Поиск@Mail.Ru

# Высокоточный метод распознавания концов предложений

@mail.ru

# Актуальность

- Графематический анализ
- Выделение контекстов словоупотребления
- Рубрицирование документов
- Построение сниппетов в поисковых системах
- Прочие задачи компьютерной лингвистики

# Постановка задачи

**Цель** - возможность автоматического определения концов предложений в web-документах.

**Метод** - разработка средства классификации знаков препинания.

**Идея** – можно ли повысить эффективность участия человека в машинном обучении?

# «Предложения» в web

Web-документы обладают большим разнообразием текстовых элементов по сравнению с традиционными текстами.

Соответственно, «класс предложений» расширяется за счет дополнительных структур:

- Заголовки таблиц
- Элементы списков
- Подписи (рисунков и т.п.)
- Прочие обозначения, содержащие знаки пунктуации

# Пример

## Элементы библиографического списка удобно рассматривать как цельные сущности.

Пример оформления:
1. Крейдлин Г.Е. Невербальная семиотика // М.: Новое литературное обозрение, 2002.
2. Якобсон Р.О. О лингвистических аспектах перевода // Вопросы теории перевода в зарубежной лингвистике. М.: 1978. С.16-24.

Остальные параметры текста (размер шрифта, интервалы, отступы, цвет и так далее) являются несущественными, так как при вёрстке все эти параметры будут приведены в соответствие со стандартами, указанными верстальщику.

# Компоненты решения

- Небольшой (около 1000 знаков препинания) размеченный вручную текст, содержащий достаточное количество «трудных случаев»
- Транслятор языка описания правил
- Средство построения классификатора
- Средство оценки контекста

# Язык описания правил

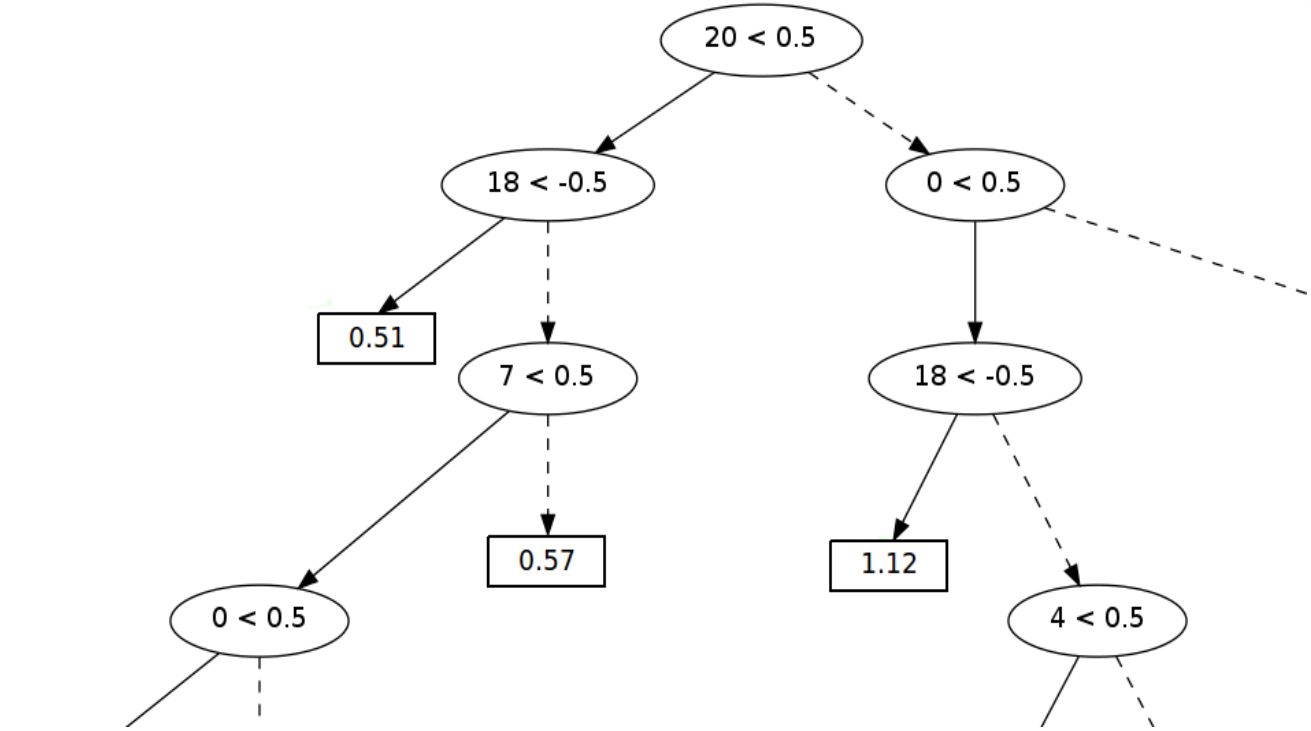
```
ABBR3a_L := RegEx(L) <- $(open)т\z

ABBR3a_R := RegEx(R) <- \A$(s)*[енк]\.

ABBR3b_L := RegEx(L) <- $(open)т\.$(s)*[енк]\z

ABBR3 := Rule() <-
        (ABBR3a_L & ABBR3a_R ) | ABBR3b_L
```

# Наиболее значимые правила

1. Тип разделителя
2. Пробел
3. Символ пунктуации
4. Цифра
5. Верхний/нижний регистр
6. Открывающая/закрывающая скобка
7. Стандартные сокращения
8. Авторские сокращения

и др.

# Фрагмент классификатора

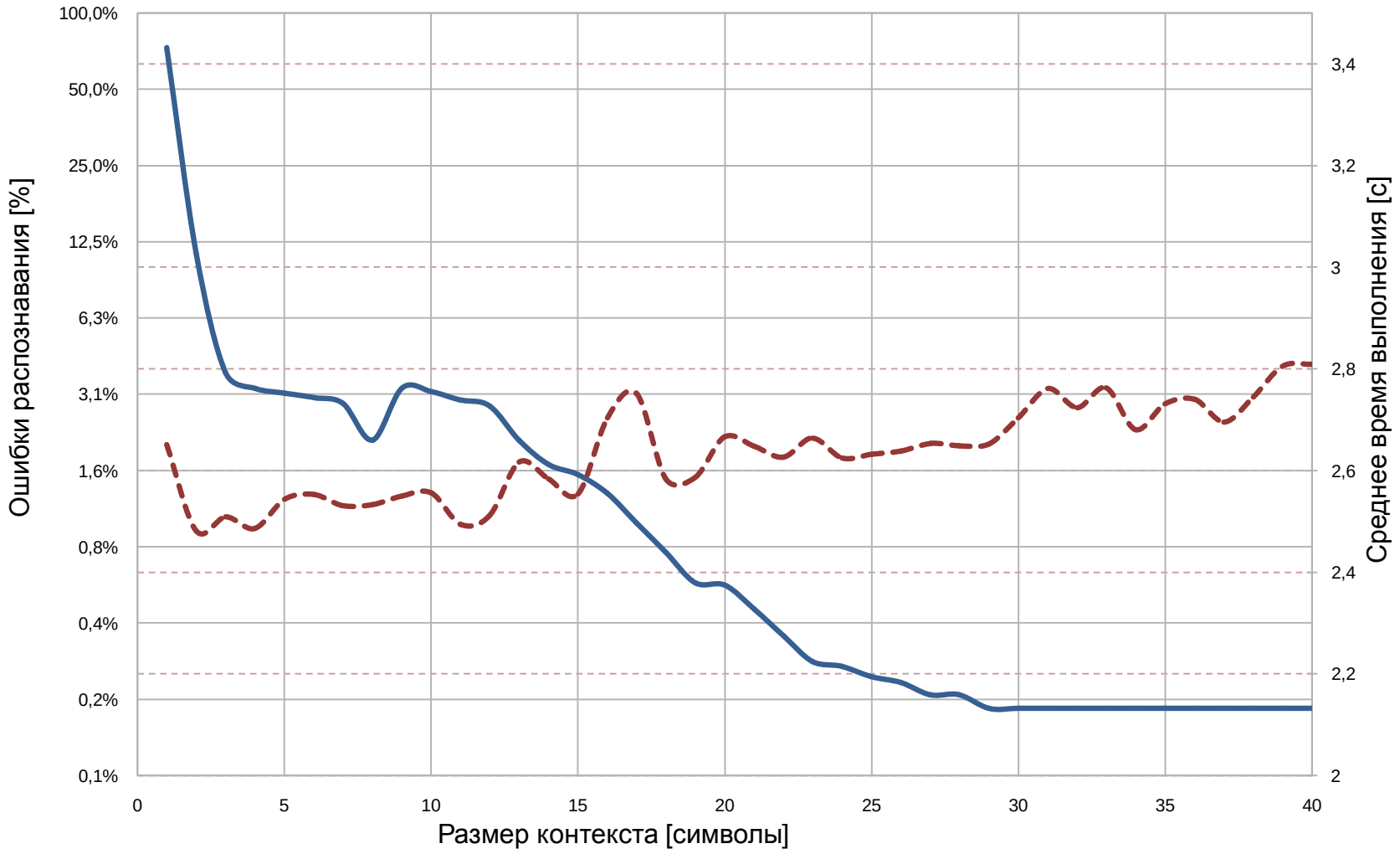



№0 - разделитель является точкой;
№4 - пробелы справа;
№7 - прописная буква слева;
№18 - многоточие;
№20 - титул справа.

# Сравнительные характеристики

| | **OpenNLP / Sentence Boundary Detector.** | | **Графематический модуль проекта «АОТ».** | | **Наш классификатор.** | |
|---|---|---|---|---|---|---|
| Алгоритм: | принцип максимальной энтропии | | эвристический | | набор правил и дерево принятия решений | |
| Число знаков в обучающей выборке: | 9820 | | Обучение не производилось. | | 9820 (та же выборка) | |
| Время выполнения: | 0,42 с | | 12,1 с | | 0,72 с | |
| **Тестовая выборка** | **Верные:** | **Неверные:** | **Верные:** | **Неверные:** | **Верные:** | **Неверные:** |
| Число разбиений: | 339 | 161 | 386 | 114 | 499 | 1 |
| Число слияний: | 455 | 45 | 462 | 38 | 497 | 3 |
| Общий процент ошибок (*Error rate*): | 41,2 % | | 30,4 % | | 0,8 % | |
| *F-мера Ван Ризбергена*: | 0,767 | | 0,835 | | 0,996 | |

# Точность и производительность



# Неразрешимый случай

*«Описание модели см. в А21. К-2301 т. IV, стр. 45. С 2003 г. изменена номенклатура.»*

# Основные результаты

- ✓ Для 99% точности достаточно 28 символов слева и 16 справа
- ✓ Практически удобный способ подбора и оценки правил
- ✓ Эффективное участие человека в процессе машинного обучения

# Благодарим за внимание!

@mail.ru