Improving back-off models with bag of words and hollow-grams

Benjamin Lecouteux, Raphaël Rubino, Georges Linarès

▶ **To cite this version:**

Benjamin Lecouteux, Raphaël Rubino, Georges Linarès. Improving back-off models with bag of words and hollow-grams. INTERSPEECH, Sep 2010, Makuhari, Japan. hal-01318103

HAL Id: hal-01318103
https://hal.science/hal-01318103

Submitted on 29 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221478122

Improving back-off models with bag of words and hollow-grams

Conference Paper · January 2010

Source: DBLP

CITATIONS
0

READS
41

3 authors, including:

Benjamin Lecouteux
Laboratoire d'Informatique de Grenoble
79 PUBLICATIONS **284** CITATIONS
SEE PROFILE

Georges Linarès
Université d´Avignon et des Pays du Vaucluse
162 PUBLICATIONS **504** CITATIONS
SEE PROFILE

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Georges Linarès
Retrieved on: 29 October 2016

Improving back-off models with bag of words and hollow-grams

Benjamin Lecouteux, Raphaël Rubino, Georges Linarès

Laboratoire Informatique d'Avignon (LIA), University of Avignon; France
{benjamin.lecouteux, raphael.rubino, georges.linares}@univ-avignon.fr

Abstract

Classical n-grams models lack robustness on unseen events. The literature suggests several smoothing methods: empirically, the most effective of these is the modified Kneser-Ney approach. We propose to improve this back-off model: our method boils down to back-off value reordering, according to the mutual information of the words, and to a new *hollow-gram* model. Results show that our back-off model yields significant improvements to the baseline, based on the modified Kneser-Ney back-off. We obtain a 0.6% absolute word error rate improvement without acoustic adaptation, and 0.4% after adaptation with a 3xRT ASR system.

Index Terms: language model, low-order interpolation, back-off

1. Introduction

Although n-gram models have demonstrated their efficiency in speech recognition systems, they are known to be correlated to the Word Error Rate (WER) in the case of unseen events [1]. The smoothing n-gram model was developed in order to solve the issues associated to the sparse training data. The back-off model addresses this problem by discounting the probability of low-count events and distributing the freed probability mass among unseen events. Many back-off methods have been proposed: additive smoothing [2], which assigns the same probability to all unseen words that follow a particular history; low-order back-off [3] and low-order interpolation [4].

In [5] the authors present an empirical study of smoothing techniques; this work shows that Kneser-Ney and their modified Kneser-Ney smoothing outperform all other algorithms. However, traditional back-off models tend to overestimate some events: [6] propose different ways to tackle this issue, by recalculating back-off and interpolating trigger pairs with the language model.

In [7], the authors address the back-off problem by carrying out the values estimation in a continuous space, allowing for a smooth interpolation of the probabilities. Then, a better generalization to unknown n-grams is expected. However required computations are prohibitive and many optimization techniques are necessary.

Other approaches are based on word similarity. The word similarity is used for language modeling by [8], as a constraint in a maximum entropy model which reduces the perplexity and the WER on the test set. However, this model is more adequate for capturing long-distance language dependencies, rather than short dependencies.

In [9], the authors use a word similarity measure for language modeling in an interpolated model, grouping similar words into classes. Another approach [10] is the back-off estimation based on hierarchical class n-grams. This method allows one to use a hierarchy of classes that lead to a better estimation of the likelihood of unseen events, based on syntactic or semantic groups. Unfortunately, class-based and similarity-based methods require prior knowledge on the words.

However, all these proposed methods do not take into account the errors generated by the Automatic Speech Recognition (ASR) system itself, and the back-off values are not context-dependent. For example, with a Kneser-Ney back-off, the unseen 3-grams: "explosion de pneu" [tire explosion] = α(explosion de)p(de pneu) and "explosion de peu" [explosion of minor] = α(explosion de)p(de peu) share the same $\alpha()$ back-off value. We assume that it is necessary to sort this type of back-off out according to their short context (3-gram).

In the approach introduced by [11], the authors propose a similarity-based model to assign probabilities to unseen 2-grams. They use similar 2-grams computed on the corpus itself, without relying on taxonomic information. Their approach allows one to improve a Katz back-off model, using for each word a list of similar words. However, some additional disambiguation process is required for unseen or low-frequency events.

In [12], the authors introduce Factored Language Models (FLM), where the probability model can be seen as a directed graphical model. A back-off graph is introduced (parallel back-off) allowing to take many paths depending of each gram. This parallel back-off is presented as a generalization of the standard back-off equation. These models allow to compute several back-off paths; then the back-off can be based on linguistic knowledge, a statistical criteria or a combination of multiple paths. Our work lies in this approach where several back-offs compete.

In this paper we propose a simple and efficient back-off model, based on word co-occurrences and distant-gram models (introduced in this paper as *hollow-gram* model). Our method can be applied easily to traditional n-gram language models. In the first section we present our approach. In the second section, we present the experimental framework. Experiments on our back-off model are presented in the last two sections. Finally, we conclude and suggest some perspectives.

2. Proposed approaches

In [1], the correlation between back-off behavior and WER is proven. The table 1 contains the WER arising from the specific back-off on training corpora. The language model coverage is obviously not the only issue, but a biased language model built with tests data improves dramatically the recognition.

Currently, in state-of-the-art ASR systems, the most popular back-off model is the (modified) Kneser-Ney, which outperforms other smoothing techniques. However, the classical smoothing algorithms do not take into account the short-context, as the example presented in introduction. We propose alternative measures allowing to determine the likelihood of a word combination.

We also propose to regulate the back-off values against the possibility of the unseen events: our method is used to check if a word sequence exists. In the case of unobserved word co-occurrences, the value is slightly altered.

The backoff is then re-estimated according to the equation:

This research is supported by the ANR (Agence Nationale de la Recherche), ASH project (ANR-09BLAN-0161-02).

$$\tilde{\alpha}(w_{i-n},..,w_i) = \alpha(w_{i-n},..,w_{i-1})^{1-\beta} p_\phi(w_{i-n},w_i)^\beta$$

Where $\tilde{\alpha}$ is the updated back-off value, α is the initial back-off value, $p_\phi(w_i, w_{i-n})$ is the smoothing back-off function based on co-occurrence (section 2.2) or *hollow-gram* (section 2.1). $p_\phi(w_{i-n}, w_i) = \delta$ if (w_i, w_{i-n}) co-occurrence is unobserved. β is an empirical fudge factor and δ is an empirical penalty value based on the binary possibility of the current n-gram. β and α are tuned with a grid based approach on the dev corpora. We introduce the *hollow-gram*-based model paradigm in the next section.

case of back-off	1-best
3-gram exists	17.7%
back-off 2-gram	28.5%
back-off 1-gram	50.0%

Table 1: WER according to the back-off level

2.1. Back-off based on *hollow-gram* model

In the case of a 3-gram model, the probability $p(w_3|w_1, w_2)$ of an unseen words sequence (w_1, w_2, w_3), is computed with the back-off value of (w_1, w_2), and the conditional probability of $p(w_3|w_2)$. This method considers an independence assumption on the co-occurrence of w_3 and w_1. We introduce a *hollow-gram* model $p_\phi(w_3|w_1)$ providing an alternative to the independence assumption, allowing relationship between short-distant words. The method is handy: a 2-gram model based on the pairs (start, end) of each 3-gram is trained. This model can be assimilated to a regular expression: $(w_1, *, w_3)$ where $*$ is an unobserved word.

In the case of an unseen event, the model backs off on a 2-gram using the back-off value: $p(w_3|w_1, w_2) = \alpha(w_1, w_2)p(w_3|w_2)$. With our hollow-gram model, the equation becomes:

$$\tilde{p}(w_3|w_1, w_2) = \alpha(w_1, w_2)^{1-\beta} p_\phi(w_3|w_1)^\beta p(w_3|w_2) \quad (1)$$

Where $\tilde{p}(w_3|w_1, w_2)$ is the resulting updated probability of the 3-gram, $\alpha(w_1, w_2)$ is the initial back-off value of the word pair (w_1, w_2), $p_\phi(w_3|w_1)$ is the hollow-gram probability and β is an empirical fudge factor. However, in our hollow-gram-based model, back-off always depends on observed events. Then, we propose a generalization to improve the prediction of the back-off on unobserved events.

2.2. Back-off using word co-occurrence

The main idea in the word co-occurrence approach is to combine a word association score to the back-off based on classical n-gram language model. A number of measures have been proposed in the past to evaluate the strength of word co-occurrences. For example, based on mutual information, conditional probabilities or standard statistical measures like the chi-square. In this work, we decide to use the likelihood ratio proposed by [13]. This measure is theoretically the most appropriate with sparse data, and can be used to find pairs of words that occur next to each other with a significantly higher frequency than it would be expected. It allows us to eliminate word frequency effects and to emphasize relations between significant word co-occurrences. The log likelihood ratio between two words w_a and w_b is described in equation 2.

$$\begin{aligned}\psi(w_a; w_b) &= \sum_{ij\in\{1;2\}} log\frac{k_{ij}N}{C_i R_j} \\ &= k_{11} log\frac{k_{11}N}{C_1R_1} + k_{12} log\frac{k_{12}N}{C_1R_2} \\ &+ k_{21} log\frac{k_{21}N}{C_2R_1} + k_{22} log\frac{k_{22}N}{C_2R_2}\end{aligned} \quad (2)$$

Where:
$C_1 = k_{11} + k_{12}, C_2 = k_{21} + k_{22}$
$R_1 = k_{11} + k_{21}, R_2 = k_{12} + k_{22}$
$N = k_{11} + k_{12} + k_{21} + k_{22}$
k_{11} = count of co-occurrences of word w_a and word w_b
k_{12} = count of occurrences of word $w_a - k_{11}$
k_{21} = count of occurrences of word $w_b - k_{11}$
k_{22} = count of tokens in the corpus $- k_{12} - k_{21} + k_{11}$

In order to count the co-occurrences of two words, we use a sliding window of a fixed size s. In our experiments, the windows size is fixed to five words. Word ordering in the window is not taken into account, it can be seen as a *bag of words*. However, we weight the count of co-occurrences of w_a and w_b by the number of words between them in the window: $\tilde{k}_{11} = \frac{1}{d(w_a;w_b)} k_{11}$.

The log likelihood ratio is computed with all the lexicon on the whole corpora. Then, values are normalized in order to compute probabilities. The obtained model $p_\psi()$ is used to weight back-off, as in equation 1. Unlike the hollow-gram model, this approach can be applied to all back-off (2-gram and 1-gram), and in the case of a full back-off behavior the updated probability of the 3-gram becomes:

$$\begin{aligned}&\tilde{p}(w_3|w_1, w_2) = \\ &\alpha(w_1, w_2)^{1-\beta} p_\psi(w_1, w_3)^\beta \alpha(w_2)^{1-\beta} p_\psi(w_2, w_3)^\beta p(w_3)\end{aligned} \quad (3)$$

Where $\alpha(w_1, w_2)$ is the initial back-off value of the words, β is an empirical fudge factor and $p_\psi()$ is the back-off smoothing function based on word co-occurrences obtained by equation 2.

2.3. Combination of word co-occurrence and hollow-gram model

Naturally, our hollow-gram model is unable to work on unseen values, but its strength lies on observed events. However, in the case of unobserved events, we propose to back-off on word co-occurrences. Then the model is based on a double back-off depending on fuzzy-observed events:

$$\tilde{\alpha}(w_1, w_2) = \begin{cases} \alpha(w_1, w_2)^{1-\beta} p_\phi(w_3|w_1)^\beta \text{ if h-gram exists} \\ \alpha(w_1, w_2)^{1-\beta} p_\psi(w_1, w_3)^\beta \text{ else} \end{cases} \quad (4)$$

2.4. Compact model approach

The drawback of the co-occurrence method is the memory consumption (the co-occurrence matrix size is $\frac{|V|^2}{2}$ where V is the vocabulary size). In order to tackle this issue, we propose a low-memory model version of our approach, based on the binary existence of words co-occurrences. If w_a and w_b have been observed in the window, the back-off is used as such, else a penalty is applied.

3. Experimental framework

3.1. The LIA broadcast news system

Experiments are carried out by using the *Laboratoire Informatique d'Avignon* (LIA) broadcast news (BN) system which was used in the ESTER evaluation campaign [14]. This system relies on the HMM-based decoder developed at the LIA: Speeral [15]. Speeral is an asynchronous decoder operating on a phoneme lattice; acoustic models are HMM-based, context dependent with cross word triphones. The language model is a classical 3-gram model, estimated on about 1,3G words from the French newspaper *Le Monde*, the ESTER broadcast news corpus (about 1M words) and Gigaword corpus, without cuts. The lexicon contains 86K words. In these experiments, the first pass is performed in 3x Real Time (RT), while the second one is assessed after an MLLR acoustic adptation in 5xRT.

3.2. The ESTER-2 corpus

The ESTER-2 corpus consists of French radio broadcasts of the Radio-France group. The training and development parts of the data set are based on the training corpus provided for the ESTER-2 (100 hours manually annotated) evaluation campaign.

We test our approach on 6 hours of speech, extracted from the ESTER-2 test set. The mean baseline WER is 30,4% in the first pass and 27.3% after the MLLR adaptation. Mean results are normalized by the number of words per show.

Shows	WER	SER	CWR
Inter (4h)	33.1	75.0	69.7
Tvme (1h)	31.3	67.6	71.2
Rfi (1h)	18.7	66.6	84

Table 2: Baseline WER, SER and CWR

Table 2 contains the baseline Word Error Rate (WER), Sentence Error Rate (SER) and Correct Word Rate (CWR). By using these three measures, we observe the global behavior of the ASR system. We present the shows separately in this Table, because each of them depicts different performance of the ASR system: journalistic speech for the *RFI* show while the *INTER* and *TVME* shows contain more spontaneous speech (phone calls).

4. Results

Throughout experiments, we present three results facets. The first one is the classical WER, while the second is the SER and the last one is the CWR. Using this new back-off, the language model dynamics is altered. In our case, we observe especially an increase of insertions. The CWR allows us to show the real rate of corrected words, and the real impact of our approach.

4.1. Back-off based on hollow-grams

This approach is based on unobserved unigram in a 3-gram $(w_1, *, w_3)$. The order of the words in the 3-gram is taken into account. This model has two advantages: it allows one to use a regular expression based model and to capture hollow-grams events into the training corpora. Results of the hollow-gram-based back-off are reported in Table 3.

Shows	WER	SER	CWR
Inter (4h)	32.8 (-0.3)	74.5 (-0.5)	70.5 (+0.8)
Tvme (1h)	31.3 (0.0)	67.0 (-0.3)	71.7 (+0.5)
Rfi (1h)	18.5 (-0.2)	65.7 (-0.9)	84.4 (+0.4)
Mean	-0.2	-0.5	+0.7

Table 3: WER, SER, CWR using the hollow-gram-based model

These first results show a slight improvement both on WER and CWR, compared to the classical back-off. We can see that global CWR improvement is twice the WER improvement. These results show that the new back-off corrects a lot of errors, but also introduce new ones.

These experiments show that a simple reordering of the back-off values allows one to improve the language model behavior: the back-off weighting disambiguates close assumptions (in terms of back-off values). However, because of the model topology, it is applied only on 2-gram back-offs. In the next section, we present the results based on word co-occurrence.

4.2. Back-off based on word co-occurrences

In these experiments, we combine the back-off values with the word co-occurrence probabilities. In practice, a word co-occurrence symmetric matrix (order is not taken into account) is built on the whole training corpora, using the method presented in section 2.2 with a window size of five words. In order to compute probabilities, we normalize the values. Then, we interpolate them with the initial modified Kneser-Ney back-off model. This model is applied both on 2-gram and unigram back-offs. Results of the co-occurrence-based back-off are reported in Table 4.

Shows	WER	SER	CWR
Inter (4h)	32.7 (-0.4)	74.5 (-0.5)	70.5 (+0.8)
TVME (1h)	31.0 (-0.3)	66.8 (-0.8)	71.8 (+0.6)
RFI (1h)	18.4 (-0.3)	65.6 (-1)	84.5 (+0.5)
Mean	-0.4	-0.6	+0.7

Table 4: WER, SER, CWR using the co-occurrence-based model

It seems that this approach leads to a higher WER and SER improvement compared to the previous one, based on the hollow-gram model. This may be due to the fact that the method is applied both on 2-gram and unigram back-offs. Once more, this handy model is able to improve a classical back-off. In the next section, we propose to combine the two methods.

4.3. Combination of word co-occurrences and hollow-grams

In these experiments, we propose to test the complementarity of the hollow-gram and the word co-occurrence models. The hollow-gram model is used if $(w_1, *, w_3)$ exists. In other cases, the co-occurrence model is applied as presented in section 2.3.

Shows	WER	SER	CWR
Inter (4h)	32.4 (-0.7)	74.6 (-0.4)	70.8 (+1.1)
Tvme (1h)	30.9 (-0.4)	67.2 (-0.4)	72.0 (+0.8)
RFI (1h)	18.4 (-0.3)	65.6 (-1.0)	84.5 (+0.5)
Mean	-0.6	-0.5	+0.9

Table 5: WER, SER, CWR using combination between co-occurrence-based back-off and hollow-gram models

Results of the combination are reported in Table 5. We observe a significant improvement of the results compared to the two previous methods individually. These results show the complementarity of the two models.

4.4. Compact model approach

The compact model depicts the binary possibility of a back-off. This binary possibility is computed from the co-occurrence matrix: if the value is not null, we consider as true the possibility

of the combination. Results of our compact model are reported in Table 6.

Shows	WER	SER	CWR
Inter (4h)	32.7(-0.3)	74.8(-0.2)	70.3(+0.6)
Tvme (1h)	31.1(-0.2)	67.5(-0.1)	71.6(+0.3)
Rfi (1h)	18.6(-0.1)	65.7(-0.9)	84.2(+0.2)
Mean	-0.25	-0.3	+0.5

Table 6: WER, SER, CWR using the compact back-off model

As expected we observe a degradation of the results compared to the combination approach. Nevertheless, with one bit per word-pair, the original back-off is slightly improved.

5. ASR system behavior and second pass

In this section, we propose to analyze the global behavior of our new language model in its best configuration. Table 7 shows insertions, substitutions and deletions on the whole test. We observe that the substitution rate is kept while the insertion rate increases highly and the deletion rate decreases. Globally, we observed that CWR improvement is twice the WER improvement: the high insertion rate shows an issue that can be handled by an accurate stop-list and a higher word penalty. Further investigations are required to understand the model behavior. This aspect show that we don't use in an optimal way our back-off model, and some tuning is possible.

After this set of experiments, we have tested several β values in order to maximize gains both on train and test data, with our co-occurrence and hollow-gram models individually. We observe important disparity on the results (sometimes slightly better than ones resulting from models combination) without stability. However, with the model combination, a better robustness of the ASR system is observed.

Show	ins	del	sub
baseline	2.73	8.58	18.86
combination	3.14	7.57	18.98
GLOBAL	+15%	-12%	+0.6%

Table 7: Behavior of insertions, substitutions and deletions

We propose to apply our back-off model on the second-pass, using the first-pass transcript (Table 5) for MLLR adaptation. Results are reported in Table 8

Shows	base$_{wer}$	BC	base$_{ser}$	BC	base$_{cwr}$	BC
Inter	30.4	29.9	73.1	72.3	72.2	73
Tvme	25.3	24.8	62.8	62.5	77.8	78.6
Rfi	17.0	16.9	64.4	64.0	85.6	86.1
Mean	-0.4		-0.6		+0.7	

Table 8: WER, SER, CWR after the MLLR adaptation on the baseline (base) and the back-off model combination (BC)

We observe an improvement on the three shows: this aspect highlights the reliability of our approach, a simple back-off values reordering can improve a Kneser-Ney based model.

6. Conclusion and future work

In this paper, we propose a new back-off model based on hollow-gram and word co-occurrence models. The proposed back-off is relevant with classical n-gram language models, without major changes. It consists in interpolating a classical (Kneser-Ney) back-off model with word co-occurrence probabilities and a model based on distant 2-grams $(w_1, *, w_2)$ (hollow-grams).

We evaluate our method on 6 hours of French broadcast news, with many strategies. Changing only back-off implementation, this method allows for a 0.6% absolute gain of WER and 0.9% of correct word rate (CWR) on the first pass without acoustic adaptation. Also, a 0.4% absolute gain of WER and 0.7% of WCR on the second pass after acoustic adaptation. Better results are obtained using both the hollow-gram and the co-occurrence models.

We also present a compact version of our model combination, with very low memory consumption: one bit for each word pair possibility. This approach leads to a slight improvement of the classical back-off.

These preliminary experiments demonstrate that the representation of unseen events is not a fully-solved issue yet. We plan to extend this approach to more sophisticated heuristics/algorithms where co-occurrence values are depending on the word distances, and hollow-gram model extended to n-grams.

7. References

[1] U. Berdy, C. Uhrik, and W. Ward, "Confidence metrics based on n-gram language model backoff behaviors," in *in Proc. EUROSPEECH*, 1997, pp. 2771–2774.

[2] G. Lidstone, "Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities," in *Transactions of the Faculty of Actuaries, 8:182-192.*, 1920.

[3] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *Acoustics, Speech, and Signal Processing*, vol. 35, pp. 400–401, 1987.

[4] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. Int Acoustics, Speech, and Signal Processing ICASSP-95. Conf*, vol. 1, 1995, pp. 181–184.

[5] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer speech and language*, vol. 13, pp. 359–394, 1999.

[6] R. Rosenfeld and X. Huang, "Improvements in stochastic language modeling," in *HLT '91: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1992, pp. 107–111.

[7] H. Schwenk and J.-L. Gauvain, "Connectionist language modeling for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, 2002.

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

[9] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 18–4, 1990.

[10] I. Zitouni, "Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition," *Computer Speech and Language*, vol. 21, pp. 88–104, 2007.

[11] I. Dagan, L. Lee, and F. C. N. Pereira, "Similarity-based models of word cooccurrence probabilities," in *Machine Learning*, 1999, pp. 34–1.

[12] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *in Proceedings of HLT/NACCL*, 2003, pp. 4–6.

[13] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, p. 74, 1993.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich trancription of french radio broadcasts," in *Interspeech*, 2009.

[15] G. Linares, P. Nocera, D. Massonie, and D. Matrouf, "The lia speech recognition system : from 10xrt to 1xrt," in *Lecture Notes in Computer Science, 4629 LNAI*, 2007, pp. pp. 302–308.