Морфологический анализ

Евгений Борисов

Морфологический анализ

Задача автоматического определения частей речи
POS (Part of Speech) tagging

предварительная обработка текста
для извлечения информации

https://ruscorpora.ru/new/corpora-morph.html

Морфологический анализ

Морфологический анализ

ЧАСТИ РЕЧИ

0+

ПЛАКАТ
Подписано в печать 09.07.2018 г. Формат 60х90/2.
Усл. печ. л. 0,5. Тираж 10 000 экз. Заказ № 129.
ООО «АЛФЕЯ» при участии ООО «ЛЕДА»
214030, РФ, г. Смоленск, ул. М. Расковой, д. 4/2.
Дистрибьютор в Беларуси: частное предприятие «Слово А»,
Беларусь, 210001, г. Витебск, ул. Комсомольская, 27/11, к. 71.
e-mail: manager@slovo-book.ru www.slovo-book.ru
Отпечатано: И. П. Лебле Ю. М.,
Россия, 150010, г. Ярославль, ул. Ярославская, д. 148-12.

САМОСТОЯТЕЛЬНЫЕ ЧАСТИ РЕЧИ

ИМЯ СУЩЕСТВИТЕЛЬНОЕ ОБОЗНАЧАЕТ ПРЕДМЕТ	**КТО? ЧТО?** СОБАКА, ДЕВОЧКА, ПТИЦА, СОЛНЦЕ, ЧУВСТВО, ПИРОГ
ИМЯ ПРИЛАГАТЕЛЬНОЕ ОБОЗНАЧАЕТ ПРИЗНАК ПРЕДМЕТА	**КАКОЙ? КАКАЯ? КАКОЕ? КАКИЕ? ЧЕЙ?** УМНЫЙ, БОЛЬШАЯ, ЖЁЛТОЕ, ДОБРЫЕ, ДЕДУШКИН
ГЛАГОЛ ОБОЗНАЧАЕТ ДЕЙСТВИЕ ИЛИ СОСТОЯНИЕ ПРЕДМЕТА	**ЧТО ДЕЛАТЬ? ЧТО СДЕЛАТЬ?** БЕГАТЬ, РИСОВАТЬ, ЖИТЬ, ИЗУЧИТЬ, РАССКАЗАТЬ
ИМЯ ЧИСЛИТЕЛЬНОЕ ОБОЗНАЧАЕТ КОЛИЧЕСТВО ИЛИ ПОРЯДОК ПРИ СЧЁТЕ	**СКОЛЬКО? КОТОРЫЙ?** ЧЕТЫРЕ, ОБА, НЕСКОЛЬКО, ПЕРВЫЙ, ДЕСЯТЫЙ
МЕСТОИМЕНИЕ УКАЗЫВАЕТ НА ПРЕДМЕТ, ПРИЗНАК, КОЛИЧЕСТВО, НЕ НАЗЫВАЯ ИХ	**КТО? ЧТО? КАКОЙ? ЧЕЙ? СКОЛЬКО? КОТОРЫЙ?** Я, ТЫ, ВЫ, МЫ, ОНИ, НИКТО, ТАКОЙ, НИЧЕЙ, НЕМНОГО, ЛЮБОЙ
НАРЕЧИЕ ОБОЗНАЧАЕТ ПРИЗНАК ДЕЙСТВИЯ	**ГДЕ? КОГДА? КУДА? ПОЧЕМУ? КАК?** БЛИЗКО, ДАВНО, НАПРАВО, СГОРЯЧА, МЕДЛЕННО

СЛУЖЕБНЫЕ ЧАСТИ РЕЧИ

ПРЕДЛОГ	**СОЮЗ**	**ЧАСТИЦА**
СЛУЖИТ ДЛЯ СВЯЗИ СЛОВ В СЛОВОСОЧЕТАНИИ В, НА, К, О, ПЕРЕД, ИЗ	СЛУЖИТ ДЛЯ СВЯЗИ ОДНОРОДНЫХ ЧЛЕНОВ ИЛИ ЧАСТЕЙ СЛОЖНОГО ПРЕДЛОЖЕНИЯ И, ИЛИ, НО, КОГДА	ВЫРАЖАЕТ ОТТЕНКИ ЗНАЧЕНИЙ СЛОВ, СЛОВОСОЧЕТАНИЙ И ПРЕДЛОЖЕНИЙ, ОБРАЗУЕТ ФОРМЫ СЛОВА НЕ, НИ, БЫ, ДА, НЕТ

ОСОБАЯ ЧАСТЬ РЕЧИ

МЕЖДОМЕТИЕ

ВЫРАЖАЕТ ЭМОЦИИ ИЛИ НОРМЫ ЭТИКЕТА

АХ, АЙ, ОЙ, АЙ, ЭЙ, УРА, УВЫ, ЗДРАВСТВУЙТЕ, СПАСИБО, ПОЖАЛУЙСТА

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

Части речи

- **S** — существительное (*яблоня, лошадь, корпус, вечность*)
- **A** — прилагательное (*коричневый, таинственный, морской*)
- **NUM** — числительное (*четыре, десять, много*)
- **ANUM** — числительное-прилагательное (*один, седьмой, восьмидесятый*)
- **V** — глагол (*пользоваться, обрабатывать*)
- **ADV** — наречие (*сгоряча, очень*)
- **PRAEDIC** — предикатив (*жаль, хорошо, пора*)
- **PARENTH** — вводное слово (*кстати, по-моему*)
- **SPRO** — местоимение-существительное (*она, что*)
- **APRO** — местоимение-прилагательное (*который, твой*)
- **ADVPRO** — местоименное наречие (*где, вот*)
- **PRAEDICPRO** — местоимение-предикатив (*некого, нечего*)
- **PR** — предлог (*под, напротив*)
- **CONJ** — союз (*и, чтобы*)
- **PART** — частица (*бы, же, пусть*)
- **INTJ** — междометие (*увы, батюшки*)

открытые классы
- существительные
- глаголы
- прилагательные

закрытые классы (новые слова не появляются)
- местоимения
- предлоги

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

https://ruscorpora.ru/new/corpora-morph.html

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

Пример разбора средствами NLTK:

'Today morning, Arthur felt very good.'

[('Today', 'NN'), ('morning', 'NN'), (',', ','), ('Arthur', 'NNP'), ('felt', 'VBD'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')]

NN	Noun, Singular.
NNP	Proper Noun, Singular.
VBD	Verb, Past Tense. (took)
RB	Adverb. (very, silently)
JJ	Adjective.

https://ruscorpora.ru/new/corpora-morph.html

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

- алгоритмы на правилах (regexp)
- алгоритмы на трансформациях
- статистические модели
- ML модели

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

алгоритмы на правилах (RegexpTagger)

```
>>> patterns = [
...     (r'.*ing$', 'VBG'),                 # gerunds
...     (r'.*ed$', 'VBD'),                  # simple past
...     (r'.*es$', 'VBZ'),                  # 3rd singular present
...     (r'.*ould$', 'MD'),                 # modals
...     (r'.*\'s$', 'NN$'),                 # possessive nouns
...     (r'.*s$', 'NNS'),                   # plural nouns
...     (r'^-?[0-9]+(\.[0-9]+)?$', 'CD'),   # cardinal numbers
...     (r'.*', 'NN')                       # nouns (default)
... ]
```

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

алгоритмы на трансформациях (Brill tagger)

последовательно подбираем правила разметки,
следующее должно улучшать результат предыдущего

пример:
1. разметить все слова как NN (существительное)
2. для окончаний '...ский' изменить тег на JJ (прилагательное)
...
n. если слово имеет приставку S и предыдущий тег X то изменить тег на Z

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

статистические модели - HMM POS Tagger

для последовательности слов w
  подобрать последовательность тэгов t
  с максимальной вероятностью

$$\hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n | w_1^n)$$

$$\hat{t}_1^n = \arg\max_{t_1^n} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)}$$

применяем формулу Байеса

$$\hat{t}_1^n = \arg\max_{t_1^n} P(w_1^n | t_1^n) P(t_1^n)$$

последовательность слов не меняется,
P(w) на максимизацию не влияет

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

статистические модели - HMM POS Tagger

$$\hat{t}_1^n = \arg\max_{t_1^n} P(w_1^n|t_1^n)P(t_1^n)$$

Дополнительные предположения

оценка слова зависит только от тега

$$P(w_1^n|t_1^n) = \prod_{i=1}^{n} P(w_i|t_i)$$

тег разметки зависит только от предыдущего тега

$$P(t_1^n) = \prod_{i=1}^{n} P(t_i|t_{i-1})$$

Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

статистические модели - HMM POS Tagger

$$\hat{t}_1^n = \arg\max_{t_1^n} P(w_1^n|t_1^n)P(t_1^n)$$

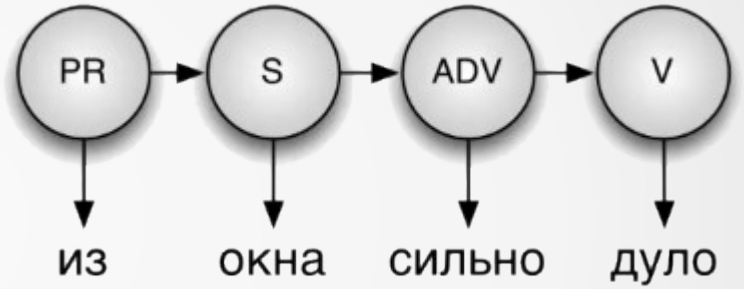
оценка слова зависит только от тега

$$P(w_1^n|t_1^n) = \prod_{i=1}^{n} P(w_i|t_i)$$

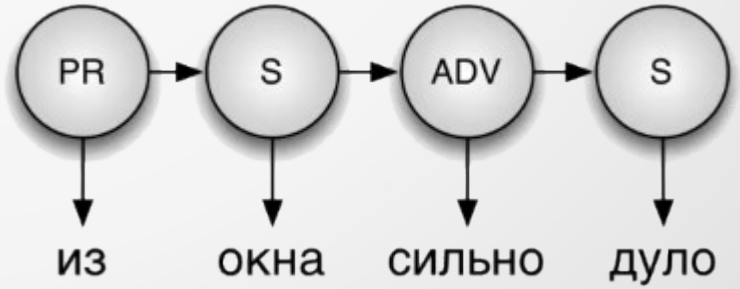
тег разметки зависит только от предыдущего тега

$$P(t_1^n) = \prod_{i=1}^{n} P(t_i|t_{i-1})$$

- выбираем наиболее вероятную последовательность тегов
- используем алгоритм Витерби
- используем сглаживание



разные варианты разметки



Морфологический анализ

Задача автоматической разметки частей речи
POS (Part of Speech) tagging

модели ML - строим классификатор

собираем размеченный датасет
[[<контекст>, слово] , метка слова]

обучаем классификатор размечать слова по контексту

Морфологический анализ

Литература

```
git clone https://github.com/mechanoid5/ml_nlp.git
```

Турдаков Д.Ю.
Основы обработки текстов. лекция 3. Разметка частей речи. ИСП РАН, 2017
https://www.youtube.com/watch?v=seAxPaKw33g

Steven Bird, Ewan Klein, and Edward Loper
Analyzing Text with the Natural Language Toolkit
https://www.nltk.org/book/

D.Jurafsky,J.H.Martin Speech and Language Processing. third edition, 2020