

# MULTILINGUAL STOCHASTIC N-GRAM CLASS LANGUAGE MODELS

Michèle Jardino

LIMSI-CNRS, ORSAY, France

## ABSTRACT

Stochastic language models are widely used in continuous speech recognition systems where *a priori* probabilities of word sequences are needed. These probabilities are usually given by n-gram word models, estimated on very large training texts. When n increases, it becomes harder to find reliable statistics, even with huge texts. Grouping words is a way to overcome this problem.

We have developed an automatic language independent classification procedure, which is able to optimize the classification of tens of millions of untagged words in less than a few hours on a Unix workstation. With this language independent approach, three corpora each containing about 30 million words of newspaper texts, in French, German and English, have been mapped into different numbers of classes. From these classifications, bi-gram and tri-gram class language models have been built. The perplexities of held-out test texts have been assessed, showing that tri-gram class models give lower values than those obtained with tri-gram word models, for the three languages.

## 1. TRAINING: AUTOMATIC CLASSIFICATION OF UNTAGGED WORDS

Starting with very large texts, n-gram word probabilities are directly evaluated from counts. Different smoothings algorithms are used to interpolate the probabilities of rare or unseen n-gram events: the best ones relate n-gram probabilities to lower order probabilities [1], [2], [3]. Nevertheless, these probabilities become less and less accurate as n increases.

A method to get correct statistics is to generalize, gathering words into classes so that the probability of a given word depends on its class and on the classes of the preceding words [4].

### 1.1. Classification criterion

If T is the training text, the classification criterion is deduced from the probability that the word string T,  $w_1 \dots w_i \dots w_N$ , exists. Reducing the context of each word to the preceding one, this probability becomes:

$$P_T^{(w)} = P(w_1) \prod_{i=2}^N P(w_i | w_{i-1})$$

Gathering words into C classes,  $P_T^{(w)}$  averages to  $P_T^{(C)}$ ,

defined as:

$$P_T^{(C)} = P(w_1) \prod_{i=2}^N P[w_i | C(w_i)] * P[C(w_i) | C(w_{i-1})]$$

where  $C(w_i)$  is the class which contains the word  $w_i$ . This expression assumes :

- only one class  $C(w_i)$  for each word  $w_i$ , but the class  $C(w_i)$  can contain more than one word;
- a word distribution within each class, which depends on the word occurrences.

It is clear that  $P_T^{(C)}$  is always lower than  $P_T^{(w)}$ , except when each vocabulary word corresponds to its own class and in this case,  $P_T^{(C)}$  equals  $P_T^{(w)}$ . Thus, for a given number of classes C, smaller than V, the size of the vocabulary, the best classification is the one for which  $P_T^{(C)}$  is the closest in value to  $P_T^{(w)}$ .

Maximizing  $P_T^{(C)}$  is strictly equivalent to minimizing the text perplexity [1] or to maximizing its average mutual information [4]. These last two values are derived from the logarithm of  $P_T$  and are easier to manipulate. We have used the second one which is especially well-adapted to classification tasks as it excludes the constant part due to the occurrences of the words in the training text, and keeps only the occurrences of the consecutive classes, taken two at a time.

### 1.2. Classification method

Classifying V vocabulary words ( $V > 20000$ ) of a text T of N words, into C classes ( $C \sim 1000$ ) is a hard computational problem, which can only be solved heuristically.

Several iterative methods have been proposed to realize automatic mappings [4] [2], but these methods depend on the order in which the classification processes was carried out. We have already shown how simulated annealing can give optimal classifications [5]. The main advantage of this algorithm is that it overcomes local solutions, so that the final result does not depend on the initial conditions or on the way the process was driven. This is an essential feature of the method, which permits us to choose less constrained initial conditions and to randomly select words and classes during the iterative process, without affecting the result.

In order to limit the computational space, our choice was to start with all words in the same class, and let randomly selected words go to randomly selected classes, the number

of classes being fixed for each process.

In our optimization process, the variation of the average mutual information between classes is computed at each step. Moving a word from one class to another one, changes at most  $4 \times C$  parameters : the occurrences of the involved consecutive classes. In this way, the classification depends both on the left and right context of the groups of words. The total number of moves, empirically determined, is two hundred times the size of the vocabulary. Usually the most frequent words are fixed after a few steps. At the end process, the perplexity has almost reached a flat minimum.

### 1.3. Classification results

Three text corpora, each containing about 30 millions of words, have been used to train classifications. The corpora are newspaper-based, and are respectively written in English (*Wall Street Journal*), French (*Le Monde*) and German (*Frankfurter Rundschau*) languages. They respectively contain 160,000, 260,000 and 620,000 different words. The vocabulary  $V$ , has been defined to the 20,000 most frequent words for each text. Words are strings between two blanks and punctuations are considered to be words. The out-of-vocabulary words are regarded as unknown words and gathered in the same class, they represent 4% of the English and French texts and 9% of the German text.

#### 1.3.1. Classification vs the number of classes

Different classifications have been made, varying  $C$ , the number of classes. The Table 1 gives perplexities of the training texts after optimization. We recall that the text perplexity can be seen as the average number of words which can follow any word sequence in the text (here sequences are reduced to one word).

C	English	French	German
1	742	545	504
500	154	113	135
1000	133	103	125
1500	121	98	119
2000	112	94	115
3000	102	89	108
20000	77	73	84

Table 1. Training text perplexities against  $C$ , the class number

The perplexity decreases from the highest value (over 500), when all words are gathered in the same class ( $C = 1$ ), to the lowest value (about 80), when each vocabulary word is alone in its own class ( $C = V = 20000$ ).

#### 1.3.2. Class contents

The analysis of the class contents shows some interesting features which have already been described for a French corpus [6]. We find the same kind of word groupings for the English and German corpora. Some examples of classes in English are listed below, with the corresponding word occurrences in the training text given in parentheses:

- classes with only one very common word
  - *punctuations*: . (914832) : (33100) ; (29752)

- *function words*: of(524012) to(511417)  
for(188099) which(50156) such(19954) ...

- syntactical classes

- *plural nouns*: businesses(5148) changes(4172)  
computers(2752) interests(2255) policies(2206)  
agencies(2000) schools(1790) owners(1744) parties(1535) ...
- *infinitive verbs*: reduce(3742) begin(2964)  
stop(1769) protect(1327) affect(1081) ...  
disclose(924) oppose(561) enable(492) propose(479) persuade(444) ...
- *past participles*: called(7912) given(3167) included(2876) approved(2752) produced(1720)  
cited(1381) ordered(1363) wrote(1353) rejected(1203) headed(1172) ...
- *comparative adjectives*: higher(9397) lower(8320)  
stronger(1535) slower(409) narrower(180) ...

- semantic classes

- *last names*: Clinton(15729) Bush(2024) Reagan(925) Kantor(505) Murdoch(256) ...
- *countries*: Japan(6841) America(5090) Germany(3008) Russia(2801) France(2312) ...

- semantic classes combined with syntactical classes

- *adverbs of position*: out(24530) ahead(2411) directly(2301) forward(1119) aside(942) ...
- *adjectives related to countries*: Canadian(5308) British(4157) Swiss(1369) Swedish(624) Belgian(375) ...

These contents reflect the short range context which has been taken into account. Other classes are not so well defined, but this mapping is the one which insures the lowest perplexity value, which is the main goal for speech recognition tasks.

## 2. EVALUATION WITH BI-GRAM AND TRI-GRAM CLASS MODELS

Classifications have been done with a bi-gram class constraint. Then, bi-gram and tri-gram class models have been built from these classifications, and compared with equivalent bi-gram and tri-gram word models.

### 2.1. Interpolation

Although there are fewer parameters to determine for an n-gram class model than for an n-gram word model, all events are not predicted by such a model and interpolations are still needed.

A fixed quantity, less than one, is deduced from all observed n-gram class occurrences and the corresponding mass is redistributed according to less specific distributions [2].

We have evaluated different kinds of redistributions: uniform among the classes; non uniform according to lower order probabilities, interpolating tri-gram probabilities with bi-gram and uni-gram probabilities (back-off).

Unknown words are treated in the same way as out-of-vocabulary words (*oov*), found in the training text; 140,000, 240,000 and 600,000 words are respectively out of the 20,000

word vocabulary in the English, French and German training texts. If  $N_T^V(ooov)$  is the number of occurrences of the out-of-vocabulary words in the training texts then the probability of an unknown word is simply assumed to be  $1/N_T^V(ooov)$ . Probabilities have been interpolated in the same way for the n-gram class and n-gram word models.

## 2.2. Assessments

In order to assess these language models, we have evaluated and compared perplexities of withheld texts, calculated from the probabilities given by the different models. The held-out texts were extracted from the same newspaper text sources than the training texts but at other dates. They contain respectively 2,500,000, 2,100,000 and 1,100,000 words for the English, French and German tests. The Table 2 gives test perplexities for bi-gram and tri-gram class models and for the bi-gram and tri-gram word models for the three languages, the interpolations are of the same kind for the word and class models, in order to allow for comparison of the results.

C	English		French		German	
	$PP_b$	$PP_t$	$PP_b$	$PP_t$	$PP_b$	$PP_t$
1	747		533		490	
500	161	115	112	87	134	102
1000	144	99	106	82	126	95
2000	130	88	102	77	121	89
3000	124	86	102	77	120	87
20000	114	111	95	91	109	100

**Table 2.** Test text perplexities against C, the class number

When words are gathered in only one class, the test perplexity depends only on the unigram probabilities and its value is the same for all orders, n, of the different n-gram models.

For bi-gram class models, identical test set perplexities are obtained using uniform and non-uniform redistributions, this is because this type of model is well-trained, with high percentages of recognized bi-grams. Nevertheless, the bi-gram word models obtained from the same training texts, give lower test perplexities, thus high-order n-gram class models are required to improve model accuracy.

We have observed a different behavior for the tri-gram class models which depends more on the kind of interpolation. The best models have been obtained with the back-off method. In this case, the test perplexities obtained with the tri-gram class models are lower than those determined with the tri-gram word models, and are closed to their limit near 2000 classes.

Thus, the tri-gram class model seems to be a good compromise between an efficient coverage of the language and sufficient accuracy in the determination of the probabilities of succession of the words. The perplexities values seem unexpectedly low, but we have verified, with the tools developed by R. Rosenfeld at CMU, that our results are similar to theirs. We think this is due to the presence of the punctuation markers in our texts.

## 3. CONCLUSION

Stochastic n-class models have been built, with a language independent approach. The test perplexities evaluated with tri-gram class models are lower than those obtained with tri-gram word models for the three studied languages, with reductions of 22% for English, 15% for French and 13% for German. Thus, these language models seem attractive for continuous speech recognition tasks. The minimum number of hypotheses required to automatically built classes, combined with an optimal algorithm, permits efficient classifications of texts containing over 30 million of words, in less than a few hours.

Furthermore, since the classification method is independent of the parameter to optimize, other classification criteria could be used; for example, classifications according to target words [7], or classifications with different contexts (right, left, mixed, with different ranges) [8]. Results can possibly give interesting linguistic analyses.

In summary, this method is attractive in two ways: as a text tagger which can give analyses combining syntax and semantics, and as a powerful tool to generate stochastic n-gram class language models for continuous speech recognition systems.

## REFERENCES

- [1] S.M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3), 1987.
- [2] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8, 1994.
- [3] P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. the estimation of powerful language models from small and large corpora. *Proceedings of ICASSP'93*, II-33, 1993.
- [4] P.F. Brown, V.J. Della Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 1992.
- [5] M. Jardino and G.Adda. Automatic word classification using simulated annealing. *Proceedings of ICASSP'93*, II-41, 1993.
- [6] M. Jardino and G.Adda. Automatic determination of a stochastic bi-gram class language model. *Proceedings of ICGP'94*, 1994.
- [7] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. *Proceedings of ACL'93*, 1993.
- [8] C. Huckle. Grouping words using statistical context. *Proceedings of ACL'95*, 1995.