Алгоритм автоматического поиска ошибок в транскрибациях

На основе марковских цепей

Авторы: Смирнова Екатерина,
Черная Анастасия


Находим биграмы-кандидаты на ошибку
• Сбор обучающего корпуса
• Марковская цепь по биграмам
Находим слова-ошибки в биграмах
• Для каждого слова в биграме проверяем, образует ли оно вероятную биграму
• Отбираем слова «без пары»
Считаем ошибкой, если можно заменить на близкое слово
• Ищем слова, близкие слову-ошибке по расстоянию Левенштейна
• Если при замене слова на близкое вероятность биграмы выше, то исходное слово – ошибка

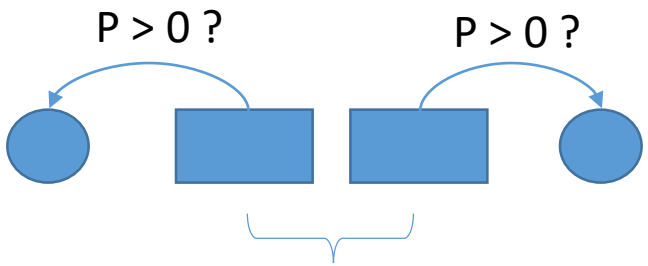
Находим биграмы-кандидаты на ошибку

- Сбор обучающего корпуса (14 264 084 токенов, объем словаря - 404 110 словоформ)
- Марковская цепь по биграмам:
  - Лемматизация
  - Отсев биграм, где одно из слов отсутствует в словаре
  - Указание порога для определения кандидата на ошибку (установлен = 0)

```
Обрабатываю ряд  35
        Всего ошибок:  36  из  147
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  38
        Всего ошибок:  24  из  85
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  39
        Всего ошибок:  16  из  80
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  42
        Всего ошибок:  11  из  67
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  45
        Всего ошибок:  24  из  88
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  46
        Всего ошибок:  18  из  70
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  52
        Всего ошибок:  36  из  128
        Количество биграм с отсутствующим в словаре словом:  2
Обрабатываю ряд  54
        Всего ошибок:  18  из  109
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  55
        Всего ошибок:  7  из  41
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  58
        Всего ошибок:  10  из  61
        Количество биграм с отсутствующим в словаре словом:  0
```

Находим слова-ошибки в биграмах

- Для каждого слова в биграме проверяем, образует ли оно другую вероятную биграму в тексте
- Отбираем слова «без пары»



Считаем ошибкой, если можно заменить на близкое слово

- Ищем слова, близкие слову-ошибке по расстоянию Левенштейна в словаре тренировочного корпуса
- Отбираем только слова с отличной от исходного слова леммой
- Заменяем исходное слово-ошибку на леммы отобранных слов
- Считаем вероятность полученной биграмы и выделяем два типа ошибок:
  - Первый тип - если при замене слова на близкое вероятность биграмы выше;
  - Второй тип – если вероятность та же или ниже (т.е. замена на схожее слово не дала результата)

	audio_ID	alphacep_transcripts	mistakes	mistakes_1st_type	mistakes_2nd_type	absent_words	bigram_mist	new_words
35	Pic-RUS_01-f_Pr-R.eaf	жилбыл один дяденька по его жены скоро должно ...	плот, ночного, манекен, стоящее, посол, варт, ...	ночного, манекен, наставь, дядечка, маменьку	плот, ночного, манекен, стоящее, посол, варт, ...		случится плот, дяденька ночного, то манекен, н...	необычный, манекенщица, настать, дядька, машенька
38	Pic-RUS_01-f_Ski-T.eaf	генин жизни одного очень увлекающийся спортом ...	генин, спортом, стал, хочется, партийный, това...	хочется, норг	генин, спортом, стал, хочется, партийный, това...		генин жизни, увлекающийся спортом, проснулся с...	хотеть, нога
39	Pic-RUS_02-f_Pr-R.eaf	однозначным был день рождения мышь решил подар...	однозначным, мышь, усмотрел, посылала, выбрал	усмотрел, посылала	однозначным, мышь, усмотрел, посылала, выбрал		однозначным был, рождения мышь, и усмотрел, в ...	смотреться, послать
42	Pic-RUS_02-f_Ski-T.eaf	этот человек встал рано утром позавтракал а по...	наложены, божественного, поранился		наложены, божественного, поранился		отправился наложены, голову божественного, сил...	
45	Pic-RUS_03-m_Ski-R.eaf	знакомым мне здесь рассказали одну смешную и п...	нагорных, проснулся, собрался, доскачет, тече,...		нагорных, проснулся, собрался, доскачет, тече,...		то нагорных, нагорных проснулся, проснулся соб...	
46	Pic-RUS_03-m_Ski-T.eaf	один чувак решил покататься на лыжах както ран...	зимним, снаряжение, слышь, переломов, перед		зимним, снаряжение, слышь, переломов, перед		ранним зимним, дядя снаряжение, горы слышь, ку...	
52	Pic-RUS_05-m_Pr-T.eaf	однажды константин решил подарить жене подарок...	однажды, константин, сумки, лампы, кастрюли, п...	сумки, ручьем, купе, нету, призадумался, братишка	однажды, константин, сумки, лампы, кастрюли, п...	сказала муциан, муциан он	однажды константин, однажды константин, стоят ...	сука, ружьё, купец, нет, задуматься, братушка
54	Pic-RUS_05-m_Ski-T.eaf	миша проснулся очень рано гдето в восемь часов...	сделав, преспокойненько, отстегнулась, перекрутил	перекрутил	сделав, преспокойненько, отстегнулась, перекрутил		утра сделав, было преспокойненько, лыжа отстег...	перекусить
55	Pic-RUS_06-f_Pr-R.eaf	один мужчина решил подарить своей жене какойни...	советовали, сувенир, маленькую, машинку	маленькую	советовали, сувенир, маленькую, машинку		понравилось советовали, ей сувенир, сувенир ма...	маленький

жилбыл один дяденька по его жены скоро должно было случиться день рождения был случится плот (*вот*) дяденька ночного (*очень долго*) мучился не знал как обычно выбрать подарок какой получше он ходил по магазинам выбирал думал чтобы мог (*ему*) купить то хотел купить сумку то он хотел купить часы то манекен но все не получалось выбрать чтонибудь стоящее ноги (*в итоге*) он отчаялся пришёл спросить у своих детей может быть они дадут какоето дельный совет дети недолго думая сказали с чего бы хотел их мало (*мама*) так как дети сюда (*всегда*) больше знают сказали купить ей машину по глупости посол (*пошел в*) варт (*авто*) салон посмотрел машины в итоге понял что всетаки наверно дорог один дядька заявил что это наставь (*достаточно*) приличную сумму стоит денег дядечка носок (*нашел*) компромисс он купил маменьку (*маленькую*) машинку подарил её собственно говоря своей жене в общемто не уверен что она была счастлива дети тоже както были смущены один дети (*дядечка*) осталась на воле (*доволен*)

- 13 ошибок

Реальные ошибки транскрибатора

Ошибки транскрибатора, найденные алгоритмом

Первая категория ошибок

Вторая категория ошибок

(-> *слово на замену*)

жилбыл один дяденька по его жены скоро должно было случиться день рождения был случится плот дяденька ночного (-> *необычный*) мучился не знал как обычно выбрать подарок какой получше он ходил по магазинам выбирал думал чтобы мог купить то хотел купить сумку то он хотел купить часы то манекен (-> *манекенщица*) но все не получалось выбрать чтонибудь стоящее ноги он отчаялся пришёл спросить у своих детей может быть они дадут какоето дельный совет дети недолго думая сказали с чего бы хотел их мало так как дети сюда больше знают сказали купить ей машину по глупости посол варт салон посмотрел машины в итоге понял что всетаки наверно дорог один дядька заявил что это наставь (-> *настать*) приличную сумму стоит денег дядечка (-> *дяденька*) носок компромисс он купил маменьку (-> *машенька*) машинку подарил её собственно говоря своей жене в общемто не уверен что она была счастлива дети тоже както были смущены один дети осталась на воле

- 13 ошибок (7 правильных)

однозначно (у одной женщины) был день рождения её муж решил подарить и (ей) подарок долго искал хозяином (ходил) в магазином смотрела (посмотрел) что есть но ничего выбрать не мог а тогда он поинтересовался у детей они предложили подарить с вами (маме) машину он приехал автосалон а выбора (выбрал) в большую красивую машину то (которую) решил повторять своей жене вот но когда он узнал этой машиной он ужаснулся решил что нет такого не может вечером когда он приехал домой выяснилось что он нашёл выход из этого положения привёз не (жене) маленькой грушой (игрушечную) машинку здесь (дети) были поражены а она (жена) была немножко мне (гневе)

- 12 ошибок

Реальные ошибки транскрибатора

Ошибки транскрибатора, найденные алгоритмом

Первая категория ошибок

Вторая категория ошибок

(-> *слово на замену*)

однозначно был день рождения её муж решил подарить и подарок долго искал хозяином в магазином смотрела что есть но ничего выбрать не мог а тогда он поинтересовался у детей они предложили подарить с вами машину он приехал автосалон а выбора в большую красивую машину то решил повторять своей жене вот но когда он узнал этой машиной он ужаснулся решил что нет такого не может вечером когда он приехал домой выяснилось что он нашёл выход из этого положения привёз не маленькой грушой (-> *игрушка*) машинку здесь были поражены а она была немножко мне

- 3 ошибки (2 правильные)

For all alphacep transcriptions:
Precision: 0.5111
Recall: 0.6571
F1: 0.5750

For file: 35_score_alpha
Precision: 0.5385
Recall: 0.5385
F1: 0.5385
For file: 38_score_alpha
Precision: 0.4615
Recall: 1.0000
F1: 0.6316
For file: 39_score_alpha
Precision: 0.8000
Recall: 0.6667
F1: 0.7273
For file: 45_score_alpha
Precision: 0.5556
Recall: 1.0000
F1: 0.7143
For file: 46_score_alpha
Precision: 0.2000
Recall: 0.2000
F1: 0.2000

For all ABK transcriptions:
Precision: 0.3636
Recall: 0.2927
F1: 0.3243

For file: 35_score_abk
Precision: 0.5000
Recall: 0.2000
F1: 0.2857
For file: 38_score_abk
Precision: 0.3636
Recall: 0.8000
F1: 0.5000
For file: 39_score_abk
Precision: 0.6667
Recall: 0.1667
F1: 0.2667
For file: 45_score_abk
Precision: 0.3333
Recall: 0.3333
F1: 0.3333
For file: 46_score_abk
Precision: 0.2222
Recall: 0.2500
F1: 0.2353

Развитие

- Расширение корпуса
- Поиск схожих слов по их фонетическому представлению
- Составление семантических цепочек на основе WordNet