Предиктивный ввод текста на основе факторной модели языка

Предиктивный ввод текста на основе факторной модели языка

**Филатов С.Ю.,** МГТУ им. Н.Э. Баумана

[serj.phil@outlook.com](mailto:serj.phil@outlook.com)

# Аннотация

Выполнен обзор существующих методов предиктивного ввода текста. Выявлены преимущества и недостатки рассматривае- мых методов и их применимость к решае- мой задаче для русского языка. На основе факторной модели языка разрабатывается метод предиктивного ввода текста для рус- ского языка. На основе сравнения разрабо- танного метода с традиционным биграмм- ным 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 мор- фологических параметров составляет около 44%. Рассмотрены возможные способы по- вышения эффективности работы предло- женного метода за счёт улучшения морфо- логического анализатора и перехода к бо- лее сложной триграммной факторной мо- дели языка.

# Введение

Различные системы автодополнения текста используются в вычислительной технике на протяжении многих лет. Основная задача та- ких систем – сокращение числа нажатий на клавиатуру, необходимых для ввода какого- то слова, например, SMS-сообщения. В по- следнее десятилетие необходимость создания качественных систем автодополнения и пре- диктивного ввода текста была обусловлена появлением и ростом популярности мобиль- ных устройств, обладающих, в основном, 12- клавшной телефонной клавиатурой либо сен- сорным экраном с небольшой диагональю. Набор текста на устройствах данной катего- рии 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 исполь- зовании частотных словарей, содержащих

слова и вероятность их встречи в тексте. Дан- ный подход показывает свою эффективность при использовании с аналитическими языка- ми, характеризуемыми малым числом слово- форм. Однако при использовании данного метода автодополнения текста с синтетиче- скими флективными языками (например, рус- ским), имеющими большое число словоформ, качество предсказания падает, поскольку ме- тод не учитывает морфологические парамет- ры вводимых слов. В традиционных методах автодополнения текста данная проблема ре- шается путём хранения вероятностей встречи кортежей из N слов, однако это влечёт повы- шенное потребление памяти, что ограничива- ет их использование в мобильных устрой- ствах.

# Способ предиктивного ввода текста на основе факторной модели языка

Факторная модель языка [Bilmes, Kirchhoff, 2003] - статистическая модель языка, разрабо- танная в 2003 году. Факторная модель языка рассматривает слово как вектор из факторов

, одним из которых выступает само слово, а остальными – его характеристи- ки важные для конкретной модели, такие как, такие так морфологические параметры, теги части речи, корни, псевдоосновы, семантиче- ский контекст употребления слова, огласовки и т.д.



В реализуемой модели языка факторами являются слово и тег, содержащий информа- цию о значениях следующих морфологиче- ских параметров:

* + грамматический род;
  + число;
  + падеж;
  + транзитивность;
  + лицо;
  + время;
  + наклонение.

Реализуемая факторная модель языка мо- жет быть описана следующим образом:

Новые информационные технологии в автоматизированных системах – 2018

# Сравнение с традиционными способами



Для разработанного способа предиктивно- го ввода были произведены изменения точно- сти предсказания и KSPC (число нажатий на клавиатуру, необходимое для ввода одного

, символа). Данные показатели были сравнены

где  – весовой коэффициент для би- граммной модели;

– значение максимальное частоты употреб- ления биграммы , при котором исполь- зуется backoff модель;

- весовой коэффициент для униграмм- ной модели,  - backoff функция, определяемая как:

  .

Для определения коэффициента α вычисля- ется величина  вероятной массы, приходя- щейся на униграммный вариант:



.

Величина определяется как



.

с показателями для традиционной биграмм- ной модели с откатом на униграммы.

Для проведения эксперимента использова- лись несколько обучающих и тестовых кор- пусов, характеристики которых приведены в таблице 1.

При проведении эксперимента число би- грамм в языковых моделях было ограничено 2500. Переобучение модели языка при прове- дении эксперимента было запрещено. В таб- лице 2 приведены показатели точности пред- сказания следующего слова для всех корпу- сов и значение KSPC.

Более высокую точность предсказания сле- дующего слова для корпуса L можно объяс- нить тем, что часть наиболее часто употреб- ляемых биграмм этого корпуса была включе- на в биграммы, используемые модулем авто- дополнения и предиктивного ввода.

Было произведено сравнение точности предсказания и значения KSPC в зависимости от числа биграмм в словаре. Число использу- емых 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 омо- нимией. |
| 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 |

Предиктивный ввод текста на основе факторной модели языка

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 |

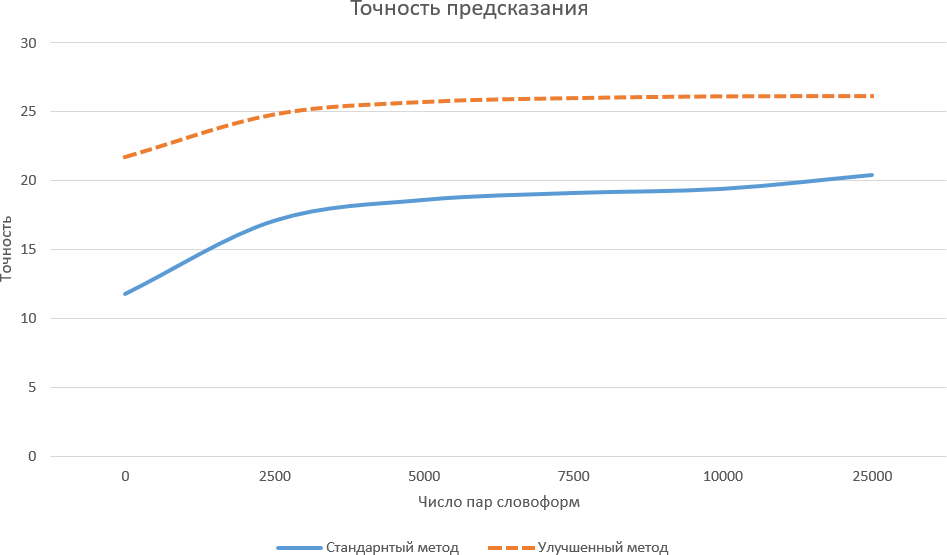


Рисунок 7 Зависимость точности прогноза от числа биграмм

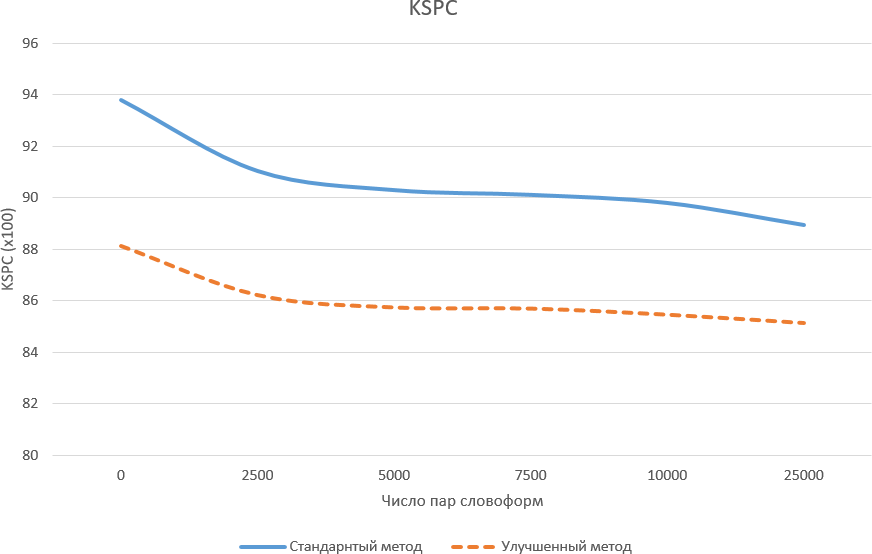


Рисунок 8 Зависимость KSPC от числа биграмм

Из графиков очевидно, что наибольший прирост качества предсказания происходит при переходе от чистой и факторной уни- граммной модели языка к простой и фактор- ной биграммной моделям с откатом. При дальнейшем увеличении числа используемых биграмм столь значительного увеличения ка- чества 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 ме- тода автодополнения на рассматриваемом новостном корпусе составляет 26% и дости-

гается при числе биграмм в 7500. Традицион- ный метод показывает максимальную точ- ность в 20.4% при числе используемых би- грамм 25000. Значения параметра KSPC для улучшенного метода на 3.8-5.7 пунктов ниже, чем у стандартного метода с откатом.

# Заключение

Разработанный метод обладает следующи- ми преимуществами по сравнению со стан- дартным:

* + большая точность предсказания для текстов различной тематики;
  + меньшее число биграмм, требуемых для достижения максимальной точности предсказания;
  + обеспечение согласованности предла- гаемых вариантов завершения по морфологи- ческим параметрам типовых конструкция в русском языке.

Основными недостатками разработанного метода являются зависимость качества его работы от обучающего набора данных и зави- симость качества работы метода от точности определения значений морфологических па- раметров введённых слов. Помимо этого, словарь, используемый в работе предложен- ного метода, имеет занимает больший объём памяти, чем словарь, используемый традици- онными биграммным методом, из-за необхо- димости дублирований слов, имеющий оди- наковое символьное представление, но раз- личные наборы значений морфологических параметров.

Для улучшения качества предсказания вво- димых слов можно предложить следующие варианты улучшения разработанного способа предиктивного ввода:

* оптимизацию перечня используемых мор- фологических параметров;
* повышение точности работы морфологи- ческого анализатора;
* переход к более сложной модели языка, учитывающей морфологические парамет- ры двух предыдущих слов.

Новые информационные технологии в автоматизированных системах – 2018

# Список литературы

Gale W. A., Sampson G. Good-Turing frequency es- timation 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 rec- ognizer //IEEE transaction on acoustics, speech, and signal processing. – 1987. – Vol. 35. – №. 3. – Pp. 400-401.

Stocky T., Faaborg A., Lieberman H. A com- monsense 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 ef- fort predictive input method based on flexible ab- breviations //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 Chap- ter of the Association for Computational Linguis- tics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003-- short papers-Volume 2. – Association for Compu- tational Linguistics, 2003. – Pp. 4-6.