Автоматический морфемный анализ слов русского языка: сравнение подходов

Выполнила: Мальтина Людмила

- Analysis of Images, Social Networks and Texts. 8th International Conference, AIST 2019 (Kazan). Мальтина Л.П., Малафеев А.Ю. Morpheme Segmentation for Russian: Evaluation of Convolutional Neural Network Models (17.07.2019 - 19.07.2019)

Подходы к морфемному анализу:

- униграммная вероятностная модель
- условные случайные поля
- свёрточные нейронные сети

Актуальность исследования

- **векторные представления внесловарных слов** (Galinsky et. al., 2017; Arefyev, Gratianova, Popov, 2018; Sadov, Kutuzov, 2018)
- **машинный перевод** (Fritzinger, Fraser, 2010)
- **распознавание речи** (Карпов, 2007)
- **информационный поиск** (Bernhard, 2006)
- **разметка корпусов** (Гришина и др., 2009; http://www.ruscorpora.ru/search-main.html)
- **проверка морфемного анализа, выполненного учащимися** (https://www.morphemeonline.ru)

Формирование выборок

неразмеченные данные на основе униграмм из НКРЯ
(http://www.ruscorpora.ru/new/corpora-freq.html)

Предобработка:

- отбор словоформ, которые содержали только буквы русского алфавита и дефисы
- приведение к нижнему регистру
- исключение незнаменательных частей речи (pymorphy2)
- объём выборки **–** 674940 словоформ

+ лемматизированная версия – 146907 лемм

Формирование выборок

размеченные данные на основе словаря А. Н. Тихонова

- словарь с полуавтоматическим указанием типов морфов (https://github.com/AlexeySorokin/NeuralMorphemeSegmentation)
- исправление неточностей разметки вручную (78 слов):

Слово	Разбор в словаре	Исправленный разбор
управляемый	у:PREF/ правл:ROOT/ я:SUFF/ **ем:ROOT**/ ый:END	у:PREF/ правл:ROOT/ я:SUFF/ **ем:SUFF**/ ый:END

- обучающая, валидационная и тестовая выборки – 40/30/30 (38368/28777/28777 слов)

Формирование выборок

размеченная выборка из слов, содержащих корни, которые отсутствуют в обучающей выборке

- отбор слов, в которых хотя бы один корень отсутствовал в обучающей выборке (https://gufo.me/dict/orthography_lopatin, https://russkiiyazyk.ru/leksika/slovar-neologizmov.html, https://github.com/kpopov94/morpheme_seq2seq)
- добавление однокоренных слов с использование сервиса (https://wordroot.ru)
- выполнение морфемного анализа этих слов

Формирование выборок

Размеченная выборка из слов, содержащих корни, которые отсутствуют в обучающей выборке

- Выборка (800 слов) включает в себя много заимствований (например, *буккроссинг*). В неё вошли:
  - термины (*аденозинтрифосфорный*)
  - неологизмы (*загуглиться*)
  - слова, образованные от имён собственных (*неогумбольдтианство*)

Характеристики 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

Униграммная вероятностная модель Morfessor

- Morfessor 2.0 (Virpioja et al., 2013; Smit et al., 2014)
- **принцип апостериорного максимума:**

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D_W) = \arg\max_{\theta} p(\theta) p(D_W \mid \theta)$$

- **функция потерь:**

$$L(\theta, D_W) = -\log p(\theta) - \log p(D_W \mid \theta)$$

- **упрощающее предположение:** как морфы в слове, так и символы, образующие морф, встречаются независимо друг от друга
- **правдоподобие данных** $p(D_W \mid \theta)$ – произведение вероятностей морфов, входящих в выбранные варианты морфемного анализа
- **априорная вероятность** $p(\theta)$ выше для словарей, которые состоят из меньшего числа морфов и содержат более короткие морфы

Униграммная вероятностная модель Morfessor

- **функция потерь для модели, использующей обучение без учителя**:

$$L(\theta, D_W) = -\log p(\theta) - \alpha \log p(D_W \mid \theta)$$

чем выше α, тем длиннее могут быть морфы

- **функция потерь для модели, использующей частичное обучение**:

$$L(\theta, D_W) = -\log p(\theta) - \alpha \log p(D_W \mid \theta) - \beta \log p(D_W \to A \mid \theta)$$

β регулирует количество разбиений для размеченных данных

Униграммная вероятностная модель Morfessor

Grid search для подбора параметров:

- лемматизация неразмеченной обучающей выборки: да/нет
- представление данных неразмеченной выборки: типы/токены
- α и β: эвристика;

$$\alpha \in \{0{,}1; 1{,}0\}$$

$$\beta \in \{1; 1000; 10000\}$$

- Всего – 28 моделей

Униграммная вероятностная модель Morfessor

Лучший результат на валидационной выборке:

- лемматизация неразмеченной обучающей выборки: нет
- представление данных неразмеченной выборки: типы

$\alpha = 0{,}1$

$\beta = 1000$

Униграммная вероятностная модель Morfessor

- hard voting classifier: наилучший результат – с 14 лучшими моделями

Результаты на валидационной выборке:

	F1-мера	Word accuracy без учёта типа морфов
одна лучшая модель Morfessor	0,9026	0,6883
hard voting classifier из 14 лучших моделей Morfessor	0,9100	0,6958

Условные случайные поля

- Ruokolainen et al., 2013

BMS-схема:

- B – begin
- M – middle
- S – single

Linear-chain CRF:

T– количество символов в слове

$y = (y_1, y_2, \ldots, y_T)$ – последовательность меток для каждого символа этого слова

$x = (x_1, x_2, \ldots, x_T)$ – словоформа, рассматриваемая как последовательность символов

t – позиция символа

w – вектор параметров модели

ф – признаковая функция, значение которой является вектором

$$p(y \mid x; w) \propto \prod_{t=2}^{T} \exp\left(w \times \phi(y_{t-1}, y_t, x, t)\right)$$

Условные случайные поля

выбор признаков эмиссии: intuition

talk + ed, play + ed, speed

- в позиции t проводится морфемная граница, если правым контекстом выступает ed:

talk + ed, play + ed, spe + ed

- в позиции t проводится морфемная граница, если правым контекстом выступает ed, но spe не является левым контекстом:

talk + ed, play + ed, speed

Условные случайные поля

- **признаки эмиссии:**

$$\{\chi_m(x,t)[y_t = y'_t] \mid m \in 1,\ldots,M, \forall y'_t \in \{B,M,S\}\}$$

набор бинарных функций $\{\chi_m(x,t)_{m=1}^{M}\}$ описывает правый и левый контекст позиции:

- *driv + ers* ?
- δ - максимальная длина символов для контекста (в примере δ = 5)
- является ли элемент из множества {*v, iv, riv, driv, <w>driv*} левым контекстом?
- является ли элемент из множества{*e, er, ers, ers</w>*} правым контекстом?
- **признаки перехода:**

$$\{[y_{t-1} = y'_{t-1}][y_t = y'_t], y'_t, y'_{t-1} \in \{B,M,S\}\}$$

Условные случайные поля

Результаты на валидационной выборке:

	F1-мера	Word accuracy без учёта типа морфов
Supervised CRF	0,9358	0,7157

Свёрточные нейронные сети

- Sorokin, Kravtsova, 2018

Типы морфов:

- PREF
- ROOT
- SUFF
- END
- POSTFIX
- LINK
- HYPH

BMES-схема:

- B – begin
- M – middle
- E – end
- S – single

- **Вход:** one-hot-encoding векторы для букв в словах
- **Выход:** распределение вероятностей классов (по расширенной BMES-схеме) для каждого символа

Сегментация *уч*:корень/*и*:суффикс/*тель*:суффикс для слова *учитель*

у	ч	и	т	е	л	ь
B-ROOT	E-ROOT	S-SUFF	B-SUFF	M-SUFF	M-SUFF	E-SUFF

Свёрточные нейронные сети

- Sorokin, Kravtsova, 2018
- **memorization:**
  - кодирование контекста каждой позиции с помощью 15-мерного вектора:
    - могут ли морфы данного типа начинаться в данной позиции? (5)
    - могут ли морфы данного типа заканчиваться в данной позиции? (5)
    - могут ли текущий символ являться морфом данного типа, состоящим из одной буквы? (5)
  - с помощью проверки условий определяется, возможна ли предсказываемая последовательность морфов
- **soft voting classifier:** ансамбль из нескольких моделей с различными случайными инициализациями

Свёрточные нейронные сети

Гиперпараметр	Рассмотренные значения
количество свёрточных слоёв	3; 4
размер окна	3; 5; 7
количество фильтров	192; 240
количество нейронов в полносвязанном выходном слое	48; 64; 96
dropout rate	0,2; 0,3; 0,4
количество моделей	1; 3
меморизация морфем	true; false
меморизация n-грамм	true; false

Свёрточные нейронные сети

	Гиперпараметры	Точность	Полнота	F1-мера	Word accuracy с учётом типа морфов	Word accuracy без учёта типа морфов
№5	свёрточные слои: 3 ширина окна: 5 фильтры: 192 полносвязанные нейроны в выходном слое: 64 dropout rate: 0,2 количество моделей: 3 меморизация: да	0,9612	0,9634	0,9623	0,8363	0,8307
№15	свёрточные слои: **4** ширина окна: 5 фильтры: **240** полносвязанные нейроны в выходном слое: 64 dropout rate: **0,4** количество моделей: 3 меморизация: да	**0,9638**	**0,9676**	**0,9657**	**0,8512**	**0,8456**

Свёрточные нейронные сети

- hard voting classifier: наилучший результат – с 4 лучшими моделями

	F1-мера	Word accuracy с учётом типа морфов
одна лучшая модель CNN	0,9657	0,8456
hard voting classifier из 4 лучших моделей CNN	0,9672	0,8510

Результаты на тестовой выборке

	Точность	Полнота	F1-мера	Word accuracy без учёта типа морфов	Word accuracy с учётом типа морфов
Morfessor	0,9143	0,9078	0,9110	0,6990	-
CRF	0,9424	0,9279	0,9351	0,7143	-
CNN	0,9666	0,9688	0,9677	0,8583	0,8522

Анализ ошибок (CNN)

Причина и количество ошибок этого типа	Пример (правильный вариант в скобках)	Комментарий к примеру
Влияние более частых морфов (28)	*с/холаст/ик/а* *(схоласт/ик/а)*	Частота морфа *с-* выше, чем частота морфа *-схоласт-*
Наличие морфов с низкой частотой в обучающей выборке (менее 15 вхождений) (24)	*делик/атес* *(деликатес)*	Корень *-деликатес-* не встречается в обучающей выборке
Опрощение (23)	*о/город/нич/еск/ий* *(огород/нич/еск/ий)*	Диахронически в слове огороднический был корень *-город-*, с точки зрения синхронии выделяется корень *-огород-*
Чередования (4)	*почт/о/обрабат/ыва/ющ/ий* *(почт/о/об/рабат/ыва/ющ/ий)*	У морфа *-работ-* есть алломорф *-рабат-*
Переразложение (1)	*кост/оч/к/а* *(кост/очк/а)*	Диахронически в слове выделяются корень *-косточ-* и суффикс *-к-*, с точки зрения синхронии есть корень *-кост-* и суффикс *-очк-*
Другое (20)	*бегл/янк/а* *(бег/л/янк/а)*	Морфы *-бег-* и *-л-* есть в обучающей выборке, а морф *-бегл-* отсутствует, но несмотря на это программа допускает ошибку

Результаты на выборке с незнакомыми корнями

	Точность	Полнота	F1-мера	Word accuracy без учёта типа морфов	Word accuracy с учётом типа морфов
Morfessor	0,5028	0,7867	0,6135	0,1850	-
CRF	0,8177	0,7751	0,7958	0,4963	-
CNN	0,8204	0,8192	0,8198	0,5687	0,5475

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*

Перспективы

- использование автоматического морфемного анализа в прикладных задачах:

1. создание эмбеддингов слов:
   - на основе сложения эмбеддингов морфов, входящих в слово
   - на основе сложения эмбеддингов однокоренных слов
   - эмбеддингов, не учитывающих морфемику и словообразование
2. сравнение полученных эмбеддингов на задачах:
   - определения семантической близости
   - классификации текстов

Список основных источников

1. Гришина, Е. А. [и др.] О задачах и методах словообразовательной разметки в корпусе текстов / Е. А. Гришина, И. Б. Иткин, О. Н. Ляшевская, М. Г. Тагабилева // Полярный вестник. – 2009. – Т. 12. – С. 5-25.
2. Карпов, А. А. Модели и программная реализация распознавания русской речи на основе морфемного анализа: дис. на соискание научной степени канд. техн. наук. — СПб, 2007. — 129 с.
3. Разбор слов по составу: [Электронный ресурс]. URL: http://www.morphemeonline.ru (дата обращения: 24.09.2018).
4. Arefyev, N. V., Gratsianova, T. Y., Popov, K. P. Morphological segmentation with sequence to sequence neural network / N. V. Arefyev, T. Y. Gratsianova, K. P. Popov // Computational linguistics and intellectual technologies: proceedings of the international conference "Dialogue 2018". – 2018. – P. 85-95.
5. Bernhard, D. Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment / D. Bernhard // Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes. – Venice, Italy, April 2006. – P. 19–23.
6. Fritzinger, F., Fraser, A. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing // Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. — 2010. — P. 224-234.

Список основных источников

7. Galinsky, R. [et al.] Morpheme level word embedding / R. Galinsky, T. Kovalenko, Ju. Yakovleva, A. Filchenkov // Artificial Intelligence and Natural Language 6th Conference, AINL 2017. – 2017. – P. 143-155.
8. Ruokolainen, T. [et al.]. Supervised morphological segmentation in a low-resource learning setting using conditional random fields / T. Ruokolainen, O. Kohonen, S. Virpioja, M. Kurimo // Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL). – 2013. – P. 29–37.
9. Sadov, M. A., Kutuzov, A. B. Use of morphology in distributional word embedding models: Russian language case / M. A. Sadov, A. B Kutuzov // Computational linguistics and intellectual technologies: proceedings of the international conference "Dialogue 2018". – 2018. URL: http://www.dialog-21.ru/media/4554/sadovmapluskutuzovab.pdf.
10. Smit, P. [et al]. Morfessor 2.0: Toolkit for statistical morphological segmentation / P. Smit, S. Virpioja, S. A. Grönroos, M. Kurimo // The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL) – 2013. – P. 21 - 24.
11. Sorokin, A., Kravtsova, A. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language / A. Sorokin, A. Kravtsova // Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol. 930. – Springer, Cham, 2018. – P. 3-10.
12. Virpioja, S. [et al.]. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline / S. Virpioja, P. Smit, S. A. Grönroos, M. Kurimo. – Report in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland, 2013. – 32 p.