УДК 004.912, 51-74

НОВЫЙ МЕТОД СГЛАЖИВАНИЯ ВЕРОЯТНОСТЕЙ

Асп. ***Датиев М.К.***, ст. н. с. ***Кулай А.Ю.***, проф. ***Датиев К.М.***
Северо-Кавказский горно-металлургический институт.
Московский институт радиотехники, электроники и автоматики

В работе рассмотрены различные методы сглаживания вероятностей, применяемые при построении n-граммных языковых моделей. Предложен новый метод – модифицированный аддитивный метод сглаживания вероятностей. Приводятся результаты экспериментов по моделированию текстов n-граммными моделями (на символах) с использованием различных методов сглаживания вероятностей: модифицированного аддитивного, абсолютного дисконтирования, Виттена-Белла, Каца, модифицированного Кнезера-Нея. Эффективность модели оценивалась по значению перплексии тестового множества.

Статистические модели языка являются необходимым инструментом для большого количества приложений, таких, как распознавание речи, машинный перевод, классификация документов, распознавание графического образа текста, информационный поиск, проверка орфографии и многие другие. Самыми популярными языковыми моделями являются *n*-граммные. При их построении, из-за недостатка материала для обучения, часто приходится решать, так называемую проблему нулевых переходов, когда в обучающем множестве не встречаются возможные *n*-граммы. Для решения этой проблемы, как правило, используются различные методы сглаживания вероятностей [1].

Одним из наиболее простых методов сглаживания, используемых на практике, является аддитивное сглаживание [2; 3], в котором условная вероятность следующего символа вычисляется по формуле:

$$p_{add}\left(w_n \mid w_1^{n-1}\right) = \frac{c\left(w_1^n\right) + \delta}{\sum_{w_n' \in A} c\left(w_1^{n-1} w_n'\right) + \delta \cdot |A|} = \frac{c\left(w_1^n\right) + \delta}{c\left(w_1^{n-1}\right) + \delta \cdot |A|}$$

,

где $c\left(w_1^n\right)$ – частота встречаемости в обучающем множестве *n*-граммы $w_1^n = w_1, ..., w_n$, $|A|$ – мощность алфавита, δ – параметр (как правило, $0 < \delta \leq 1$).

Джеффрис [2] и Лидстоун [3] предлагали использовать δ=1. Вычисленные таким образом оценки вероятностей совпадают с байесовскими оценками.

В работе Гейла и Черча [4] показано, что аддитивное сглаживание, вообще говоря, менее эффективно, чем другие, рассматривавшиеся в исследовании методы сглаживания вероятностей при $n \geq 2$ (особенно при больших значениях *n*). Основным его недостатком является неиспользование информации о распределении *n*-грамм более низких порядков. Для устранения данного недостатка предлагается следующая модификация:

$$p_{\text{smooth}}\left(w_n / w_1^{n-1}\right) = \frac{c\left(w_1^{n-1} w_n\right)}{c\left(w_1^{n-1}\right) + |A|} + \frac{|A|}{c\left(w_1^{n-1}\right) + |A|} \cdot p_{\text{smooth}}\left(w_n / w_2^{n-1}\right)$$

,

здесь $p_{\text{smooth}}\left(w_n / w_2^{n-1}\right)$ – сглаженная вероятность при условии наблюдения «укороченной» истории, в качестве *1*-граммной модели используются байесовские оценки вероятностей.

$$p_{\text{smooth}}\left(w_1\right) = \frac{c\left(w_1\right) + 1}{\sum_{w_1'} c\left(w_1'\right) + |A|}.$$

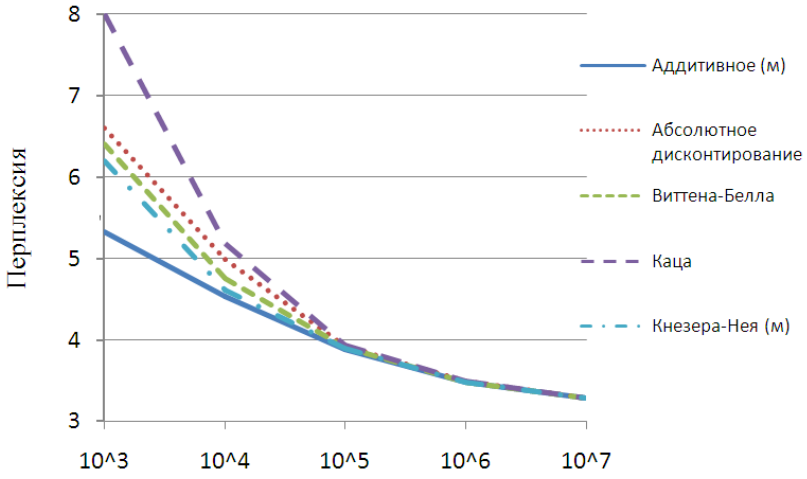
Проведены эксперименты по моделированию текстов на английском языке *n*-граммными моделями (на

символах) с использованием модифицированного аддитивного сглаживания и наиболее популярных методов сглаживания: абсолютного дисконтирования [5], Виттена-Белла [6], Каца [7] и Кнезера-Нея (модифицированного) [1]. Для экспериментов использовали современные газетные публикации по общественно-политической тематике. Эффективность модели оценивали по значению перплексии тестового множества (чем лучше модель, тем ниже перплексия) [8]:

$$perplexity = 2^{-\frac{1}{L}\cdot\sum_{i=1}^{N}\log P_M(T_i)},$$

где $T=\{T_1, T_2, \ldots, T_N\}$ – это тестовые данные общим объемом L, M – используемая языковая модель.

Далее представлены результаты экспериментов для *4*-граммных моделей.



Перплексия *4*-граммных моделей с различными методами сглаживания
в зависимости от объема обучающего множества
для английского языка.

В проведенных экспериментах предлагаемый модифицированный метод аддитивного сглаживания вероятностей оказался наиболее эффективным по сравнению с рядом широко распространенных методов сглаживания, особенно при небольших объемах обучающего множества.

ЛИТЕРАТУРА

1. *Chen S.F. and Goodman J.* An Empirical Study of Smoothing Techniques for Language Modeling. // Computer science group, Harvard University, Cambridge, Massachusetts, TR-8-98, August, 1998.

2. *Jeffreys H.* Theory of Probability. // Clarendon Press, Oxford, second edition, 1948.

3. *Lidstone G.J.* Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. // Transactions of the Faculty of Actuaries, 8: 182-192, 1920.

4. *Gale W.A. and Church K.W.* What's wrong with adding one? // In N. Oostdijk and P. de Haan, editors, Corpus-Based Research into Language. Rodolpi, Amsterdam, 1994.

5. *Ney H., Essen U. and Kneser R.* On structuring probabilistic dependences in stochastic language modeling. // Computer Speech and Language, 8:1–38, 1994.

6. *Bell T.C., Cleary J.G. and Witten I.H.* Text Compression. // Prentice Hall, Englewood Cliffs, N.J., 1990.

7. *Katz S.M.* 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):400–401, March 1987.

8. *Bahl L.R., Baker J.K., Jelinek F. and Mercer R.L.* Perplexity – a measure of the difficulty of speech recognition tasks. // Program of the 94th Meeting of the Acoustical Society of America J. Acoust. Soc. Am., 62:S63, 1977. Suppl. no. 1.