gains if we use 1000 hours of speech and 100 million words of text. For experiments of that scale to be feasible, we will have to prune the set of phrasal cohorts, which are usually very large because of the symmetric definition. We plan to report such results in future publications.

6. Conclusion

In this paper, we present a detailed study of unsupervised discriminative language modeling using the phrasal cohort technique described in [1]. We extend the approach to the fully unsupervised scenario where cohorts have to be extracted from *untranscribed* speech. We show with extensive experiments that even without transcribed data to derive phrase-level confusions, we are still able to achieve over 40% of the gains of the supervised approach with the proposed technique.

7. Acknowledgements

The authors would like to thank the entire "confusion modeling" team of the 2011 JHU CLSP summer workshop for creating the various resources used by this work. This research was partially supported by the National Science Foundation Grant No 0963898 and Grant No 0964102, and also by the JHU Human Language Technology Center of Excellence.

8. References

[1] Sagae, K., Lehr, M., Prud'hommeaux, E., Xu, P., Glenn, N., Karakos, D., Khudanpur, S., Roark, B., Saraclar, M., Shafran, I., Bikel, D., Callison-Burch, C., Cao, Y., Hall, K., Hasler, E., Koehn, P., Lopez, A., Post, M. and Riley, D., "Hallucinated n-best lists for discriminative language modeling", in Proc. ICASSP, 2012.

[2] Kuo, J., Fosler-Lussier, E., Jiang, H. and Lee, C., "Discriminative training of language models for speech recognition", in Proc. ICASSP, 2002.

[3] Roark, B., Saraclar, M., Collins, M. and Johnson, M., "Discriminative language modeling with conditional random elds and the perceptron algorithm", in Proc. ACL, 2004

[4] Roark, B., Saraclar, M. and Collins, M., "Discriminative syntactic language modeling for speech recognition", in Proc. ACL, 2005.

[5] Roark, B., Saraclar, M. and Collins, M., "Discriminative n-gram language modeling", in Computer Speech and Language, vol. 21, pp. 373-392, 2007.

[6] Kurata, G., Itoh, N. and Nishimura M., Acoustically discriminative training for language models, in Proc. ICASSP, 2009.

[7] Jyothi, P. and Fosler-Lussier, E., Discriminative language modeling using simulated ASR errors, in Proc. Interspeech, 2010.

[8] Huang, J., Li, X. and Acero, A., Discriminative Training Methods for Language Models Using Conditional Entropy Criteria, in Proc. ICASSP, 2010.

[9] Kurata, G., Itoh, N. and Nishimura M., Training of errorcorrective model for ASR without using audio data, in Proc. ICASSP, 2011.

[10] Xu, P., Karakos, D. and Khudanpur, S., Self-supervised discriminative training of statistical language models, in Proc. IEEE ASRU, 2009.