PROBABILITY SMOOTHING

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

DEFINITION

Probability smoothing is a language modeling technique that assigns some non-zero probability to events that were unseen in the training data. This has the effect that the probability mass is divided over more events, hence the probability distribution becomes more *smooth*.

MAIN TEXT

Smoothing overcomes the so-called *sparse data problem*, that is, many events that are plausible in reality are not found in the data used to estimate probabilities. When using maximum likelihood estimates, unseen events are assigned zero probability. In case of information retrieval, most events are unseen in the data, even if simple unigram language models are used (see N-GRAM MODELS): Documents are relatively short (say on average several hundreds of words), whereas the vocabulary is typically big (maybe millions of words), so the vast majority of words does not occur in the document. A small document about "information retrieval" might not mention the word "search", but that does not mean it is not relevant to the query "text search". The sparse data problem is the reason that it is hard for information retrieval systems to obtain high recall values without degrading values for precision, and smoothing is a means to increase recall (possibly degrading precision in the process). Many approaches to smoothing are proposed in the field of automatic speech recognition [1]. A smoothing method may be as simple so-called Laplace smoothing, which adds an extra count to every possible word. The following equations show respectively (8) the unsmoothed, or maximum likelihood estimate, (9) Laplace smoothing, (10) Linear interpolation smoothing, and (11) Dirichlet smoothing [3]:

$$P_{ML}(T{=}t|D{=}d) = tf(t,d) \;/\; \textstyle\sum_{t'} tf(t',d) \tag{8}$$

$$P_{LP}(T{=}t|D{=}d) = (tf(t,d)+1) \;/\; \textstyle\sum_{t'}(tf(t',d)+1) \tag{9}$$

$$P_{LI}(T{=}t|D{=}d) = \lambda P_{ML}(T{=}t|D{=}d) + (1{-}\lambda)P_{ML}(T{=}t|C) \tag{10}$$

$$P_{Di}(T{=}t|D{=}d) = (tf(t,d) + \mu P_{ML}(T{=}t|C)) \;/\; ((\textstyle\sum_{t'} tf(t',d)) + \mu) \tag{11}$$

Here, $tf(t,d)$ is the frequency of occurrence of the term t in the document d, and $P_{ML}(T|C)$ is the probability of a term occurring in the entire collection C. Both linear interpolation smoothing (see also the entry LANGUAGE MODELS) and Dirichlet smoothing assign a probability proportional to the term occurrence in the collection to unseen terms. Here, λ ($0 < \lambda < 1$) and μ ($\mu > 0$) are unknown parameters that should be tuned to optimize retrieval effectiveness. Linear interpolation smoothing has the same effect on all documents, whereas Dirichlet smoothing has a relatively big effect on small documents, but a relatively small effect on bigger documents. Many smoothed estimators used for language models in information retrieval (including Laplace and Dirichlet smoothing) are approximations to the *Bayesian predictive distribution* [2].

CROSS REFERENCE

LANGUAGE MODELS, N-GRAM MODELS

RECOMMENDED READING

[1] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology. Harvard University, August 1998.

[2] Hugo Zaragoza, Djoerd Hiemstra, Michael Tipping, and Stephen Robertson. Bayesian Extension to the Language Model for Ad Hoc Information Retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4-9, 2003.

[3] ChengXiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS) 22*(2), pages 179-214, 2004.