Предиктивный ввод текста на основе факторной модели языка

Филатов С.Ю., МГТУ им. Н.Э. Баумана
serj.phil@outlook.com

Аннотация

Выполнен обзор существующих методов предиктивного ввода текста. Выявлены преимущества и недостатки рассматриваемых методов и их применимость к решаемой задаче для русского языка. На основе факторной модели языка разрабатывается метод предиктивного ввода текста для русского языка. На основе сравнения разработанного метода с традиционным биграммным подходом показано превосходство разработанного алгоритма, обеспечившего снижение числа нажатий на клавиши для ввода одного символа на 5% и повышение точности прогноза окончания вводимого слова на 4%. Точность предсказания морфологических параметров составляет около 44%. Рассмотрены возможные способы повышения эффективности работы предложенного метода за счёт улучшения морфологического анализатора и перехода к более сложной триграммной факторной модели языка.

1 Введение

Различные системы автодополнения текста используются в вычислительной технике на протяжении многих лет. Основная задача таких систем – сокращение числа нажатий на клавиатуру, необходимых для ввода какого-то слова, например, SMS-сообщения. В последнее десятилетие необходимость создания качественных систем автодополнения и предиктивного ввода текста была обусловлена появлением и ростом популярности мобильных устройств, обладающих, в основном, 12-клавшной телефонной клавиатурой либо сенсорным экраном с небольшой диагональю. Набор текста на устройствах данной категории является затруднительным из-за малого физического размера клавиш и, зачастую, назначением ввода нескольких различных символов на одну клавишу.

Используемые подходы [Gale, Sampson, 1995], [Katz, 1987], [Stocky, Faaborg, Lieberman, 2004], [Sandnes, 2015] к решению проблемы автодополнения текста основаны на использовании частотных словарей, содержащих слова и вероятность их встречи в тексте. Данный подход показывает свою эффективность при использовании с аналитическими языками, характеризуемыми малым числом словоформ. Однако при использовании данного метода автодополнения текста с синтетическими флективными языками (например, русским), имеющими большое число словоформ, качество предсказания падает, поскольку метод не учитывает морфологические параметры вводимых слов. В традиционных методах автодополнения текста данная проблема решается путём хранения вероятностей встречи кортежей из N слов, однако это влечёт повышенное потребление памяти, что ограничивает их использование в мобильных устройствах.

2 Способ предиктивного ввода текста на основе факторной модели языка

Факторная модель языка [Bilmes, Kirchhoff, 2003] - статистическая модель языка, разработанная в 2003 году. Факторная модель языка рассматривает слово как вектор из k факторов $w = \{f^1, \dots, f^k\}$, одним из которых выступает само слово, а остальными – его характеристики важные для конкретной модели, такие как, такие так морфологические параметры, теги части речи, корни, псевдоосновы, семантический контекст употребления слова, огласовки и т.д.

В реализуемой модели языка факторами являются слово и тег, содержащий информацию о значениях следующих морфологических параметров:

- грамматический род;
- число;
- падеж;
- транзитивность;
- лицо;
- время;
- наклонение.

Реализуемая факторная модель языка может быть описана следующим образом:

$$P(s_i|s_{i-1}t_it_{i-1}) =$$

$$\begin{cases} d_{s_is_{i-1}} \frac{C(s_i,s_{i-1})}{C(s_{i-1})}, \text{при } C(s_i,s_{i-1}) > k \\ \alpha_{s_is_{i-1}t_it_{i-1}} g(s_i,t_i,t_{i-1}), \text{ при } C(s_i,s_{i-1}) \le k \end{cases}$$

,

где $d_{s_is_{i-1}}$ – весовой коэффициент для биграммной модели;

k – значение максимальное частоты употребления биграммы $s_{i-1}s_i$, при котором используется backoff модель;

$\alpha_{s_is_{i-1}}$- весовой коэффициент для униграммной модели, $g(s_i,t_i,t_{i-1})$ - backoff функция, определяемая как:

$$g(s_i,t_i,t_{i-1}) = P(s_i)P(t_i|\,t_{i-1}).$$

Для определения коэффициента α вычисляется величина β вероятной массы, приходящейся на униграммный вариант:

$$\beta_{s_is_{i-1}} = 1 - \sum_{s_i:C(s_i,s_{i-1})>k} d_{s_is_{i-1}} P_{MLE}(s_i|s_{i-1})$$

.

Величина α определяется как

$$\alpha_{s_is_{i-1}t_it_{i-1}} = \frac{\beta_{s_is_{i-1}}}{\sum_{s_i:C(s_is_{i-1})\le k} g(s_i,t_i,t_{i-1})}.$$

3 Сравнение с традиционными способами

Для разработанного способа предиктивного ввода были произведены изменения точности предсказания и KSPC (число нажатий на клавиатуру, необходимое для ввода одного символа). Данные показатели были сравнены с показателями для традиционной биграммной модели с откатом на униграммы.

Для проведения эксперимента использовались несколько обучающих и тестовых корпусов, характеристики которых приведены в таблице 1.

При проведении эксперимента число биграмм в языковых моделях было ограничено 2500. Переобучение модели языка при проведении эксперимента было запрещено. В таблице 2 приведены показатели точности предсказания следующего слова для всех корпусов и значение KSPC.

Более высокую точность предсказания следующего слова для корпуса L можно объяснить тем, что часть наиболее часто употребляемых биграмм этого корпуса была включена в биграммы, используемые модулем автодополнения и предиктивного ввода.

Было произведено сравнение точности предсказания и значения KSPC в зависимости от числа биграмм в словаре. Число используемых биграмм при этом составляло 0, 2500, 5000, 7500, 10000 и 25000.

Таблица 1. Используемые корпуса текстов

Название корпуса	**Объём корпус (число биграмм)**	**Источник корпуса**	**Входит в обучающий корпус**	**Примечания**
I	68321	OpenCorpora (сеть Internet)	Да	Подмножество корпуса OpenCorpora со снятой омонимией.
L	348351	Художественная литература	Да	Проза
W	700516	Художественная литература	Нет	Проза
P	84646	Художественная литература	Нет	Поэзия
N	2175221	Новости	Нет	Новостные статьи с ресурса Lenta.ru

Таблица 2. Количественные показатели работы метода

Корпус	Метод	Точность	KSPC (x100)
I	Стандартный	18%	89.8
I	Улучшенный	21%	87.7
L	Стандартный	17.2%	91.8
L	Улучшенный	28.8%	85.4
W	Стандартный	15.9%	92.5
W	Улучшенный	26.6%	86.8
P	Стандартный	12.6%	93.9
P	Улучшенный	21.1%	90
N	Стандартный	17%	91
N	Улучшенный	24.6%	86.4

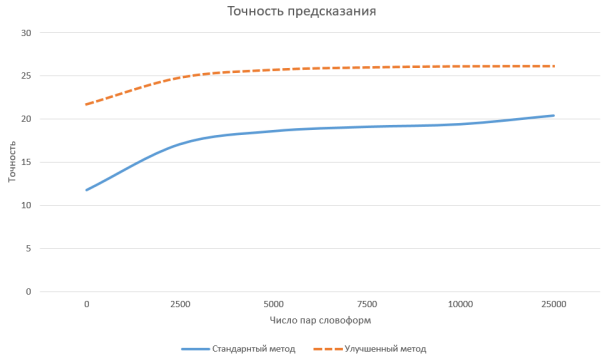

Рисунок 7 Зависимость точности прогноза от числа биграмм

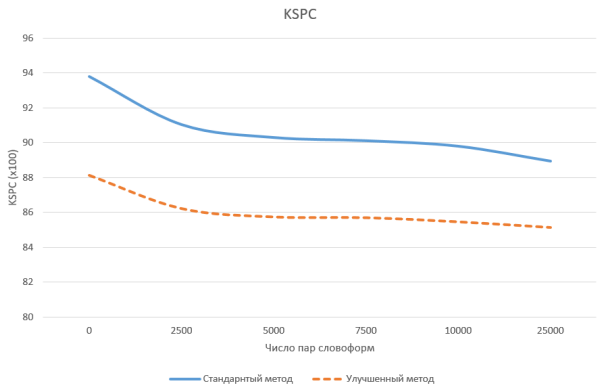

Рисунок 8 Зависимость KSPC от числа биграмм

Из графиков очевидно, что наибольший прирост качества предсказания происходит при переходе от чистой и факторной униграммной модели языка к простой и факторной биграммной моделям с откатом. При дальнейшем увеличении числа используемых биграмм столь значительного увеличения качества предсказания не происходит. Прирост при переходе от 10000 до 25000 биграмм для улучшенного метода привёл к росту качества предсказания всего на 0.01%. Максимальная точность предсказания для улучшенного метода автодополнения на рассматриваемом новостном корпусе составляет 26% и достигается при числе биграмм в 7500. Традиционный метод показывает максимальную точность в 20.4% при числе используемых биграмм 25000. Значения параметра KSPC для улучшенного метода на 3.8-5.7 пунктов ниже, чем у стандартного метода с откатом.

4 Заключение

Разработанный метод обладает следующими преимуществами по сравнению со стандартным:

- большая точность предсказания для текстов различной тематики;
- меньшее число биграмм, требуемых для достижения максимальной точности предсказания;
- обеспечение согласованности предлагаемых вариантов завершения по морфологическим параметрам типовых конструкция в русском языке.

Основными недостатками разработанного метода являются зависимость качества его работы от обучающего набора данных и зависимость качества работы метода от точности определения значений морфологических параметров введённых слов. Помимо этого, словарь, используемый в работе предложенного метода, имеет занимает больший объём памяти, чем словарь, используемый традиционными биграммным методом, из-за необходимости дублирований слов, имеющий одинаковое символьное представление, но различные наборы значений морфологических параметров.

Для улучшения качества предсказания вводимых слов можно предложить следующие варианты улучшения разработанного способа предиктивного ввода:

- оптимизацию перечня используемых морфологических параметров;
- повышение точности работы морфологического анализатора;
- переход к более сложной модели языка, учитывающей морфологические параметры двух предыдущих слов.

Список литературы

Gale W. A., Sampson G. Good-Turing frequency estimation without tears //Journal of Quantitative Linguistics. – 1995. – Vol. 2. – №. 3. – Pp. 217-237.

Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer //IEEE transaction on acoustics, speech, and signal processing. – 1987. – Vol. 35. – №. 3. – Pp. 400-401.

Stocky T., Faaborg A., Lieberman H. A commonsense approach to predictive text entry //CHI'04 Extended Abstracts on Human Factors in Computing Systems. – ACM, 2004. – Pp. 1163-1166.

Sandnes F. E. Reflective text entry: a simple low effort predictive input method based on flexible abbreviations //Procedia Computer Science. – 2015. – Vol. 67. – Pp. 105-112.

Bilmes J. A., Kirchhoff K. Factored language models and generalized parallel backoff //Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2. – Association for Computational Linguistics, 2003. – Pp. 4-6.