АЛГОРИТМ АВТОМАТИЗИРОВАННОГО РАЗРЕШЕНИЯ АНАФОРЫ МЕСТОИМЕНИЙ ТРЕТЬЕГО ЛИЦА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

AUTOMATED THIRD PERSON ANAPHORA RESOLUTION ALGORITHM ON THE BASIS OF MACHINE LEARNING METHODS

П.В. Толпегин *(pavel@tolpegin.ru)*
Д.П. Ветров *(vetrovd@yandex.ru)*
Д.А. Кропотов *(dkropotov@yandex.ru)*

Вычислительный центр им. А.А. Дородницына РАН

Рассматривается один из подходов к автоматизированной расстановке анафоры местоимений третьего лица. Правила референции были получены с помощью применения машинного обучения. Апробация показала точность более 60%.

Введение

Разрешение анафоры является одной из центральных проблем в задаче анализа русскоязычных естественно-языковых (ЕЯ) текстов. В современной литературе также встречаются: прагматический анализ или референционный анализ. Задача этого вида анализа текстов заключается в связывании построенных на этапе семантического или синтаксического анализа графов в единую когнитивную карту по особым правилам.

Анафора является весьма широким понятием, которое скрупулезно исследуется в работах ученых-лингвистов (Н.Д. Арутюновой, Т.В. Булыгиной, Дж. Гандел, А.А. Кибрика, Л.Н. Иорданской, Дж. Николс, Е.В. Падучевой, А.С. Чехова, А.Д. Шмелева и др.). В нашем случае исследуется одна из возможностей автоматизированного разрешения анафоры местоимений третьего лица.

Особый интерес системы автоматизированной расстановки референции для местоимений представляют при проектировании систем машинного перевода. Кроме того, они могут быть полезны для расширения смыслового представления текста (например, в модели «Смысл <-> Текст» [1], а также в модуле первичного семантического анализа [2].)

В качестве материала для исследований были выбраны выдержки из новостных лент: фиксировались естественные данные, которые опираются не только на морфологические и синтаксические признаки, построенные не по шаблонным моделям анафоры местоимений. В последнее время, в большей степени становятся популярными работы (Г.Хирст, Ш.Лаппин, Р.Митков, М.Поэсио и др.) по созданию ЕЯ-корпусов, размеченных на предмет референции для западноевропейских языков. К сожалению, Национальный корпус русского языка (НКРЯ) в настоящее время не имеет анафорической разметки.

В настоящей работе для разрешения анафоры применена методология машинного обучения по прецедентам, позволяющая извлечь скрытые закономерности, содержащиеся в наборах данных [5]. Основной гипотезой при проведении исследований являлось предположение о возможности разрешения анафор в большинстве случаев по некоторым формальным признакам.

Технология решения

Для реализации поставленной задачи был вручную размечен корпус новостных текстов объемом 3 Мбайта на предмет референции местоимений третьего лица.

Представляемый для ручной разметки анафоры текст проходил предварительную автоматизированную морфологическую, синтаксическую и первичную семантическую разметку при помощи программных решений А.В. Сокирко и И.М. Ножова [3,4].

Эксперт-лингвист производил ручную разметку анафоры местоимений третьего лица, ассоциируя местоимение с встречающимся ранее существительным или местоимением (в случае комплексной референции).

В качестве функции, отвечающей за референциальный выбор, была выбрана модель, получающая в качестве входных параметров признаки текущей анафоры и признаки гипотетических антецедентов, согласованных с анафорой в роде и числе и встречавшихся ранее в тексте. Алгоритм функции анализирует полученные данные и выдает в качестве ответа нечеткую лингвистическую оценку, исходя из значения которой можно судить о том, реферируют ли между собой выбранные в паре существительное и местоимение или нет. Более подробно механизм принятия решения описан в главе «Структура решающего правила».

Признаковое пространство

В качестве признаков были выбраны:

1) число имен собственных между анафорой и антецедентом;
2) количество предложений, разделяющих анафору и антецедент;
3) стоит ли антецедент в именительном падеже:
4) является ли антецедент именем собственным;
5) количество существительных и местоимений, расположенных в предложениях между рассматриваемыми анафорой и антецедентом;
6) совпадает ли падеж анафоры и антецедента;
7) статистическая информация о том, в каком сегменте предложения располагается антецедент – насколько ближе к началу;
8) статистическая информация о том, в каком сегменте предложения располагается анафора – насколько ближе к началу;
9) количество анафор, реферирующих с текущим антецедентом по данным ручной разметки, расположенных между анафорой и антецедентом;
10) число глаголов в сегменте, содержащем антецедент;
11) число причастий и деепричастий в сегменте, содержащем антецедент;
12) число местоименных прилагательных и союзов в сегменте, содержащем антецедент;
13) число существительных в именительном падеже в сегменте, содержащем антецедент;
14) род, падеж и число анафоры и антецедента (в виде бинарных признаков);

Общий объем обучающей выборки составил более 8000 записей.

Для дальнейшего анализа статистических сведений были задействованы методы распознавания образов.

Структура решающего правила

Для принятия решения использовалась система алгоритмов опорных векторов (Support Vector Machines) [6]. Эти алгоритмы хорошо зарекомендовали себя при решении большого количества практических задач, связанных с анализом данных. Классический метод опорных векторов, предназначенный для обработки независимых прецедентов, был видоизменен с учетом специфики задачи. В нашем случае в качестве прецедента выступает пара анафора-антецедент, которая принадлежит к одному из двух классов в зависимости от наличия в ней референции. Очевидно, что среди гипотетических антецедентов, предшествующих данной анафоре, найдется, по крайней мере, один, связанный с ней референцией. Для того чтобы избежать неоднозначности, можно полагать, что такой объект единственен (в самом деле, связанность анафоры хотя бы с одним антецедентом, например, с ближайшим, в силу транзитивности отношения автоматически гарантирует ее связность со всеми остальными гипотетическими антецедентами, связанными с данным). Для разрешения анафоры использовался следующий алгоритм.

1) Фиксируем очередную анафору и присваиваем n=1.
2) Переходим к следующему, n-му гипотетическому антецеденту, согласованному в роде и числе, расположенному ранее в тексте начиная с ближайшего к рассматриваемой анафоре.
3) Для пары анафора-антецедент запускаем метод опорных векторов n-го уровня.
4) Выход метода опорных векторов y_n преобразуем в оценку степени уверенности в том, что данная пара связана референциальной связью по следующей формуле

$$p_n = \frac{1}{1+\exp(\lambda y_n)}.$$

5) Увеличиваем n на единицу. Если n<N, то переход к шагу 2, иначе - к шагу 6.
6) Получившиеся оценки степени уверенности умножаем на релаксационные коэффициенты r_n, отражающие априорные знания о статистическом распределении реферируемых пар

$$s_n = p_n r_n .$$

7) Связываем рассматриваемую анафору с m-ым антецедентом, где $m = \arg\max_n s_n$. Затем переход к шагу 1.

Метод опорных векторов n-го уровня обучался по таким парам анафора-антецедент, в которых между ними было ровно n-1 существительных или местоимений, согласованных с анафорой в роде и числе (т.е. n-1 потенциальных антецедентов для рассматриваемой анафоры). Пары, в которых имела место референция, были отнесены к первому классу, а остальные – ко второму. Далее для такой обучающей выборки запускался

стандартный метод опорных векторов, разделявший два класса. Для новой пары объектов с вектором признаков x, выход метода определяется как

$$y(x) = \sum_{i=1}^{k} w_i K(x_i, x) + w_0 ,$$

где w_i, w_0 - веса алгоритма, которые находятся в процессе обучения, $K(x', x'')$ - ядровая функция, обеспечивающая нелинейность получаемой границы между классами. В наших экспериментах использовался наиболее распространенный вид функции $K(x', x'') = \exp(-\gamma \|x' - x''\|^2)$, где параметр γ находился в процессе кросс-валидации [5].

Исходя из априорных знаний известно, что чаще всего анафоры реферируются с ближайшим гипотетическим антецедентом и что вероятность реферируемости падает с ростом n. В качестве коэффициентов релаксации были выбраны регуляризованные частоты встречаемости реферируемых пар n-го уровня относительно всех пар n-го уровня, вычислявшиеся по формуле

$$r_n = \theta \frac{1}{N} + (1 - \theta) \nu_n$$

Коэффициент регуляризации θ также подбирался с помощью кросс-валидации. С учетом количественного распределения реферируемых анафор, значение N в настоящем исследовании было выбрано равным 4. Доля анафор, ближайшие гипотетические антецеденты которых, находятся вне зоны анализа составляет менее 10%.

Данные о выборках, использовавшихся для обучения и контроля приведены в таблице 1.

	Обучающая совокупность	Контрольная совокупность
Общее количество пар	3936	3927
Количество связанных пар	230	207
Количество анафор	205	163
Точность референции	81%	62%

Таблица 1. Данные о выборках

Точность работы алгоритма составила 62% правильно отреферированных анафор. В большинстве случаев для референции анафоры имелось не менее трех вариантов (в некоторых случаях до 30), удовлетворяющих формальным критериям (согласованность в роде и числе). Столь высокий результат можно объяснить сочетанием удачного подбора признаков, грамотным выбором метода распознавания и адекватной структурой решающего правила.

Заключение

Полученный результат можно считать вполне удовлетворительным. Полученная погрешность в автоматизированной расстановке анафоры местоимений третьего лица свидетельствует не столько об ошибке, накопившейся при использовании модулей графематического, морфологического, синтаксического и первичного семантического анализа, сколько о том, что авторы анализируемых текстов использовали семантические средства языка для передачи смысла референции. Иначе говоря, для одинаковых синтаксических конструкций могло существовать более одного случая с несовпадением антецедента.

Но исследования на этом не прекращаются. Убедившись в эффективности результатов исследований, авторы планируют расширить область анализа в пользу личных, возвратных, притяжательных и, в особенности, указательных местоимений. Последние представляют собой особый интерес не столько для машинного перевода, сколько для построения когнитивной карты текста.

Пример более комплексной референции. Профессор Рубинштейн рассказывал о новом методе машинного обучения на семинаре. Об этом я услышал впервые.(1)

При более комплексном подходе к разрешению анафоры в ЕЯ-тексте можно говорить на примере словаря [7], который мог бы лечь в основу базы знаний «О Мире». Таким образом, межклаузная референция могла бы также проводиться на основе родовых, видовых, меронимических, других структурных и синонимических связях.

Список литературы

1. *Мельчук И.А.* Опыт теории лингвистических моделей "Смысл <-> Текст" // М., 1974.

2. *Сокирко А.В.* Первичный семантический анализ // АОТ :: Технологии :: Первичный семантический анализ: http://www.aot.ru/docs/seman.html (17 октября 2005 г.)

3. *Сокирко А.В.* Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) // Диссертация на соискание ученой степени кандидата технических наук.: М. 2001.

4. *Ножов И.М.* Морфологическая и синтаксическая обработка текста(модели и программы) // Диссертация на соискание ученой степени кандидата технических наук.: М. 2003.

5. *Vapnik V.* Statistical Learning Theory // Wiley, 1998.

6. *Burges C.J.C.* A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery 2, 121-167, 1998.

7. *Баранов О.С.* Идеографический словарь русского языка // М.: ЭТС 1996.