Алгоритм автоматического поиска ошибок в транскрибациях

на основе марковских цепей и векторной модели

Авторы:

Смирнова Екатерина, Черная Анастасия

ДПО “Компьютерная лингвистика”, ВШЭ, 2020

Цель проекта:

разработать алгоритм автоматического поиска ошибок в текстах распознанной речи.

Характеристика данных проекта

Корпус транскрибаций для тестирования алгоритма:

24 транскрибации аудиозаписей рассказов о подарках и катании на лыжах проекта SpokenCorpora, распознанных с помощью модели Alphacep технологии Kaldi (3775 слов)

43 транскрибации аудиозаписей из того же корпуса, распознанных с помощью модели от сервиса АБК с использованием технологии Kaldi (4166 слов)

Корпус для обучения модели на основе марковских цепей:

1. Открытый корпус русского языка проекта OpenCorpora + корпус субтитров к фильмам и сериалам + корпус сказок для детей и взрослых с сайта Kaggle (31 747 468 - токенов, объем словаря - 794 924 словоформ)

2. Открытый корпус русскоязычных диалогов, извлеченных из художественной литературы, документалистики и пр. + корпус сказок для детей и взрослых (общий объем 14 274 009 токенов, 402 925 словоформ)

Модели и словари:

Словарь частотных лемм Ляшевской и Шарова
Словарь вводных слов и частотных дискурсивных маркеров
Векторная модель, обученная на корпусе Taiga с функциональными словами (объем словаря - 249 946 слов)

Используемые библиотеки

- sklearn
- gensim
- nltk
- pymorphy2
- opencorpora
- fuzzywuzzy
- pandas
- fonetika

Кратко об используемых системах распознавания

Достоинства:

- Открытый код, свободный доступ
- Самая высокая точность распознавания WER (в сравнении с HTK, CMU Sphinx, Julius, iAtros, RWTH ASR)
- Предобученные модели для русского языка: alphacep и АБК

Недостатки:

- Проигрывает конкурентам в скорости
- Предобученные модели должны быть дообучены на своих данных
- Документация сложна для неопытных пользователей

1. Краткая классификация ошибок в тестируемых транскрибациях:

Ошибки в членении фонетических слов на слова на письме:

- **Несколько слов вместо одного:**

  слез фромм кожа (с леспромхоза), а не (они), сучка трубам (сучкорубом), гидра геологическая (гидрогеологическая); дети осталась на воле (довольны);

  работал прицеп щекам (прицепщиком)

- **Одно слово вместо нескольких:**

  слышь (с лыж); тоесть (то есть), наложены (на лыжный), нагорных (на горных), чтобы (что бы), варт салон (в автосалон), адам (а там); задачу (на дачу); поехали загреба (за грибами)

2. Краткая классификация ошибок в тестируемых транскрибациях:

- **Неверно распознаны звуки в начальной позиции:**

пот (вот), том (потом); убил (вбил); глухой (бухой); стал (встал)

- **Неверно распознаны звуки в середине слова:**

  с выпиской (с выпивкой); немножко выбил (выпил); казаться (кататься);
  жил-был один молодой человек да одевайся (дядя Вася); комбайн надувался (назывался); не пети (пейте) алкоголь

- **Неверно распознаны звуки в конечной позиции:**

слышь (с лыж), еле (ели); вода (вот), к (ка [пойду-ка]), но(ну), тут(то), пьяно(пьяный); перелома (переломал); ноги (в итоге); в горе (с горы); дети осталась на воле (остались довольны));
загреба (за грибами)

3. Краткая классификация ошибок в тестируемых транскрибациях:

“И”-“Е”

загреба (за грибами); загребай (за грибами); пришлось коли (Коле); а не (они)

“О”- “А”

а не (они); нападал (наподдал); гидра геологическая (гидрогеологическая)

4. Краткая классификация ошибок в тестируемых транскрибациях:

Опущены слова:

(часто опускаются короткие слова, когда один звук - это одно слово)

он приехал (в) автосалон; он ходит (в) гипсе; пошел (в) детский мир;
приехала скорая помощь (и) его в гибнут кротова ; но когда он узнал (цену) этой машиной;
решил ещё раз покататься на лыжах (покатался) он не очень удачно

Неверно распознанные предлоги:

по (у) его жены; в горе (с горы); задачу (на дачу); в анкете (на банкете); мы ездили с друзьями в воскресенье погреба (за грибами)

Лишние распознанные звуки:

пошёл кататься (с) на лыжах; он (в) упал с горы; в день рождения своей жены (и) он пошел; через час (с) у него

Модули алгоритма

Модуль предобработки корпусов для обучения

Модуль предобработки корпуса транскрибаций

Модуль поиска ошибок на основе марковской цепи

Модуль поиска ложно найденных ошибок на основе Word2Vec

Модуль предобработки корпусов для обучения

Модуль предобработки корпусов для обучения:

- токенизация
- удаление пунктуации и других нерелевантных для задачи токенов
- конвертация цифр в слова
- замена ‘ё' на 'е'
- приведение слов к нижнему регистру
- лемматизация
- составление биграмм

Модуль предобработки корпуса транскрибаций

Модуль предобработки транскрибаций:

- замена ‘ё’ на ‘е’
- вставка дефисов для корпуса Alphacep
- Исключение из обработки вводных слов и дискурсивных маркеров
- составление биграмм

Модуль поиска ошибок на основе марковской цепи

1

Близость слов для замены - классическое расстояние Левенштейна

- Марковская цепь по биграмам

Находим биграмы-кандидаты на ошибку

Находим слова-ошибки в биграмах

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

- Ищем слова, близкие слову-ошибке по расстоянию Левенштейна
- Если при замене слова на близкое вероятность биграмы выше, то исходное слово – ошибка

Считаем ошибкой, если можно заменить на близкое слово

Модуль поиска ошибок на основе марковской цепи

2

Близость слов для замены - расстояние Левенштейна по фонетическим представлениям слов

- Марковская цепь по биграмам

Находим биграмы-кандидаты на ошибку

Находим слова-ошибки в биграмах

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

- Ищем слова, **фонетически** близкие слову-ошибке по расстоянию Левенштейна
- Если при замене слова на близкое вероятность биграмы выше, то исходное слово – ошибка

Считаем ошибкой, если можно заменить на близкое слово

Более подробно про поиск ошибок на основе марковской цепи с использованием классического расстояния Левенштейна:

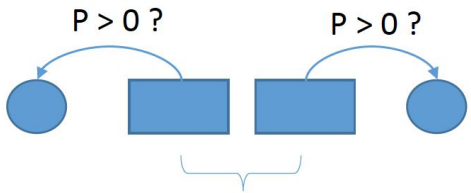
Находим биграмы-кандидаты на ошибку

Марковская цепь по биграмам:

- Лемматизация
- Отсев биграм, где одно из слов отсутствует в словаре тренировочного корпуса
- Указание порога для определения кандидата на ошибку (по умолчанию установлен 0)
- Отсев вводных и дискурсивных слов (по умолчанию)

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/словаре фонетических представлений 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)

Считаем ошибкой, если можно заменить на близкое слово

БЛИЗОСТЬ

Классическое расстояние Левенштейна

```
{'завтраку': ['завтра',
              'завтрак',
              'завтраком',
              'завтракал',
              'завтраке',
              'завтрака',
              'завтракаю',
              'завтраку',
              'завтракай',
              'завтраки',
              'завтраму',
              'завтак',
              'завтраках',
              'завтракая']}
```

Расстояние Левенштейна по фонетическим представлениям

```
{'завтраку':  ['завтракаю',
               'завтраку',
               'отроку']}
```

Вот так происходит запуск алгоритма на тестовых транскрибациях:

[*]:
```
%time
trans_alpha['mistakes'] = ''
trans_alpha['bigram_mist'] = ''
trans_alpha['new_words'] = ''
trans_alpha['mistakes_1st_type'] = ''
trans_alpha['mistakes_2nd_type'] = ''
trans_alpha['absent_words'] = ''


for row in trans_alpha.alphacep_transcripts.index:
    print('Обрабатываю ряд ', row)
    text = insert_hyphen_in_text(trans_alpha.alphacep_transcripts[row], hyphened_words)
    mistakes, bigrams_prob_dict, no_such_word = find_mistakes(text, cfreq_all, cprob_all)
    true_mistakes = find_true_mistakes(mistakes, bigrams_prob_dict)
    leven_mist = find_variation(vocab, true_mistakes)

    first_mistakes, second_mistakes, new_words = categorize_mistakes(leven_mist, true_mistakes, cprob_all)

    trans_alpha['mistakes'][row] = ', '.join([key for key in true_mistakes])
    trans_alpha['bigram_mist'][row] = ', '.join([true_mistakes[key] for key in true_mistakes])
    trans_alpha['new_words'][row] =', '.join(new_words)
    trans_alpha['absent_words'][row] = ', '.join(no_such_word)
    trans_alpha['mistakes_1st_type'][row] = ', '.join(first_mistakes)
    trans_alpha['mistakes_2nd_type'][row] = ', '.join(second_mistakes)
```

```
CPU times: user 7 µs, sys: 1e+03 ns, total: 8 µs
Wall time: 31.9 µs
Обрабатываю ряд  35
        Всего ошибок:  24  из  137
        Количество биграм с отсутствующим в словаре словом:  2
Обрабатываю ряд  38
        Всего ошибок:  15  из  81
        Количество биграм с отсутствующим в словаре словом:  3
Обрабатываю ряд  39
        Всего ошибок:  12  из  78
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  42
        Всего ошибок:  5  из  65
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  45
        Всего ошибок:  19  из  83
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  46
        Всего ошибок:  16  из  68
        Количество биграм с отсутствующим в словаре словом:  0
Обрабатываю ряд  52
        Всего ошибок:  31  из  112
        Количество биграм с отсутствующим в словаре словом:  2
Обрабатываю ряд  54
        Всего ошибок:  15  из  106
        Количество биграм с отсутствующим в словаре словом:  2
```

Результат работы алгоритма сохраняется в таблицу:

	audio_ID	alphacep_transcripts	mistakes	bigram_mist	new_words	mistakes_1st_type	mistakes_2nd_type	absent_words
35	Pic-RUS_01-f_Pr-R.zip	жилбыл один дяденька по его жены скоро должно ...	дяденька, плот, ночного, посол, салон, наставь...	один дяденька, случится плот, дяденька ночного...	дядька, настаивать	дяденька, наставь	дяденька, плот, ночного, посол, салон, наставь...	посол варт, варт салон
38	Pic-RUS_01-f_Ski-T.zip	генин жизни одного очень увлекающийся спортом ...	хочется, партийный, товарищ, тоесть, природу, ...	катался хочется, хочется партийный, партийный ...	хотеть	хочется	хочется, партийный, товарищ, тоесть, природу, ...	генин жизни, и норг, норг тоесть
39	Pic-RUS_02-f_Pr-R.zip	однозначным был день рождения мышь решил подар...	однозначным, мышь, посылала	однозначным был, рождения мышь, в посылала	однозначно, присылать	однозначным, посылала	однозначным, мышь, посылала	
42	Pic-RUS_02-f_Ski-T.zip	этот человек встал рано утром позавтракал а по...	наложены, божественного	отправился наложены, голову божественного			наложены, божественного	
45	Pic-RUS_03-m_Ski-R.zip	знакомым мне здесь рассказали одну смешную и п...	нагорных, доскачет, поехать, наложением, слышь	лыжах нагорных, дороге доскачет, позвать поеха...	горный	нагорных	нагорных, доскачет, поехать, наложением, слышь	
46	Pic-RUS_03-m_Ski-T.zip	один чувак решил покататься на лыжах както ран...	ранним, зимним, слышь, перед	лыжах ранним, ранним зимним, горы слышь, алког...	слышать	слышь	ранним, зимним, слышь, перед	
52	Pic-RUS_05-m_Pr-T.zip	однажды константин решил подарить жене подарок...	однажды, константин, сумки, лампы, кастрюли, к...	однажды константин, однажды константин, стоят ...	сумочка, лить, нет, задуматься	сумки, лье, нету, призадумался	однажды, константин, сумки, лампы, кастрюли, к...	сказала муциан, муциан он
54	Pic-RUS_05-m_Ski-T.zip	миша проснулся очень рано гдето в восемь часов...	миша, иза, лыжа, отстегнулась, перекрутил	миша проснулся, горку иза, одна лыжа, лыжа отс...	прикрутить	перекрутил	миша, иза, лыжа, отстегнулась, перекрутил	было преспокойненько, преспокойненько отправиться
55	Pic-RUS_06-f_Pr-R.zip	один мужчина решил подарить своей жене какойни...	сувенир	ей сувенир			сувенир	
58	Pic-RUS_06-f_Ski-T.zip	жилбыл один дядечка один раз он проснулся утр...	дядечка	один дядечка	дядька	дядечка	дядечка	

Как отработал алгоритм на одном из примеров:

жилбыл один дяденька по его жены скоро должно было случиться день рождения был случится плот дяденька ночного 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 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 приличную сумму стоит денег дядечка носок компромисс он купил маменьку 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

Реальные ошибки транскрибатора

15 ошибок

Ошибки транскрибатора, найденные алгоритмом

первая категория ошибок

вторая категория ошибок

слово отсутствует в словаре тренировочного корпуса

Правильно найдено **8** из **15** ошибок.
4 ошибки определены неверно.

жилбыл один дяденька (--> дядька) по его жены скоро должно было случиться день рождения был случится плот дяденька ночного 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 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 (--> настаивать) приличную сумму стоит денег дядечка носок компромисс он купил маменьку 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

Автоматическое исправление ошибок: предложенные замены

Классическое расстояние Левенштейна

8 правильных исправления из 193 замен

- сумерки -> сумка
- выбивать -> выпивать
- таку -> такой
- вращаться -> возвращаться
- грушой -> игрушка
- выбили -> выпили
- выпевать -> выпивать
- Полночь -> полно

Расстояние Левенштейна по фонетическим представлениям

4 правильных исправления из 57 замен

- сумерки -> сумка
- свиток -> света
- таку -> такой
- тако -> так

Модуль поиска ложно найденных ошибок на основе Word2Vec

Словарь вида: {`слово-ошибка: `биграма со словом-ошибкой`}, полученный с помощью марковской цепи

Находим 10 ближайших семантических ассоциатов к слову-ошибке в биграме со словом-ошибкой

- Проставляем частеречный тэг у слова-ошибки
- Лемматизируем каждое слово в биграмме
- Используя векторную модель Word2Vec и косинусную близость векторов, рассматриваем только те семантические ассоциаты, которые не совпадают по форме с нашим словом-ошибкой и имеют такой же частеречный тэг
- Убираем частеречный тэг у подходящего кандидата
- Проверяем, первое или второе слово в биграмме ошибочное
- Проверяем вероятность подходящих ассоциатов в модели марковской цепи

Находим такого кандидата среди 10, который образует биграму, вероятность которой больше 0 в марковской цепи

Если мы нашли ассоциата, который образовал “вероятную” биграму в модели на основе марковс цепей, то мы исключаем из словаря ошибок такой ключ

Получаем словарь истинных ошибок вида: {ошибка: биграма со словом-ошибкой} исключая исходную ошибку из марковской цепи

Про векторную модель:

Находим 10 ближайших семантических ассоциатов к слову-ошибке в биграме со словом-ошибкой

Модель Word2Vec:

- предобученная на корпусе Taiga с функциональными словами
- Алгоритм обучения: Continuous Skipgram
- Размер окна - 5
- Размер обучающего корпуса - почти 5 млрд слов
- Частеречная разметка Universal POS-tags

Оценка модуля с word2vec на одном из примеров

жилбыл один **дяденька** по его жены скоро должно было случиться день рождения был случится плот дяденька ночного 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 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** (--> настаивать) приличную сумму стоит денег дядечка носок компромисс он купил маменьку 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

верно исключено из ошибок: 2
не должно было исключаться из ошибок: 1
должно быть исключено из ошибок, но не исключено: 2

Правильно найдено **6** из **15** ошибок

3 ложно опознанные ошибки

Оценка качества работы алгоритма

Разметка **20** транскрибаций (по 10 для каждой модели транскрибатора - alphacep и abk)

<u>Использованы следующие метрики:</u>

- **Precision**
- **Recall**
- **F1**
- **Accuracy**

	A	B	C	D
		word	true	pred
0		однознач	1	1
1		был	0	0
2		день	0	0
3		рождения	0	0
4		мышь	1	1
5		решил	0	0
6		подарить	0	0
7		ей	0	0
8		подарок	0	0
9		долго	0	0
10		искал	0	0
11		ходила	0	0
12		в	0	0
13		магазин	0	0
14		и	0	0
15		усмотрел	1	1

Сравнительная таблица для оценки качества алгоритма с разными модулями

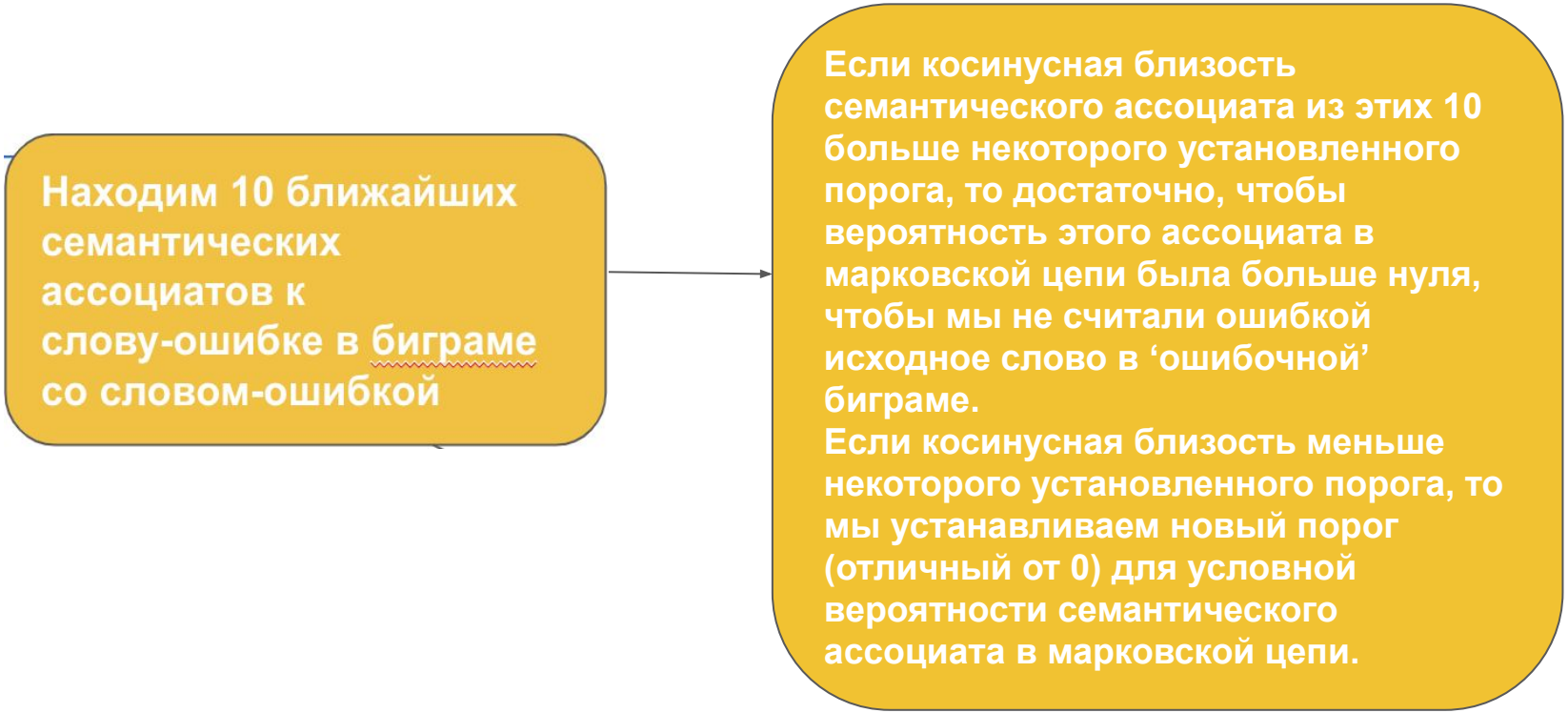
Модель транскрибатора	**Марковские цепи и Левенштейн (корпус 1)**	**+ Векторная модель**	**Марковские цепи и Левенштейн (корпус 2)**
alphacep	Precision: 0.6022 Recall: 0.3889 F1: 0.4726 Accuracy: 0.8751	Precision: 0.6301 Recall: 0.3194 F1: 0.4240 Accuracy: 0.8751 **После изменений числовых порогов:** Precision: 0.6301 Recall: 0.4107 F1: 0.4973 Accuracy: 0.9071	Precision: 0.5345 Recall: 0.4429 F1: 0.4844 Accuracy: 0.8681
abk	Precision: 0.4493 Recall: 0.2138 F1: 0.2897 Accuracy: 0.8433	Precision: 0.4912 Recall: 0.1931 F1: 0.2772 Accuracy: 0.8495 **После изменений числовых порогов:** Precision: 0.4833 Recall: 0.2000 F1: 0.2829 Accuracy: 0.8485	Precision: 0.4337 Recall: 0.2466 F1: 0.3144 Accuracy: 0.8381

Возможное дальнейшее развитие проекта

1. Оптимизация модуля поиска ложно найденных ошибок на основе модели word2vec:
   - применить иные векторные модели
   - ввести числовые пороги для косинусной близости семантических ассоциатов слову-ошибке, а также порог для вероятности семантического ассоциата в марковской цепи
2. Возможность особым образом учитывать частотные, определяющие тематику тестируемого корпуса слова
3. Возможность учитывать возможное ошибочное членение фонетических слов (“варт салон”->”в автосалон”)
4. Улучшение модуля исправления ошибок

Оптимизация модуля поиска ложно найденных ошибок на основе модели word2vec:

- порог для косинусной близости семантических ассоциатов слову-ошибке
- порог для вероятности семантического ассоциата в марковской цепи



Оптимизация модуля поиска ложно найденных ошибок на основе модели word2vec:

Результаты эксперимента с числовыми порогами

prob_threshold = 0.0008

cos_similarity_threshold = 0.69

До оптимизации:
Precision: 0.7500
Recall: 0.3750
F1: 0.5000
Accuracy: 0.9161

После оптимизации:
Precision: 0.7778
Recall: 0.4375
F1: 0.5600
Accuracy: 0.9231

жилбыл один **дяденька** по его жены скоро должно было случиться день рождения был случится плот дяденька ночного 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 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** (--> настаивать) приличную сумму стоит денег дядечка носок компромисс он купил маменьку 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

верно исключено из ошибок: 2

не должно было исключаться из ошибок: 0

должно быть исключено из ошибок, но не исключено: 2

Правильно найдено **6** из **15** ошибок

2 ложно опознанные ошибки

Алгоритм будет оптимизироваться...

Спасибо за внимание!

Репозиторий проекта находится по адресу:
https://github.com/smekur/Spoken_Corpora_with_Kaldi/tree/master/mistakes_search

Электронная почта Смирновой Е.: ekanerina@yandex.com

Электронная почта Черной А.: khokhlova_as@mail.ru