DATA-DRIVEN METHODS FOR ANAPHORA RESOLUTION OF RUSSIAN TEXTS

Каменская М.А., Российский университет дружбы народов
Смирнов И.В., Институт системного анализа РАН
Храмоин И.В., Институт системного анализа РАН

Международная конференция по компьютерной лингвистике «Диалог»
Бекасово, 4-8 июня 2014

Проблема разрешения анафоры для русского языка

- Теоретических исследований больше, чем практических.
- Мало общедоступных корпусов, размеченных на предмет анафоры.

Подходы к решению задачи для английского языка

- Подход, основанный на системе правил
  - *Winograd (1972), Wilks (1975), Hobbs (1976)*
  - *Rich and LuperFoy (1988), Carbonell (1988)*
- Методы машинного обучения
  - *Mitkov R. (1994, 1996), Connolly and Burger and Day (1994)*
  - *Ponzetto S. P., Strube M. (2006)*
  - *Kong F. (2008)*
  - *Huang Z. (2009)*
  - *Лукашевич Н. В. (2011)*
  - *Кибрик А.А. (2012)*
- CoNLL-2011 Shared task
  - *Lee H.*
  - *Zhou H.*

Подходы к решению задачи для русского языка

- Кибрик (1996)
  - Теоретические работы. Раскрывает теоретические аспекты явления референции и приводит ряд лингвистических признаков, отражающих природу анафоры.
- Толпегин (2006)
  - Алгоритм построения статистической модели, разрешающей анафору личных местоимений третьего лица, методами машинного обучения.
- Мальковский (2013)
  - Метод разрешения местоименной анафоры на основе результатов морфологического и синтаксического анализа, а также информации о сочетаемости слов.

Задачи исследования

- Изучить референцию личных местоимений третьего лица, указательных, возвратных местоимений.
- Исследовать влияние семантических признаков на качество разрешения анафоры русскоязычных текстов.
- Сравнить статистический метод машинного обучения и метод, основанный на правилах.

Постановка задачи разрешения анафоры

- Задача разрешения местоименной анафоры сводится к задаче распознавания правильных пар «анафор-антецедент» на основе анализа прецедентов.
- Множество прецедентов строится по размеченному корпусу. Содержит множество положительных примеров пар «анафор-антецедент» и отрицательных примеров пар.
- Гипотетический антецедент может отстоять в тексте слева не далее, чем на заданное количество слов, которое зависит от корпуса и определяется эмпирическим путем.
- Каждый обучающий пример представляется набором признаков и значений.

Основные этапы решения задачи

- Ручная разметка корпуса текстов на предмет референции местоимений третьего лица.
- Выбор признакового пространства.
- Построение обучающей выборки по размеченному корпусу.
- Применение метода обучения к обучающей выборке.
- Распознавание правильных пар на тестовом множестве.

Признаковое пространство

Морфологические и синтаксические признаки:

1. Род, число, падеж и одушевленность анафора в виде бинарных признаков
2. Род, число, падеж и одушевленность антецедента в виде бинарных признаков
3. Совпадает ли значение признака одушевленности анафора и антецедента
4. Количество предложений, разделяющих анафор и антецедент
5. Количество слов, расположенных в предложениях между анафором и рассматриваемым антецедентом
6. Количество гипотетических антецедентов, расположенных между анафором и рассматриваемым антецедентом
7. Количество существительных, расположенных в предложениях между анафором и рассматриваемым антецедентом.
8. В какой синтаксической связи состоят антецедент и анафор

Признаковое пространство

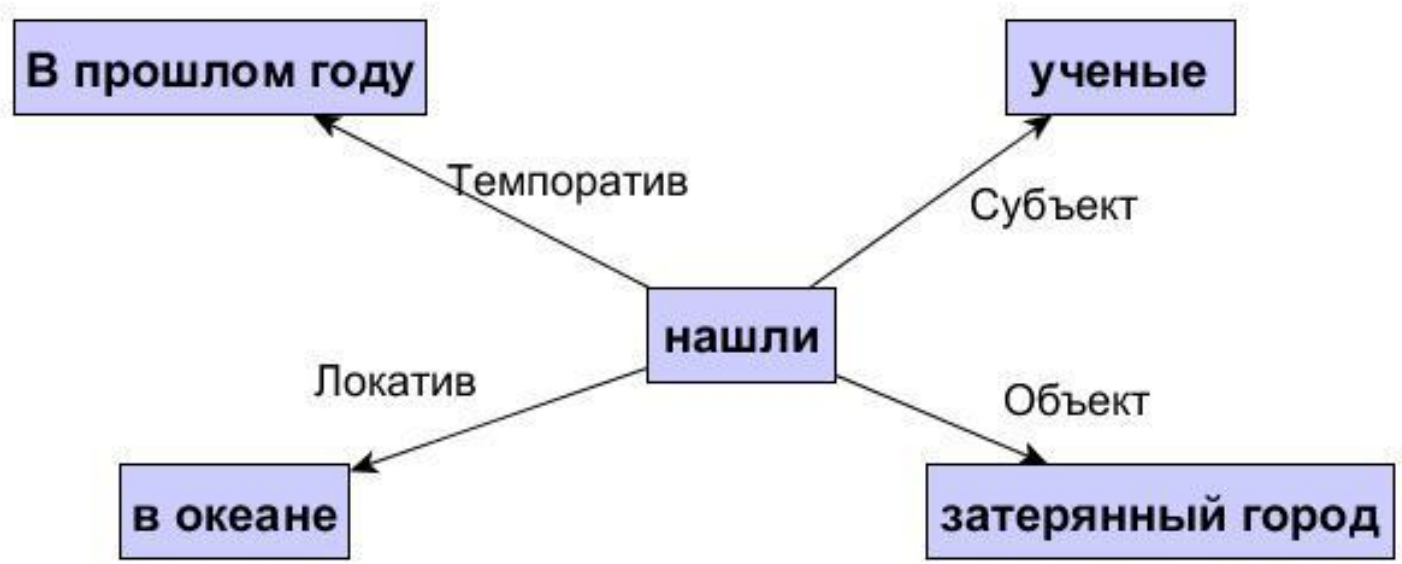
Семантические признаки:

9. Семантические роли анафора [Смирнов и др., 2014]
10. Семантические роли антецедента
11. Комбинация категориально-семантического класса предиката анафора и категориально-семантического класса антецедента
12. Комбинация категориально-семантического класса предиката анафора и категориально-семантического класса предиката антецедента

Семантические роли

Пример:

В прошлом году ученые нашли в океане затерянный город.



Алгоритм построения обучающей выборки по размеченному корпусу

1. Выбираем в тексте первую размеченную пару «анафор-антецедент».
2. Отыскиваем по тексту справа от анафора все существительные или местоимения, для которых ранее найден антецедент, согласующиеся с анафором в роде и числе. Область поиска ограничивается заранее заданным количеством слов.
3. Все найденные на шаге 2 существительные и местоимения становятся отрицательными примерами.
4. Шаги 1-3 повторяем до тех пор, пока не исчерпаем все размеченные пары «анафор-антецедент».

Методы обучения и классификации

Сравнивались 2 метода:

- статистический метод обучения, основанный на машине опорных векторов (реализация - LibSVM)
- индуктивный метод, основанный на построении деревьев решений (RepTree)

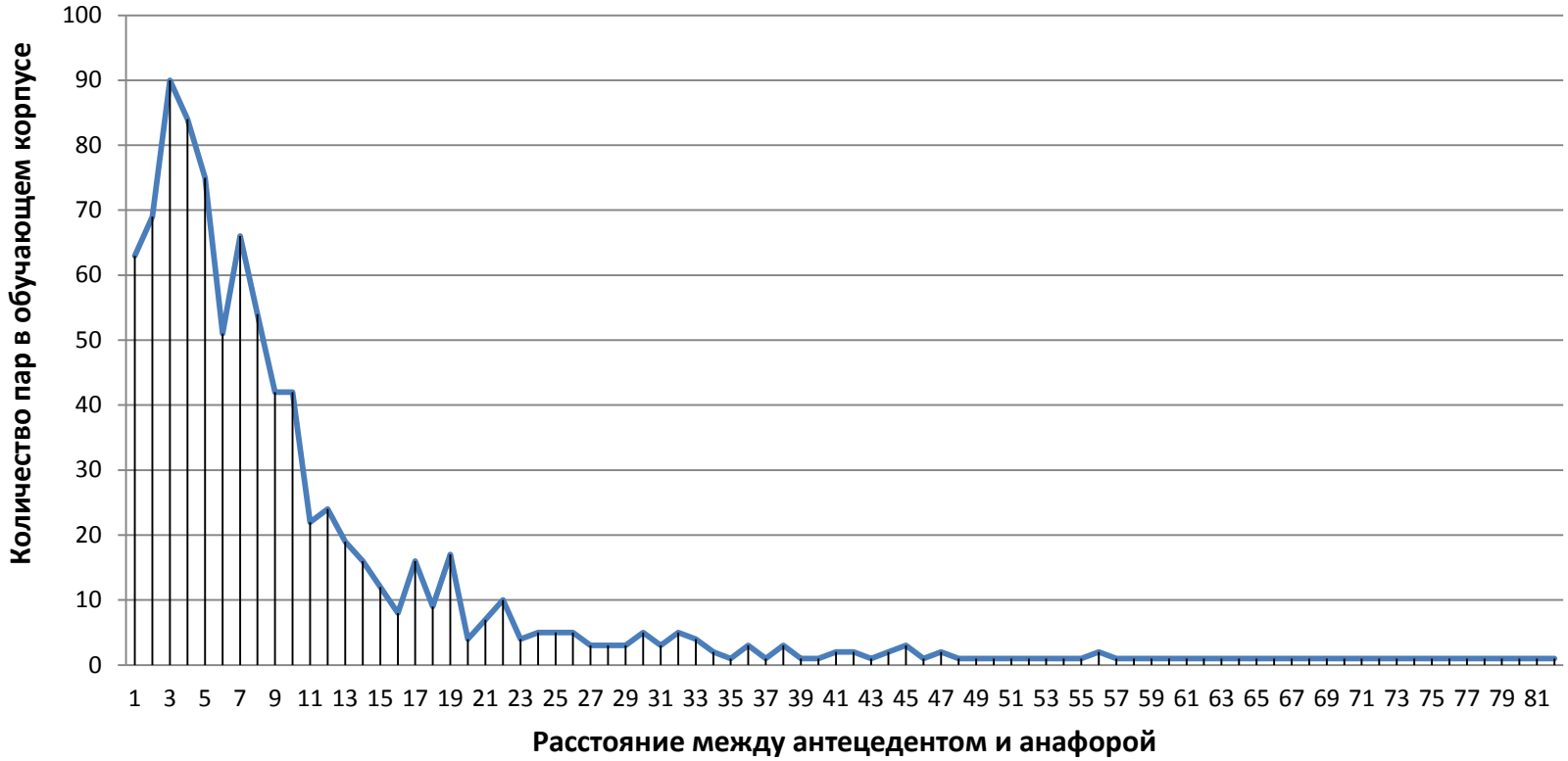
Алгоритм разрешения анафоры

1. Выбираем в тексте первый анафор, не состоящий в связи с каким-либо антецедентом. Если анафор не найден, алгоритм завершает свою работу.
2. Отыскиваем по тексту справа от анафора все существительные или местоимения, для которых уже найден антецедент, согласующиеся с текущим анафором в роде и числе, и добавляем их в список гипотетических антецедентов.
3. Для каждого местоимения в списке гипотетических антецедентов заменяем текущее значение категориального класса на значение категориального класса его антецедента.
4. Каждую пару «анафор-гипотетический антецедент» подаем на вход методу классификации и получаем на выходе вероятность того, что гипотетический антецедент является реальным антецедентом.
5. Выбираем антецедент, который с наибольшей вероятностью соотносится с анафором. Связываем выбранный антецедент с анафором. Переходим к шагу 1.

Корпуса текстов

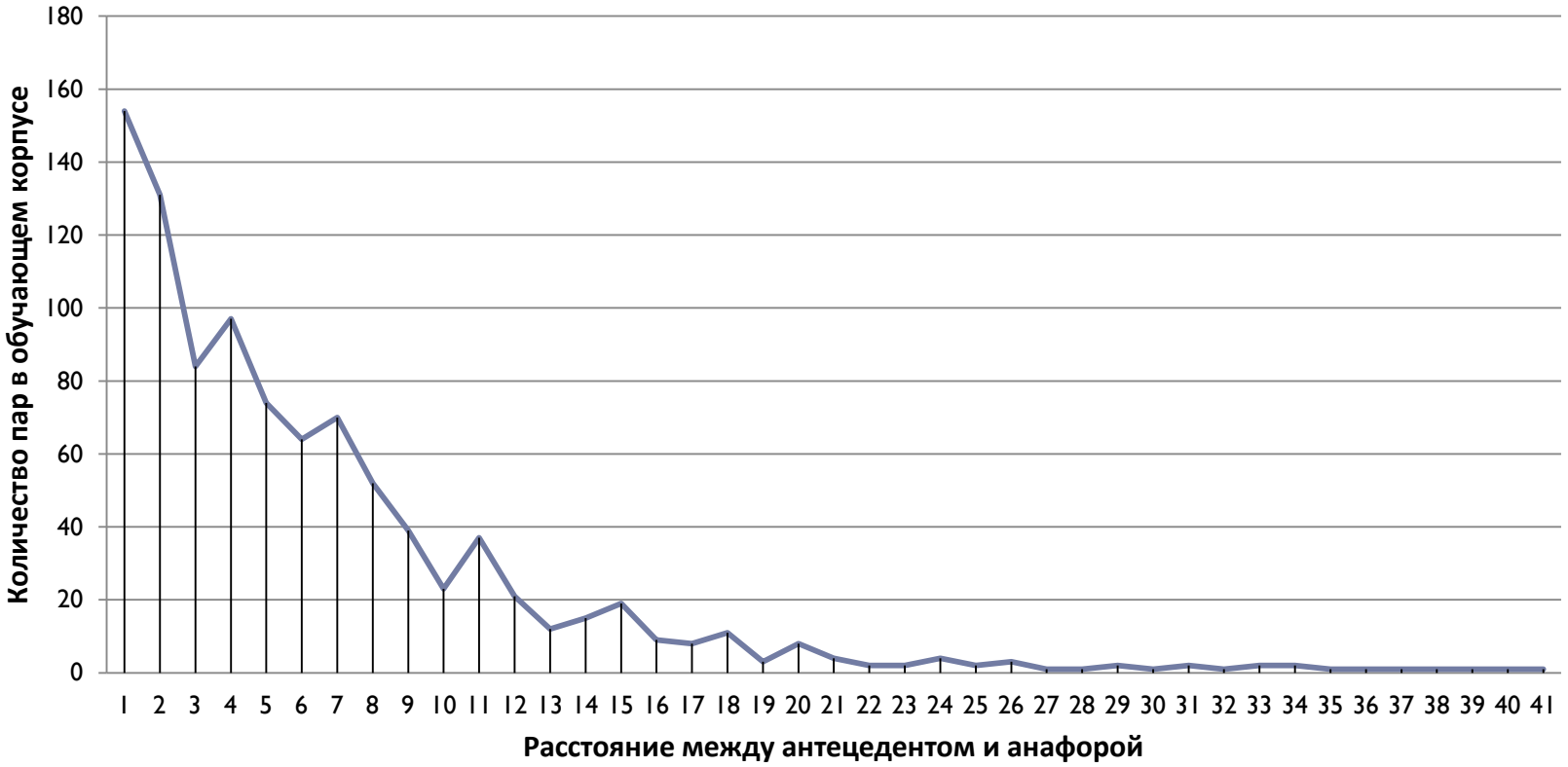
- Coprpus-1: мы разметили 15 текстов из библиотеки Мошкова + 35 текстов из корпуса СинТагРус: 910 анафорических пар
- Coprpus-2 предоставлен в качестве обучающего корпуса организаторами Форума по оценке систем лингвистического анализа текстов, проводимого в рамках конференции Диалог-2014: 92 текста и 967 анафорических пар
- Coprpus-3 объединяет corpus-1 и corpus-2.

Расстояние в словах между анафором и антецедентом в corpus-1



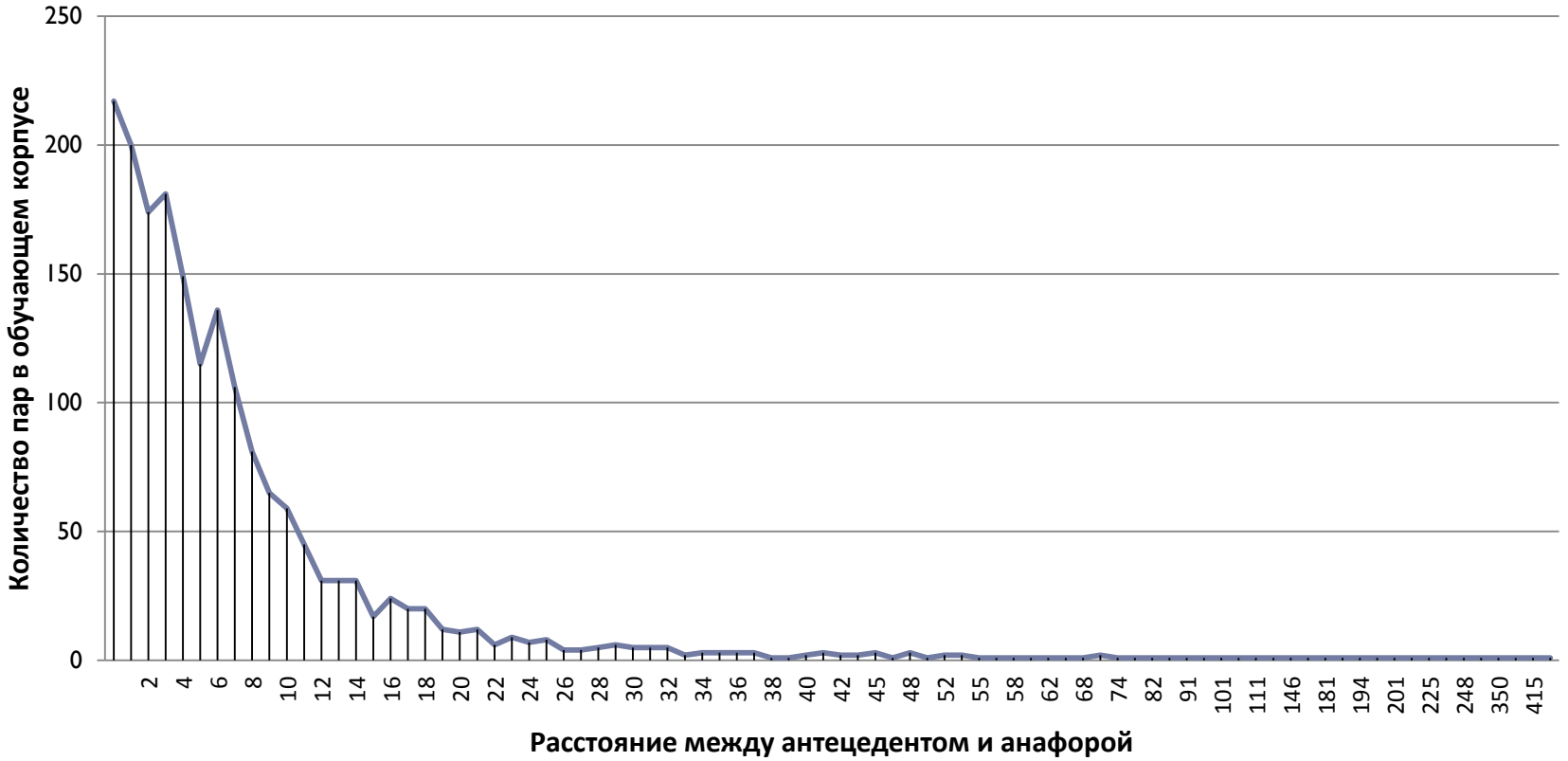
90% антецедентов отстает от анафора не более чем на 25 слов

Расстояние в словах между анафором и антецедентом в corpus-2



90% антецедентов отстает от анафора не более чем на 14 слов

Расстояние в словах между анафором и антецедентом в corpus-3



90% антецедентов отстает от анафора не более чем на 18 слов

Эксперименты. Признаки

Для определения влияния семантических признаков на точность разрешения анафоры проводилось несколько экспериментов по обучению на различных наборах признаков:

- FS-1: морфологические и синтаксические признаки
- FS-2: морфологические, синтаксические и семантические признаки

Эксперименты. Оценки

Использовались следующие оценки:

- SCORE-1 - оценка точности распознавания как положительных, так и отрицательных примеров анафорических пар (на обучающем множестве)
- SCORE-2 - оценка точности классификации пар только с действительным антецедентом (оценка точности разрешения анафоры)

Результаты экспериментов. CORPUS-1

Наборы признаков	SVM	REPTree
SCORE-1		
FS-1	0.811	0.773
FS-2	**0.821**	0.789
SCORE-2		
FS-1	0.473	0.484
FS-2	**0.539**	0.529

Результаты экспериментов. CORPUS-2

Наборы признаков	SVM	REPTree
SCORE-1		
FS-1	0.746	0.746
FS-2	**0.771**	0.747
SCORE-2		
FS-1	0.603	0.592
FS-2	**0.610**	0.609

Результаты экспериментов. CORPUS-3

Наборы признаков	SVM	REPTree
SCORE-1		
FS-1	0.766	0.634
FS-2	**0.781**	0.689
SCORE-2		
FS-1	0.571	0.548
FS-2	**0.579**	0.553

Выводы

- Метод опорных векторов показал по сравнению с деревьями решений результаты лучшие на величину от 0.1% до 13.2% точности.
- Обучение с набором семантических признаков показало повышение точности обучения по сравнению с набором без семантических признаков на величину от 0.1% до 6.6%.
- Наилучший достигнутый результат точности разрешения анафоры - 61%. Получен на Corpus-2 методом SVM на наборе признаков FS-2.

Дальнейшие работы

- Расширение пространства признаков.
- Совершенствование метода распознавания потенциальных антецедентов.

Спасибо за внимание!

Каменская М.А. ma_kamenskaya@mail.ru
Смирнов И.В. ivs@isa.ru
Храмоин И.В. hramoin@isa.ru