N-GRAM MODELS

Djoerd Hiemstra
University of Twente
http://www.cs.utwente.nl/~hiemstra

DEFINITION
In language modeling, n-gram models are probabilistic models of text that use some limited amount of history, or word dependencies, where n refers to the number of words that participate in the dependence relation.

MAIN TEXT
In automatic speech recognition, n-grams are important to model some of the structural usage of natural language, i.e. the model uses word dependencies to assign a higher probability to "how are you today" than to "are how today you", although both phrases contain the exact same words. If used in information retrieval, simple unigram language models (n-gram models with $n = 1$), i.e., models that do not use term dependencies, result in good quality retrieval in many studies. The use of bigram models (n-gram models with $n = 2$) would allow the system to model direct term dependencies, and treat the occurrence of "New York" differently from separate occurrences of "New" and "York", possibly improving retrieval performance. The use of trigram models would allow the system to find direct occurrences of "New York metro", etc. The following equations contain respectively (12) a unigram model, (13) a bigram model, and (14) a trigram model:

$$P(T_1, T_2, \cdots T_n|D) = P(T_1|D)P(T_2|D)\cdots P(T_n|D) \quad (12)$$

$$P(T_1, T_2, \cdots T_n|D) = P(T_1|D)P(T_2|T_1, D)\cdots P(T_n|T_{n-1}, D) \quad (13)$$

$$P(T_1, T_2, \cdots T_n|D) = P(T_1|D)P(T_2|T_1, D)P(T_3|T_1, T_2, D)\cdots P(T_n|T_{n-2}, T_{n-1}, D) \quad (14)$$

The use of n-gram models increases the number of parameters to be estimated exponentially with n, so special care has to be taken to smooth the bigram or trigram probabilities (see PROBABILITY SMOOTHING). Several studies have shown small but significant improvements of using bigrams if smoothing parameters are properly tuned [2, 3]. Improvements of the use of n-grams and other term dependencies seem to be bigger on large data sets [1].

CROSS REFERENCE
LANGUAGE MODELS, PROBABILITY SMOOTHING

RECOMMENDED READING

[1] Donald Metzler and W. Bruce Croft. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th ACM Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 472–479, 2005.
[2] David R.H. Miller, Tim Leek, and Richard M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 214–221, 1999.
[3] Fei Song and W. Bruce Croft. A General Language Model for Information Retrieval. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 4–9, 1999.