NFS

► Storage Protocols

NF-SS

► Normal Form ORA-SS Schema Diagrams

N-Gram Models

Djoerd Hiemstra
University of Twente, AE Enschede, The Netherlands

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.

Key Points

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 (1) a unigram model, (2) a bigram model, and (3) a trigram model:

$$P(T_1, T_2, \cdots T_n|D) = P(T_1|D)P(T_2|D) \cdots P(T_n|D) \quad (1)$$

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

$$P(T_1, T_2, \cdots T_n|D) = P(T_1|D)P(T_2|T_1, D)P(T_3|T_1, T_2, D) \cdots P(T_n|T_{n-2}, T_{n-1}, D) \quad (3)$$

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. 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-references

► Language Models
► Probability Smoothing

Recommended Reading

1. Metzler D. and Bruce Croft W. A Markov random field model for term dependencies. In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2005, pp. 472–479.
2. Miller D.R.H., Leek T., and Schwartz R.M. A hidden Markov model information retrieval system. In Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1999, pp. 214–221.
3. Song F. and Bruce Croft W. A general language model for information retrieval. In Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1999, pp. 4–9.

NIAM

► Object-Role Modeling

NN Classification

► Nearest Neighbor Classification

NN Query

► Nearest Neighbor Query
► Nearest Neighbor Query in Spatio-temporal Databases