Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2017»

Москва, 31 мая — 3 июня 2017

# АВТОМАТИЗАЦИЯ ПОСТРОЕНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ РАСПОЗНАВАНИЯ СПОНТАННОЙ РУССКОЙ РЕЧИ

**Белик В. В.** (ogibbion@gmail.com),
**Бирин Д. А.** (d.birin@kvant-rdi.spb.ru)

ФГУП «НИИ «КВАНТ», Санкт-Петербург, Россия

# AUTOMATIC DEVELOPMENT OF LANGUAGE MODELS FOR SPONTANEOUS RUSSIAN SPEECH RECOGNITION

**Belik V. V.** (ogibbion@gmail.com),
**Birin D. A.** (d.birin@kvant-rdi.spb.ru)

Federal State Unitary Enterprise “Research Institute ‘KVANT’”,
Saint-Petersburg, Russia

The paper considers automatization in the development of language models. While trying to create a language model which could be used in Russian spontaneous speech recognition systems we faced the fact that the are neither corpora appropriate for the task nor automatic tools that could make data processing less time-consuming. As a result we implemented a toolkit that allows using one and the same system a) to process text files including correction of misspellings and misprints, and deletion of useless information, corpus segmentation, b) to construct frequency lists and n-gram files and finally c) to make, unite and test language models. In this paper we describe each step of the work of the system in detail and provide information on evaluation and testing. The use of our system allowed us to process corpus of news messages containing more than 1,000,000,000 words. We also processed 3 corpora of smaller size. The results show that the use of our toolkit guarantees considerable improvement of the resulting perplexity median achieved in course of testing the final language model.

**Key words:** text corpora, n-gram, language model, text processing, speech recognition, perplexity

## 1. Введение

Статистические модели применяются для задач анализа звукового сигнала более полувека и давно стали неотъемлемой частью систем распознавания речи, а развитие корпусных исследований предоставляет массу разнообразного материала для их построения. Однако в том, что касается использования языковых моделей при распознавании русской речи, многие исследователи отмечают, что свободный порядок слов и большое количество словоформ не позволяет применять статистические методы языкового моделирования с тем же успехом, что и в случае большинства европейских языков [Kanevsky et al., 1996].

- Построение языковых моделей (далее -из текстовых данных удаляются недопустимые элементы;
- корпус разделяется на обучающую и тестовую части;
- по обучающей части корпуса строятся частотные словари;
- по обучающей части корпуса строится файл n-грамм;
- на основе файла n-грамм, полученного на предыдущем шаге (и, опционально, словаря, полученного на третьем шаге), ЯМ;
- новый файл ЯМ может быть слит с одним из файлов, полученных в ходе предыдущих запусков;
- полученная ЯМ проходит тестирование.

## 2. Подготовительный этап: формирование корпуса

Разработанное программное обеспечение позволяет адаптировать любой текстовый корпус для подачи на вход системе построения ЯМ. При поиске материалов для формирования ЯМ нами были рассмотрены следующие корпусы:

- Национальный корпус русского языка;
- Хельсинский аннотированный корпус (ХАНКО);
- Корпусы русского языка университета г. Лидс, созданные в 2000-е годы в Центре переводческих исследований С. А. Шаровым (http://corpus.leeds.ac.uk/ruscorpora.html);
- Корпус Библиотеки Мошкова, созданный А. Сокирко по текстам из библиотеки Мошкова;
- Корпусы в системе Sketch Engine (http://sketchengine.co.uk/), созданные английской лингвистической службой Lexical Computing Ltd;
- Araneum Russicum.

Среди данных корпусов нас, прежде всего, интересовал русский корпус Sketch Engine объемом 20 миллиардов словоупотреблений и корпус Araneum Russicum, созданные из текстов Интернета по технологии Wacky, использующей принцип «плетения паутины», или «краулинг ползания» по Интернету [Khokhlova, 2016]. Однако от их использования было решено отказаться, так как в корпусах отсутствует жанровая метаразметка и невозможно утверждать что-либо об их сбалансированности [Захаров, 2015, с. 7]. Ни один из прочих доступных текстовых корпусов не дает достаточного объема данных для

повышения точности распознавания спонтанной речи. Основная часть современных русских корпусов была сформирована полностью или в большей части на основе письменных источников, отличающихся от устных как структурно, так и лексически. Также стоит отметить, что даже при формировании корпусов на основе устных источников (как, например, в случае с устной частью «Национального корпуса русского языка») зачастую удаляется важная для распознавания информация, как то ошибки, оговорки, повторы, паузы хезитации. Неоднозначно обстоит дело и с обозначением границ фраз.

Стоит также упомянуть о корпусе «Один речевой день», содержащим около более 1200 часов звучания и 130 информантов [Bogdanova-Beglarian, 2017]. К сожалению, на данный момент аннотирование корпуса не завершено и к нему не предоставляется свободный доступ.

Создание достаточного по объему текстового корпуса с необходимыми характеристиками является крайне трудоемкой задачей, так как работа по сбору не поддается полной автоматизации. Таким образом, было принято решение создавать ЯМ на основе нескольких подкорпусов, слитых в соответствии с данными, полученными по результатам их тестирования. В ходе работы были подготовлены 4 подкорпуса, отражающих спонтанную либо псевдоспонтанную речь:

- субтитры;
- форумы;
- речь.

Первые два подкорпуса были собраны по открытым источникам при помощи специального программного обеспечения, позволяющего ограничивать кроллинг заданными через регулярные выражения ссылками, третий был получен вручную посредством написания подстрочников к аудиозаписям, полученным из открытых источников. Тестирование работоспособности системы построения ЯМ и ее нагрузочное тестирование проводилось по корпусу новостных текстов, собранному из открытых источников и включающему около 35 млн предложений (более миллиарда слов).

Текущая версия объединенного корпуса, предназначенного для построения ЯМ, используемых при распознавании спонтанной речи, содержит около 9 млн предложений (58 млн словоупотреблений и более миллиона уникальных словоупотреблений). При этом подстрочники спонтанной речи составляют около 10 %, субтитры — 50 % и форумы — 40 % . Проверенный словарь корпуса насчитывает около 400 тысяч единиц и обеспечивает покрытие тестовых данных более чем на 98 %.

Вся дальнейшая обработка корпусов автоматизирована, представлена в виде одного проекта и может быть инициирована последовательным запуском шагов на кластере серверов. При этом формируется очередь запуска, и каждый следующий шаг запускается по завершении работы предыдущего.

Для запуска автоматизированной системы обработки требуется преобразование входных файлов в корпус (это могут быть подстрочники речевых сообщений, логи чатов, приложений для обмена сообщениями и пр.). При преобразовании из исходного файла выделяется текст. При добавлении текста в корпус проверяется его уникальность. Входными аргументами являются: путь к папке

с текстовыми файлами, путь к текстовому корпусу. Файлы распределяются по разделам. Каждый раздел представляет собой набор архивов, включающий 100 000 документов. Характеристики каждого раздела включают в себя описание документов с указанием источника и даты добавления в корпус.

## 3. Первый этап: предобработка корпуса

Для запуска задачи предобработки необходим текстовый корпус и проверенные словари имен нарицательных, имен собственных, аббревиатур, заполнителей, автоисправлений. Данные словари выполняют следующие функции:

- исправление регистра;
- ёфикация;
- исправление ошибок и опечаток;
- удаление из выходных словарей технической информации;
- объединение устойчивых словосочетаний в одну графическую единицу (функция опциональна).

Правила предобработки, представленные в виде модулей, подключаемых к соответствующему шагу проекта построения ЯМ, основаны на применении регулярных выражений и последовательно проводят следующие операции:

- заменяют случайные вставки букв латинского алфавита на кириллические символы;
- исправляют регистр, проводят ёфикацию;
- удаляют недопустимые символы и комбинации символов;
- отделяют знаки препинания пробелами;
- разделяют текст на отдельные предложения, каждое из которых занимает отдельную строку;
- исправляют ошибки и опечатки.

На выходе шага мы получаем чистый текстовый корпус и автоматически составленные словари, включающие словоформы, отсутствующие в базовом словаре. (common.txt, содержащий имена нарицательные, proper.txt, содержащий имена собственные, unknown.txt, содержащий словоформы, которые не удалось причислить к двум вышеуказанным категориям из-за незначительного статистического разрыва между написанием их со строчной и прописной букв, trash.txt, содержащий неприемлемые для данного языка словоформы).

Правила предобработки корпуса были сформулированы посредством анализа текстовых данных, получаемых на выходе шага предобработки. На данный момент при обработке текстовых данных подключается около 80 правил. Введение каждого нового правила тестировалось при помощи запуска обработки и анализа как выходных текстов, так и частотных словарей. В первую очередь были сформулированы правила, позволяющие удалить из текста все недопустимые символы и комбинации символов. Далее в ходе анализа частотных словарей (начиная с наиболее частотных словоформ и спускаясь вниз по частотным спискам) и частично очищенных текстовых данных были сформулированы правила автоисправлений (ошибок, опечаток, регистра и пр.),

после чего было начато формирование справочных списков, позволяющих массово применять правила. Первые версии списков были составлены на основе доступных орфографических, частотных и толковых словарей.

В дальнейшем списки расширялись за счет анализа частотных словарей создаваемых текстовых корпусов; ошибки с верными вариантами добавлялись в словарь автоисправлений (в котором на данный момент учтены все ошибки и опечатки с частотностью по собранным корпусам 15 и более).

## 4. Второй этап: сегментация

После нормализации текстов корпуса он разделяется на две части: обучающую и тестовую. Соотношение объемов частей выставляется вручную. По умолчанию тестовая часть составляет 1%. Обучающая часть используется далее при построении частотных словарей и файла n-грамм, тестовая часть используется при проверке и оценке корпуса.

## 5. Третий этап: построение частотных словарей

Частотные словари строятся на основе обучающей части корпуса. Сначала строится словарь с покрытием 100%.

После этого может быть запущена дополнительная чистка словаря, при которой из него удаляются все недопустимые символы и сочетания символов (иноязычные, цифровые и цифробуквенные формы и частоты, символы и сочетания символов, входящие в словарь заполнителей). Так как чистка предшествует построению файла n-грамм, он также не включает словоформ на языках, отличных от основного языка документа.

Далее сохраняется указанное количество вариантов частотного словаря, отличающихся друг от друга покрытием (по умолчанию — 5 словарей с покрытием 95–99% и шагом в 1%), программное обеспечение позволяет также отсекать полный словарь корпуса по частоте. Дополнительно строится словарь, включающий только проверенные словоформы (словоформы, совпадающие с единицами, входящими в список проверенных словоформ).

Входными данными для этого шага являются: обработанная на шаге чистки обучающая часть текстового корпуса и словарь заполнителей fillers.txt, включающий наборы символов, не учитываемых при построении словарей. Для запуска шага необходимо определить язык и имя корпуса. По завершении обработки мы получаем очищенные словари, частотные словари по каждой части корпуса, кривые покрытия, отражающие зависимость размера словаря от объема корпуса (для имен нарицательных, имен собственных и всех словоформ). Стабилизация кривых рассматривается как первичный показатель достаточности объема корпуса.

## 6. Четвертый этап: построение файла n-грамм

После того как отработал шаг «Построение частотных словарей» запускается шаг «Подсчет n-грамм». На этапе предварительной обработки в текстовые файлы добавляются маркеры начала и конца предложения. Таким образом, подсчет словоформ и n-грамм идет для каждой строки отдельно. При построении файла n-грамм по умолчанию подсчитывается количество уни-, би-, три- и тетраграмм по обучающей части корпуса. Далее происходит слияние файлов и формируется общий файл n-грамм, объединяющий данные по всем частям корпуса.

На вход шага поступает обучающая часть чистого текстового корпуса. Для запуска шага необходимо определить язык корпуса, имя корпуса, словарь (применение словаря опционально; словарь ограничивает список слов, добавляемых в файлы n-грамм; словоформы, отсутствующие в словаре, в файле n-грамм заменяются тегом <UNK>). По завершении обработки мы получаем файл n-грамм заданного порядка и файл словаря, которые могут быть использованы при построении ЯМ.

## 7. Пятый этап: построение ЯМ

ЯМ строится на основе файла n-грамм и словаря, полученных на предыдущем шаге. Словарь определяет, n-граммы, включающие какие словоформы, попадут в ЯМ. N-граммы со словами, отсутствующими в словаре, не учитываются при формировании ЯМ.

По файлу n-грамм вычисляется десятеричный логарифм вероятности для каждой уни-, би- три- и тетраграммы (на этом шаге может корректироваться степень модели, так, например, мы можем получить биграммную ЯМ на основе файла, в котором присутствуют уни- би- и триграммы).

Входными данными для этого шага являются файл n-грамм и файл словаря. Для запуска шага необходимо определить язык корпуса, имя корпуса и тип сглаживания (по умолчанию используется сглаживание Гуда-Тьюринга; помимо этого в дополнительных параметрах запуска могут быть выбраны сглаживание Кнесера-Нея или сглаживание Виттена-Белла) и порядок ЯМ. Также в дополнительных параметрах может быть указана минимальная частотность для n-грамм каждого порядка, учитываемых при расчете логарифмов вероятностей. По завершении обработки мы получаем ЯМ.

## 8. Шестой этап: слияние ЯМ

При слиянии из двух полученных ранее файлов ЯМ формируется один. При этом все логарифмы вероятностей пересчитываются в соответствии с заданным нормировочным коэффициентом. Так, например, создав 2 файла моделей по корпусу субтитров и корпусу Интернет-форумов на данном этапе мы можем слить их в один более универсальный.

Входными данными для этого шага служат 2 из файлов ЯМ, полученных при предыдущих запусках шага построения ЯМ. Перед запуском необходимо

определить нормировочный коэффициент для слияния (определяющий при слиянии вес первого, основного, из заданных файлов; вес второго будет равен единице за минусом веса первого) и пути к файлам ЯМ. Помимо этого может быть проведено слияние ЯМ по результатам тестирования. При данном варианте слияния нормировочный коэффициент не указывается вручную, а подбирается по лучшему значению медианы коэффициента связности, полученному на указанном тесте.

На выходе шага мы получаем объединённый файл ЯМ с заданными характеристиками.

## 9. Седьмой этап: тестирование ЯМ

При оценке качества (тестировании) ЯМ используется тестовая часть корпуса, полученная на втором шаге. В имени теста присутствует название корпуса, например, или test-subtitles. Помимо этого при тестировании ЯМ русского языка используется несколько наборов файлов, основную часть которых составляют текстовые файлы новостных сообщений и подстрочники речевых сообщений (спонтанная речь, телефонные переговоры, новости).

Тесты оцениваются по двум показателям: медиана пропуска и медиана коэффициента связности. Для каждого документа подсчитывается процент слов, отсутствующих в словаре, на основании чего подсчитывается медиана пропуска для теста в целом.

Медиана коэффициента связности (*median perplexity*) вычисляется как:

$$2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i)}$$

где $N$ — число элементов в примере, а $q(x_i)$ — вероятность тестовых событий.

Каждый из тестов представляет собой корпус со списком входящих в него текстовых файлов и файл настроек, включающий название теста, его краткое описание и ожидаемое значение медианы логарифма коэффициента связности.

По завершении тестирования мы получаем для каждого теста: файл report.txt (содержащий статистические данные по каждому тексту: количество слов, процент уни- би- и три- и тетраграмм, абсолютное и относительное значения пропуска и логарифма коэффициента связности и пр., а также минимальное, максимальное и среднее значение медианы для теста в целом) и файл report.xml (в котором частично дублируются данные файла report.txt, относящиеся ко всем текстам теста; в файле сохраняются только данные значений пропуска и логарифма коэффициента связности). Кроме того, составляется отчет о тестировании, содержащий краткую информацию по всем тестам: медиану связности, медиану пропуска и время выполнения.

Порог связности задается как лучшее значение, полученное на основе тестирования предыдущих ЯМ. Статус «Success» присваивается, если полученное значение медианы связности не превысило заданного порога.

## 10. Результаты тестирования системы построения ЯМ и полученные модели

Нагрузочное тестирование системы построения ЯМ проводилось на новостном корпусе объемом 4,5 миллиона документов. При построении ЯМ применялось сглаживание Гуда-Тьюринга; триграммы с частотой 1 отбрасывались. На каждом этапе была проведена выборочная проверка результатов. В ходе тестирования была произведена обработка корпуса от чистки до построения 3-граммной ЯМ (также было протестировано построение 2- и 4-граммных файлов n-грамм и ЯМ). Полный цикл обработки корпуса занимает около 55 часов (от 43 до 58 по результатам 5 запусков). В результате обработки корпуса была получена ЯМ со следующими характеристиками:

- проверенный словник ЯМ включает 869 456 единицы;
- количество би- и триграмм составляет 84 623 335 и 103 634 294 соответственно.

По результатам тестирования ЯМ на случайно отобранных подстрочниках новостных сообщений была получена медиана коэффициента связности, равная 88 (что на 12 % лучше данных, которые были получены Сиу Манхунгом на материале английского языка [Manhung, 2000]); пропуск составил 0,2 %. При этом ЯМ, построенная на необработанном корпусе, дала медиану коэффициента связности превышающую 700.

Таким образом, была доказана работоспособность системы на корпусных данных объемом более 4,5 миллионов документов (более миллиарда лексических единиц).

Работы по созданию ЯМ, предназначенной для повышения качества распознавания спонтанной русской речи, велись при помощи созданной автоматизированной системы. Все полученные ЯМ строились на основе проверенного словаря, дающего по тестовым данным пропуск равный 2 %. При тестировании первой версии ЯМ, сформированной путем последовательного запуска шагов «Построение файлов n-грамм» и «Построение ЯМ» был получен коэффициент связности, превышающий 2000 (т. е. ЯМ была основана на текстах, в которых были отделены знаки препинания и исправлено около 30 наиболее частотных ошибок; полной нормализации текстов не проводилось). ЯМ, построенная при запуске всех шагов, позволила снизить коэффициент связности (повысить качество ЯМ) до 377.

Текущая ЯМ была построенная по результатам отработки всех шагов на каждом из трех подкорпусов, а далее слита из трех ЯМ с нормировочными коэффициентами, автоматически определенными по тестовым данным. При тестировании была получена медиана связности, равная 295 (что, безусловно, значительно хуже данных, полученных по результатам тестирования ЯМ, построенной на новостных данных, однако приближается к результатам, полученным А. Б. Холоденко на модели лемм [Холоденко, 2002]). Таким образом, автоматизация обработки позволила нам улучшить первоначальную ЯМ более чем в 6 раз.

## Литература


1. *Zakharov V. P.* (2015), Evaluation of Russian Internet corpora [Otsenka kachestva Internet-korpusov russkogo jazika], available at: http://corpora.phil.spbu.ru/Works2015/Zakharov.pdf.
2. *Khokhlova M.* (2016), A survey of Large Russian Corpora, available at: http://openbooks.ifmo.ru/ru/file/4106/4106.pdf.
3. *D. Kanevsky, M. Monkowsky, J. Sedivy* (1996). Large Vocabulary Speaker-Independent Continuous Speech Recognition in Russian Language. Proceedings of the Conference. SPECOM'96, St. Petersburg, pp. 117–121.
4. *Bogdanova-Beglarian N. V., Blinova O. V., Martynenko, G. Ya., Sherstinova, T. Yu.* (2017). Corpus "One day of speech" in Exploratory Study on Sociolinguistic Variation of Spoken Russian [Corpus "Odin rechevoj den" v issledovanijah sotsiolijgvisticheskoj variativnosti russkoj razgovornoj rechi]. Proceeding of the workshop AR$^3$, pp. 14–20.
5. *S. Manhung et al.* (2000), "Integrating a context-dependent phrase grammar in the variable n-gram framework". Proceeding of ICASSP, pp. 1643–1646.
6. *Holodenko A. B.* (2002), Developing statistical language models for the systems of Russian speech recognition [O postrojeniji statisticheskih jazikovih modelej dlja system raspoznavanija russkoj rechi], Intellectual Systems [Intellektual'nije sistemi], Vol. 6, publication 1–4, pp. 381–394.