Class Phrase Models For Language Modeling

Klaus Ries
ries@cs.cmu.edu

Finn Dag Buø
finndag@ira.uka.de

Alex Waibel
ahw@cs.cmu.edu

Interactive System Labs
Carnegie Mellon University, USA
University of Karlsruhe, Germany

ABSTRACT

Previous attempts to automatically determine multi-words as the basic unit for language modeling have been successful for extending bigram models [10, 9, 2, 8] to improve the perplexity of the language model and/or the word accuracy of the speech decoder. However, none of these techniques gave improvements over the trigram model so far, except for the rather controlled ATIS task [8]. We therefore propose an algorithm, that minimizes the perplexity improvement of a bigram model directly. The new algorithm is able to reduce the trigram perplexity and also achieves word accuracy improvements in the Verbmobil task. It is the natural counterpart of successful word classification algorithms for language modeling [4, 7] that minimize the leaving-one-out bigram perplexity. We also give some details on the usage of class finding techniques and m-gram models, which can be crucial to successful applications of this technique.

1.   Introduction

The selection of a basic unit for language modeling is not necessarily naturally given. In languages such as English and German, which are the focus of this investigation, the word level seems to be a useful abstraction. For Asian languages such as Chinese, Korean and Japanese, however, the basic basic unit is usually chosen at a subword level. The automatic selection of basic units has the advantage, that the bias of simple segmentation criteria is relaxed and important longer units are modeled explicitly. We select the basic unit by successive joins of basic units, and we start with English resp. German words. This has the following applications:

- the fixed context of the language model is enhanced dynamically depending on the length of the basic units
- fixed expressions are very likely to have pronunciations different from the individual words (e.g.*going_to, you_know, you_all*)
- the output of the speech decoder contains more linguistic information than the word string

The first item has been of much help to bigram models in the past, and a lot of researchers reported improvements in this arena. The second application could be realized by introducing specialized pronunciation variants for basic units like *going_to* instead of merely concatenating the pronunciations of *going* and *to*. This could be achieved by manual dictionary modification, dictionary learning or clustering of senonens. The third application is still very speculative: [6] used mutual information to find linguistically motivated segments, [1] calls for grammar inference methods to find simple syntactical finite state grammars.

Since the successive joins of basic units produces a possibly large number of types of basic units, the data sparseness problem becomes more serious. One approach to overcome this problem is to use classes of words and to use these word classes as the basic units to join. This is also the approach we want to follow here, though we find little evidence, that searching for phrases of words can be improved by searching for phrases of word classes for the purpose of language modeling in speech recognition.

2.   The Bigram Leaving-One-Out Perplexity Criterion

The objective of the phrase finding procedure is to find a pair of basic units, that cooccur frequently, such that joining all occurrences in the corpus is a useful operation. After a pair is selected we replace all occurrences of that pair by a new phrase symbol throughout the corpus. In the past most implementations of this idea made use of measures of cooccurrence (except for [2]), that have been useful in other domains, and the pair is chosen by maximization on that criterion. Well known measures are

- mutual information [6] **MI**
- frequency $p(w_1, w_2)$
- iterative marking frequency [9]
- backward bigram **BB**: $p(w_1|w_2)$
- backward perplexity **BP**: $p(w_1, w_2) \cdot \log(p(w_1|w_2))$
- Suhotin's measure [11], see also [9]

In contrast to these criteria one can try to maximize the desired criterion directly, which is the perplexity. The

maximum likelihood estimate of the bigram probability $\prod_{i=1}^{n} p(w_i|w_{i-1})$ of the training set is:

$$F'_{ML} = \prod_{i=1}^{n} \frac{N(w_i, w_{i-1})}{N(w_{i-1})} = \frac{\prod_{w,w'} N(w, w')^{N(w,w')}}{\prod_w N(w)^{N(w)}}$$

Taking the logarithm and rearranging the term we get:

$$F_{ML} = \sum_{w,w'} N(w, w') \cdot \log(N(w, w')) - \sum_w N(w) \cdot \log N(w)$$

The probabilities should be determined on a separate cross validation set and we will therefore minimize the leaving-one-out bigram perplexity of the resulting model along the lines of [4, 7]:

$$F_{LO} = \sum_{w,w',N(w,w')>1} N(w, w') \cdot \log(N(w, w') - 1 - b)$$
$$+ \; n_1 \cdot \log \frac{(n_+ - 1)b}{n_0 + 1} \; - \sum_w N(w) \cdot \log(N(w) - 1)$$

where b is an absolute discounting factor, $N(\cdot, \cdot)$ is the bigram table, n_1 is the number of bigrams occuring exactly once, n_+ is the number of bigrams occuring at least once and n_0 is the number of bigrams not occuring in the corpus. F_{LO} can be calculated for the original corpus as well as the corpus with the selected pair $\langle A, B \rangle$ joined. Since we are in general most interested in the change of F_{LO} after joining A and B to $\langle A, B \rangle$ relative to the old corpus we call this quantity $\Delta_{\langle A,B\rangle} F_{LO}$. F_{LO} as stated above is not a valid measure unless $N(w) > 1$ for all w which is wrong for most corpora and would require smoothing. For all practical purposes we are only interested in $\Delta_{\langle A,B\rangle} F_{LO}$ and this term drops out for almost all w.

One could of course also attempt to minimize the corresponding m-gram perplexity for $m > 2$, but for reasons of computational tractability we attempt the bigram case only. The monogram case of this criterion is very similar to the multigram model [2] using the viterbi-assumption, however, the model evaluation of [2] is not done using the convenient leaving-one-out criterion. The bigram leaving-one-out perplexity criterion (**PP**) can also reflect information, which is obtained from the context of a phrase. Traditional criteria for grammar inference evaluate just the gain of a rule to the constituents used for the join, whereas **PP** applies a simple but effective statistical model to measure local effects on neighboring words. Noting that $\Delta_{\langle A,B\rangle} F_{LO}$ also allows us to reject word pairs from being considered as candidates for a possible join, we can still maximize a different measure, say **X**. The resulting measure will be called **hybrid-X**.

Under the assumption that $A \neq B$ we can simply go through both the bi- and trigram table once and calculate $\Delta_{\langle A,B\rangle} F_{LO}$ for all $\langle A, B \rangle$. A similar technique was applied in many implementations of [4] and elaborated in [7]. Furthermore the trigram table can be calculated incrementally after a pair $\langle A, B \rangle$ is joined from the trigram table for all trigrams that contain $\langle A, B \rangle$. One small sacrifice of this procedure is, that the bigram prediction of $\langle A, B \rangle$ after $\langle A, B \rangle$ is made as $p(\langle A, B \rangle | B)$. To show the principle we ignore the more tedious cases where we have to update n_+, n_0 or n_1 and also ignore the $\sum_w N(w) \cdot \log(N(w) - 1)$ term.

We initialize

$$\Delta_{\langle A,B\rangle} F_{LO} := -N(A, B) \cdot \log(N(A, B) - 1 - b)$$

For each trigram w_1, w_2, w_3 in the corpus we have to add to $\Delta_{\langle w_1,w_2\rangle} F_{LO}$ (and similarly $\Delta_{\langle w_2,w_3\rangle} F_{LO}$) the following terms:

1. New model, bigram $\langle w_1, w_2 \rangle, w_3$:

$$N(w_1, w_2, w_3) \cdot \log(N(w_1, w_2, w_3) - 1 - b)$$

2. New model, bigram w_2, w_3:

$$\Delta N(w_2, w_3) \cdot \log(\Delta N(w_2, w_3) - 1 - b)$$

where $\Delta N(w_2, w_3) := N(w_2, w_3) - N(w_1, w_2, w_3)$

3. Old model, bigram w_2, w_3:

$$-N(w_2, w_3) \cdot \log(N(w_2, w_3) - 1 - b)$$

The leaving-one-out criterion does not dictate the phrase finding procedure we described above. For the corpora we worked with, however, this technique was sufficiently fast. A procedure with possible applications to very large corpora like Wall Street Journal should not try to scan the whole corpus for each phrase. In the spirit of the iterative marking frequency [9] a framework, that scans the corpus less frequently, could look like:

1. Find a potential large (ranked) list of candidate phrases according to $\Delta_{\langle A,B\rangle} F_{LO}$ or some other criterion.

2. Calculate a bigram table of the corpus, where this list was used to join basic units.

3. Calculate ΔF_{LO} for all splits of the phrases.

4. Exclude those phrases that did not improve the perplexity and calculate a ranked list of the phrases according to ΔF_{LO}. Goto 2 or 5.

5. Use the list calculated in step 4 and join this list of phrases in the corpus. Add this list to the already found phrases. Make this corpus the current corpus and goto 1 or **STOP**.

The crucial point is the calculation of ΔF_{LO} in step 3. Since we have to calculate ΔF_{LO} for all possible ways of splitting the phrase it is convenient to restrict ourselves to pairs of word. The calculation can be done by just examining the bigram table in a fashion similar as shown above.
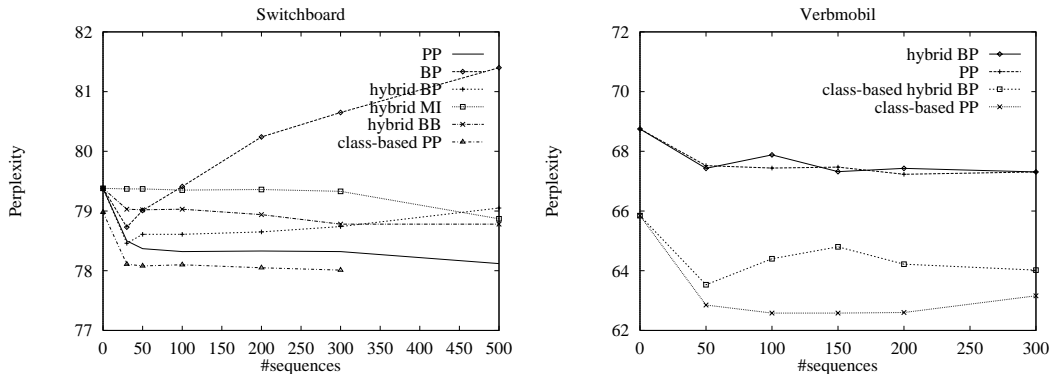
Figure 1: Perplexity results on Switchboard and Verbmobil: The two graphs show results using different phrase finding criteria on word and class-phrases for the Switchboard and Verbmobil corpora. The newly proposed **PP** compares very favorable. For the small Verbmobil corpus the class-phrases show a much smoother plot than the word-phrases.

3. Data Driven Word Classification

The words were classified using unsupervised word classification according to the bigram perplexity criterion [4]. Many authors either use a fixed number of classes as [4, 7] or let the criterion decide, how many classes to choose. In the current formulation of [4], the model prior is a uniform distribution. We added a Gaussian prior on the number of classes we found, since in most cases the optimal number of classes for the trigram model is higher than the one chosen by the uniform prior. We also added a phase, that also allows two clusters to be merged.

4. M-gram Training and Decoding

To use the class-phrase model in the decoder we have to include all phrases of words $\langle w_1, \ldots, w_l \rangle$, such that w_1 is in class c_1, w_2 is in class c_2, etc. for all phrases $\langle c_1, \ldots, c_l \rangle$ in the class-phrase model to the decoder dictionary and language model vocabulary. All word-phrases $\langle w_1, \ldots, w_l \rangle$, that belong to $\langle c_1, \ldots, c_l \rangle$, belong to one class that could be denoted with the label $\langle c_1, \ldots, c_l \rangle$. To train the class-phrase-trigram model we join all word-phrases that can be joined using the class-phrases. One could then simply train a trigram model on this corpus without classes, with the original classes, just with the classes of phrases or with classes of words and phrases. In the calculation of the class based trigram model one has to calculate $p(w|c)$. For classes of phrases this quantity can either be estimated from the data directly or be calculated as $p(\langle w_1, \ldots, w_l \rangle | \langle c_1, \ldots, c_l \rangle) = p(w_1|c_1) \cdot \ldots \cdot p(w_l|c_l)$. A linear interpolation scheme could be used to combine these different models.

5. Experiments

We will present experiments of the Switchboard and the Verbmobil corpus. The Switchboard corpus is a collection of English spontaneous dialogs between 2 unknown parties via telephone with a pregiven topic out of a selection of 70

topics. The training corpus is roughly 2 million words long. The Verbmobil corpus we used for training contains 278.000 words and is a collection of spontaneous German appointment negotiations. Naturally one would expect the corpus with less data, the Verbmobil corpus, to profit more from class based methods. Another expectation would be, that the more restricted domain, again Verbmobil, will profit more from phrase models than the less restricted one. In preexperiments only **MI**, **BB**, **BP** and their hybrid variants as well as **PP** delivered competitive performance. To train the trigram model we used an improved backoff-model [5].

In figure 1 the perplexity results on the Switchboard corpus are shown. As one can see, the perplexity criterion performs the best among all criteria and for the BP criterion one can observe, that using the hybrid model considerably restricts the problems of the original criterion. The class-based PP model shows, that the introduction of classes does not change the shape of the curve and preserves the advantages of the class based model. Not shown in figure 1 is an interpolation experiment, where the class-based class-phrase model is interpolated with the corresponding class-phrase trigram model without classes. Interpolating a class-based and a non-class-based model without phrases, which themselves have perplexities of 79.38 resp. 78.98, yields a model with a perplexity of 77.61. The perplexity of the class-phrase model had been reduced by interpolating with a model without classes from 78.01 to 76.81. However, we achieved roughly the same performance using a model based on word-phrases using a class-based trigram model (class-based word-phrase model).

For the Verbmobil corpus we found no significant improvement from interpolating class-based and non-class-based models. However, the class based model is far better than the standard model and it would therefore be very favorable to use this in conjunction with the phrase model. The qualitative result, that the **PP** criterion is superior to the other models in terms of perplexity, is showing again. In figure 2 the first 50 word-phrases found in the Verbmobil corpus can

be seen. On Verbmobil we were also able to improve the word accuracy of the decoder: The standard trigram model achieves 70.5%, the phrase model without classes achieves 71.4% and the class-based trigram model without phrases achieves 71.5%. If we use a class-based word-phrase trigram model we achieve a word accuracy of 72.1%. However we have not been able to produce good word accuracy results for a class-phrase model on the Verbmobil corpus. One variation we tested was using a small but accurate set of word classes. These automatically derived classes encoded days of the week, months, ordinal numbers, **morning/afternoon**, two variations of **before** and two noise words. The only class-phrases containing non-single word classes were of the types **monday_the** and **eightteenth_and**. This type of phrases was not found in the word-phrases at all.

hab'_ich bin_ich Name_ist w"urd'_ich mit_Ihnen bis_zum
bei_Ihnen in_der wir_uns E_R wir_das Ihnen_das
ich_w"urde kann_ich halten_Sie wir_k"onnten hier_ist
lassen_Sie sagen_wir L_E den_ganzen wir_'s N_I h"att'_ich
habe_ich w"ar'_'s da_bei_Ihnen mir_aus neun_bis
wir_m"ussen h"atte_ich wir_k"onnen h"atten_Sie
U_E treffen_wir_uns E_H treffen_uns T_N wir_sollten
vierzehn_bis Sie_mir es_geht ich_Sie ein_un' f"ur_mich
ich_k"onnte A_L w"urde_ich mu"s_ich f"ur_ein

Figure 2: Word-phrases in Verbmobil: The first 50 phrases found in the Verbmobil corpus according to the **PP** criterion are shown. The vocabulary of the Verbmobil corpus itself contains some phrases such as **Acht-Uhr-Termin** (eight o'clock appointment) and **herzlichen_Dank** (thanks a lot).

6.   Conclusion and Future Research

We have shown that the leaving-one-out bigram perplexity criterion is effective in reducing the perplexity and superior to other criteria proposed so far and we have shown an effective procedure to calculate it. Using this we can turn improvements in perplexity into improvements in word accuracy on the Verbmobil corpus. The combination of class-based and phrase models has proven to combine well. However we have found only little evidence that searching for class-phrases instead of word-phrases is helpful in terms of perplexity and we haven't been able to achieve good word accuracy results with this model. We have also seen, that the class-phrases are not just a smoothing technique to find all important word-phrases but rather find different phrases. In similar experiments we have applied word-phrase models on a corpus of spontaneous Spanish appointment negotiations and found similar perplexity and word accuracy results for the word-phrase model. We have investigated the use of word-phrase and class-phrase models for the Switchboard corpus as well, however a similar reduction in perplexity could not be turned into word accuracy improvements. The main reason for this might be found in the higher regularity of the Verbmobil task and the lower word accuracy rates of current Switchboard speech decoders.

Finally we have proposed a framework to use this criterion on very large corpora like Wall Street Journal. The application of phrase m-gram models on very large corpora seems to be promising, since simply using fixed length m-gram models with $m > 3$ may be less appropriate than the more dynamic notion of context achievable with phrases. Another application of this criterion could be in the inference of syntactical grammars, that would be based on a very large corpus of word tags. A pilot experiment on a tagged Verbmobil corpus has shown that we are able to produce similar perplexity improvements on this type of corpus as well. Yet another application would be a **hybrid-salience** model, where the phrases are used to enhance the salience of the text [3].

7.   Acknowledgments

8.   REFERENCES

1. Steven Abney. *Corpus-Based Methods in Language and Speech*, chapter Part-Of-Speech Tagging and Partial Parsing. ELSNET. Kluwer Academic Publishers, Dordrecht, 1996.

2. Sabine Deligne and Frederic Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigram. In *ICASSP*, 1995.

3. Allen Gorin. On automated language acquisition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

4. Reinhard Kneser and Herman Ney. Improved clustering techniques for class-based statistical language modeling. In *Eurospeech*, Berlin, Germany, 1993.

5. Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *ICASSP*, 1995.

6. David M. Magerman and Mitchell P. Marcus. Distituent parsing and grammar induction. pages 122a–122e.

7. Sven Martin, Joerg Liebermann, and Hermann Ney. Algorithms for bigram and trigram clustering. In *Eurospeech*, 1995.

8. Michael K McCandless and James R Glass. Empirical acquisition of language models for speech recognition. In *ICSLP*, Yokohama, Japan, 1994.

9. Klaus Ries, Finn Dag Buø, and Ye-Yi Wang. Improved language modeling by unsupervised acquisition of structure. In *ICASSP*, 1995.

10. B. Suhm and A. Waibel. Towards better language models for spontaneous speech. In *ICSLP*, Yokohama, Japan, 1994.

11. B. V. Suhotin. Methode de dechiffrage, outil de recherche en linguistique. *TA Informationes*, 2:3–43, 1973.