Метод выделения именованных сущностей на основе Википедии

Ткаченко Максим Владиславович
545 группа

Математико-механический факультет
Санкт-Петербургского государственного университета

Научный руководитель:
д.ф.-м.н., проф. Новиков Б.А.

Рецензент:
к.ф.-м.н., доц. Барашев Д.В.

Июнь 8, 2011

Введение в область

- Выделение имен объектов в тексте и определение их типа
- Типы сущностей: имена людей (PER), названия мест (LOC), организаций (ORG) и разное (MISC)

Пример:

Prime Minister [PER Benjamin Netanyahu] said
[LOC Israel] ...

Приложения

- Автоматическое извлечение структурированной информации из текста
- Упрощение обработки больших коллекций данных

Википедия и именованные сущности

- Около 73% статей Википедии описывают именованные сущности ($\sim$ 2,2 млн.)
- Энциклопедия содержит обширную информацию о синонимии и об омонимии терминов

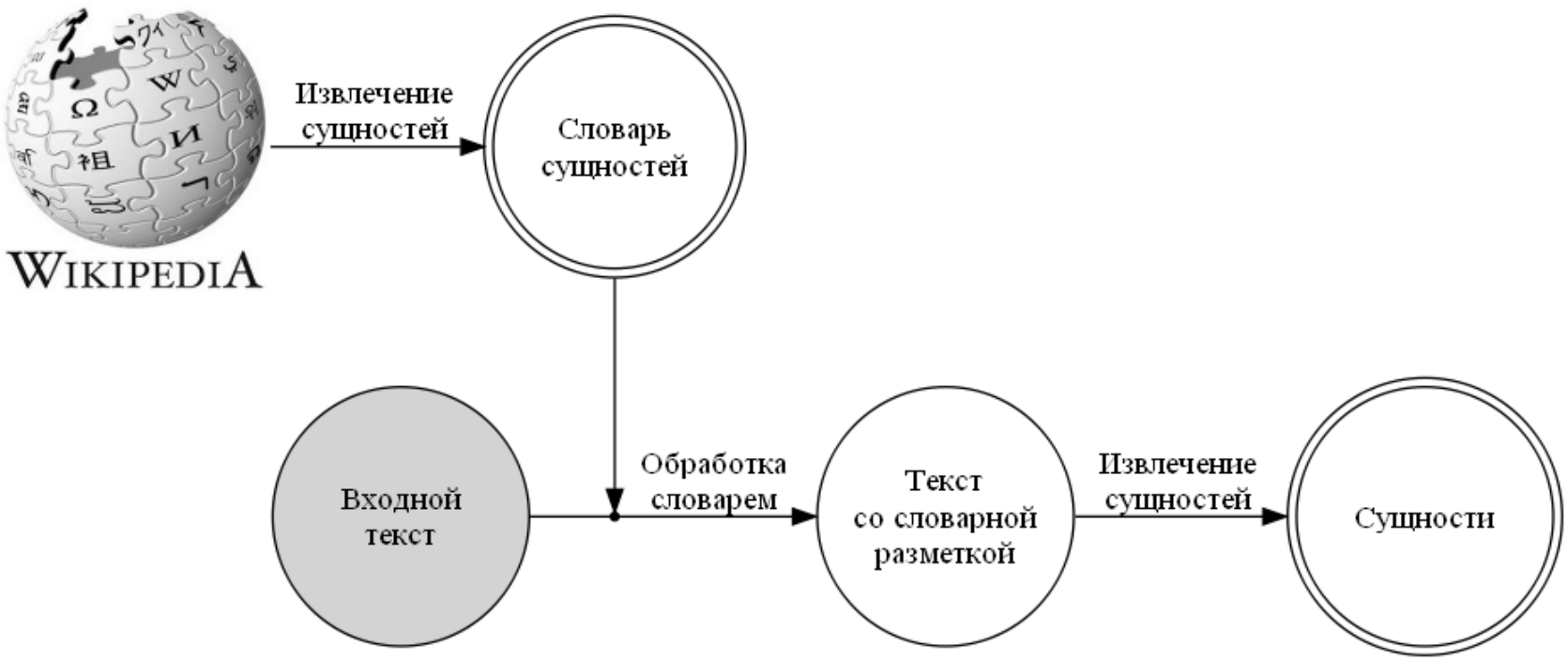
- Предыдущие работы:
  - Использование определений из Википедии - улучшение на 1.6%
  - Использование Википедии как дополнительной тренировочной коллекции - нет улучшения

Задачи

- Предложить и проанализировать новый подход к использованию Википедии в задаче выделения сущностей
- Создать и протестировать компоненту выделения сущностей для четырех классов

Метод

- Создать словарь сущностей на основе Википедии (Заголовок $\rightarrow$ Класс)
- Разметить классами все вхождения заголовков в тексте
- Использовать систему выделения сущностей на основе машинного обучения



Классификация Википедии

Классы:

- 15 типов сущностей: люди, места, организации, растения, животные, ...
- OTHER - не сущности

Подход:

- Методы машинного обучения: наивный байесовский классификатор (NB) и метод опорных векторов (SVM)
- Признаки: заголовок статьи, определение, категории, страницы-списки, шаблоны

Результаты классификации Википедии (1)

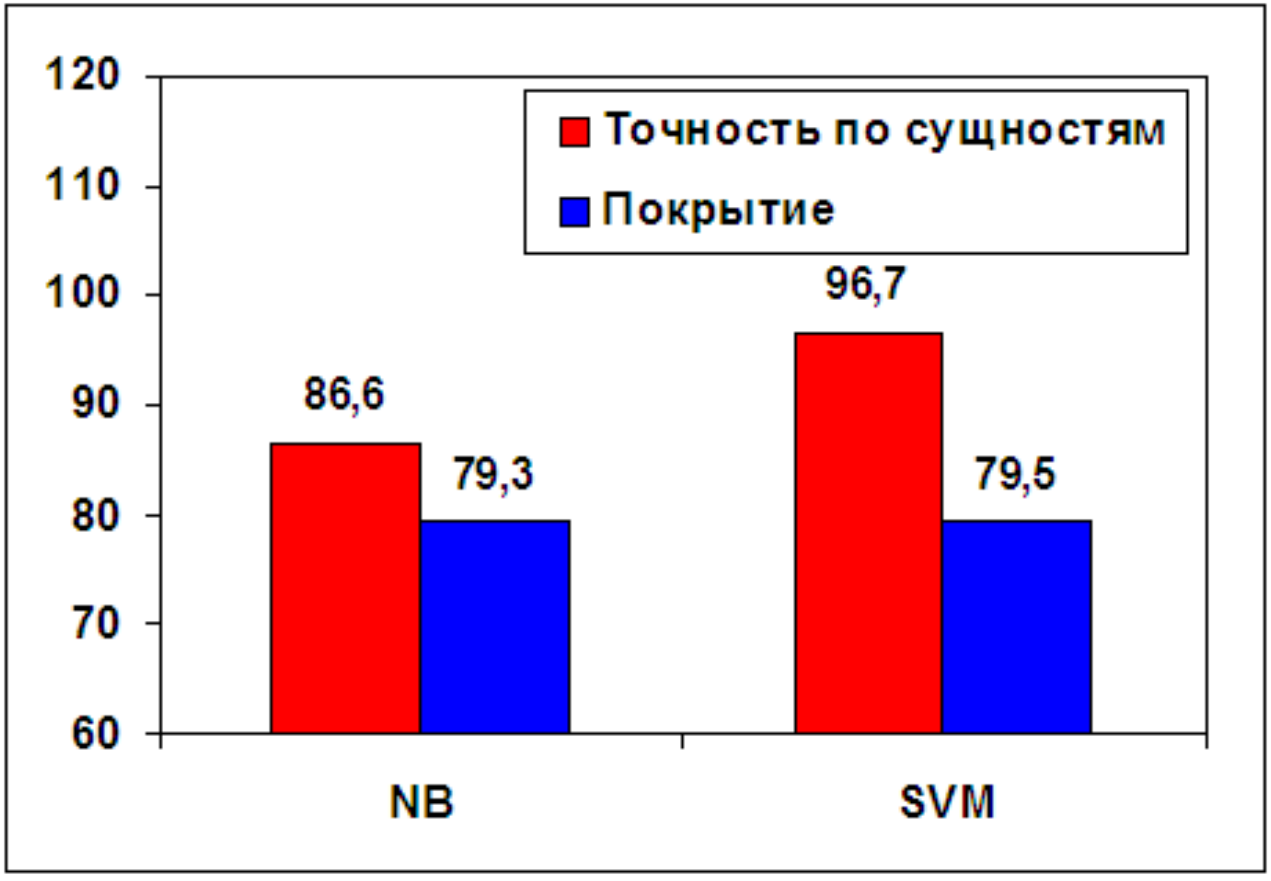
Результаты были проверены на случайной выборке:

- Тренировочное множество - 1357 стр.
- Тестовое множество - 680 стр.

Метрики:

- ТОЧНОСТЬ ПО СУЩ. = вероятность того, что классификатор правильно определит класс сущности, описываемой в статье
- ПОКРЫТИЕ = вероятность того, что статья, описывающая сущность, не будет определена как OTHER

Результаты классификации Википедии (2)



Результаты тестирования на случайной выборке страниц Википедии

Система выделения сущностей

- Модель условных случайных полей (conditional random fields)
- Базовые признаки:
  - Слова
  - Части речи
  - Формы слов
    - iPhone $\rightarrow$ xXxx, 12-month $\rightarrow$ 00-xx
  - Префиксы и суффиксы текущего слова длин от 2 до 5 символов
  - Является ли слово началом предложения
- Дополнительный признак: разметка текста с помощью Википедии

Тестирование (1)

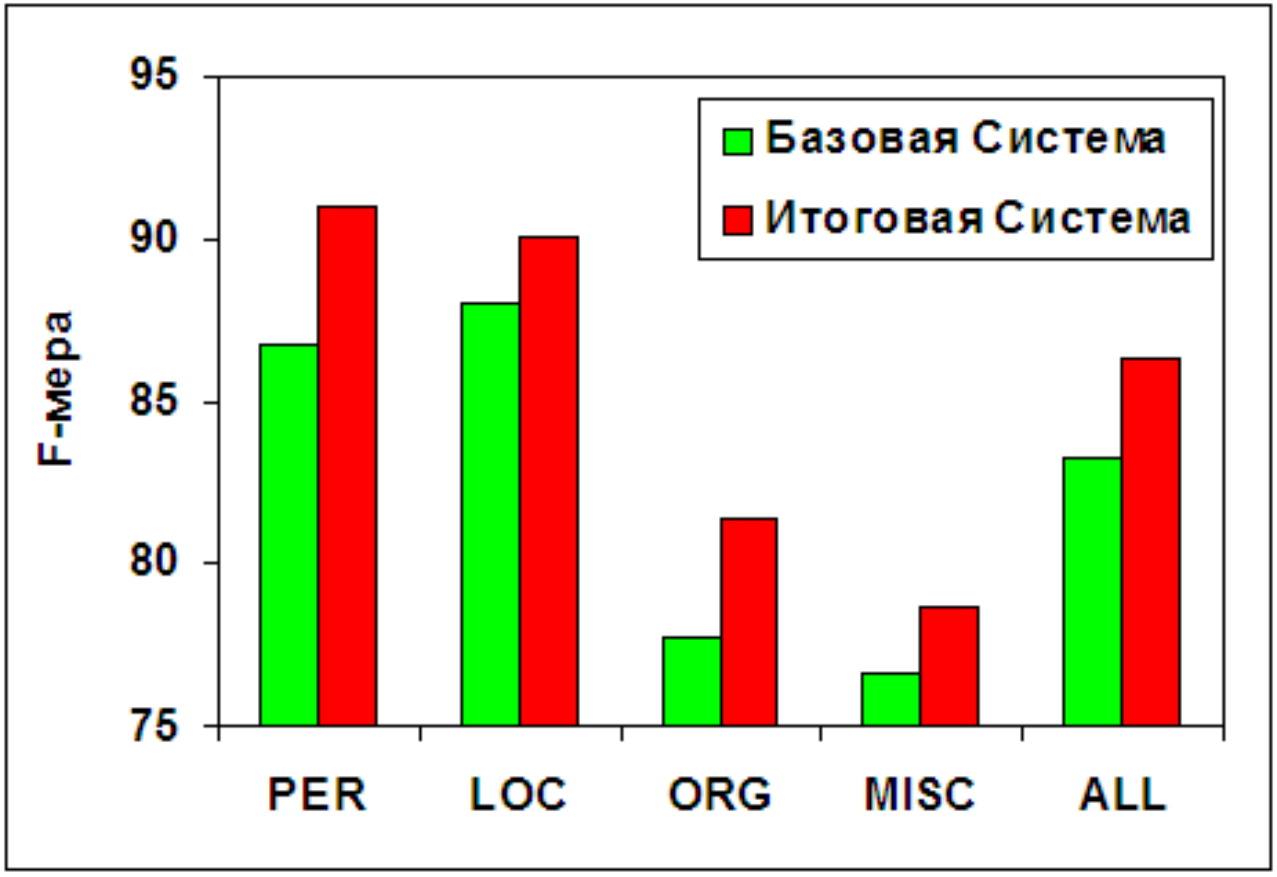
Результаты были проверены на коллекции CoNLL-2003:

- Классы: люди (PER), места (LOC), организации (ORG), разное (MISC)
- Тренировочное множество - 220 тыс. токенов
- Тестовое множество - 50 тыс. токенов

Метрики:

- ТОЧНОСТЬ (P) $= \frac{\text{Число верно выделенных сущностей}}{\text{Число всех выделенных сущностей}}$
- ПОЛНОТА (R) $= \frac{\text{Число верно выделенных сущностей}}{\text{Число сущностей в коллекции}}$
- F-МЕРА $= 2PR/(P+R)$

Тестирование (2)



Результаты тестирования на коллекции CoNLL-2003

Результаты

- Предложен подход к использованию Википедии в задаче выделения сущностей (улучшение на 3%)
- Получена разметка Википедии по 15 классам именованных сущностей
- Создана и протестирована компонента выделения сущностей для четырех классов