



# **Лексика и морфологический анализ**

Евгений Борисов

# Морфологический анализ

## Уровни сложности при автоматической обработке текстов

**Прагматика** - отдельные слова и устойчивые словосочетания

**Семантика** - смыслы последовательностей слов

**Синтаксис** - правила формирования последовательностей слов

**Морфология** - отдельные слова и устойчивые словосочетания

# Морфологический анализ

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

цель: предварительная обработка текста для извлечения информации

# Морфологический анализ



# Морфологический анализ

<b>ЧАСТИ РЕЧИ</b> <b>САМОСТОЯТЕЛЬНЫЕ ЧАСТИ РЕЧИ</b>			
<b>ИМЯ СУЩЕСТВИТЕЛЬНОЕ</b> ОБОЗНАЧАЕТ ПРЕДМЕТ		<b>КТО? ЧТО?</b> СОБАКА, ДЕВОЧКА, ПТИЦА, СОЛНЦЕ, ЧУВСТВО, ПИРОГ	
<b>ИМЯ ПРИЛАГАТЕЛЬНОЕ</b> ОБОЗНАЧАЕТ ПРИЗНАК ПРЕДМЕТА		<b>КАКОЙ? КАКАЯ? КАКОЕ? КАКИЕ? ЧЕЙ?</b> УМНЫЙ, БОЛЬШАЯ, ЖЁЛТОЕ, ДОБРЫЕ, ДЕДУШКИН	
<b>ГЛАГОЛ</b> ОБОЗНАЧАЕТ ДЕЙСТВИЕ ИЛИ СОСТОЯНИЕ ПРЕДМЕТА		<b>ЧТО ДЕЛАТЬ? ЧТО СДЕЛАТЬ?</b> БЕГАТЬ, РИСОВАТЬ, ЖИТЬ, ИЗУЧИТЬ, РАССКАЗАТЬ	
<b>ИМЯ ЧИСЛИТЕЛЬНОЕ</b> ОБОЗНАЧАЕТ КОЛИЧЕСТВО ИЛИ ПОРЯДОК ПРИ СЧЁТЕ		<b>СКОЛЬКО? КОТОРЫЙ?</b> ЧЕТЫРЕ, ОБА, НЕСКОЛЬКО, ПЕРВЫЙ, ДЕСЯТЫЙ	
<b>МЕСТОИМЕНИЕ</b> УКАЗЫВАЕТ НА ПРЕДМЕТ, ПРИЗНАК, КОЛИЧЕСТВО, НЕ НАЗЫВАЯ ИХ		<b>КТО? ЧТО? КАКОЙ? ЧЕЙ? СКОЛЬКО? КОТОРЫЙ?</b> Я, ТЫ, ВЫ, МЫ, ОНИ, НИКТО, ТАКОЙ, НИЧЕЙ, НЕМНОГО, ЛЮБОЙ	
<b>НАРЕЧИЕ</b> ОБОЗНАЧАЕТ ПРИЗНАК ДЕЙСТВИЯ		<b>ГДЕ? КОГДА? КУДА? ПОЧЕМУ? КАК?</b> БЛИЗКО, ДАВНО, НАПРАВО, СГОРЯЧА, МЕДЛЕННО	
<b>СЛУЖЕБНЫЕ ЧАСТИ РЕЧИ</b>		<b>ОСОБАЯ ЧАСТЬ РЕЧИ</b>	
<b>ПРЕДЛОГ</b> СЛУЖИТ ДЛЯ СВЯЗИ СЛОВ В СЛОВСОЧЕТАНИИ В, НА, К, О, ПЕРЕД, ИЗ	<b>СОЮЗ</b> СЛУЖИТ ДЛЯ СВЯЗИ ОДНОРОДНЫХ ЧЛЕНОВ ИЛИ ЧАСТЕЙ СЛОЖНОГО ПРЕДЛОЖЕНИЯ И, ИЛИ, НО, КОГДА	<b>ЧАСТИЦА</b> ВЫРАЖАЕТ ОТТЕНКИ ЗНАЧЕНИЙ СЛОВ, СЛОВСОЧЕТАНИЙ И ПРЕДЛОЖЕНИЙ, ОБРАЗУЕТ ФОРМЫ СЛОВА НЕ, НИ, БЫ, ДА, НЕТ	<b>МЕЖДОМЕТИЕ</b> ВЫРАЖАЕТ ЭМОЦИИ ИЛИ НОРМЫ ЭТИКЕТА АХ, АЙ, ОЙ, АЙ, ЭЙ, УРА, УВЫ, ЗДРАВСТВУЙТЕ, СПАСИБО, ПОЖАЛУЙСТА

0+

ПЛАКАТ  
Подготовлено к печати 09.07.2018 г. Формат 60х90 см.  
Усл. печ. л. 0,5. Тираж 10 000 экз. Заказ № 123.  
ООО «АЛФА» при участии ООО «ЛЕДА»  
214020, Рязань, ул. М. Горького, д. 4-2.  
Дистрибутор в Беларуси: частное предприятие «Слово А»,  
Беларусь, 210001, г. Витебск, ул. Комсомольская, 27111, к. 71.  
E-mail: [plakat@alfo-book.ru](mailto:plakat@alfo-book.ru) [www.alfo-book.ru](http://www.alfo-book.ru)  
Отпечатано: И. П. Лебедев Ю. М.  
Россия, 150010, г. Ярославль, ул. Ярославская, д. 148-12

© «Слово А», 2018  
© Оформление: ООО «Леда», 2018  
© Оформление: ООО «Алфа», 2018





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

# 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.

# Морфологический анализ

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

- алгоритмы на правилах (regex)
- 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

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

$$\hat{t}_1^n = \arg \max_{t_1^n} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)} \quad \text{применяем формулу Байеса}$$

$$\hat{t}_1^n = \arg \max_{t_1^n} P(w_1^n | t_1^n) P(t_1^n) \quad \text{последовательность слов не меняется, } P(w) \text{ на максимизацию не влияет}$$

# 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

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

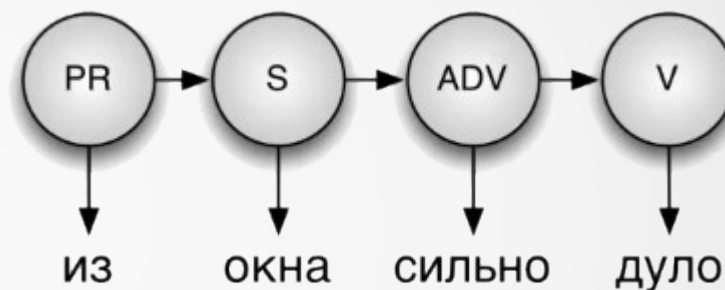
оценка слова зависит только от тега

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

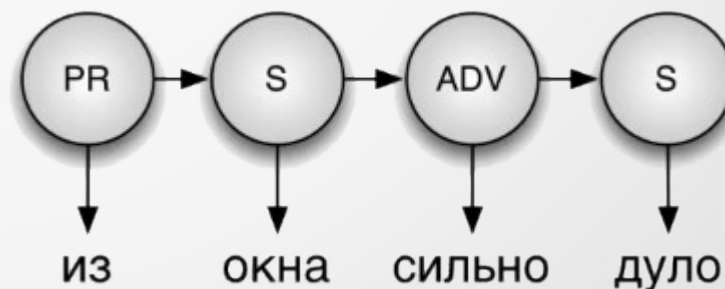
тег разметки зависит только от предыдущего тега

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

- 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](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