IMPROVING N-GRAM MODELING USING DISTANCE-RELATED UNIT ASSOCIATION MAXIMUM ENTROPY LANGUAGE MODELING

Shuwu Zhang†, Harald Singer†, Dekai Wu‡, Yoshinori Sagisaka†
†ATR Interpreting Telecommunication Research Labs.Kyoto,Japan,619-0288
{szhang,singer,sagisaka}@itl.atr.co.jp
‡The HongKong University of Science and Technology,HongKong
dekai@cs.ust.hk

ABSTRACT

In this paper, a distance-related unit association maximum entropy (DUAME) language modeling is proposed. This approach can model an event (unit subsequence) using the co-occurrence of full distance unit association (UA) features so that it is able to pursue a functional approximation to higher order N-gram with significantly less memory requirement. A smoothing strategy related to this modeling will also be discussed. Preliminary experimental results have shown that DUAME modeling is comparable to conventional N-gram modeling in perplexity with significantly small number of parameters.

1. Introduction

Language modeling is an important component of automatic speech recognition and translation. Currently, N-gram modeling is the most prominent approach in language modeling. Since it regards language sequence as a discrete stochastic process, N-gram is able to efficiently characterize possible combination of linguistic pieces as a Markov model. Thus, this approach can predict a possible succeeding language unit by utilizing only a few immediate preceding units. However, since features (language constraints) used in this kind of model are constrained to possible subsets of successive unit sequence, it is inflexible to incorporate different types of features and it is infeasible in terms of memory to extend to longer distance features.

Maximum entropy (ME) language modeling approaches have been discussed in many studies [1] [2] [3] [4] [5] [6] [7]. Most schemes with ME exploit longer distance features or language dependencies. These types of features are often fragmentary. Thus, ME model has to be interpolated with conventional N-gram models as a compensator of the N-gram. The smoothing problem in ME modeling has not been well mentioned in past publications yet.

In this paper, 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 simulates an event (unit subsequence) using the co-occurrence of different distance unit association (UA) features. Thus, it is able to pursue functional approximation to higher order N-gram and requires less memory. In section 2, we introduce this modeling approach. We also discuss the smoothing strategy related to DUAME modeling in this section. Preliminary experimental results are given in section 3.

2. DUAME Language Modeling

2.1. Principle of Maximum Entropy

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.

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

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

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

where $\tilde{p}(H_i)$ is the observed probability of context H_i in the training set. We can assume that this feature ex-

pectation will be approximated to its target expectation

$$m(h_i, x_i) \approx E(p(h_i, x_i)) = a(h_i, x_i). \quad (3)$$

It has been shown [8][2] that the optimal maximum likelihood exponential model $m^*(S)$ is identical to the maximum entropy model

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

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 Model

The basic principle of DUAME modeling was introduced in [9]. 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 $< g_d(h, x), a_d(h, x), m_d(h, x), \alpha_d(h, x) >$

where 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 $\alpha_d(h, x)$ is the factor of $g_d(h, x)$.

Thus, for an event $< h_1, ..., h_n, x >$ with contextual window length n, we can simulate it using the co-occurrence of n different distance unit association features instead of the longer N-gram features.

$$\alpha_n(h_1, x)\alpha_{n-1}(h_2, x)....\alpha_1(h_n, x) \Rightarrow \alpha_{n-gram}(h_1, ..., h_n, x)$$

These features with distance attributes are independent of each other and their co-occurrence is a precise expression for this event.

And, $\alpha_d(h, x)$ is an adjustable factor, it pursues a trade-off to affect many different events and features. It is impossible for a real N-gram to use the same way because corresponding $c(h_i, x)$ in N-gram is a real count value and are counted repeatedly between different events.

Thus, a general expression of distance-related unit association ME modeling can be written as the following.

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

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

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

$$\mathrm{m(x|h_1, ..., h_n)} = \frac{\mathrm{r(h_1, ..., h_n, x)}}{\mathrm{Z(h_1, ..., Z_n)}} = \frac{\prod_{\mathrm{i=1}}^{\mathrm{n}} \alpha_{\mathrm{n-i+1}}^{\mathrm{g_{n-i+1}(h_i, x)}}(\mathrm{h_i, x})}{\sum_{\mathrm{x_i \in V}} \mathrm{r(h_1, ..., h_n, x_i)}}. \quad (6)$$

Specifically, for the triplet event $< h_1, h_2, x >$ with distance n=2,

$$m(x|h_1, h_2) = \begin{cases} \frac{\alpha_2(h_1, x) * \alpha_1(h_2, x)}{Z(h_1, h_2)} & \text{if } g_2(h_1, x), g_1(h_2, x) \\ \frac{\alpha_2(h_1, x)}{Z(h_1, h_2)} & \text{if only } g_2(h_1, x) \\ \frac{\alpha_1(h_2, x)}{Z(h_1, h_2)} & \text{if only } \exists g_1(h_2, x) \\ \frac{\alpha_{uni}(x)}{Z(h_1, h_2)} & \text{if only } \exists g_{uni}(x) \end{cases} \quad (7)$$

The function of DUAME is similar to a backoff N-gram, because DUAME model can model N-gram features with the co-occurrence of full distance unit association (UA) features and smooth the distribution of an unobserved event with the co-occurrence of decreasing UA features.

For distance-n DUAME model, possible memory requirement is less than $V^n + n * V^2 + V$, i.e., it takes an order of V^n memory. Here, V is denoted as the vocabulary size, n is the length of the contextual window, V^n is for total context $Z(h_1, ..., h_n)$, and $n * V^2 + V$ is for storing unit association feature factors. For a distance-2 DUAME model, the possible memory requirement is less than $3V^2 + V$.

Correspondingly, an identical N-gram takes an order of V^{n+1} memory. Therefore, the distance-n unit association ME modeling requires far less memory than the N-gram modeling.

2.3. Smoothing Strategy in DUAME Modeling

According to the general ME expression, the normalization factor $Z(h_1, ..., h_n) = \sum_{x \in V} r(h_1, ..., h_n, x)$ in (6) should be calculated over all related events including observed and unobserved events. In addition, the feature expectation $m(h_i, x_i)$ in (2) is a summation over all possible contexts H_i even though some events do not occur in the training data. The advantages with this way are that the distribution of the unobserved event can be smoothed automatically with decreasing features, which are assumed to be activable in this unobserved event and the normalization condition under limited contextual window can be ensured definitely. However, since unobserved events are taken into summation with the same weight as observed events, it is possible to cause exaggerated evaluation for these unobserved events.

At present, there is no ideal smoothing technique for maximum entropy modeling. Here, we propose to use the normal smoothing strategies in the maximum likelihood (ML) N-gram modeling for the ME modeling. In N-gram modeling, one smoothing approach is an absolute discounting backoff approach, which was proposed in [10]. With this approach, a small constant d can be extracted from a significantly smaller amount of N-gram count. Then the conditional probability is calculated with

$$p(z|y) = \begin{cases} \frac{c(y,z) - d}{c(y)} & \text{if } c(yz) > k \\ d * \frac{\#(y, \_)}{c(y)} * p(z) & \text{otherwise.} \end{cases} \quad (8)$$

where $c(.)$ means the count of occurrences and $\#(.)$ denotes the number of different occurrences. In addition,

a suggested optimal discount ratio is:

$$d = \frac{\#(c(y,\_) = 1)}{\#(c(y,\_) = 1) + 2 * \#(c(y,\_) = 2)}.$$

Similarly, for DUAME modeling, we can set one discount ratio for each context $H_i =< h_1, ..., h_n >$ with

$$d_{H_i} = \frac{\#(c(H_i,\_) = 1)}{\#(c(H_i,\_) = 1) + 2 * \#(c(H_i,\_) = 2)} \quad (9)$$

and

$$\theta_{H_i} = d_{H_i} * \frac{\#(H_i,\_)}{\sum_x c(H_i, x)}. \quad (10)$$

Thus, (6) can be changed into (11)

$$\mathrm{m(x|h_1, ..., h_n)} = \begin{cases} \frac{r(h_1,...,h_n,x) - d_{H_i}}{Z(h_1,...,h_n)} & \text{if } \exists < h_1, ..., h_n, x > \\ \theta_{H_i} \frac{r(h_1,...,h_n,x)}{Z(h_1,...,h_n)} & \text{otherwise.} \end{cases} \quad (11)$$

This change should be imported into both estimation and evaluation in DUAME modeling.

2.4. Issues on DUAME Estimation with GIS Algorithm

As in conventional ME modeling, the generalized iterative scaling (GIS) algorithm [8] has been used to train the DUAME model. Two criteria can be used to verify convergence of the DUAME model. One is standard perplexity

$$PP = 2^{-\frac{1}{n}\log P(S)}. \quad (12)$$

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

$$D(a, m) = \sum_{g_i(h,x)} a_i(h, x) log(\frac{a_i(h, x)}{m_i(h, x)}). \quad (13)$$

This criterion is used to verify holistic discrimination between target expectation $a_i(h, x)$ and feature expectation $m_i(h, x)$ over all features.

In general, for a feature $g_i(h, x)$, the ratio $\beta = \frac{a_i(h,x)}{m_i(h,x)}$ can be approximated to 1 with a finite amount of vibrations (as shown in Figure 1). Although the convergence of GIS algorithm has been proved [8][2], an inappropriate initialization of the feature parameter $\alpha_i(h, x)$ or an extreme lower target expectation $a_i(h, x)$ still could cause some accidental disturbances that make increment β skip out of the normal vibration area; this disturbance would be unrecoverable. In this case, we use a simple attenuator to filter unexpected increment in each iteration.

$$\beta_i^t = \begin{cases} \varphi_i \frac{a_i}{m_i^t} & \text{if } \frac{a_i}{m_i^t} \gg \max[\beta_i^0, ..., \beta_i^{t-1}] \\ \sigma_i \frac{a_i}{m_i^t} & \text{if } \frac{a_i}{m_i^t} \ll \min[\beta_i^0, ..., \beta_i^{t-1}] \\ \frac{a_i}{m_i^t} & \text{otherwise} \end{cases} \quad (14)$$

where φ_i and σ_i are the adjustable damping and enlarging coefficients, respectively. Experiments have shown that this is helpful in ensuring the convergence of the model estimation.

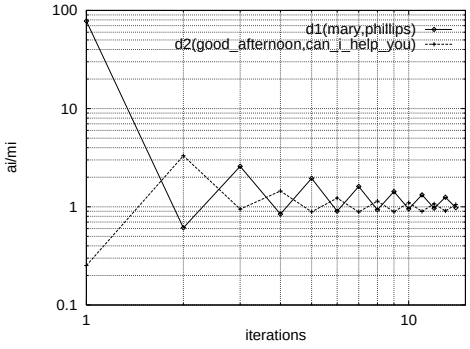

Figure 1. Illustration of feature expectation $m_i(h, x)$ approximated to target expectation $a_i(h, x)$ with vibrations

3. Preliminary Experiments on DUAME Language Modeling

Based on two corpora, the ATR travel arrangement task English (ATR-TAT) corpus [11] 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	> 0	8580	88752	117331	128822
-TAT	> 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 [12]. 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 [13].

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.

Some experiments on feature selection were also performed. As mentioned in section 2.4, some features with extreme lower probability can cause diverging estimations. Therefore, it is necessary to select more reliable features for robust modeling. In our preliminary experiments, we chose features simply by setting some thresholds of count for different types of features. Table 4 shows that the reliable feature selection was helpful for the improvement of the model.

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

Table 4. Improvement of DUAME with feature selection

model	feature cutoffs $T_{d3}\_T_{d2}\_T_{d1}$	ATR-TAT	WSJ
3-DUAME	0_0_0	46.12	240.67
	5_5_5	45.83	
	10_5_2	44.67	232.21
2-DUAME	0_0_0	43.47	220.14
	0_5_5	43.12	
	0_10_5	42.20	214.42
	0_100_5	38.43	

4. Discussion and Conclusion

We have presented a new distance-related unit association maximum entropy language modeling approach. This approach has obvious advantages such as less memory requirement and the ability of functional approximation to higher order N-gram. So, DUAME modeling will be useful for improving the current N-gram language modeling in speech recognition.

As in conventional ME modeling, a disadvantage of DUAME modeling is that training takes a long time when the vocabulary size is large. Although some computational tricks mentioned in [4] have been adopted, we still need a few weeks to train the DUAME model on the WSJ corpus. In future works, we will address this problem, and we will also study some more effective approaches for feature selection. Meanwhile, we will continue investigating the improvement of speech recognition accuracy with this model.

References

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

[2] 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.

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

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

[5] A. Stolcke, C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, and D. Wu. Dependency language modeling. Technical report, Workshop96 Project Report, Johns Hopkins Univ., 1996.

[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] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 1996.

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

[9] 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*, March,1999.

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

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

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

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