Modeling of term-distance and term-occurrence information for improving n-gram language model performance

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

In this paper, we explore the use of distance and co-occurrence information of word-pairs for language modeling. We attempt to extract this information from history-contexts of up to ten words in size, and found it complements well the *n*-gram model, which inherently suffers from data scarcity in learning long history-contexts. Evaluated on the WSJ corpus, bigram and trigram model perplexity were reduced up to 23.5% and 14.0%, respectively. Compared to the distant bigram, we show that word-pairs can be more effectively modeled in terms of both distance and occurrence.

1 Introduction

Language models have been extensively studied in natural language processing. The role of a language model is to measure how probably a (target) word would occur based on some given evidence extracted from the history-context. The commonly used *n*-gram model (Bahl et al. 1983) takes the immediately preceding history-word sequence, of length $n-1$, as the evidence for prediction. Although *n*-gram models are simple and effective, modeling long history-contexts lead to severe data scarcity problems. Hence, the context length is commonly limited to as short as three, i.e. the trigram model, and any useful information beyond this window is neglected.

In this work, we explore the possibility of modeling the presence of a history-word in terms of: (1) the distance and (2) the co-occurrence, with a target-word. These two attributes will be exploited and modeled independently from each other, i.e. the distance is described regardless the actual frequency of the history-word, while the co-occurrence is described regardless the actual position of the history-word. We refer to these two attributes as the term-distance (TD) and the term-occurrence (TO) components, respectively.

The rest of this paper is structured as follows. The following section presents the most relevant related works. Section 3 introduces and motivates our proposed approach. Section 4 presents in detail the derivation of both TD and TO model components. Section 5 presents some perplexity evaluation results. Finally, section 6 presents our conclusions and proposed future work.

2 Related Work

The distant bigram model (Huang et.al 1993, Simon et al. 1997, Brun et al. 2007) disassembles the *n*-gram into $(n-1)$ word-pairs, such that each pair is modeled by a distance-k bigram model, where $1 \le k \le n-1$. Each distance-k bigram model predicts the target-word based on the occurrence of a history-word located k positions behind.

Zhou & Lua (1998) enhanced the effectiveness of the model by filtering out those word-pairs exhibiting low correlation, so that only the well associated distant bigrams are retained. This approach is referred to as the distance-dependent trigger model, and is similar to the earlier proposed trigger model (Lau et al. 1993, Rosenfeld 1996) that relies on the bigrams of arbitrary distance, i.e. distance-independent.

Latent-semantic language model approaches (Bellegarda 1998, Coccaro 2005) weight word counts with TFIDF to highlight their semantic importance towards the prediction. In this type of approach, count statistics are accumulated from long contexts, typically beyond ten to twenty words. In order to confine the complexity introduced by such long contexts, word ordering is ignored (i.e. bag-of-words paradigm).

Other approaches such as the class-based language model (Brown 1992, Kneser & Ney 1993)

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 233–237,
Sofia, Bulgaria, August 4-9 2013.

use POS or POS-like classes of the history-words for prediction. The structured language model (Chelba & Jelinek 2000) determines the "heads" in the history-context by using a parsing tree. There are also works on skipping irrelevant history-words in order to reveal more informative *n*-grams (Siu & Ostendorf 2000, Guthrie et al. 2006). Cache language models exploit temporal word frequencies in the history (Kuhn & Mori 1990, Clarkson & Robinson 1997).

3 Motivation of the Proposed Approach

The attributes of distance and co-occurrence are exploited and modeled differently in each language modeling approach. In the *n*-gram model, for example, these two attributes are jointly taken into account in the ordered word-sequence. Consequently, the *n*-gram model can only be effectively implemented within a short history-context (e.g. of size of three or four).

Both, the conventional trigger model and the latent-semantic model capture the co-occurrence information while ignoring the distance information. It is reasonable to assume that distance information at far contexts is less likely to be informative and, hence, can be discarded. However, intermediate distances beyond the *n*-gram model limits can be very useful and should not be discarded.

On the other hand, distant-bigram models and distance-dependent trigger models make use of both, distance and co-occurrence, information up to window sizes of ten to twenty. They achieve this by compromising inter-dependencies among history-words (i.e. the context is represented as separated word-pairs). However, similarly to *n*-gram models, distance and co-occurrence information are implicitly tied within the word-pairs.

In our proposed approach, we attempt to exploit the TD and TO attributes, separately, to incorporate distant context information into the *n*-gram, as a remedy to the data scarcity problem when learning the far context.

4 Language Modeling with TD and TO

A language model estimates word probabilities given their 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 its corresponding history. Let the word located at i^{th} position, w_i, be the target-word and its preceding word-sequence $w_{i-1}^{i-n+1} = (w_{i-n+1} \dots w_{i-2} w_{i-1})$ of length $n-1$, be its history-context. Also, in order to alleviate the data scarcity problem, we assume the occurrences of the history-words to be independent from each other, conditioned to the occurrence of the target-word w_i, i.e. $w_{i-k} \perp w_{i-l} | w_i$, where $w_{i-k}, w_{i-l} \in h$, and $k \neq l$. The probability can then 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 that w_{i-k} is the word at position k^{th}.

4.1 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} at the k^{th} position behind the target-word, as the joint of two events: i) the word w_{i-k} occurs within the history-context: $w_{i-k} \in h$, and ii) it occurs at distance k from the target-word: $\Delta(w_{i-k}) = k$, ($\Delta = k$ for brevity); i.e. $(h_k = w_{i-k}) \equiv (w_{i-k} \in h) \cap (\Delta = k)$.

Thus, 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. This can be rewritten in terms of the likelihood function of the distance event (i.e. $\Delta = k$) and the occurrence event (i.e. $w_{i-k} \in h$), where both of them can be modeled and exploited separately, as follows:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{\begin{bmatrix} P(t = w_i) \\ \prod_{k=1}^{n-1} P(\Delta = k | w_{i-k} \in h, t = w_i) \\ \prod_{k=1}^{n-1} P(w_{i-k} \in h | t = w_i) \end{bmatrix}}{Z(h)} \quad (3)$$

The formulation above yields three terms, referred to as the prior, the TD likelihood, and the TO likelihood, respectively.

In Eq.3, we have decoupled 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:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{\begin{bmatrix} P(t = w_i)^{\beta_n} \\ (\prod_{k=1}^{n-1} P(\Delta = k | w_{i-k} \in h, t = w_i))^{\beta_d} \\ (\prod_{k=1}^{n-1} P(w_{i-k} \in h | t = w_i))^{\beta_o} \end{bmatrix}}{Z(h)} \quad (4)$$

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

Notice that the model depicted in Eq.4 is the log-linear interpolation (Klakow 1998) of these models. The prior, which is usually implemented as a unigram model, can be also replaced with a higher order *n*-gram model as, for instance, the bigram model:

$$P\left(t = w_i | h = w_{i-1}^{i-n+1}\right) \approx \frac{\begin{bmatrix} P(t = w_i | h = w_{i-1})^{\beta_n} \\ (\prod_{k=1}^{n-1} P(\Delta = k | w_{i-k} \in h, t = w_i))^{\beta_d} \\ (\prod_{k=1}^{n-1} P(w_{i-k} \in h | t = w_i))^{\beta_o} \end{bmatrix}}{Z(h)} \quad (5)$$

Replacing the unigram model with a higher order *n*-gram model is important to compensate the damage incurred by the conditional independence assumption made earlier.

4.2 Term-Distance Model Component

Basically, the TD likelihood measures how likely a given word-pair would be separated by a given distance. So, word-pairs possessing consistent separation distances will favor this likelihood. The TD likelihood for a distance k given the co-occurrence of the word-pair (w_{i-k}, w_i) can be estimated from counts as follows:

$$P(\Delta = k | w_{i-k} \in h, t = w_i) = \frac{\mathrm{C}(w_{i-k} \in h, t = w_i, \Delta = k)}{\mathrm{C}(w_{i-k} \in h, t = w_i)} \quad (6)$$

The above formulation of the TD likelihood requires smoothing for resolving two problems: i) a word-pair at a particular distance has a zero count, i.e. $\mathrm{C}(w_{i-k} \in h, t = w_i, \Delta = k) = 0$, 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 $\mathrm{C}(w_{i-k} \in h, t = w_i) = 0$, which results in a division by zero.

For the first problem, we have attempted to redistribute the counts among the word-pairs at different distances (as observed within the window). We assumed that the counts of word-pairs are smooth in the distance domain and that the influence of a word decays as the distance increases. Accordingly, we used a weighted moving-average filter for performing the smoothing. Similar approaches have also been used in other works (Coccaro 2005, Lv & Zhai 2009). Notice, however, that this strategy is different from other conventional smoothing techniques (Chen & Goodman 1996), which rely mainly on the count-of-count statistics for re-estimating and smoothing the original counts.

For the second problem, when a word-pair was not seen at any distance (within the window), we arbitrarily assigned a small probability value, $P(\Delta = k | w_{i-k} \in h, t = w_i) = 0.01$, to provide a slight chance for such a word-pair (w_{i-k}, w_i) to occur at close distances.

4.3 Term-Occurrence Model Component

During the decoupling operation (from Eq.2 to Eq.3), the TD model held only the distance information while the count information has been ignored. Notice the normalization of word-pair counts in Eq.6.

As a complement to the TD model, the TO model focuses on co-occurrence, and holds only count information. As the distance information is captured by the TD model, the co-occurrence count captured by the TO model is independent from the given word-pair distance.

In fact, the TO model is closely related to the trigger language model (Rosenfeld 1996), as the prediction of the target-word (the triggered word) is based on the presence of a history-word (the trigger). However, differently from the trigger model, the TO model considers all the word-pairs without filtering out the weak associated ones. Additionally, the TO model takes into account multiple co-occurrences of the same history-word within the window, while the trigger model would count them only once (i.e. considers binary counts).

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

$$P(w_{i-k} \in h | t = w_i) = \frac{\mathrm{C}(w_{i-k} \in h, t = w_i)}{\mathrm{C}(t = w_i)} \quad (7)$$

When a word-pair did not co-occur (within the observation window), we assigned a small probability value, $P(w_{i-k} \in h | t = w_i) = 0.01$, to provide a slight chance for the history word to occur within the history-context of the target word.

5 Perplexity Evaluation

A perplexity test was run on the BLLIP WSJ corpus (Charniak 2000) with the standard 5K vocabulary. The entire WSJ '87 data (740K sentences 18M words) was used as train-set to train the *n*-gram, TD, and TO models. The dev-set and the test-set, each comprising 500 sentences and about 12K terms, were selected randomly from WSJ '88 data. We used them for parameter fine-tuning and performance evaluation.

5.1 Capturing Distant Information

In this experiment, we assessed the effectiveness of the TD and TO components in reducing the *n*-gram's perplexity. Following Eq.5, we interpolated *n*-gram models (of orders from two to six) with the TD, TO, and the both of them (referred to as TD-TO model).

By using the dev-set, optimal interpolation weights (i.e. β_n, β_d, and β_o) for the three combinations (*n*-gram with TD, TO, and TD-TO) were computed. The resulting interpolation weights were as follows: *n*-gram with TD = (0.85, 0.15), *n*-gram with TO = (0.85, 0.15), and *n*-gram with TD-TO = (0.80, 0.07, 0.13).

The history-context window sizes were optimized too. Optimal sizes resulted to be 7, 5 and 8 for TD, TO, and TD-TO models, respectively. In fact, we observed that the performance is quite robust with respect to the window's length. Deviating about two words from the optimum length only worsens the perplexity less than 1%.

Baseline models, in each case, are standard *n*-gram models with modified Kneser-Ney interpolation (Chen 1996). The test-set results are depicted in Table 1.

N	NG	NG-TD	Red. (%)	NG-TO	Red. (%)	NG-TDTO	Red. (%)
2	151.7	134.5	11.3	119.9	21.0	116.0	23.5
3	99.2	92.9	6.3	86.7	12.6	85.3	14.0
4	91.8	86.1	6.2	81.4	11.3	80.1	12.7
5	90.1	84.7	6.0	80.2	11.0	79.0	12.3
6	89.7	84.4	5.9	79.9	10.9	78.7	12.2

Table 1. Perplexities of the *n*-gram model (NG) of order (*N*) two to six and their combinations with the TD, TO, and TD-TO models.

As seen from the table, for lower order *n*-gram models, the complementary information captured by the TD and TO components reduced the perplexity up to 23.5% and 14.0%, for bigram and trigram models, respectively. Higher order *n*-gram models, e.g. hexagram, observe history-contexts of similar lengths as the ones observed by the TD, TO, and TD-TO models. Due to the incapability of *n*-grams to model long history-contexts, the TD and TO components are still effective in helping to enhance the prediction. Similar results were obtained by using the standard back-off model (Katz 1987) as baseline.

5.2 Benefit of Decoupling Distant-Bigram

In this second experiment, we examined whether the proposed decoupling procedure leads to better modeling of word-pairs compared to the distant bigram model. Here we compare the perplexity of both, the distance-*k* bigram model and distance-*k* TD model (for values of *k* ranging from two to ten), when combined with a standard bigram model.

In order to make a fair comparison, without taking into account smoothing effects, we trained both models with raw counts and evaluated their perplexities over the train-set (so that no zero-probability will be encountered). The results are depicted in Table 2.

k	2	4	6	8	10
DBG	105.7	112.5	114.4	115.9	116.8
TD	98.5	106.6	109.1	111.0	112.2

Table 2. Perplexities of the distant bigram (DBG) and TD models when interpolated with a standard bigram model.

The results from Table 2 show that the TD component complements the bigram model better than the distant bigram itself. Firstly, these results suggest that the distance information (as modeled by the TD) offers better cue than the count information (as modeled by the distant bigram) to complement the *n*-gram model.

The normalization of distant bigram counts, as indicated in Eq.6, aims at highlighting the information provided by the relative positions of words in the history-context. This has been shown to be an effective manner to exploit the far context. By also considering the results in Table 1, we can deduce that better performance can be obtained when the TO attribute is also involved. Overall, decoupling the word history-context into the TD and TO components offers a good approach to enhance language modeling.

6 Conclusions

We have proposed a new approach to compute the *n*-gram probabilities, based on the TD and TO model components. Evaluated on the WSJ corpus, the proposed TD and TO models reduced the bigram's and trigram's perplexities up to 23.5% and 14.0%, respectively. We have shown the advantages of modeling word-pairs with TD and TO, as compared to the distant bigram.

As future work, we plan to explore the usefulness of the proposed model components in actual natural language processing applications such as machine translation and speech recognition. Additionally, we also plan to develop a more principled framework for dealing with TD smoothing.

References

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

Bellegarda, J. R. 1998. A multispan language modeling framework for larfge vocabulary speech recognition. IEEE Trans. on Speech and Audio Processing, 6(5): 456-467.

Brown, P.F. 1992 Class-based n-gram models of natural language. Computational Linguistics, 18: 467-479.

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

Cavnar, W.B. & Trenkle, J.M. 1994. N-gram-based text categorization. Proc. SDAIR-94, pp.161-175.

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

Chen, S.F. & Goodman, J. 1996. An empirical study of smoothing techniques for language modeling. In. Proc. ACL '96, pp. 310-318.

Chelba, C. & Jelinek, F. 2000. Structured language modeling. Computer Speech & Language, 14: 283-332.

Clarkson, P.R. & Robinson, A.J. 1997. Language model adaptation using mixtures and an exponentially decaying cache. In Proc. ICASSP-97, pp.799-802.

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

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. 2006. A closer look at skip-gram modelling. In Proc. LREC-2006, pp.1222-1225.

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

Katz, S.M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. on Acoustics, Speech, & Signal Processing, 35:400-401.

Klakow, D. 1998. Log-linear interpolation of language model. In Proc. ICSLP 1998, pp.1-4.

Kneser, R. & Ney, H. 1993. Improving clustering techniques for class-based statistical language modeling. In Proc. EUROSPEECH '93, pp.973-976.

Kuhn, R. & Mori, R.D. 1990. A cache-based natural language model for speech recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 12(6): 570-583.

Lau, R. et al. 1993. Trigger-based language models: a maximum-entropy approach. In Proc. ICASSP-94, pp.45-48.

Lv Y. & Zhai C. 2009. Positional language models for information retrieval. In Proc. SIGIR'09, pp.299-306.

Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modelling. Computer Speech and Language, 10: 187-228.

Simons, M., Ney, H. & Martin S.C. 1997. Distant bigram language modelling using maximum entropy. In Proc. ICASSP-97, pp.787-790.

Siu, M. & Ostendorf, M. 2000. Variable n-grams and extensions for conversational speech language modeling. IEEE Trans. on Speech and Audio Processing, 8(1): 63-75.

Zhou G. & Lua K.T. 1998. Word association and MI-trigger-based language modeling. In Proc. COLING-ACL, 1465-1471.