COMPARING RNNS AND LOG-LINEAR INTERPOLATION OF IMPROVED SKIP-MODEL ON FOUR BABEL LANGUAGES: CANTONESE, PASHTO, TAGALOG, TURKISH

Mittul Singh, Dietrich Klakow

Spoken Language Systems, Saarland University, Germany
{mittul.singh, dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

Recurrent neural networks (RNNs) are a very recent technique to model long range dependencies in natural languages. They have clearly outperformed trigrams and other more advanced language modeling techniques by using non-linearly modeling long range dependencies. An alternative is to use log-linear interpolation of skip models (i.e. skip bigrams and skip trigrams). The method as such has been published earlier. In this paper we investigate the impact of different smoothing techniques on the skip models as a measure of their overall performance. One option is to use automatically trained distance clusters (both hard and soft) to increase robustness and to combat sparseness in the skip model. We also investigate alternative smoothing techniques on word level. For skip bigrams when skipping a small number of words Kneser-Ney smoothing (KN) is advantageous. For a larger number of words being skipped Dirichlet smoothing performs better. In order to exploit the advantages of both KN and Dirichlet smoothing we propose a new unified smoothing technique. Experiments are performed on four Babel languages: Cantonese, Pashto, Tagalog and Turkish. RNNs and log-linearly interpolated skip models are on par if the skip models are trained with standard smoothing techniques. Using the improved smoothing of the skip models along with distance clusters, we can clearly outperform RNNs by about 8-11 % in perplexity across all four languages.

Index Terms— RNNs, log-linear interpolation, skip models, smoothing, under researched languages

The work was supported by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction and Software Cluster project EMERGENT (under grant number 01IC10S01O). This work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

1. INTRODUCTION

Statistical language models, essential in speech recognition, machine translation and information retrieval, broadly defines a language modeling task as the prediction of the next word given a history of words (context). Models like n-gram language models are able to capture the regularities (sparseness) of the language whereas word cluster based models better describe the topical information of the data.

Most of speech research until now has focused on a few popular languages like English, Mandarin Chinese and Modern Arabic. Even though under-researched languages like Cantonese, Pashto, Tagalog and Turkish are the *lingua franca* of millions and have a growing set of web pages, not a lot of effort has gone towards developing methods that model these languages.

A major drawback of these techniques has been the limit on context size. Recently, [1, 2] have presented language models which overcome the limitations on context size and hence, model long range dependencies efficiently. Whereas [1] used recurrent neural network (RNN) based language models to capture long range dependencies, in [2] log-linearly interpolated skip n-gram model was used. RNN's connectionist language model outperformed state-of-the-art n-gram models, but had higher training complexity than standard n-grams. Skip n-gram based language models with a training complexity similar to that of an n-gram model also showed on-par performance to RNNs.

In this paper, we compare these two techniques for four under-researched language datasets obtained under the IARPA's Babel program. Such a comprehensive comparison has not been previously done on these languages. We investigate the application of smoothing techniques to skip n-grams for modeling long range dependencies. To improve performance of skip n-grams we experiment with Kneser-Ney [3] and Dirichlet [4] smoothing techniques. In [2], standard smoothing techniques were applied to the skip model. In contrast, we propose a new unification of Kneser-Ney and Dirichlet smoothing technique for the skip models. Quite similar to [2], we also apply word clusters to skip n-grams to alleviate the sparseness problem. However, we use them at word level instead of skip n-gram level in the skip model.

In our experiments with language datasets available under the Babel program, we found that the Kneser-Ney technique worked better for small context skip n-gram model, whereas larger context skip model performed better with Dirichlet smoothing technique. Hence, we combined these two techniques to produce a unified smoothing technique. Applying this smoothing technique to skip model, we outperform the ones smoothed with standard techniques and the RNN based language model. Adding word cluster information to skip model with improved smoothing further improved the performance by 4 %.

The rest of the paper is organized as follows. Section 2 describes the unification of smoothing techniques on bigrams and distance bigrams in detail and gives a brief discussion on different cluster models that have been used in this paper to form the combined model. Section 3 describes the different language corpora, the experimental setup and the results thus obtained. We conclude by discussing our results in section 4.

2. MODEL DESCRIPTION

We want to compare two methods that go beyond the trigram: recurrent neural networks (RNN) and the log-linear interpolation of skip bigrams and skip trigrams. In the past years, [1] used RNNs based language model to capture long range dependencies. This model did not have any limits on the size of the context. Recurrent connections of the neural network allowed it to cycle context information for an arbitrarily long time and provide contexts of arbitrary lengths. This led the RNN based language models to show great improvement in performance over the then state-of-the-art language models.

An alternative is the use of skip models constructed using log-linear interpolation [2]. Unlike RNNs, in this model all long range dependencies are enumerated explicitly using skip bigrams and trigrams. Normally these skip models are smoothed using standard off-the shelf smoothing techniques like the Absolute Discounting variant suggested by Kneser and Ney [3] or the Dirichlet smoothing [4]. Dirichlet smoothing is very successful over long contexts and frequently used in information retrieval applications [5].

In our experiments, we found that the constituents of skip n-grams: bigrams performed well with Kneser-Ney smoothing whereas distance bigrams performed well with Dirichlet smoothing. Hence, we tried a unification of these techniques which improved over the RNN based language model. Section 2.1 describes this unification of smoothing techniques in detail. To further improve the performance of the skip n-gram model, we applied them in a word cluster based language framework. Section 2.2 briefly discusses the clustering methods used to develop the above mentioned application. Section 2.3 details the combination of these techniques generated to form word-cluster based modified skip n-gram model.

Method	$p(w\|h)$
Dirichlet	$\frac{c(w;h)+\mu p_{BG}(w\|h)}{c(h)+\mu}$
Kneser-Ney	$\frac{\max(c(w;h)-\delta,0)}{c(h)} + \frac{\delta c_u(h)}{c(h)} p_{BG}(w\|h)$
Jelinek Mercer	$(1-\lambda)p_1(w\|h) + \lambda p_2(w\|h)$

Table 1. Summary of the smoothing methods used to smooth bigrams and distance bigrams. Here p_{BG} is the background language model to which the smoothing methods backoff.

2.1. Smoothing Techniques For Distance Bigrams And Bigrams

We propose to combine the Kneser-Ney smoothed skip model (p_{KN}) and the Dirichlet smoothed skip model (p_{Dir}). Table 1 gives a summary of the various smoothing techniques used. A combination of these techniques is carried out using the Jelinek-Mercer interpolation method [6]. The unified smoothing based language model (*UniSt*) thus obtained is described as:

$$p_{UniSt}(w|h) = \frac{\lambda_1 p_{Dir}(w|h) + \lambda_2 p_{KN}(w|h) + p_{BG}(w|h)}{\lambda_1 + \lambda_2 + 1}$$

A unified smoothing performed in such a manner shows performance gains when compared to those obtained by individual smoothing. This is also clear from the results of the experiments as discussed in Section 3.2.

2.2. Word Clustering Algorithms

Word clustering techniques group words together on the basis of a notion of their similar context. With respect to words, this theme can be represented by syntactical or semantical features. Thus, clustering techniques provide useful context information which when combined with skip bigram models alleviates the problem of sparseness. The models thus formed show enhanced performance over non cluster based models.

In this paper, we employ Brown's clustering technique [7] and aspect model [8] based word clusters to provide topical information. Brown's clustering technique can be referred to as a hard clustering technique, which assigns explicit membership to each word. In contrast to hard clustering, soft clustering relaxes the explicit membership assigned by hard clustering. Thus one word can belong to more than one cluster in soft clustering. This enables soft clustering to encode more information than hard clustering. However, this algorithm starts to overfit for a large number of clusters. To overcome this problem of overfitting we modify the EM algorithm by additively smoothing [9] the individual E and M steps described as follows:
E-Step:

$$p(l|w_1, w_2) = \frac{p(w_2|l)p(l|w_1) + \eta}{\sum_{k=1}^{C} p(w_2|k)p(k|w_1) + \eta C}$$

M-Step:

$$p(w_2|l) = \frac{\sum_{w_1 \in V} c(w_2; w_1)p(l|w_1, w_2) + |V|\eta}{\sum_{w_2'} \sum_{w_1 \in V} c(w_2'; w_1)p(l|w_1, w_2') + |V|^2\eta}$$

$$p(l|w_1) = \frac{\sum_{w_2 \in V} c(w_2; w_1)p(l|w_1, w_2) + |V|\eta}{\sum_{w_2} \sum_{w_1' \in V} c(w_2; w_1')p(l|w_1', w_2) + |V|^2\eta}$$

	Cantonese	Pashto	Tagalog	Turkish
Training data				
Words	410536	379596	320841	329380
Sentences	34890	28180	32851	41668
Test data				
Words	23071	19261	14670	15020
Sentences	1768	1313	1298	1784
Vocabulary	9932	9361	13431	23794

Table 2. A statistical summary of language datasets. (Vocabulary is measured in number of words)

where $c(w_2; w_1)$ denotes the frequency of bigram pair $(w_1; w_2)$ in the corpus, l is the hidden variable varying from 1 to C and $|V|$ is the vocabulary size. The additive smoothing of individual steps is controlled by the parameter η. In our experiments, using a smaller value of η (in range of 10^{-6}) improved the generalization performance and avoided overfitting.

2.3. The Combined Language Model

The skip bigrams can now be reformulated by applying the techniques already explained in section 2.1 and 2.2. The skip bigram's constituent models: bigrams ($p(w_1|w_2)$) and distance bigrams $\{p(w_1|w_j) : j = 2, 3, \ldots, n\}$, are individually smoothed using the p_{UniSt} (see section 2.1). Simultaneously, word clusters are evaluated over bigrams and distance bigrams (see section 2.2). The *UniSt* bigrams and distance bigrams are then combined with their clustering based counterparts ($p_{soft}(w|h), p_{hard}(w|h)$) through the Jelinek-Mercer interpolation method ($p_c(w|h)$), described as follows:

$$p_c(w_i|w_j) = \sigma_1 p_{UniSt}(w_i|w_j) + \sigma_2 p_{soft}(w_i|w_j) + \sigma_3 p_{hard}(w_i|w_j)$$

where $\sum_{i=1}^{3} \sigma_i = 1$. A final log-linear interpolation yields the following modified skip (*MS*) bigram model:

$$p_{MS}(w_1|h) = \frac{1}{Z_\lambda(h)} p_c(w_1|w_2)^{\lambda_u} \times \prod_{i=2}^{n} \left(\frac{p_c(w_1|w_i)}{p(w_1)} \right)^{\lambda_i}$$

where λ_is are the log-linear interpolation parameters.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We used Cantonese, Pashto, Tagalog and Turkish language datasets in our experiments. The datasets include transcriptions of phone conversations collected under the IARPA Babel Program language collection releases babel101-v0.4c, babel104b-v0.4aY, babel106b-v0.2f and babel105-v0.5. Cantonese is a particular dialect of Chinese spoken in large parts of southern China. It is segmented on a character level. The Pashto language (also known as Afghani and Pathani) is mainly spoken in Afghanistan and Pakistan. Tagalog is one of the main languages spoken in Philipines, and Turkish is the predominantly spoken in Turkey with smaller groups located in Europe and central Asia.

To evaluate language models these datasets were divided into training and test data. We used perplexity as a performance measure. First 200 words of the test set were used as the development set to tune the parameters involved in language models. Training and test corpus size and respective vocabulary sizes are summarized in Table 2.

3.2. Smoothing Techniques For Skip Bigrams

We use the unified smoothing technique (*UniSt*) described in Section 2.1 to compare the performance of bigrams and distance bigrams over a window of five previous words on the Turkish language dataset. Figure 1 shows the variation of performance of distance bigrams on the test set for different distances. As can be seen from the figure, for a smaller context we note that the bigrams and distance bigrams smoothed using the Kneser-Ney smoothing (*KN*) generally performs better than the ones smoothed using the Dirichlet technique (*Dir*). However, for larger contexts *Dir* ($d > 1$)performs better than *KN*. The combined smoothing technique (*UniSt*) takes advantage of both these methods and is thus able to outperform its component smoothing techniques. A similar trend was observed for the other languages demonstrating the robustness of the technique.

3.3. Results

To evaluate performance of the various methods discussed in this paper, we use the language datasets from Cantonese, Pashto, Tagalog and Turkish. A Kneser-Ney smoothed trigram (KN3) is used as the baseline for comparisons with other language models. We report the perplexity results of the *RNN* based language model and the skip model smoothed using Kneser-Ney smoothing (*SkipKN*). We compare these with the unified smoothed skip model (*SkipUniSt*). For further comparisons, we construct another two versions of the modified

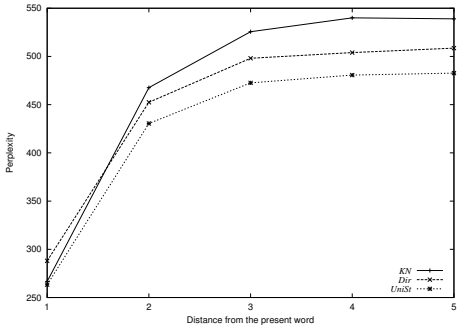

Fig. 1. Variation of perplexity for different distances of the distance bigram

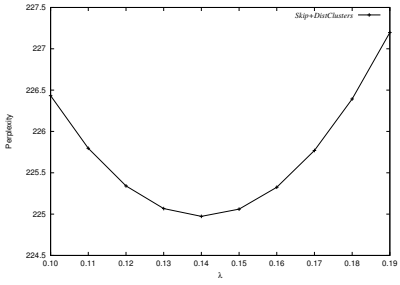

Fig. 2. Variation of test set perplexity of *Skip+DistClusters* with log-linear interpolation parameter. The least value is observed at $\lambda_4 = 0.138$

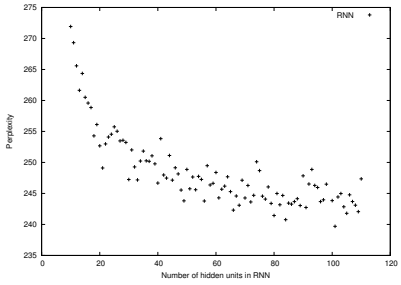

Fig. 3. Variation of test set perplexity for a particular instance of *RNN* for different number of hidden units

skip model by adding cluster information to its constituent models: one includes cluster information only in the distance bigrams (*Skip+DistClusters*) and the other adds word clusters to both bigrams and distance bigrams (*Skip+Clusters*).

We observe the follwoing trends: Both *SkipKN* and *RNN* show on-par results on the various datasets used. *SkipUniSt* outperforms the baseline (*KN3*) on different language datasets by about 8-16 %. In comparison to *RNN*, it gives a performance improvement of 3-8 %, whereas with respect to *SkipKN* it shows an improvement of about 5 %. Adding cluster information to unified-smoothed skip model's component distance bigrams only improves its performance by 1 %. Additionally, if the word clusters are combined with both the component models of skip n-gram an improvement of 4 % is observed. We summarize these results on various languages in table 3.

LM	Cantonese	Pashto	Tagalog	Turkish
KN3	85.206	125.853	144.83	266.322
RNN	85.012	122.432	129.150	245.016
SkipKN	83.457	124.657	130.031	236.871
SkipUniSt	78.563	118.405	123.957	224.277
Skip+DistClusters	77.586	116.382	122.558	223.255
Skip+Clusters	**75.481**	**112.529**	**117.456**	**216.711**

Table 3. Numerical results for different models over Babel languages

3.4. Sensitivity to variation in the meta parameters

We look at the variation of perplexity with meta parameters of both the skip model and RNN based language model. Figure 2 and 3 show the variation of perplexity as a function of each of these methods' meta parameters. As seen in figure 2 and 3, we observe that perplexity varies smoothly with λ_4 for the skip model, whereas the RNN based model shows large variations even for small changes in its number of hidden units. This makes the skip model easier to tune than the RNN based model. Tuning the RNN's meta parameters on the development set can be done using a grid-search based algorithm. However, even this might not be enough to obtain a good performance on the test set.

4. CONCLUSION

Our proposed unified-smoothed skip model was able to outperform state-of-the-art language models. Moreover, it outperformed a recurrent neural network based language model. A simple unification of the smoothing techniques gave 3-8 % improvement over a connectionist based language model across all the four Babel languages. We achieved this by using a lower training complexity model than RNN.

We proposed a unified smoothing technique for skip models. It combines the advantages of both Kneser-Ney and Dirichlet smoothing techniques resulting in enhanced performance of the skip n-gram model. We also applied the word cluster information at a word level to this model which further improved its performance. An addition of word clusters to only the distance bigrams in the skip model showed a minor improvement, whereas adding them to both the bigrams and distance bigrams showed a greater improvement. This suggests that cluster information can be better modeled by small context cluster models.

Sensitivity analysis over meta parameters of skip n-grams and RNN based language mode showed that the former is more robust towards small changes in parameters than the latter one. Smooth variation of perplexity in skip models also makes them easier to tune than a RNN based technique.

Further work might involve combining the skip models with RNNs and their evaluation on standard speech recognition tasks.

Acknowledgements

We would like to thank Mayank Kumar and Michael Wiegand for useful discussions and their suggestions, which helped improve the paper.

5. REFERENCES

[1] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur, "Recurrent neural network based language model," In *INTERSPEECH* [1], pp. 1045–1048.

[2] Saeedeh Momtazi, Friedrich Faubel, and Dietrich Klakow, "Within and across sentence boundary language model," In *INTERSPEECH* [2], pp. 1800–1803.

[3] Hermann Ney, Ute Essen, and Reinhard Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech & Language*, vol. 8, no. 1, pp. 1 – 38, 1994.

[4] David J.C. MacKay and Linda C. Bauman Peto, "A hierarchical dirichlet language model," *Natural Language Engineering*, vol. 1, pp. 1–19, 1994.

[5] Chengxiang Zhai and John Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, Apr. 2004.

[6] Fred Jelinek and Robert L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Proceedings, Workshop on Pattern Recognition in Practice*, Edzard S. Gelsema and Laveen N. Kanal, Eds., pp. 381–397. North Holland, Amsterdam, 1980.

[7] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[8] Lawrence Saul and Fernando Pereira, "Aggregate and Mixed-Order markov models for statistical language processing," in *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Claire Cardie and Ralph Weischedel, Eds., pp. 81–89. Association for Computational Linguistics, Somerset, New Jersey, 1997.

[9] George James Lidstone, "Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities," *Transactions of the Faculty of Actuaries*, vol. 8, pp. 182–192, 1920.

[10] Dietrich Klakow, "Log-linear interpolation of language models," In *ICSLP* [10].