TR-IT-0333

Distance-related Unit Association Maximum Entropy (DUAME) Language Modeling

張 樹武
Shuwu Zhang

January,2000

In this report, we proposed a distance-related unit association maximum entropy (DUAME) language modeling. In comparison with conventional N-gram modeling, some major characteristics of DUAME modeling are: 1). Instead of longer N-gram, it can simulate an event (unit subsequence) using the exponential co-occurrence of full distance unit association (UA) features. Thus, it is functional comparable to higher order N-gram, 2). DUAME modeling can smooth the distribution of an partially unobserved event with the exponential co-occurrence of decreasing UA features. It is more accurate to predict this part of events compared to conventional backoff or interpolation smoothing in N-gram modeling, 3). Because all UA features in DUAME model are relevant to only two units, it takes much less memory requirement for storing feature parameters and it is more available in terms of memory to exploit longer distance language correlations compared to longer order N-gram features. Preliminary experimental results have shown that DUAME modeling is very useful for improving the current N-gram language modeling in speech recognition.

Contents

1 Introduction

Language modeling is an important component of automatic speech recognition. Currently, N-gram modeling [1] has become the dominant approach to language modeling, because it can be used to predict a possible succeeding language unit by utilizing only a few immediate preceding units. It can also be well integrated in real speech recognition. Based on advanced studies on language modeling, traditional N-gram has been refined and been improved continually [3] [5] [8] [15]. However, because of the limitation of classical Markov theory, it is inflexible to incorporate more types of potential language features because of the limitation of the classical Markov theory. Furthermore, it is infeasible in terms of memory to exploit longer distance language correlations.

Maximum entropy (ME) approach is an available modeling approach for combining various language constraints [14] [10] [13] [11] [6] [12] [19]. Under the ME framework, each group of knowledge sources is regarded as a set of features and a correspondence can be made to a set of probability functions. ME solution is to pursue the function which has the highest entropy. ME thus allows integration of various different knowledge sources, each retaining their special feature representation, into a uniform framework.

In this report, we present an independent distance related unit association maximum entropy language modeling approach. We call it DUAME modeling for short. In comparison with N-gram modeling, DUAME modeling have many prospective advantages such as functional comparable to higher order N-gram, more accurate prediction for partially unobserved event, much less memory requirement, more available in terms of memory to exploit longer distance language correlations and more feasible to integrate with other types of language features or language modeling approaches under maximum entropy framework. Some experimental results have shown that DUAME modeling is very helpful for improving the current N-gram language modeling in speech recognition.

2 Distance-related Unit Association Maximum Entropy (DUAME) Language Modeling

2.1 Principle of Maximum Entropy Modeling

Based on a given language sequence $S =< x_1, ..., x_z >$, we can define the followings.

Definition I: an event $< H_i, x_i >$ is a contextual window $H_i =< x_{i-n+1}, ..., x_{i-1} >$ succeeding the current unit x_i.

All such events constitute the event space $\mathcal{E} < H, x >$. For a special language corpus, events appearing in the corpus are called observed events, and the others can be classified into the unobserved event set.

Definition II: a feature $g(h_i, x_i)$ is a kind of global or local description for some events with the attribute set $< g(h_i, x_i), a_i, m_i, \alpha_i >$.

where, $g(h_i, x_i)$ is an indicator of the feature, h_i is a subset of the context under contextual window H_i, x_i is the current language unit, a_i is its target expectation, correspondingly, m_i refers to its feature expectation, and $\alpha_i = e^{\lambda_i}$ is an exponential feature factor related to the probability of the feature. A feature set $\mathcal{G}$ can be extracted from the observed event set.

Based on the above definitions of event and feature, an exponential probability distribution can be used to evaluate the language sequence.

$$m^*(S) = argmax_{m \in \mathcal{M}} \prod_{i=1}^{z} m(x_i|H_i). \tag{1}$$

where,

$$m(x_i|H_i) = \frac{r(H_i, x_i)}{Z(H_i)}$$

is an exponential probability given context H_i.

$$r(H_i, x_i) = \prod_{t=1}^{k} e^{\lambda_i g_t(h_i, x_i)} = \prod_{t=1}^{k} \alpha_t^{g_t(h_i, x_i)}$$

is the multiplication of feature factors activated in event $< H_i, x_i >$,
and

$$Z(H_i) = \sum_{x \in V} r(H_i, x)$$

is used for normalization.

For each feature $g(h_i, x_i)$, there is a corresponding exponential expectation

$$m(h_i, x_i) = \sum_{H_i \supseteq h_i} \tilde{p}(H_i) * m(x_i|H_i) \tag{2}$$

where $\tilde{p}(H_i)$ is the observed probability of context H_i in the training set. We can assume that this feature expectation will be approximated to its target expectation

$$m(h_i, x_i) \approx E(p(h_i, x_i)) = a(h_i, x_i). \tag{3}$$

It has been shown [2][11] that the optimal maximum likelihood exponential model $m^*(S)$ is identical to the maximum entropy model

$$p^*(S) = argmax_{p \in \mathcal{H}} - \sum_{<H_i,x_i>} p(H_i,x_i)\log p(H_i,x_i). \tag{4}$$

Therefore, the maximum entropy distribution $p^*(S)$ can be replaced with the maximum likelihood exponential distribution $m^*(S)$ to estimate and evaluate the maximum entropy of a language sequence.

2.2 General Expression of DUAME modeling

The basic principle of DUAME modeling was introduced in [17] and [18]. In DUAME modeling, a feature is defined as a distance related unit association.

Definition III: A distance-related unit association feature is a span distance unit pair (h, x) with the attribute $< g_d(h,x), a_d(h,x), m_d(h,x), \alpha_d(h,x) >$.

Here h denotes a contextual unit within a limited window,

x is the current unit (class,word, or phrase),

d refers to the distance between h and x, $a_d(h,x)$ denotes the target expectation of feature $g_d(h,x)$,

$m_d(h,x)$ is its feature expectation,

and $\alpha_d(h,x) = e^{\lambda_d(h,x)}$ is an exponential factor of feature $g_d(h,x)$.

Based on this definition, a general expression of distance-related unit association maximum entropy modeling can be written as follows.

For a given unit sequence $S =< x_1, x_2, ..., x_z >$ and contextual window n,

$$m^*_{n-duame}(S) = argmax_{m \in \mathcal{M}} \prod_{i=1}^{z} m(x_i|h_n, ..., h_1), \tag{5}$$

and with event $< h_n, ..., h_1, x >$,

$$m(x|h_n, ..., h_1) = \frac{r(h_n, ..., h_1, x)}{Z(h_n, ..., h_1)} = \frac{\prod_{j=1}^{n} \alpha_j(h_j,x)^{g_j(h_j,x)}}{\sum_{x_t \in V} r(h_n, ..., h_1, x_t)}. \tag{6}$$

Here,

$$m(x|h_n, ..., h_1)$$

is an exponential probability given context $< h_n, ..., h_1 >$,

$$r(h_n, ..., h_1, x) = \prod_{j=1}^{n} \alpha_j(h_j,x)^{g_j(h_j,x)}$$

is the summation of exponential feature factors activated in event $< h_n, ..., h_1, x >$, and

$$Z(h_n, ..., h_1) = \sum_{x_t \in V} r(h_n, ..., h_1, x_t)$$

is for normalization.

2.3 Major Characteristics of DUAME Modeling

In comparison with conventional N-gram modeling, DUAME language modeling has following major

1. Functional comparable to higher order N-gram:
   Instead of conventional N-gram, DUAME modeling can model N-gram event with the exponential co-occurrence of full distance unit association (UA) features. Thus, it is functional comparable to higher order N-gram modeling.

2. More Accurate Prediction:
   DUAME modeling can smooth the distribution of an partially unobserved event with the exponential co-occurrence of decreasing UA features. It is more accurate to predict this part of events compared to conventional backoff or interpolation smoothing in N-gram modeling.

3. Much less memory cost compared to the corresponding order N-gram:
   For the distance-n DUAME model, the possible memory requirement is less than $O(V^n + n \times V^2 + V)$. Here, V is denoted as the vocabulary size, and n is the length of the contextual window. Correspondingly, an identical order N-gram takes an order of V^{n+1} memory. Therefore, the DUAME model requires far less memory than the N-gram model.

4. Be feasible to be extended to longer history event:
   Because all UA features in DUAME are relevant to only two units, it is more feasible in terms of memory to exploit longer distance language correlations compared to longer order N-gram features.

5. Be available to be integrated with different types features:
   A DUAME model can be estimated by maximum entropy approach, it is more available to integrate with other types of language features or language modeling approaches under uniform ME framework.

2.4 Estimation of DUAME model

As in conventional maximum entropy modeling, the DUAME model can be trained with an improved generalized iterative scaling (GIS) algorithm (see Algorithm I).

Two criteria can be used to verify convergence of the DUAME model. One is standard perplexity (Figure 1)

$$PP = exp^{-\frac{1}{N}\ln P(S)}. \tag{7}$$

The other one is Kullback-Leibler divergence (minimum discrimination information). (Figure 2)

$$D(a||m) = \sum_{g_i(h,x)} a_i(h,x)log(\frac{a_i(h,x)}{m_i(h,x)}). \tag{8}$$

This criterion is employed to verify holistic discrimination between target expectation $a_i(h,x)$ and feature expectation $m_i(h,x)$ over all features. Theoretically, its value should be approximated to 0 with convergence.

Because the number of activated features for the event is inconstant in DUAME model, the increment of each feature must be computed numerically for ensuring convergence monotonically. An effective way of doing this is by simplified Newton method (Algorithm II). With an appropriate initialized value β_0 and suitable attention paid to the function $f(\beta)$, the increment β can be gotten by the solution of equation $f(\beta) = 0$ iteratively.

Input:

contextual window length: n,

the number of units: m,

contexts: $H_1, ..., H_z$,

UA features: $g = g_1, ..., g_k$,

initialized feature parameters: $\alpha_1, ..., \alpha_k$,

and target expectation of UA features: $a_1, ..., a_k$

Output:

feature parameters $\alpha_1, ..., \alpha_k$

normalized context factors $Z(H_1), ..., Z(H_z)$

Algorithm:

```
Initialization: for $i = 1$ to $k$ $[\alpha_i = 1]$
Until convergence
  Initializing feature expectation: for $i = 1$ to $k$ $[m_i = 0]$
  for $i = 1$ to $z$
    for $j = 1$ to $m$
      checking # of activated features in the event and
      computing event expectation
      $m(H_i, x_j) = P(H_i)\frac{r(H_i,x_j)}{Z(H_i)}$
      for each activated feature $g_l$, accumulating
       feature expectation $m_{g_l,\#} + = m(H_i, x_j)$
  for $i = 1$ to $k$
      solving increment $\beta_i$ from polynomial equation
      $\sum_{j=1}^{n} m_{i,j}(\beta_i)^j = a_i$
      computing feature gain: $\alpha_i^{t+1} = \alpha_i^t \beta_i$
return all feature parameters $\alpha$
```

Algorithm I: Improved GIS for DUAME Estimation

Input:

$f(\beta) = m_n\beta_n + ... + m_1\beta_1 + a$ for all $m_i \geq 0$,

where a refers to the target expectation of the feature.

initial point $\beta_0 = 1$

Output:

increment β where $f(\beta) = 0$

Algorithm:

```
Initialization: $\beta = \beta_0$
Until $|f(\beta)| \leq \epsilon$
    $\beta = \beta - \frac{f(\beta)}{f'(\beta)}$
return $\beta$
```

Algorithm II: Simplified Newton's Method for the solution of increment β

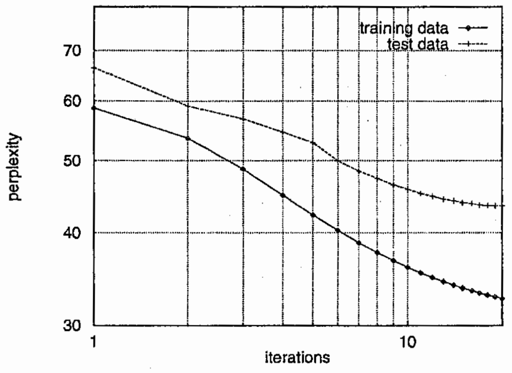

Figure 1: Convergence of DUAME estimation with regard to perplexity

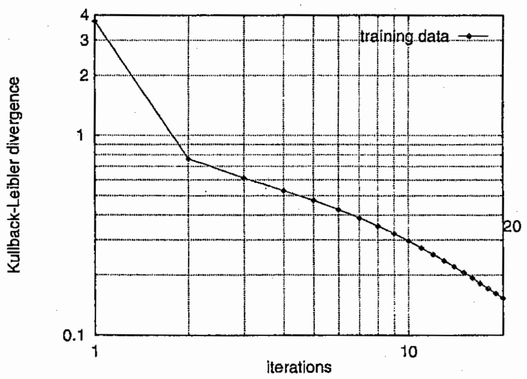

Figure 2: Convergence of DUAME estimation with regard to Kullback-Leibler divergence

3 Experiments

3.1 Experiments on DUAME Language Modeling

Based on two corpora, the ATR travel arrangement task English (ATR-TAT) corpus [4] and the 1994 Wall Street Journal (WSJ'94) corpus (Table 1), we conducted experiments on the DUAME modeling.

Table 1: Corpora

	Corpus	
Name	ATR-TAT	WSJ
Size	1M	10M
Style	spoken	written

Table 2: Feature extraction and selection

		unit association (UA) features			
	cutoff	d0 UA	d1 UA	d2 UA	d3 UA
ATR -TAT	> 0	8580	88752	117331	128822
	> 2		24876		
	> 5		12596	12131	11184
	> 10			6337	6390
WSJ	> 0		1332554	1933543	2154013
	> 2		505862		
	> 5	26680		184314	
	> 10				80008

We compared the perplexities of the distance-2 and distance-3 DUAME model with the maximum likelihood (ML) 2-ram, 3-gram, and 4-gram. In the ATR-TAT corpus, the basic language unit was defined as the extended word obtained with the variable N-gram language modeling tool [7]. In the WSJ corpus, the basic language unit was the normal word. The perplexities for various N-gram models were calculated with the CMU-Cambridge language modeling toolkit [9].

Table 3 shows that 2-DUAME models with both corpora resulted in substantial improvement in 2-gram models and were comparable to corresponding 3-gram models in perplexity. However, we did not expect that the perplexities of both 3-DUAME and 4-gram would be a little higher than 2-DUAME and 3-gram, respectively. This could be due to sparse data as a result of an insufficient corpus.

3.2 Recognition Improvement with Composite N-gram and DUAME modeling

Based on ATRSPREC spontaneous multilingual speech recognition system, we conducted recognition experiments with composite N-gram and DUAME models.

The baseline language model in ATRSPREC is composite N-gram model. It has already been applied in ATR speech recognition system for a few years. In the previous version,

Table 3: Comparison of N-gram and N-DUAME in perplexity

model	ATR-TAT	WSJ
LN 4-gram	45.67	235.96
3-DUAME	46.12	240.67
LN 3gram	44.95	223.22
2-DUAME	43.47	220.14
LN 2gram	59.50	270.53

*LN:linear discounting backoff

units were grouped and classes were split incrementally by minimizing the total entropy [7]. Currently, an improved version named multi-class composite N-gram [16] is being used. This modeling approach is able to assign each unit with bidirectional classes by splitting unit neighboring characteristics into preceding and succeeding connections, thereby more efficient and reliable clustering information can be obtained.

Based on ATR travel arrangement task (ATR-TAT) English corpus, we first trained a composite class bigram model. Based on the extended definition of language unit and class (Table 4) from composite bigram modeling. a 2-DUAME model was trained iteratively.

Table 4: Re-definition of unit and class by composite N-gram modeling

	initial def.	extended def.
subwords/words	11594	11588
compound words	18	303
pron. variants	18050	23511
classes	76	1069

In recognition, we applied composite class bigram model in the first pass integrated decoding for generating a candidate word lattice. In the second pass search, this lattice was rescored by 2-DUAME evaluation for getting optimal N-best result. Meanwhile, we also compared composite class bigram with word-based bigram and POS-based bigram, and 2-DUAME with composite class trigram in both perplexity and recognition accuracy (see Table 5).

Table 5 shows that composite class bigram resulted in 6.17% reduction of word error rate (WER) in recognition than word bigram. Because of only 76 part-of-speech classes information, the performance of POS-based bigram is the worst one. And, 2-DUAME model has yielded WER reduction of 10.09% over composite class bigram. We can thereby say that 2-DUAME model has significant performance improvement than corresponding bigram model.

Actually, rescoring in the second pass search was constrained in a made lattice, the function of 2-DUAME has been definitely restricted on the current procedure of recognition. So, we wish to integrate 2-DUAME model into the first pass decoding in future works.

Table 5: Recognition Accuracy Improvement with DUAME language modeling

Training set: ATR-TAT corpus with 985,245 tokens
Test set: 253 utterances including 2476 tokens

Model	PP	WER(%)	Parameters
POS 2-gram	113.10	48.65	1.7×10^4
word 2-gram	32.27	32.74	1.3×10^8
composite 2-gram	**26.70**	**30.72**	1.1×10^6
2-DUAME	**22.42**	**27.62**	3.3×10^6
composite 3-gram	21.10	-	1.2×10^9

4 Discussions

In this report, we have presented an independent distance related unit association maximum entropy language modeling approach for improving current N-gram modeling. This approach have many prospective advantages such as functional comparable to higher order N-gram, more accurate prediction for partially unobserved event, much less memory requirement, more available in terms of memory to exploit longer distance language correlations and more feasible to integrate with other types of language features or language modeling approaches under maximum entropy framework. Some experimental results have shown that DUAME modeling is very helpful for improving the current N-gram language modeling in speech recognition.

Furthermore, based on the principle of hierarchical language modeling, We will investigate better way to integrate DUAME model into first pass decoding. we will also continue to study the combined modeling to exploit linguistic correlations for complementing underlying DUAME model.

参考文献

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Aalysis and Machine Intelligence*, pages 179–190, 1983.

[2] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. In *The Annals of Mathematical Statistics*, pages 1470–1480, 1972.

[3] P. F. Brown et al. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479, 1992.

[4] T. Morimoto et.al. Speech and language database for speech translation research. In *ICSLP'94*, pages 1791–1794, 1994.

[5] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, pages 400–401, 1987.

[6] R. Lau, R. Rosenfeld, and S. Roukos. Adaptive language modeling using the maximum entropy principle. In *Proceedings of the Human Language Technology Workshop*, pages 108–113, 1993.

[7] H. Masataki and Y. Sagisaka. Variable order ngram generation by word-class splitting and consecutive word grouping. In *ICASSP'96*, pages 188–191, 1996.

[8] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[9] P.Clarkson and R.Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Eurospeech'97*, pages 2707–2710, 1997.

[10] S.A. Della Pietra, V.J. Della Pietra, R. Mercer, and S. Roukos. Adaptive language modeling using minimum discriminant estimation. In *ICASSP'92*, pages I–633–636, 1992.

[11] S.D. Pietra, V.D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Analysis And Machine Intelligence*, 19:1–13, 1997.

[12] V.J. Della Pietra, A.L. Berger, and S.A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, pages 39–71, 1996.

[13] E. S. Ristad. A maximum entropy modeling toolkit. Technical report, Dept. of Computer Sciences,Princeton University, January 1997.

[14] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, pages 187–228, 1996.

[15] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. on information theory*, 37:1085–1093, 1991.

[16] H. Yamamoto and Y. Sagisaka. Multi-class composite n-gram based on connection direction. In *ICASSP'99*, pages 533–536, 1999.

[17] S. Zhang, H. Singer, and Y. Sagisaka. Distance-related unit association maximum entropy language modeling. In *The 1999 Spring meeting of the acoustical society of Japan*, pages 43–44, March,1999.

[18] S. Zhang, H. Singer, D. Wu, and Y. Sagisaka. Improving n-gram modeling using distance-related unit association maximum entropy language modeling. In *Eurospeech'99*, September,1999.

[19] X. Zhou, S. Chen, and R. Rosenfeld. Linguistic features for whole sentence maximum entropy language models. In *EUROSPEECH'99*, Sep. 1999.