Проект “ANTIDICT”

База морфологических структур, допустимых в современном русском языке

Студенты:
Алексей Доркин
Владислава Смирнова
Антон Вахранев

Куратор: Варвара Магомедова

О чем проект

– изучение необычных слов, которые можно встретить в интернете (50 млн словоформ из ГИКРЯ)

– классификация слов по большим категориям (например, заимствования, экспрессивные формы, слова со случайными и с намеренными ошибками и искажениями)

– классификация слов по подкатегориям (например, заимствования, слова с заимствованными морфемами, слова, образованные от русских корней)

– морфологический анализ слов

Что мы успели сделать

Работа над заимствованиями:

– декомпозировали 1 цикл работы и определились с разделением обязанностей и планом работы

– познакомились с fasttext

– начали придумывать формальные критерии для определения заимствованных слов, определять теги для разметки и готовить размеченные вручную наборы для обучения модели (возможно, понадобится помощь добровольцев)

– выкачали словарь заимствованных слов Дьякова и начали работу с ним

– сделали анализ слов с размноженными буквами и склеили их с помощью mystem

Что нам предстоит сделать в ближайшее время

1. Составить список тегов для заимствований
2. Скачать словарь Зализняка со словоформами
3. Доработать словарь Дьякова
4. Подготовить наборы данных для fasttext и составить словарь заимствований
5. Обучить fasttext различать заимствования и подкатегории заимствований

Пока что мы работаем с леммами, но если у нас получится, то мы сгенерируем всю словоизменительную парадигму для заимствований и попробуем поработать с ней

Данные

Пример слов из нашего дата сета

огурытсы
убиватели
полуфоточных
околобожественные
греческих
закрытымиглазами
артустановками
физлицомпредпринимателем
толстинг
поджержи
кдчдзи
мосарабских
поворяться
оплошсти
Ремонтама
Видикону
Гизиопу
Багчевский
укушалсо
фразопотопом
мегазлые
Толстопузики
обмакдональдщинность
слоняга
красунчег
Овнюки

Данные

Слова, проанализированные mystem

Опечатки mystem понимает не всегда:

сурьъезный, [{'analysis': [{'lex': 'сурьъезный', 'wt': 0.3818474906, 'qual': 'bastard', 'gr': 'A=(вин,ед,полн, муж,неод|им,ед,полн,муж)'}], 'text': 'сурьъезный'}, {'text': '\n’}]

красаавчик, [{'analysis': [{'lex': 'красаавчик', 'wt': 0.2980500193, 'qual': 'bastard', 'gr': 'S,муж,неод=(вин, ед|им,ед)'}], 'text': 'красаавчик'}, {'text': '\n’}]

электроной, [{'analysis': [{'lex': 'электрона', 'wt': 0.7518655792, 'qual': 'bastard', 'gr': 'S,жен,од=твор,ед'}], 'text': 'электроной'}, {'text': '\n'}]

итаалия, [{'analysis': [{'lex': 'итаалия', 'wt': 0.3106070811, 'qual': 'bastard', 'gr': 'S,мж,од=им,ед'}], 'text': 'итаалия'}, {'text': '\n'}]

революцыя, [{'analysis': [], 'text': 'революцыя'}, {'text': '\n'}]

Некоторые слова после склеивания оказались словарными:

уважаемый, [{'analysis': [{'lex': 'уважаемый', 'wt': 0.9776687473, 'gr': 'A=(вин,ед,полн,муж,неод|им,ед, полн,муж)'}], 'text': 'уважаемый'}, {'text': '\n'}]

Что такое fastText?

FastText – это бесплатная, легковесная библиотека с открытым исходным кодом, которая позволяет пользователям обучать различные текстовые модели и классификаторы.

Статья

ИСПОЛЬЗОВАНИЕ МОРФОЛОГИИ В ДИСТРИБУТИВНЫХ СЕМАНТИЧЕСКИХ МОДЕЛЯХ: ЭКСПЕРИМЕНТЫ С РУССКИМ ЯЗЫКОМ

Садов М. А. (mikeabyrvalg5@gmail.com) – НИУ ВШЭ, Москва, Россия; Кутузов А. Б. (andreku@ifi.uio.no) – Университет Осло, Осло, Норвегия.

Два подхода к подсловам

В статье сравнили два подхода к включению информации о подсловах в эмбеддинговые (векторные) модели для русских слов:

1. морфологический подход, включающий две модели SkipGram;

2. символьный n-граммовый подход на основе модели fastText.

Немецкое слово «Abwasserbehandlungsanlange» означает «станция очистки сточных вод» и является конкатенацией соответствующих слов. Даже в английском языке, например, слово «subword» можно разбить на «sub» + «word». Этот приём, кстати, может помочь с проблемой OOV (out of vocabulary) — в обучающей выборке может не быть слова «subtask», но поскольку «subtask» = «sub» + «task», а эти слова (подслова), допустим, есть в обучении, мы можем надеяться, что сеть будет правильно обрабатывать / генерировать и слово «subtask».

На чём обучали?

Все модели прошли обучение на НКРЯ. Кроме того, был составлен датасет сходств редких и мультиморфных слов для русского языка, и модели обучили ещё и на них.

Результаты

После оценки было обнаружено, что модель fastText показала отличные результаты, несмотря на то, что у неё было больше слов OOV для предсказания. А обе модели SkipGram не смогли обучиться представлению слов OOV.

Model name	P	p-value	OOV words	total words
SkipGram+Morfessor	-0.1647	0.5134	100%	18
fastText	0.7176	0.0007	100%	18

Table 3: Spearman's rank correlation values (ρ) between the human experts scores and models estimations on the common set of OOV words (Morfessor experimental set-up).

Model name	P	p-value	OOV words	total words
SkipGram+CRF-based	-0.3244	0.189	100%	18
fastText	0.7176	0.0007	100%	18

Table 3.2: Spearman's rank correlation values (ρ) between the human experts scores and models estimations on the common set of OOV words (CRF-based model experimental set-up).

Датасет

Полученный датасет сходства слов имеет хороший показатель меры согласованности между аннотаторами ($\alpha = 0{,}648$), что позволяет утверждать, что он подходит для измерения производительности моделей векторного представления слов на редких и мультиморфных словах.

Morfessor Baseline

Были также выявлены несколько недостатков в алгоритме Morfessor Baseline: был сформулирован вывод, что производительность инструмента для сложных слов с фузией сравнительно низкая, а сам алгоритм имеет слишком много ограничений, чтобы считаться надежными.

violating morpho-tactic rules: for example, it suggested the prefix "по" in non-word-initial positions ("вал/ян/оса/по/ж/ник", "felt boots maker"). The other limitation came from the fusion nature of Russian: our Morfessor model performed well on the words with complex but consecutive morphology (for example, "*полу/рас/па/вший/ся*", "half-disintegrated"), but predictably made mistakes in the cases with fusion ("*сте/речь*", "to guard"). The last typical error of our Morfessor model was over-segmentation of strings which include frequent morphemes ("*ра/б*", "slave").

Их планы после статьи

В будущем планируется более детально изучить, как различные подходы морфосегментации влияют на производительность моделей, обученных на морфемах. Также было изъявлено желание расширить датасет о сходстве редких и мультиморфных слов для русского языка.

Репозиторий

https://github.com/wksmirnowa/antidict