Обзор некоторых статистических моделей естественных языков*

Е. А. Будников
unicorn1992@bk.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В работе производится обзор и сравнение следующих моделей натурального языка: n-граммы, n-граммы на классах, дисконтная модель. В первой части работы будет проведён обзор основной литературы по данной тематике, во второй части будут введены основные понятия и описаны сами методы.

Ключевые слова: языковая модель, униграмма, биграмма, дисконтная модель.

Введение

В задачах, связанных с распознаванием языков, мы часто сталкиваемся с проблемой распознавыния строки слов, прошедших через зашумлённый канал. Чтобы эффективней решать эту проблему, необходимо уметь оценивать априорную вероятность появления тех или иных слов. Метод n-грамм описывается в [1, 2, 3, 4]. Метод n-грамм на классах (*Class-Based n-gram Models*) подробно описан в [5, 1]. Про дисконтную модель (*discounting*) можно почитать в [3, 1, 4].

Пусть $W = w_1 w_2 \dots w_k$ — строка, которую подают на вход зашумлённого канала. Роль такого канала могут исполнять радиоэфир или человек, который переводит строку на другой язык. На выходе получим сигнал Y. По этому выходу необходимо восстановить исходную строку. Вообще говоря, многие строки W могут на выходе приводить к сигналу Y, но чтобы минимизировать вероятность ошибки, необходимо взять такую строку $\hat{\mathrm{W}}$, апостеорная вероятность которой $\Pr(W|Y)$ максимальна. При фиксированном выходе Y эта задача эквивалентна максимизации совместной плотности строки W и выхода Y $\Pr(W, Y)$. Но при этом по формуле Байеса получим:

$$\Pr(W, Y) = \Pr(Y|W) \cdot \Pr(W). \tag{1}$$

Получили разбиение большой задачи на две подзадачи. Мы будем заниматься нахождением второго множителя. Будем обозначать $w_i^j = w_i w_{i+1} \dots w_j$. При таких обозначениях $W \equiv w_1^k$.

$$\Pr(w_1^k) = \Pr(w_k|w_1^{k-1}) \cdot \Pr(w_{k-1}|w_1^{k-2}) \cdot \ \dots \ \cdot \Pr(w_2|w_1) \cdot \Pr(w_1) \tag{2}$$

Определение 1. *Моделью естественного языка назовём функцию*

$$f : R^P \times R^N \to R^k,$$

где R^P — пространство параметров (оценок вероятностей), R^N — пространство акустических входов, R^k — пространство прогнозов

Качество модели будем оценивать по значению *перплексии* [2].

Научные руководители: В. Я. Чучупал, В. В. Стрижов

Определение 2. *Перплексией назовём следующую величину:*

$$PP = \Pr(w_1 w_2 \dots w_k)^{\frac{1}{k}}.$$

Чем меньше перплексия, тем лучше модель.

Модель n-грамм

Одной из основных проблем, возникающих при решении задачи, является огромное количество параметров, поэтому методы во многом направлены на то, чтобы уменьшить число параметров.

В методе n-грам мы считаем две предыстории одинаковыми, если они оканчиваются на одинаковые $n-1$ слов. Другими словами,

Определение 3. *Модель естественного языка называется моделью на n-граммах, если для параметров модели выполнено условие:*

$$\Pr(w_k|w_1^{k-1}) = \Pr(w_k|w_{k-n+1}^{k-1}). \tag{3}$$

Если словарь содержит V слов, то 1-граммы (или *униграммы*) порождают модель, имеющую $V-1$ независимых параметров: V параметров $\Pr(w_i)$ связаны равенством

$$\sum_{i=1}^{V} \Pr(\tilde{w}_i) = 1, \tag{4}$$

где $\tilde{w}_i$ — слова из словаря. 2-граммы (или *биграммы*) порождают V^2-1 независимых параметров: $V(V-1)$, имеющих форму $\Pr(w_2|w_1)$, и $V-1$, имеющих форму $\Pr(w)$. По индукции легко показать, что модель n-грамм содержит V^n-1 параметров. Действительно, $V^{n-1}(V-1)$ параметров, имеющих форму $\Pr(w_n|w_1^{n-1})$, и $V^{n-1}-1$ параметров, имеющих форму $\Pr(w_{n-1}|w_1^{n-2})$(по предположению индукции). Всего V^n-1.

Настраивать параметры модели будем по тексту T, который называется *обучающим текстом*, в процессе, который называется *обучением*. Пусть $C(\mathbf{w})$ — число, означающее, сколько раз строка $\mathbf{w}$ встретилась в обучающем тексте. Тогда в случае *униграмм* максимум правдоподобия для параметра $\Pr(w)$ достигается при $\Pr(w) = \frac{C(w)}{T}$. Для случая n-грамм имеет место быть такой результат:

$$\Pr(w_n|w_1^{n-1}) = \frac{C(w_1^{n-1}w_n)}{\sum_w C(w_1^{n-1}w)}. \tag{5}$$

Модель n-грамм на классах

Чем больше значение n, тем точнее модель. Но в условиях ограниченной обучающей выборки текстов с ростом n доверие к полученной модели должно падать. Необходимо умешьшать число параметров, стараясь не терять значительно точности. Например, можно оценивать вероятность появления не отдельного слова, а некоторой группы слов. Один из способов сделать — использовать n-граммы на классах.

Совершенно ясно, что некоторые слова могут иметь похожие распределения вероятностей. Например, понятно, что слова «Пятница» и «Среда» имеют похожие распределения. Но не одинаковые. Вряд ли мы услышим где-то в офисе радостное восклицание «Слава Богу, наконец-то среда!» или станем беспокоиться о среде, выпавшей на тринадцатое число. Но тем не менее, объединение слов в классы представляется очень удачной идеей.

Пусть существует некоторая функция $\pi : \Omega \to G$, где Ω — множество слов, словарь, а G — множество классов слов. Тогда обозначим $Pr(w|g)$ встречаемость слова w в классе g, а $\Pr(g_n|g_1^{n-1})$ — вероятность встретить слово из класса g_n после последовательности слов, имеющих форму $g_1 g_2 \dots g_{n-1}$.

Тогда для биграмм имеют место следующие соотношения:

$$\Pr(w_i|w_{i-1}) = \Pr(w_i, \pi(w_i)|w_{i-1}, \pi(w_{i-1})) = \tag{6}$$

$$= \Pr(w_i|\pi(w_i), \pi(w_{i-1}), w_{i-1}) \cdot \Pr(\pi(w_i)|\pi(w_{i-1}), w_{i-1}).$$

Для общего случая n-грамм введём

Определение 4. *Модель n-грамм назовём моделью n-грамм на классах, если выполняется гипотеза:* $\Pr(w_k|w_1^{k-1}) = \Pr(w_k|g)\Pr(g_k|g_1^{k-1})$,*где* $k = 1, \dots, n$.

Опишем теперь один алгоритм построения функции π на примере биграмм. Пусть $T = (t_1, t_2, \dots, t_T)$ — обучающая выборка, причём все слова содержатся в словаре V. Функция правдоподобия тогда равна

$$L(T) = \Pr(T) = \prod_{x,y \in V} \Pr(x|y)^{C(x,y)}, \tag{7}$$

где x,y — слова из словаря, причём y предшествует x, а $C(x, y)$ показывает, сколько раз последовательность слов «yx» встретилась в обучающей выборке T.

Для удобства будем использовать логарифм функции правдоподобия вместо самой функции:

$$\log L(T) = \sum_{x,y \in V} C(x, y) \cdot \log \Pr(x|y). \tag{8}$$

Из данного выше определения модели n-грамм на классах заключаем, что максимум правдоподобия для биграмм достигается при

$$\Pr(w_i|w_{i-1}) = \frac{C(w_i)}{C(\pi(w_i))} \cdot \frac{C(\pi(w_i), \pi(w_{i-1}))}{C(\pi(w_{i-1}))}, \tag{9}$$

где $C(w_i)$ — число раз, которые слово w_i встретилось в обучающей выборке, а $C(\pi(w))$ — число раз, которые слова из класса $\pi(w)$ встретились в выборке, аналогично $C(\pi(w_x), \pi(w_y))$ — число пар вида «$\pi(w_y)\pi(w_x)$», встретившиеся в выборке.

Подставим теперь это выражение в функцию правдоподобия и преобразуем:

$$\log L(T) = \sum_{x,y \in V} C(x, y) \cdot \log \left(\frac{C(x)}{C(\pi(x))} \cdot \frac{C(\pi(x), \pi(y))}{C(\pi(y))} \right) \tag{10}$$

$$= \sum_{x,y \in V} C(x, y) \cdot \log \left(\frac{C(x)}{C(\pi(x))} \right) + \sum_{x,y \in V} C(x, y) \cdot \log \left(\frac{C(\pi(x), \pi(y))}{C(\pi(y))} \right)$$

$$= \sum_{x \in V} C(x) \cdot \log \left(\frac{C(x)}{C(\pi(x))} \right) + \sum_{g,h \in G} C(g, h) \cdot \log \left(\frac{C(g, h)}{C(h)} \right)$$

$$= \sum_{x \in V} C(x) \cdot \log C(x) - \sum_{x \in V} C(x) \cdot \log C(\pi(x))$$

$$+\sum_{g,h\in G} C(g,h)\cdot\log C(g,h)-\sum_{g,h\in G} C(g,h)\cdot\log C(h)$$

$$=\sum_{x\in V} C(x)\cdot\log C(x)+\sum_{g,h\in G} C(g,h)\cdot\log C(g,h)$$

$$-2\sum_{g\in G} C(g)\cdot\log C(g),$$

где (g,h) — некоторая последовательность классов «hg».

Теперь вы заметим, что первое слагаемое не зависит от выбора функции π. Поэтому его рассматривать необязательно, когда мы будем оптимизировать π. Будем максимизировать функцию

$$F_\pi=\sum_{g,h\in G} C(g,h)\cdot\log C(g,h)-2\sum_{g\in G} C(g)\cdot\log C(g). \tag{11}$$

Приведём теперь алгоритм оптимизации функции π. Перед запуском алгоритма определяется число классов.

1: **для всех** $w\in\Omega$
2: $G(w)=1$ // *инициализация*
3: **для** $i=1\ldots n$
4: **повторять**
5: **для всех** $c\in G$
6: Переместить слово w в класс c, запомнив его предыдущий класс
7: Вычислить изменения F_π для этого перемещения в c. Переместить слово w назад в его предыдущий класс
8: Переместить слово w в класс, кторый больше всего увеличивает F_π, или никуда не перемещать, если увеличения ни на каком перемещении не происходит
9: **пока** s

Дисконтная модель (*discounting*)

Рассмотрим событие S, которое встретилось s раз, а общее количество наблюдений A. Тогда оценка вероятности S по принципу наибольшего правдоподобия будет равна

$$\Pr(S)=\frac{s}{A}. \tag{12}$$

Но тогда, в соответствии с этим принципом, событиям, которые не были встречены среди обучающего текста T, будут приписаны нулевые вероятности, а значит, будучи встреченными на тесте, они никогда не будут распознаны. Чтобы справиться с этой проблемой, можно поступить следующим способом. В оценке вероятности события вместо числа s брать

$$s'=d_s\cdot s, \tag{13}$$

где d_s — множитель, зависящий от числа раз, которые событие встретилось в обучающем тексте. Тогда получим дисконтную оценку вероятности события A:

$$\Pr{}_{discount}(S)=\frac{s'}{A}=\frac{d_s\cdot s}{A}. \tag{14}$$

Различные дисконтные методы различаются стратегией выбора d_s.

Обозначим c_s число всех событий которые встретились в процессе обучения ровно s раз. Тогда общее число наблюдений $A = \sum_{s \geqslant 1} c_s \cdot s$. Получается, что таким образом мы перераспределили оценки вероятности между событиями и оставили на все не встретившиеся в обучении слова $1 - \frac{1}{A} \sum_{s \geqslant 1} d_s \cdot c_s \cdot s$. Если c_0 — число таких событий, то оценка вероятности каждого из них равна

$$\frac{1}{c_0} \left(1 - \frac{1}{A} \sum_{s \geqslant 1} d_s \cdot c_s \cdot s \right). \tag{15}$$

Дисконтная модель Гуда-Тьюринга (Good-Turing). В статье[6] предлагается следующая стратегия выбора множителя:

$$d_s = (s+1) \frac{c_{s+1}}{s \cdot c_s}. \tag{16}$$

Эта стратегия называется оценкой Гуда-Тьюринга. Несмотря на очевидную простоту стратегии, у неё есть существенный недостаток: она проваливается в случае, если $c_a = 0$ для некоторого a и существует $b > a$, такой, что $c_b \neq 0$. Решение этой проблемы было предложено в [7]. Пусть есть некое, достаточно большое число k, такое что все оценки вероятностей событий, встретившихся в процессе обучения более k раз, признаем надёжными. При этом d_s будет выглядеть так:

$$d_s = \begin{cases} \frac{(s+1)\frac{c_{s+1}}{s \cdot c_s} - (k+1)\frac{c_{k+1}}{c_1}}{1 - (k+1)\frac{c_{k+1}}{c_1}} & , \ 1 \leqslant s \leqslant k \\ 1 & , \ s > k \end{cases} \tag{17}$$

Этот метод тоже нестабильный, так как возможны ситуации, когда $d_s < 0$.

Модель абсолютного уменьшения (Absolute discounting). Одной из альтернатив модели Гуда-Тьюринга является модель абсолютного уменьшения[8]. В этой модели происходит уменьшение числа a для каждого события на фиксированное число m.

$$d_s = \frac{s - m}{s}. \tag{18}$$

Для того чтобы уменьшение суммарной вероятности было таким же, как в модели Гуда-Тьюринга, необходимо:

$$m = \frac{c_1}{\sum_{s \geqslant 1} c_s}. \tag{19}$$

Заключение

Работа не описывает весь перечень методов, использующихся при моделировании естественных языков. Представленные в работе алгоритмы описывают по сути один подход, скорее дополняют друг друга, нежели конкурируют. К достоинствам метода n-грамм стоит отнести простоту реализации и интуитивную понятность подхода. Метод n-грамм на классах развивают идею уменьшения количества параметров модели, при этом, конечно, теряя в качестве прогнозирования. Дисконтная модель исправляет существенный недостаток моделей на n-граммах — нулевые значения параметров, которые приводят к невозможности прогнозирования последовательностей, которые могут в жизни, но не встретились в обучении. Модель Гуда-Тьюринга, как уже отмечалось выше, обладает простой реализацией

и прозрачной интерпретацией, но, к сожалению, является при этом неустойчивой. Модель абсолютного уменьшения исправляет этот недостаток.

Литература

[1] Huang X., Acero A., Hon H.-W. *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development* /Prentice Hall PTR, 2001.

[2] Jelinek F. *Statistical Methods for Speech Recognition.* //The MIT Press, Cambridge, Massachusetts, 1997.

[3] Gotoh Y., Renals S. *Statistical language modelling.* //In Steve Renals and Gregory Grefenstette, editors, ELSNET Summer School, volume 2705 of Lecture Notes in Computer Science, pp. 78 –105, Springer, 2000.

[4] Young S., Bloothooft G., editors *Corpus-Based Methods in Language and Speech Processing* /Kluwer Academic Publishers, Dordrecht, 1997.

[5] Brown P. F., Della Pietra V. J., deSouza P. V., Mercer R. L. *Class-based n-gram models of natural language.* //Proceedings of the IBM Natural Language ITL, pp. 283 –298, Paris, France, March 1990.

[6] Good I. J. *The population frequencies of species and the estimation of population parameters.* //Biometrika, vol. 40(3, 4):pp. 237 –264, 1953.

[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, vol. 35(3):pp. 400 – 401, March 1987.

[8] Ney H., Essen U., Kneser R. *On structuring probabilistic dependencies in stochastic language modelling.* //Computer Speech and Language, vol. 8:pp. 1 –38, 1994.