О.Н. Ляшевская
O.N. Lyashevskaya

ПАРАЛЛЕЛЬНЫЙ КОРПУС АВТОМАТИЧЕСКИХ И РУЧНЫХ РАСШИФРОВОК УСТНОЙ РУССКОЙ РЕЧИ[1]

A PARALLEL CORPUS OF AUTOMATIC AND MANUAL TRANSCRIPTS OF RUSSIAN SPOKEN DISCOURSE

Аннотация. В докладе представлен проект исследовательского корпуса параллельных автоматических и ручных расшифровок устной речи. Автоматическая расшифровка получена с помощью сервиса SpeechKit Cloud (Яндекс); выравнивание автоматической и ручной аннотаций производилось на уровне условной диалогической единицы, условного слова и слога (в автоматическом режиме с последующей ручной коррекцией). Обсуждаются проблемные случаи выравнивания и вопросы аннотации ошибок распознавания.

Ключевые слова. Корпус оценки качества, лингвистика ошибок, автоматическое распознавание речи, параллельная разметка.

Abstract. The paper outlines the project of an evaluation corpus with parallel automatic and manual transcripts of Russian oral speech. Automatic speech recognition was provided by SpeechKit Cloud (Yandex); alignment of the automatic and manual annotations was performed at the level of discourse units, word-to-word and syllable-to-syllable matching. We discuss some problematic cases of alignment as well as annotation of ASR errors.

Keywords. Evaluation corpora, linguistics of errors, automatic speech recognition, parallel annotation.

[1] Работа выполнена при поддержке гранта РФФИ № 15-07-09306.

1. Компьютер как “несовершенный” носитель языка и “автор” корпуса ошибок

Слабым звеном корпусного подхода к описанию языка долгое время считалось отсутствие negative evidence, т.е. отрицательных примеров, в материале корпусов. Вместе с тем, с ростом популярности так называемых учебных корпусов (learner corpora) пришло понимание, что именно они обеспечивают богатый эмпирический материал, восполняющий лакуну корпусных примеров с ошибками. Ценность этих данных в том, что это не искусственно построенные корреляты к “идеальным” предложениям, а спонтанно порожденные высказывания, которые носителями языка оцениваются как содержащие грамматические, прагматические, орфографические, орфоэпические и т.п. ошибки.

Как правило, авторами текстов в корпусах ошибок являются а) дети, усваивающие первый язык; б) дети и взрослые, усваивающие вторые языки; в) носители так называемого эритажного языка, т.е. потомки эмигрантов, выросшие в иноязычной среде и слабо владеющие языком родителей и дедов; г) неопытные переводчики; д) студенты, практикующиеся в малоосвоенном жанре академического письма и т.д. Таких носителей языка принято называть “несовершенными” (или более политкорректно, “нестандартными”) говорящими. Однако компьютерно-лингвистические системы – автоматические переводчики, модули генерации новостей, системы автоматического распознавания и порождения речи – которые задуманы как имитаторы речевой деятельности человека, также легко отнести к классу “несовершенных” носителей языка. Сравним, например, две расшифровки аудиозаписи, сделанные носителем языка (1) и компьютером (2):

(1) когда казалось, о каком там искусстве говорить

(2) как оказалось из катакомб об искусстве говорить

Несмотря на то, что технологии автоматического распознавания активно развиваются, дополняют акустическую модель параметрами синтаксической, семантической, доменной и т.п. дистрибуции (см., например, свежий обзор и ссылки в [Важенина и др. 2014]), очевидно, что полнота и точность письменной кодировки речевого сигнала все еще далека от идеала. Вместе с тем, вполне можно представить, что аналогичную (2) расшифровку мог бы сделать и студент-иностранец, и даже носитель языка, оказавшийся в шумной обстановке.

В этой работе мы хотим привлечь внимание лингвистов-теоретиков к тому интересному факту, что, казалось бы, чисто технические корпусы, которые создаются компьютерными лингвистами для оценки качества систем обработки и понимания текста, могут оказаться полезным ресурсом для изучения ошибок. В частности, корпус расшифровок устной речи, о котором далее пойдет речь, здесь не исключение. Немаловажно, что в отличие от экзаменационных сочинений и учебных переводов, непосредственные результаты аудирования (в виде транскриптов прослушанных человеком звуковых текстов) малодоступны, а автоматические транскрипты позволяют в какой-то мере компенсировать их недостаток. Кроме того, как кажется, ответы компьютерных систем, созданные с помощью вероятностных моделей, могут натолкнуть исследователей на важные теоретические вопросы о том, как устроено восприятие и понимание речи человеком.

2. Проект корпуса оценки качества автоматического распознавания русской речи

Проект оценки качества автоматического распознавания устной русской речи задуман как очередной блок инициативы RU-EVAL – форума по оценке качества систем автоматической обработки текста [Ляшевская и др. 2010; Толдова и др. 2012, 2014). Форум был организован разработчиками Национального корпуса русского языка с целью обсуждения состоя-

ния отрасли и качества компьютерных систем и электронных ресурсов, разрабатываемых исследовательскими и коммерческими лабораториями, а также частными энтузиастами. Одним из важных результатов работы форума стало создание корпусов с параллельной разметкой, отражающей ответы участников на ту или иную задачу автоматической обработки текстового корпуса. Отдельным слоем наносится ручная разметка, так называемый «золотой стандарт», относительно которого измеряется качество других аннотаций.

Пилотный проект корпуса оценки качества распознавания речи содержит звучащие файлы в wav-формате и два блока транскриптов с разметкой, ручной и автоматический. Ручная аннотация произведена студентами московских вузов, с последующей редакторской проверкой. Автоматический блок на пилотной стадии содержит только один вариант компьютерной аннотации, полученный с помощью сервиса SpeechKit Cloud (Яндекс).

Акустическая модель SpeechKit вычисляет распределение вероятностей для примерно 4000 сенонов (начальных, серединных или конечных сегментов фонемы) в каждом 20 мс фрейме, на который разбивается голосовой поток с использованием комбинации скрытых Марковских моделей и глубокой нейронной сети прямого действия. Вторая, языковая модель ранжирует гипотезы с использованием рекуррентной нейронной сети. По оценкам разработчиков, полнота распознавания составляет 82% (общая точность в доступных источниках не указана). Поскольку изначально система разрабатывалась под задачи голосового распознавания поисковых запросов, она хорошо работает с короткими монологическими диктовками; точность распознавания цифр и географических объектов составляет 95%. По нашим эмпирическим оценкам, в длинных записях качество распознавания существенно падает при фоновом шуме, а также в диалоге (например, если говорящие сидят на разном расстоянии от микрофона).

Аннотация транскриптов включает:

а) лексико-грамматическую разметку (часть речи, лемма, грамматические характеристики, анализатор Mystem 3.0);

б) разметку вокалической и орфоэпической структуры;

в) словообразовательную разметку (автоматически в объеме словообразовательного словаря НКРЯ);

г) разметку несоответствий между автоматическим и ручным транскриптами и классов этих ошибок.

Согласно практике корпусов оценки качества, (а–в) предполагает быструю автоматическую аннотацию, требования к ее полноте и точности диктуются исключительно задачами обнаружения несоответствий. Разметка типа (г) в пилотном корпусе осуществляется, естественно, вручную.

Выравнивание ручных и автоматических транскриптов на уровне условных диалогических единиц производится с помощью системы HunAlign, работающей под GUI “Евклид” (разработчики Т.Архангельский, М.Кружков, Е.Лощилова), с последующей ручной коррекцией результатов. Далее, на уровне условного слова и слога выравнивание проводится с помощью специально созданного скрипта. Привязка фрагментов письменных транскриптов к временным меткам на звуковой дорожке не проводится.

Тестовый материал в пилотном корпусе состоит из записей бытовой речи, радиопередач, докладов на конференции, социологического интервью (35 записей, из них 11 монологов, 19 диалогов и 5 полилогов, общий объем звучания около 4,5 ч.).

3. Типология несоответствий

Типология несоответствий между ручным (Р) и автоматическим (А) вариантами транскриптов ориентирована, во–первых, на оценку полноты автоматического распознавания, во–вторых, на критерии акустического подобия и текстовой

связности, и в–третьих, на многофакторную оценку точности в соответствии с типами компонентов разметки (а–в).

Объем найденных несоответствий в первую очередь зависит от качества выравнивания. Сложность выравнивания связана с тем, что разбиение текста на условные диалогические единицы в транскрипте А и транскрипте Р по определению разнится: автоматическая модель руководствуется скорее паузами, тогда как носители языка ориентируются более на границы клауз и логическую завершенность фрагментов высказывания. Работа модуля выравнивания затруднена и в том случае, если полнота расшифровки А низка (в одном из наших тестовых файлов она составила 17%). Выравнивание на пословном уровне также может быть неточным, если распознана только часть условной дискурсивной единицы, однако все же зависимость от полноты разметки здесь не такая сильная.

Основной единицей, для которой оценивается качество соответствий, является word error rate, т.е. мы оцениваем, имеется ли соответствие слову W_P из транскрипта Р в транскрипте А и какова точность этого соответствия. Границы слова определяются фонетически, с допустимостью замен в падежах и предлогах, см. ниже.

Таблица 1. Оценка ошибок при распознавании форм фамилии *Мейерхольд.*

Кол-во	**Ручной вариант**	**Авт. вариант**	**R**	**Type**	**P**
3	Мейерхольд	мейерхольд	1	–	1
1	Мейерхольд	мр в холле	1	A	0.88
1	Мейерхольд	верховой	1	A	0.78
0	Мейерхольд	<НЕТ>	0	–	0
3	Мейерхольда	мейерхольда	1	–	1
1	Мейерхольду	немировича	1	C	0.41

Оценка полноты (R) имеет три градации: 1 – 0,5 – 0.

Акустический тип несоответствия (Type) предполагает, что модель не смогла выбрать из фонетически близких альтернатив правильный вариант; ошибка текстовой связности предполагает, что модель выбирала из частотных паттернов контекстной сочетаемости, и этот выбор оказался неверным (ср., например, ошибочную расшифровку *немировича* вместо *Мейерхольду* в контексте *Станиславского и* __ [1]). В табл. 1 эти типы ошибок кодируются как A и C, соответственно.

Точность распознавания (P) расчитывается согласно коммуникативному принципу. Наиболее мягко штрафуются нарушения предложно-падежной структуры и опущение частиц и союзов, так как они редко ведут к потере информации при передаче сообщения. Поскольку окончания слов в беглой речи значительно редуцируются, необходимость восстановления этой информации для понимания высказывания в режиме реального времени может подвергаться сомнению. Часто наблюдаемые опущения и замены предлогов, союзов и частиц могут быть связаны как с объективными трудностями их распознавания, так и с технологиями ранжирования гипотез, в которых дистрибуция частотных строевых элементов может не учитываться.

Наиболее жестко штрафуются расшифровки с длинными цепочками потерь (несколько неправильно распознанных слов подряд) и нарушениями словных границ. Помимо этого, в качестве штрафов вычисляется расстояние между слоговыми структурами (включая схему ударения), вокалическими структурами, консонантными структурами. В то же время, словообразовательные замены и замены с сохранением части речи получают льготные бонусы. Формула взвешивания, безусловно,

[1] Типы ошибок не являются взаимоисключающими, ср. ошибочную расшифровку *власенко* вместо *классика* в контексте __, *Чайковский*.

должна еще будет уточняться на более представительном корпусном материале.

Литература

1. *Ляшевская О., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С., Савчук С., Коваль С.* и др. (2010), Оценка методов автоматического анализа текста: морфологические парсеры русского языка. Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог», вып. 9 (16), с. 318–326.

2. *SpeechKit Cloud*: https://tech.yandex.ru/speechkit/cloud/.

3. *Толдова С., Ройтберг А., Ладыгина А., Васильева М., Азеркович И., Курзуков М., Сим Г., Горшков Д., Иванова А., Недолужко А., Гришина Ю.* (2014), RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог», вып. 13 (20), с. 681–695.

4. *Толдова С.Ю., Соколова Е.Г., Астафьева И., Гарейшина А., Королева А., Привознов Д., Сидорова Е., Тупикина Л., Ляшевская О.Н.* (2012), Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», вып. 11 (18), с. 797–809.

5. *Важенина Д.А., Кипяткова И.С., Марков К., Карпов А.А.* (2014), Методика выбора фонемного набора для автоматического распознавания русской речи. Труды СПИИРАН, вып. 5.36, с. 92–113.

References

1. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A. et al.* (2010), Otsenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka [NLP evaluation: Russian morphological parsers], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 9 (16), pp. 318–326.

2. *SpeechKit Cloud*: https://tech.yandex.ru/speechkit/cloud/.

3. *Toldova S., Roytberg A., Ladygina A., Vasilyeva M., Azerkovich I., Kurzukov M., Sim G., Gorshkov D., Ivanova A., Nedoluzhko A., Grishina Yu.* (2014), RU–EVAL–2014: Evaluating anaphora and coreference resolution for Russian // Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 13 (20), pp. 681–695.

4. *Toldova S., Sokolova E., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O.* (2012), Otsenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011–2012: Russian syntactic parsers]. Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 11 (18), pp. 797–809.

5. *Vazhenina D.A., Kipyatkova I.S., Markov K., Karpov A.A.* (2014), Metodika vybora fonemnogo nabora dlja avtomaticheskogo raspoznavanija russkoj rechi [Technique for Phoneme Set Selection for Automatic Russian Speech Recognition]. In: Trudy SPIIRAN [Papers of SPIIRAN], Vol. 5.36, pp. 92–113.

Ольга Ляшевская / Olga Lyashevskaya

НИУ Высшая школа экономики, Институт русского языка им. В.В.Виноградова РАН (Россия) / National Research University Higher School of Economics, Vinogradov Institute of the Russian Language RAS (Russia).

E-mail: olesar@yandex.ru