IMPROVING LANGUAGE MODELING BY USING DISTANCE AND CO-OCCURRENCE INFORMATION OF WORD-PAIRS AND ITS APPLICATION TO LVCSR

Tze Yuang Chong[1,2], Rafael E. Banchs[3], Eng Siong Chng[1,2], Haizhou Li[1,2,3]

[1]Temasek Laboratories, Nanyang Technological University, Singapore 639798
[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798
[3]Institute for Infocomm Research, Singapore 138632
tychong@ntu.edu.sg,rembanchs@i2r.a-star.edu.sg,aseschng@ntu.edu.sg,hli@i2r.a-star.edu.sg

ABSTRACT

This paper reports our study in exploiting the distance and co-occurrence information of word-pairs to improve the *n*-gram language model. We used these two types of information for modeling the distant context, up to history length of ten. Also we show that the proposed model provides complementary information about the *n*-gram's context that is unable to be captured by the *n*-gram model due to data scarcity. Evaluated on the WSJ and SWB-1 corpora, the proposed model reduced the trigram perplexity up to 11.2% and 6.5% respectively. In an N-best re-ranking task of the Aurora-4 database, our model aided a hexagram model to perform ~9% relatively better in terms of WER.

Index Terms— Term-distance, term-occurrence, language model, speech recognition

1. INTRODUCTION

The *n*-gram model [1] has been the de-facto standard for language modeling and has been the cornerstone for many natural language processing tasks. The *n*-gram model uses the occurrence of the immediate $(n-1)$ word sequence in the history to predict the target-word. The computation is straightforward, but the *n*-gram model inherently suffers from data scarcity. In practice, the history length is usually limited to only three or four words while sacrificing the context located beyond this narrow window.

The scarcity problem is due to the exponentially growth of number of parameters as the history context expands. In order to utilize longer context efficiently, many alternatives have been proposed to approximate the ordered word sequence in the history context to a more manageable form. The distant bigram [2, 3, 4], for example, dissembles the *n*-gram into $(n-1)$ word-pairs, and each word-pair forms a distance-k bigram model, i.e. predicting target-word based on a history-word located at k distance behind ($1 \leq k \leq n-1$). In order to make use of the entire history context, the distant bigram probabilities are usually linearly combined to yield an averaged probability. As the degree of association between a history-word and a target-word depends on many factors, the combination of the distant bigrams (e.g. linear interpolation or maximum entropy) requires a set of weighting coefficient. In general, the weight depends on the distance, i.e. models of the same distance share the same weight [2]. Other more complex weighting schemes, e.g. applying at the level of word-pair or history-word, lead to some non-trivial adaptation procedures [3, 4].

In another approach, the well associated word-pairs are kept by filtering out those less associative ones from the inventory, measured with some metrics, e.g. mutual information [5]. This approach can be thought of as a special case of the distant bigram model with zero-or-one weighting, i.e. one for well-associated word-pair and vice-versa. But it is better known as the distance-dependent trigger model, not to be confused with the earlier proposed trigger model which uses history-words of arbitrary location in the history for prediction [6, 7]. As compared to the distance-dependent model, the original trigger model may be relatively insusceptible to data scarcity as the position of the history-word is ignored.

The latent semantic model, similar to the trigger model, ignores the word ordering in the history, and predicts target-word based on only the co-occurrence information, i.e. bag-of-words [8, 9]. Without the position information, the history context can be extended farther down to cover the entire document. Also, TFIDF factor is used to scale the word counts in order to highlight the semantic importance of the history-words towards the prediction.

There have been other attempts to capture the long-span context, such as the structure language model [10] uses the parse-tree to determine the "heads" in the history-context that are well associate to the target-word. The skip-gram model [11, 12] reveals the long-span *n*-grams by bypassing irrelevant words. The cache model [13, 14] accumulates the temporal words frequencies. The topic-based model [15, 16] exploits the topical information in the history-context. The connectionist approach [17, 18] uses neural network to learn the long-span regularities.

For approaches such as the distant bigram model and the distance-dependent trigger model, the probability depends on the occurrence of history-word in a specific position, i.e. the occurrence and the position information are coupled. On the other hand, the trigger model and the latent

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

semantic model discard the position information and exploit only the occurrences information. We suggest that, rather than "hardly" decide if to retain the position information, these two types of information (occurrence and position) shall be coordinated in a more flexible manner, e.g. combined with a weighting scheme. Moreover, we believe that the position information, which reflects the syntactic structure of a language, is an important cue for language modeling. Although it is reasonable to ignore the word position from the far context, the syntactic structure near the target-word is informative. But the *n*-gram model might not effectively capture this information due to data scarcity.

In our previous work [19], term-distance (TO) and term-occurrence (TO) models have been proposed and have been shown to reduce the trigram's perplexity up to 14%. Given a history context, the TD and TO models exploit the relationship between target-word and history-words in terms of distance (i.e. position) and co-occurrence. The TD and TO information are *decoupled* from the word-pair (i.e. target-word and history-word). When combining with the *n*-gram model, the TD and TO models will be weighted, as correspond to tuning the degree of how informative the syntactic and semantic component in a language.

In this paper, we investigate the applicability of the TD and TO models in improving the performance of a speech recognition system. Also we examine the proposed model on conversational dataset besides of the WSJ corpus [19].

Next section discusses the formulation of the TD and TO models. Section 3 presents the perplexity evaluation result on the WSJ and SWB corpora, followed by the speech recognition re-ranking task result in Section 4. Finally we conclude this work and suggest avenues for future work.

2. TERM-DISTANCE AND TERM-OCCURRENCE LANGUAGE MODELS

A language model estimates the probability of a word given its history, i.e. $P\left(t = w_i | h = w_{i-1}^{i-n+1}\right)$, where t denotes the target-word and h denotes the history. Let the word located at i^{th} position, w_i be the target-word, and the preceding $n-1$ words, i.e. $w_{i-1}^{i-n+1} = (w_{i-n+1} \dots w_{i-2} w_{i-1})$ be the history. In order to reach farther history context, we assume the occurrences of the history-words to be independent from each other, conditioned to the occurrence of the target-word, i.e. $w_{i-k} \perp w_{i-l} | w_i$, where $w_{i-k}, w_{i-l} \in h$, and $k \neq l$. Thus the probability can be approximated as:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{P(t = w_i) \prod_{k=1}^{n-1} P(h_k = w_{i-k} | t = w_i)}{Z(h)} \quad (1)$$

where $Z(h)$ is a normalizing term, and $h_k = w_{i-k}$ indicates w_{i-k} as the history-word in k^{th} position. The conditional independence assumption allows the *n*-gram probability to be approximated jointly by distance-*k* bigram's likelihoods, i.e. $P(h_k = w_{i-k} | t = w_i)$.

2.1 Derivation of the TD-TO model

In order to define the TD and TO components for language modeling, we express the observation of an arbitrary history-word, w_{i-k} in k^{th} position behind a target-word, w_i, as the joint of two events: i) word w_{i-k} occurs within the history-context, i.e. $w_{i-k} \in h$, and ii) distance k to the target-word, i.e. $\Delta(w_{i-k}) = k$, ($\Delta = k$ for brevity). Thus, $(h_k = w_{i-k}, t = w_i) \equiv (w_{i-k} \in h) \cap (\Delta = k) \cap (t = w_i)$.

Therefore, the probability in Eq.1 can be written as:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{P(t = w_i) \prod_{k=1}^{n-1} P(w_{i-k} \in h, \Delta = k | t = w_i)}{Z(h)} \quad (2)$$

where the likelihood $P(w_{i-k} \in h, \Delta = k | t = w_i)$ measures how likely the joint event $(w_{i-k} \in h,\ \Delta = k)$ would be observed given the target-word w_i. It can be expressed in terms of product of the likelihoods of distance event (i.e. $\Delta = k$) and occurrence event (i.e. $w_{i-k} \in h$), as follows:

$$P(w_{i-k} \in h, \Delta = k | t = w_i) = P(\Delta = k | w_{i-k} \in h, t = w_i) P(w_{i-k} \in h | t = w_i) \quad (3)$$

The likelihood functions are referred to as TD likelihood and TO likelihood, respectively. For brevity, they are denoted as $D_{w_i, w_{i-k}}(k)$ and $C_{w_i}(w_{i-k})$. Hence,

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{P_U(w_i) \prod_{k=1}^{n-1} D_{w_i, w_{i-k}}(k) \prod_{k=1}^{n-1} C_{w_i}(w_{i-k})}{Z(h)} \quad (4)$$

where $P_U(\cdot)$ is the prior probability.

In Eq.4, we *decouple* the observation of a word-pair into the events of distance and co-occurrence. This allows for independently modeling and exploiting them. In order to control their contributions towards the final prediction of the target-word, we weight these components:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{P_U(w_i)^{\beta_n} \prod_{k=1}^{n-1} D_{w_i, w_{i-k}}(k)^{\beta_d} \prod_{k=1}^{n-1} C_{w_i}(w_{i-k})^{\beta_c}}{Z(h)} \quad (5)$$

where β_n, β_d, and β_c are the weights for the prior, TD and TO models, respectively.

Notice that the model depicted in Eq.5 is the log-linear interpolation [20] of these models. The prior, which is usually a unigram model, is replaced with a higher order *n*-gram model, e.g. the trigram model $P_T\left(w_i | w_{i-1}^{i-2}\right)$:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{P_T\left(w_i | w_{i-1}^{i-2}\right)^{\beta_n} \prod_{k=1}^{n-1} D_{w_i, w_{i-k}}(k)^{\beta_d} \prod_{k=1}^{n-1} C_{w_i}(w_{i-k})^{\beta_c}}{Z(h)} \quad (6)$$

Interpolating with higher order *n*-gram is important to compensate the damage incurred by the conditional independence assumption made in Eq.1.

2.2 Term-distance model component

Basically, the TD likelihood measures how likely a distance would separate a given word-pair. So, word-pairs possessing consistent separation distances will favor this likelihood. The TD likelihood can be estimated from counts as follows.

$$D_{w_i,w_{i-k}}(k) = \frac{\#(w_{i-k} \in h, t = w_i, \Delta = k)}{\#(w_{i-k} \in h, t = w_i)} \quad (7)$$

The above formulation requires smoothing for resolving two problems: i) a word-pair at a particular distance has a zero count, i.e. $\#(w_{i-k} \in h, t = w_i, \Delta = k) = 0$, which results in a zero probability, and ii) a word-pair is not seen at any distance within the observation window, i.e. zero co-occurrence $\#(w_{i-k} \in h, t = w_i) = 0$, which results in a division by zero. Preliminary, we use some ad-hoc manners to assign small probabilities to these likelihoods [19].

2.3 Term-occurrence model component

During the decoupling operation (from Eq.2 to Eq.4), the TD model kept the distance information but ignored the count information (i.e. the word-pair counts is normalized as shown in Eq.7). As a complement to the TD model, the TO model focuses on the co-occurrence, and captures only the count information. Since the distance information is kept by the TD model, the co-occurrence count held by the TO model is independent from the word-pair distance.

The word-pairs that frequently co-occur with arbitrary distances (within an observation window) would favor the TO likelihood. It can be estimated from counts as:

$$C_{w_i}(w_{i-k}) = \frac{\#(w_{i-k} \in h, t = w_i)}{\#(t = w_i)} \quad (7)$$

Similar to the TD likelihood, for unseen word-pair, i.e. $\#(w_{i-k} \in h, t = w_i) = 0$ a small probability value is assigned to the TO likelihood [19].

3. PERPLEXITY EVALUATION

Perplexity evaluation was conducted on the BLLIP's Wall Street Journal (WSJ) corpus [21] and the Switchboard-1 (SWB) corpus [22]. For the WSJ corpus, the train-set included the entire `87 subset (140K sentences 18M words), while the dev-set and test-set were selected randomly from `88 subset (500 sentences 15K words). For the SWB corpus, the train-set contained 309K sentences (4M words) selected randomly from the corpus, while the rest of the data is divided equally as dev-set and test-set (2.6K sentences 34K words each). The train-set was used to train the *n*-gram, TD, and TO models, while the dev-set is used to adapt the interpolation weights, i.e. β_n, β_d, and β_c (see Eq.5).

In both experiments, the vocabulary was selected randomly from the train-set such that it yielded about 5% of OOV rate on the dev-set. The vocabulary size is 19K for WSJ corpus, and 4K for SWB corpus. In this experiment, the *n*-gram model was smoothed with Kneser-Ney method.

3.1 TOTD models of different context length

In this experiment, we combined a trigram model with the TDTO model of history length 1−10 and the perplexity reductions on the WSJ and SWB corpora are shown in Fig.1. The perplexities of the WSJ and SWB corpora are represented separately by the left and right axes. The horizontal axis represents the history length of the TDTO model, i.e. length of zero denotes a plain trigram perplexity.

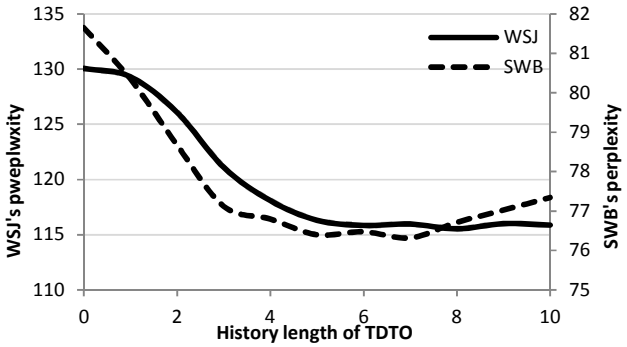

Fig.1. TDTO model reduced the trigram perplexities up to 11.2% on the WSJ and 6.5% on the SWB corpus.

As shown in Fig.1, the TDTO model reduced the trigram perplexity on both corpora, for WSJ from 130.1 to 115.5 (11.2%) and for SWB from 81.2 to 76.3 (6.5%), with the history length of 7 and 8, respectively. When the TD and TO information were added gradually as the history context expanded from length 1−5, the trigram perplexities were constantly reduced. And when farther context was included, there was no significant reduction as the information captured was noisier. The noise is due to the distance and co-occurrence of word-pairs that are not associated to the target-word. In this study, the model is grossly weighted at global level (Eq.6), any noisy TD or TD would be assigned the same weight as the informative one. Such noise further deteriorated the model, in the case of SWB, led to perplexity increase after history length of 8. As the conversation is usually more spontaneous as compared to the newswire text that has been properly planned, the syntactic information, in particular, is seldom carried across a lengthy context.

Fig.2 shows the weight settings for the *n*-gram, TD, and TO models that yielded the optimum perplexities at different TDTO's history length (Fig.1). These weight settings can be interpreted as the amount of information carried by each model. As both corpora contain different type of text, i.e. newswire text and conversation transcript, the results reflect the amount of syntactic and semantic information contributed by the TD and TO models to the *n*-gram model.

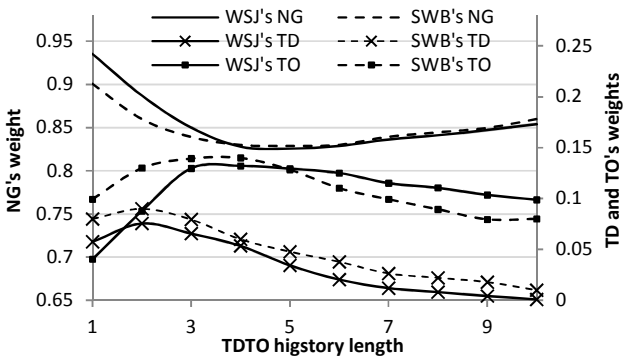

Fig.2. The optimum weights for the *n*-gram, TD, and TO models that correspond to the optimum perplexities in Fig.1.

Generally, the amount of complementary TD and TO information gradually decreased with the history length. For TD, which models the syntactic structure is ineffective at the far context, as has been shown by the gradually decreasing plots that hits 0 at length 10. The TO plot is also declining due to the non-semantic word-pairs, where there is a need to apply TFIDF [8, 9] or binary count [6].

From the TO's plots, we say that the TO information in the near context (history length < 5) is more useful to the SWB text. Higher amount of TO information is required to compensate the comparatively weak SWB's *n*-gram model.

3.2 TDTO complements the *n*-gram model

The plots in Fig.1 also suggest that the TDTO model is capable to improve the *n*-gram model even when its context is equivalent or shorter than the *n*-gram's context, e.g. the trigram perplexity was reduced by the TDTO model of history length two. This observation shows that the TDTO model provides complementary information about the word sequence that is unable to be captured by the *n*-gram model due to scarcity issue. For example, the context beyond an unseen or rare word will be abandoned by the *n*-gram model (i.e. back-off), whereas the TDTO model can still derive information from such context for prediction.

On the WSJ corpus, we combined the *n*-gram models of order 1–6, with the TDTO model of history length 5. The result shows that the complementary TDTO information reduced the hexagram perplexity up to 9%.

Table 2. Perplexities of the *n*-gram before (PPL_{NG}) and after (PPL_{IN}) interpolating with the TDTO model. The TDTO model (history length of 5) is combined with *n*-gram model of order 1–6.

Order	1	2	3	4	5	6
PPL_{NG}	1072.8	197.9	130.1	119.7	118.2	118.0
PPL_{IN}	350.2	159.7	116.3	108.8	107.6	107.4
Red. (%)	67.4	19.3	10.6	9.1	9.0	9.0

6. N-BEST RERANKING LVCSR TASK

The speech recognition was tested on Aurora-4 corpus [23]. The train-set contains 7,138 clean utterances and the test-set used in this experiment contains 330 clean utterances. The standard 5K vocabulary was used.

The acoustic model used 39-dimension MFCCs feature built on 3-state triphone HMMs. Here, the TDTO model was applied for N-best re-ranking task: we generated 200 best hypotheses from the decoder and recomputed the language model score by using the *n*-gram model with and without the TDTO model. We conducted a two-fold cross validation with the 330-utterance test-set: one subset was used to fine-tune the rescoring parameters (i.e. the grammar factor and the insertion penalty) and the other subset was used to evaluate the WER, and vice-versa in the second fold.

In order to observe the gain contributed by the TDTO model, we first computed the WER by using the *n*-gram model, of order 2 and 6 in this experiment, as the baselines. Then we combined the *n*-gram model with the TDTO model of history length 5. Both the *n*-gram and TDTO models were trained by using the WSJ corpus [21], similar as previous experiment. The results are as follows.

Table 3. WER produced by the *n*-gram model (NG) with and without the TDTO model and the relative improvement gained.

	WER		Rel.Imp. (%)
	NG	NG+TDTO	
Bigram	8.69	8.37	3.67
Hexagram	6.88	6.26	8.92

The results show that the perplexity reduction gained by the TDTO model (as shown in previous section) has been translated as WER improvement in this task. For the bigram model, the TDTO model provided the distant information and yielded 3.67% WER improvement. For the hexagram model, as the TDTO model shared the common 5-word context, the 8.92% WER improvement gain shows the effectiveness of the complementary TD and TO information provided by the TDTO model. This outcome is coherent to the perplexity evaluation result discussed in Section 3.2.

7. CONCLUSION

We have presented exploiting TD and TO information to complement the *n*-gram model with syntactic and semantic information. Besides capturing information from far context, the TD and TO models provide addition information from the *n*-gram's context which is unable to be modeled by the *n*-gram due to data scarcity issues. Evaluated on the WSJ and SWB corpora, TDTO model has reduced the trigram's perplexity up to 11.2% and 6.5%; while on the Aurora-4 re-ranking task, improved the WER of the bigram and hexagram models up to 3.67% and 8.92%, relatively.

As future work, we see the need to develop a more principled weights scheme. The TD and TO probabilities shall be weighted at finer level, such as POS, words, POS-pairs, or word-pairs. Such efforts would, besides improving the model, help to interpret the regularity of distance and co-occurrence of word-pairs in different language at finer level.

12. REFERENCES

[1] Bahl, L., Jelinek, F. & Mercer, R., "A statistical approach to continuous speech recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 5, pp. 179-190, 1983.

[2] Huang, X. et al., "The SPHINX-II speech recognition system: an overview", Computer Speech and Language, vol. 2, pp. 137-148, 1993.

[3] Simons, M., Ney, H. & Martin S. C., "Distant bigram language modelling using maximum entropy," in Proc. ICASSP, 1997, pp. 787-790, 1997.

[4] Brun, A., Langlois, D. & Smaili, K., "Improving language models by using distant information," in Proc. ISSPA, pp. 1-4, 2007.

[5] Zhou, G. & Lua, K.T., "Word association and MI-trigger-based language modeling," in Proc. COLING-ACL, pp. 1465-1471, 1998.

[6] Lau, R. et al., "Trigger-based language models: a maximum-entropy approach," in Proc. ICASSP, pp. 45-48, 1994.

[7] Rosenfeld, R., "A maximum entropy approach to adaptive statistical language modeling," Computer Speech and Language, vol. 10, pp. 187-228, 1996.

[8] Bellegarda, J. R., "A multispan language modeling framework for large vocabulary speech recognition," IEEE Trans. on Speech and Audio Processing, vol. 6, no. 5, pp. 456-467, 1998.

[9] Coccaro, N., "Latent semantic analysis as a tool to improve automatic speech recognition performance," Doctoral Dissertation, University of Colorado, Boulder, CO, USA, 2005.

[10] Chelba, C. & Jelinek, F., "Structured language modeling," Computer Speech & Language, vol. 14, pp. 283-332, 2000.

[11] Siu, M. & Ostendorf, M., "Variable *n*-grams and extensions for conversational speech language modeling," IEEE Trans. on Speech and Audio Processing, vol. 8, no. 1, pp. 63-75, 2000.

[12] Guthrie, D., et al., "A closer look at skip-gram modeling," in Proc. LREC, pp. 1222-1225, 2006.

[13] Kuhn, R. & Mori, R. D., "A cache-based natural language model for speech recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 12, no. 6, pp. 570-583, 1990.

[14] Clarkson, P. R. & Robinson, A. J., "Language model adaptation using mixtures and an exponentially decaying cache," in Proc. ICASSP, pp. 799-802, 1997.

[15] Mrva, D. & Woodland P. C., "A PLSA-based language model for conversational telephone speech," in Proc. ICSLP, pp. 2257-2260, 2004.

[16] Chien, J.-T. & Chueh, C.-H., "Latent dirichlet language model for speech recognition," in Proc. SLT, pp. 201-204, 2008.

[17] Bengio, Y. et al., "A neural probabilistic language model," Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003.

[18] Mikolov, T., "Extensions of recurrent neural network language model," in Proc. ICASSP, pp. 5528-5531, 2011.

[19] Chong, T. Y., Rafael E. Banchs, Chng, E. S., & Li, H., "Modeling of term-distance and term-occurrence information for improving n-gram language model performance," in Proc. ACL, pp. 233-237, 2013.

[20] Klakow, D., "Log-linear interpolation of language model," in Proc. ICSLP, pp. 1-4, 1998.

[21] Charniak, E., et al., "BLLIP 1987-89 WSJ Corpus Release 1," Linguistic Data Consortium, Philadelphia, 2000.

[22] Godfrey, J. J. & and Holliman, E., "Switchboard-1 Release 2," Linguistic Data Consortium, Philadelphia, 1997.

[23] Parihar, N. & Picone, J., "Aurora working group: DSR front-end LVCSR evaluation AU/384/02," Aurora Working Group, European Telecommunication Standards Institute, 2002.