Автоматический морфемный анализ слов русского языка. Эмбеддинги на основе однокоренных слов

ВЫПОЛНИЛА: МАЛЬТИНА ЛЮДМИЛА

Подходы к автоматическому морфемному анализу

Машинное обучение:

- **без учителя:**
  - **использование предсказуемости сегмента** (Harris, 1970; Bernhard, 2006)
  - **моделирование последовательности морфов** (униграммная вероятностная модель) (Creutz, Lagus, 2005; Virpioja et al., 2013; Smit et al., 2014)
- **с учителем:**
  - **условные случайные поля** (Ruokolainen et al., 2013)
  - **свёрточные нейронные сети** (Sorokin, Kravtsova, 2018)
  - **частичное обучение: модификации метода условных случайных полей** (Ruokolainen et al., 2014)

Использованные данные

- **униграммы из Национального корпуса русского языка** – неразмеченные данные (674 940 словоформ, 146 907 лемм)
- **данные на основе морфемно-орфографического словаря А. Н. Тихонова** (95 922 слов) – размеченные данные
- **редкие слова** (800 слов) – размеченные данные
  - заимствования (*буккроссинг*)
  - термины (*аденозинтрифосфорный*)
  - неологизмы (*загуглиться*)
  - слова, производные от имён собственных (*неогумбольдтианство*)

Создание выборки с редкими словами

Из словарей* были выбраны слова, в которых хотя бы один корень отсутствовал в обучающей выборке

Добавлены однокоренные слова

Выполнен морфемный анализ слов вручную

*Словарь сленга в сфере информационных технологий [https://github.com/kpopov94/morpheme_seq2seq/blob/master/slang.in]

Словарь неологизмов [https://russkiiyazyk.ru/leksika/slovar-neologizmov.html]

Орфографический словарь В. В. Лопатина [https://gufo.me/dict/orthography_lopatin]

Характеристика размеченных выборок

Выборка	Среднее кол-во морфов в слове	Доля префиксов	Доля корней	Доля суффиксов	Доля окончаний	Доля интерфиксов	Доля постфиксов
Обучающая	3,823	0,114	0,319	0,367	0,137	0,036	0,028
Валидационная	3,836	0,116	0,318	0,367	0,135	0,036	0,029
Тестовая	3,829	0,116	0,318	0,366	0,136	0,036	0,028
Слова с незнакомыми корнями	2,726	0,022	0,436	0,377	0,145	0,012	0,006

Результаты на данных из словаря А.Н.Тихонова

Модель	Точность	Полнота	F-мера	Доля полностью верно разобранных слов
Morfessor	0,9143	0,9078	0,9110	0,6990
Условные случайные поля	0,9424	0,9279	0,9351	0,7143
Свёрточные нейронные сети	**0,9666**	**0,9688**	**0,9677**	**0,8583**

Лингвистическая классификация ошибок

Причина ошибки	Пример
Влияние более частотных морфов	**с/холаст/ик/а** **(схоласт/ик/а)**
Наличие морфов с низкой частотой	**мотодивизи/я** **(мото/дивизи/я)**
Опрощение	**о/город/нич/еск/ий** **(огород/нич/еск/ий)**
Морфологические чередования	**почт/о/обрабат/ыва/ющ/ий** **(почт/о/об/рабат/ыва/ющ/ий)**
Переразложение	**кост/оч/к/а** (ср. костъка) **(кост/очк/а)**
Другое	**мног/о/лет/ник** **(мног/о/лет/н/ик)**

Автоматический морфемный анализ редких слов

Модель	Точность	Полнота	F-мера	Доля полностью верно разобранных слов
Morfessor	0,5028	0,7867	0,6135	0,1850
Условные случайные поля	0,8177	0,7751	0,7958	0,4963
Свёрточные нейронные сети	**0,8204**	**0,8192**	**0,8198**	**0,5687**

CNN: высокое качество, если аффиксы имеют высокую частоту:

- постфикс *-ся*
- суффиксы *-ть-, -вш-, -и-, -изм-, -ист-, -ова-*
- префиксы *рас-, за-*

качество ниже, если аффиксы имеют низкую частоту:

- префикс *ре-*
- суффикс *-инг*

Эмбеддинги на основе однокоренных слов

1. Морфемный анализ словаря word2vec модели (CNN-модель для морфемного анализа)
2. Морфемный анализ OOV-слов (CNN-модель для морфемного анализа)
3. Получение эмбеддингов для OOV-слов на основе слов, которые являются для них однокоренными и есть в словаре word2vec модели
4. Применение полученных эмбеддингов для определения семантической близости

Датасет для оценки семантической близости

- [Sadov, Kutuzov, 2018]: **104 пары редких многоморфемных слов:**
  - частота < 1000 по Национальному корпусу русского языка
  - в слове не менее 5 морфем (Morfessor)
  - слова в каждой паре относятся к одной и той же части речи

Слово 1	Слово 2	Семантическая близость
франкфурт-на-майне	франкфурт	7.53846153846
основополагающий	полный	3.38461538462
плексигласовый	химический	3.46153846154
неподтвержденный	качественный	1.0
древневерхненемецкий	немецкий	5.23076923077

Способы получения эмбеддингов на основе однокоренных слов

- **Word2vec + averaged:** усреднение эмбеддингов всех однокоренных слов
- **Word2vec + frequency weighted:** использование всех однокоренных слов с частотными весами
- **Word2vec + probability weighted:** использование всех однокоренных слов с весами-вероятностями того, что в данном слове был верно выделен корень
- **Word2vec + averaged with top frequencies:** на каждый корень слова берётся по 5 однокоренных слов с наибольшей частотой, эмбеддинги выбранных слов усредняются
- **Word2vec + averaged with top probabilities:** на каждый корень слова берётся по 5 однокоренных слов с наибольшей вероятностью того, что в данном слове был верно выделен корень, эмбеддинги выбранных слов усредняются

Способы получения эмбеддингов на основе однокоренных слов

- **Word2vec + averaged with top morphemic F1 and frequency:** на каждый корень слова берётся по 3 однокоренных слова с наибольшим значением «морфемной F1» и наибольшей частотой, эмбеддинги выбранных слов усредняются
- **Word2vec + averaged with top morphemic F1 and probability:** на каждый корень слова берётся по 3 однокоренных слова с наибольшим значением «морфемной F1» и наибольшей вероятностью того, что в данном слове был верно выделен корень, эмбеддинги выбранных слов усредняются

Morphemic F1

- **True positive:** кол-во морфов, совпадающих в исходном слове и однокоренном для него. Корень берётся с весом 1, префикс, суффикс, постфикс – с весом 0.3, окончания и соединительные гласные не учитываются
- **False positive:** кол-во морфов, которые есть в однокоренном слове, но отсутствуют в исходном. Веса берутся аналогично
- **True negative:** кол-во морфов, которые есть в исходном слове, но отсутствуют в однокоренном для него. Веса берутся аналогично

Эмбеддинги на основе однокоренных слов для определения семантической близости

Эмбеддинги	Корреляция Спирмана с экспертными оценками	P-value	Доля OOV-пар
Word2vec baseline	0.4179	1.0140e-05	0.2885
FastText	0.6402	2.5429e-13	0.0000
Word2vec + averaged	0.6842	1.1794e-15	0.0288
Word2vec + frequency weighted	0.7031	8.8012e-17	0.0288
Word2vec + probability weighted	0.6815	1.6899e-15	0.0288
Word2vec + averaged with top frequencies	**0.7132**	2.0110e-17	0.0288
Word2vec + averaged with top probabilities	0.6964	2.2539e-16	0.0288
Word2vec + averaged with top morphemic F1 and frequency	0.7122	2.3187e-17	0.0288
Word2vec + averaged with top morphemic F1 and probability	0.7121	2.3633e-17	0.0288

Эмбеддинги на основе однокоренных слов: ближайшие слова

Word2vec + averaged with top frequencies	
древневерхненемецкий_ADJ	древние_NOUN 0.6765 древний_ADJ 0.662 древность_NOUN 0.6428 этруск_NOUN 0.6337 древнеегипетский_ADJ 0.6083
камнесамоцветный_ADJ	цветок_NOUN 0.6572 многоцветный_ADJ 0.5702 цвета_NOUN 0.5656 полудрагоценный_ADJ 0.5539 цветочек_NOUN 0.5538

Эмбеддинги на основе однокоренных слов: ближайшие слова

Word2vec + averaged with top probabilities	
древневерхненемецкий_ADJ	древнегерманский_ADJ 0.6932 древнееврейский_ADJ 0.6884 старонемецкий_ADJ 0.6526 арамейский_ADJ 0.6517 древнегреческий_ADJ 0.6511
камнесамоцветный_ADJ	раннеперестроечный_ADJ 0.5753 шумозаградительный_ADJ 0.5599 камнеобработка_NOUN 0.5273 цветовод_NOUN 0.5235 цветоводческий_ADJ 0.5217

Эмбеддинги на основе однокоренных слов: ближайшие слова

Word2vec + averaged with top morphemic F1 and frequency	
древневерхненемецкий_ADJ	древний_ADJ 0.6939 дохристианский_ADJ 0.6483 шумерский_ADJ 0.643 архаический_ADJ 0.6386 сарматский_ADJ 0.6382
камнесамоцветный_ADJ	самоцветный_ADJ 0.7339 самоцвет_NOUN 0.7006 яшма_NOUN 0.6413 камнерезный_ADJ 0.6346 полудрагоценный_ADJ 0.6264

Эмбеддинги на основе однокоренных слов: ближайшие слова

Word2vec + averaged with top morphemic F1 and probability	
древневерхненемецкий_ADJ	нижненемецкий_ADJ 0.6422 раннеславянский_ADJ 0.6245 арамейский_ADJ 0.6182 сарматский_ADJ 0.6093 древна_NOUN 0.6074
камнесамоцветный_ADJ	самоцветный_ADJ 0.7339 самоцвет_NOUN 0.7006 яшма_NOUN 0.6413 камнерезный_ADJ 0.6346 полудрагоценный_ADJ 0.6264