Моделирование границ высказываний в спонтанной монологической речи

Степихов Антон Анатольевич¹ Шипило Александр Владимирович²

¹Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург,
Россия

²Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург,
Россия

Sentence boundary modelling in spontaneous monologue

Stepikhov Anton¹ Shipilo Alexander²

¹Saint Petersburg State University (SPSU), St. Petersburg, Russia

²Saint Petersburg State University (SPSU), St. Petersburg, Russia

Аннотация

В докладе представлены первые результаты статистического моделирования границ предложений в спонтанной речи. Материалом исследования является корпус русских спонтанных монологов, содержащий информацию об экспертном членении текстов на предложения и морфологическую разметку. Целью анализа было выявление степени точности n-граммной модели автоматического определения межфразовых границ, в основе которой лежит экспертное членение спонтанной речи.

Abstract

The paper describes the first results of statistical modelling of sentence boundaries in Russian spontaneous speech. The study is based on the corpus of Russian spontaneous monologue which contains information about inter-

annotator agreement in sentence boundary detection and morphological annotation. The analysis aimed to explore the accuracy of an n-gram sentence boundary detection model based on expert manual annotation of unscripted speech.

Ключевые слова: сегментация, членение, спонтанная речь, моделирование, русский язык.

Keywords: segmentation, sentence boundary detection, spontaneous speech, modelling, Russian.

1. Введение

Спустя более чем полвека изучения русской спонтанной речи проблема ее членения на предложения остается одной из ключевых как для лингвистики, так и для компьютерных наук. Проблема заключается в том, что устная речь в отличие от письменного

текста не содержит пунктуационных знаков и поэтому не может быть подвергнута однозначной сегментации. Сложность членения спонтанного текста на фразы связана как с самим характером неподготовленной речи, изобилующей различными сбоями в процессе говорения, так и с языковыми особенностями, например, широким распространением в устной речи бессоюзия.

Тем не менее, информация о межфразовых границах важна для анализа текста как лингвистами, так и автоматическими системами. Наличие границ между предложениями оптимизирует работу информационно-поисковых систем, улучшает автоматическую морфологическую и синтаксическую разметку, автоматическое реферирование и перевод, а также облегчает восприятие человеком автоматически распознанной речи, представленной в виде текста [Shriberg 2005].

Сведения о межфразовых границах получают путем ручной или автоматической сегментации орфографической расшифровки устного текста. Ручная сегментация основывается на экспертной разметке, автоматическая — на компьютерных моделях. Большинство из них стремится воспроизвести ручную разметку, которая считается своего рода «золотым стандартом» для определения межфразовых границ в спонтанной речи.

В последние двадцать лет были разработаны разнообразные модели автоматического определения межфразовых границ. В их основе лежат различные методы и принципы статистического моделирования, такие как *n*-граммные языковые модели, байесовская вероятность, скрытые марковские модели, деревья принятия

решений, максимальная энтропия, бустинг, линейная и логистическая регрессия, условные случайные поля, метод опорных векторов, глубокие нейронные сети, метод *k* ближайших соседей и др. (см., напр., [Magimai-Doss et al. 2007; Liu et al. 2010]). Перечисленные модели обычно основаны как на текстовой (лексической, грамматической и т.д.), так и на просодической информации (паузация, темп, мелодика, интенсивность и др.), при этом существуют и такие подходы к моделированию границ между предложениями, которые опираются только на просодические или текстовые параметры.

В настоящей статье приводятся первые результаты разработки модели автоматического определения границ между предложениями в русской спонтанной монологической речи. Данная модель основывается на текстовой информации и строится на частеречных *n*-граммах, вероятность которых в тексте определяется на базе байесовского алгоритма.

2. Материал и методика исследования

2.1. Корпус

Материалом для исследования послужил корпус русских спонтанных монологов, который состоит из 160 текстов разных типов (около 55000 словоформ, 9 часов звучания), полученных от 32 информантов (по 5 текстов от каждого).

Для записи спонтанной речи информантам предлагалось прочитать и пересказать рассказ И. А. Бунина «Лапти» и фрагмент его рассказа «Антоновские яблоки» (сюжетный и описательный тексты), описать цикл рисунков Х. Бидструпа «Шляпа» и

картину В. ван Гога «Хижины», а также рассказать о своем свободном времени или образе жизни.

Корпус сбалансирован по возрасту, полу и профессии информанта. С целью достижения максимальной естественности речи в условиях записи для нее подбирались люди, хорошо знакомые с исследователем (подробнее о корпусе см. [Stepikhov 2013a]).

2.2. Экспертное членение

Записанные тексты были подвергнуты орфографической расшифровке, в которой отсутствовали знаки препинания. Затем расшифровки предлагались группе из 20 экспертов-филологов (студентов и преподавателей) для членения их на предложения. В качестве маркера границы высказывания предлагалось использовать точку либо любой другой знак, удобный для эксперта (например, косую черту). Для того чтобы эксперты при анализе текста могли сосредоточить свое внимание на его грамматике и семантике, в тексте не содержалось графических обозначений заполненных пауз хезитации, а также других ремарок, описывающих процесс порождения речи ([вздох], [смех] и т.п.). Время для выполнения задания не ограничивалось. В результате проведенного эксперимента было получено 20 вариантов синтаксического членения каждого текста.

3. Анализ данных и результаты исследования

3.1. Анализ экспертного членения и морфологическая разметка

В ходе анализа экспертной разметки было вычислено количество пограничных совпадений (далее –

КПС), т. е. количество экспертов, которые отметили межфразовую границу в том или ином месте. КПС может принимать значения от 0 (граница не обозначена ни одним из экспертов) до 20 (полное совпадение в оценке места границы). Ниже приводится фрагмент размеченного текста после подсчета КПС:

что значит свободное время 1 у студентов 16 у студентов постоянно свободное время 18 учебный процесс и есть свободное время 14 вот 13 ну как 3 приходишь после занятий 1 уставший 2 делаешь себе покушать 9 родителей нету 13 сделал себе покушать 7 поел 11 вот 6 пошел естественно к компьютеру 2 проверил почту 2 посмотрел кто тебе позвонил 1 кто не позвонил 14 хотя дни поразному проходят 15 порой 3 вот 1 Миша зайдет 11 с Мишей посидим поболтаем 12 он там посидит 1 поиграется у меня в компьютер 10 вот 8.

По итогам данного анализа границей предложения считалось:

- место, отмеченное 60% экспертов и более (от 12 до 20 человек);
- позиция после последнего слова в зоне границы между высказываниями с теми же пороговыми значениями (о зонах границ между высказываниями см. [Stepikhov 2013b]).

Кроме экспертной аннотации, была выполнена морфологическая разметка корпусных данных. С этой целью был использован морфоанализатор Sketch Engine [Kilgariff et al. 2014].

3.2. Статистическое моделирование

Корпусные данные были в дальнейшем подвергнуты статистическому моделированию. Его

целью было выявление степени точности модели автоматического определения межфразовых границ, в основе которой лежит экспертная разметка спонтанной речи.

В качестве базовой модели была выбрана n-граммная модель, алгоритм которой основывался на частеречной характеристике словоформ на границах предложений, полученных в результате экспертного членения монологов.

Сопоставлялись биграммная и триграммная модели членения текста.

Для моделирования и проверки точности модели корпус был разделен на две части. Обучающая выборка, на базе которой проходила тренировка статистической модели, составила 80% корпуса. Тестовая выборка, по которой оценивалась работа модели, составила 20% корпусных данных.

Работа модели оценивалась на основе стандартных для обработки естественных языков показателей: общая точность (accuracy), точность (precision), полнота (recall) и F-мера.

Биграммная модель показала следующие результаты:

- общая точность – 0,63;
- точность – 0,14;
- полнота – 0,76;
- F-мера – 0,23.

Результаты модели, основанной на сочетаниях триграмм:

- общая точность – 0,68;
- точность – 0,14;
- полнота – 0,67;
- F-мера – 0,24.

Полученные результаты выявили необходимость учета в модели синтаксической информации, а также тестирования моделей, основанных на иных статистических алгоритмах.

Благодарность. Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 15-04-00165.

Литература

- Shriberg E.* How people really talk and why engineers should care // Proc. Interspeech 2005. Pp. 1791-1794.
- Magimai-Doss M., Hakkani-Tür D., Çetin Ö., Shriberg E., Fung J., Mirghafori N.* Entropy based classifier combination for sentence segmentation // Proc. ICASSP 2007. Vol. 4. Pp. 180-192.
- Liu Y.-F., Tseng S.-C., Jang J.-S. R., Chen C.-H. A.* Coping imbalanced prosodic unit boundary detection with linguistically motivated prosodic features // Proc. Interspeech 2010. Pp. 1417-1420.
- Stepikhov A.* Analysis of expert manual annotation of the Russian spontaneous monologue: Evidence from sentence boundary detection // SPECOM 2013. LNCS (LNAI). Vol. 8113. 2013a. Pp. 33–40.
- Stepikhov A.* Resolving ambiguities in sentence boundary detection in Russian spontaneous speech // TSD 2013. LNCS (LNAI). Vol. 8082. 2013b. Pp. 426–433.
- Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* The Sketch Engine: ten years on // Lexicography ASIALEX. Vol. 1. 2014. Pp. 7–36. См. также: <https://www.sketchengine.co.uk> (Дата обращения: 28.06.2016).

References