МЕТОДЫ РАЗРЕШЕНИЯ МЕСТОИМЕННОЙ АНАФОРЫ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Каменская М.А., Храмоин И.В.

Российский университет дружбы народов, ma_kamenskaya@mail.ru, ivan_khramoin@yahoo.com

Доклад посвящен исследованию методов автоматического разрешения анафоры местоимений третьего лица. Целями исследования являются сравнение двух методов разрешения анафоры и оценка влияния семантических признаков на точность разрешения анафоры.

Ключевые слова: разрешение анафоры, машинное обучение, метод опорных векторов, деревья решений.

Введение

Лингвистический анализ текстов на естественном языке представляет интерес в задачах проектирования систем машинного перевода, информационного поиска и извлечения информации. Наибольшую трудность при анализе текста вызывает анализ связей между словами, в частности референциальных связей. Референциально связанными называются слова, которые описывают один и тот же объект реального мира. В данном исследовании рассматривается задача разрешения местоименной анафоры, то есть задача поиска референциальных связей, в которых участвуют личные местоимения третьего лица, возвратные и указательные местоимения. Именная группа, на которую ссылается анафор (т.е. местоимение), называется антецедентом.

Разрешение анафоры является важным этапом в автоматической обработке естественно-языковых текстов, повышающим качество понимания текста. Задача автоматического разрешения анафоры для русского языка в настоящее время активно исследуется, но соотношение точности и полноты существующих методов не позволяет считать задачу решенной. В работе [1] исследовано влияние семантических признаков на качество разрешения анафоры для английского языка. Исследование показало, что использование семантических ролей повышает точность автоматической расстановки анафорических связей на величину от 0.1% до 5.6%.

Целью исследования является оценка влияния семантических признаков на точность разрешения анафоры для русского языка и сравнение статистического метода обучения, основанного на машине опорных векторов, и индуктивного метода, основанного на построении деревьев решений.

Методы

В данном исследовании задача разрешения местоименной анафоры сводится к задаче распознавания правильных пар «анафор-антецедент» на основе анализа прецедентов и делится на два этапа: этап обучения и этап разрешения анафоры.

На этапе обучения выявляются закономерности, позволяющие в дальнейшем классифицировать пары «анафор-гипотетический антецедент» как пары с действительной анафорической связью. Гипотетическим антецедентом считается имя существительное или местоимение, для которого уже установлена анафорическая связь, согласованные с анафором по числу и роду. Множество обучающих примеров строится по размеченному корпусу текстов и содержит множество положительных примеров пар «анафор-антецедент», между которыми действительно существует анафорическая связь, и отрицательных примеров анафорических пар, где вторым компонентом является гипотетический антецедент, не имеющий связи с анафором.

Каждый обучающий пример представляется набором морфологических, синтаксических и семантических признаков:

Морфологические и синтаксические

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.

Семантические

9. семантические роли анафора;
10. семантические роли антецедента;
11. категориально-семантический класс антецедента;
12. комбинация категориально-семантического класса предиката анафора и категориально-семантического класса антецедента;
13. комбинация категориально-семантического класса предиката анафора и категориально-семантического класса предиката антецедента.

Значения признаков вычислялись на основе результатов морфологического, синтаксического и семантического анализа текстов с помощью лингвистического анализатора ИСА РАН [2].Методы установления семантических ролей и категориально-семантических классов описаны в работе [3].

На этапе разрешения анафоры происходит поиск анафоров и гипотетических антецедентов. Затем каждая пара классифицируется как анафорическая или неанафорическая с помощью модели, полученной на этапе обучения.

В качестве методов классификации использовались метод опорных векторов SVM [4] и метод построения деревьев решений REPTree [5].

Для обучения и проверки методов использовалось несколько корпусов, вручную размеченных на наличие анафоры. Первый корпус включал 17 текстов из библиотеки Мошкова, 34 текста из корпуса СинТагРус и содержал 910 анафорических пар. Второй корпус был предоставлен в качестве обучающего корпуса организаторами Форума по оценке систем лингвистического анализа текстов, проводимого в рамках конференции Диалог-2014. Он содержал 92 текста и 967 анафорических пар.

Экспериментальное исследование

Для определения влияния семантических признаков на точность разрешения анафоры проводилось несколько экспериментов по обучению на различных наборах признаков. Первый набор включал только морфологические и синтаксические признаки, а во втором наборе к ним добавлялись семантические признаки.

В табл. 1 представлены результаты эксперимента на первом корпусе, в табл. 2 – на втором корпусе. Точность установления анафорических пар рассчитывалась по следующей схеме: первая оценка SCORE-1 касалась точности распознавания как положительных, так и отрицательных примеров анафорических пар (т.е. это классификация по двум классам – правильная пара, неправильная пара), вторая оценка SCORE-2 касалась точности классификации пар только с действительным антецедентом (т.е. она показывает собственно точность разрешения анафоры).

Таблица 1. Точность разрешения анафоры для различных методов и наборов признаков на первом корпусе

Наборы признаков	SVM	REPTree
SCORE-1		
Морфологические и синтаксические	0.811	0.773
+ семантические	0.821	0.789

SCORE-2		
Морфологические и синтаксические	0.473	0.484
+ семантические	0.539	0.529

Таблица 2. Точность разрешения анафоры для различных методов и наборов признаков на втором корпусе

Наборы признаков	SVM	REPTree
SCORE-1		
Морфологические и синтаксические	0.746	0.746
+ семантические	0.771	0.747
SCORE-2		
Морфологические и синтаксические	0.603	0.592
+ семантические	0.61	0.609

Выводы

Результаты экспериментов показали, что оба используемых метода обучения выдают приемлемые результаты. Метод опорных векторов во всех экспериментах, за исключением двух, показал по сравнению с деревьями решений результаты лучшие на величину от 0.1% до 3.8% точности. Обучение с набором семантических признаков для всех методов во всех экспериментах показало повышение точности обучения по сравнению с набором без семантических признаков на величину от 0.1% до 6.6%.

Наилучший результат точности разрешения анафоры в 61% был достигнут на втором корпусе методом SVM с использованием семантической группы признаков.

Таким образом, по результатам экспериментов метод опорных векторов выдаёт лучшие результаты по сравнению с результатами метода построения деревьев решений. Также можно сделать вывод, что семантические признаки улучшают качество разрешения местоименной анафоры.

Литература

1. *Ponzetto S.P., Strube M.* Semantic role labeling for coreference resolution // EACL '06 Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. – Pp. 143-146.

2. *Осипов Г.С., Смирнов И.В., Тихомиров И.А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения //Журнал "Искусственный интеллект и принятие решений". Номер 2-2008. - С. 3-10.

3. *Смирнов И.В., Шелманов А.О., Кузнецова Е.С., Храмоин И.В.* Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов // Искусственный интеллект и принятие решений. М.: ИСА РАН – 2014 – №1 – с. 95-108.

4. LIBSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

5. Weka 3: Data Mining Software in Java. University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/

MACHINE LEARNING METHODS FOR ANAPHORA RESOLUTION

Kamenskaya M.A., Khramoin I.V.

Peoples' Friendship University of Russia, ma_kamenskaya@mail.ru, ivan_khramoin@yahoo.com

This paper presents the methods of ahaphora resolution based on machine learning. The question of effect of semantic role lableling is also raised.

Key words: anaphora resolution, machine learning, support vector machine, decision trees.