Автоматическое разрешение анафоры в русском языке

Гуреенкова Ольга
Новосибирский Государственный Университет
ООО «Экспасофт»

Навигация

1. Определения
2. Постановка задачи
3. Особенности данных
4. Признаки
5. Алгоритм
6. Результаты
7. Схема работы алгоритма

1 Определения

Кореференция

Языковое явление, при котором два или более выражения в тексте относятся к одному и тому же объекту внеязыковой действительности (референту).

Анафора

Частный случай кореференции. Кореференция, при которой определяющая сущность в тексте предшествует остальным кореферентным сущностям. Чаще всего «остальными сущностями» являются **местоимения**.

Антецедент

Существительное, предшествующее анафорическому местоимению.

Анафор

Анафорическое местоимение, относящееся к антецеденту.

2 Постановка задачи

Дана матрица объектов-признаков X: $\{x_1, \ldots, x_n\}$, где каждый объект является парой антецедент-анафора. Матрице сопоставлен вектор ответов y, компоненты которого принимают следующие значения:

$$y_i = \begin{cases} 1, \text{если объект } \textit{является} \text{ верной парой анафор - антецедент} \\ 0, \text{иначе} \end{cases}$$

Необходимо предсказывать верный ответ на неизвестных объектах. Таким образом, задачу можно рассматривать как задачу **бинарной классификации**.

Ограничения

1. Рассматриваем только **личные местоимения**.
2. Оценка качества алгоритма проводится только для тех случаев, где расстояние от антецедента до анафора **не превышает 3 предложения**, включая предложение с антецедентом.

3 Особенности данных

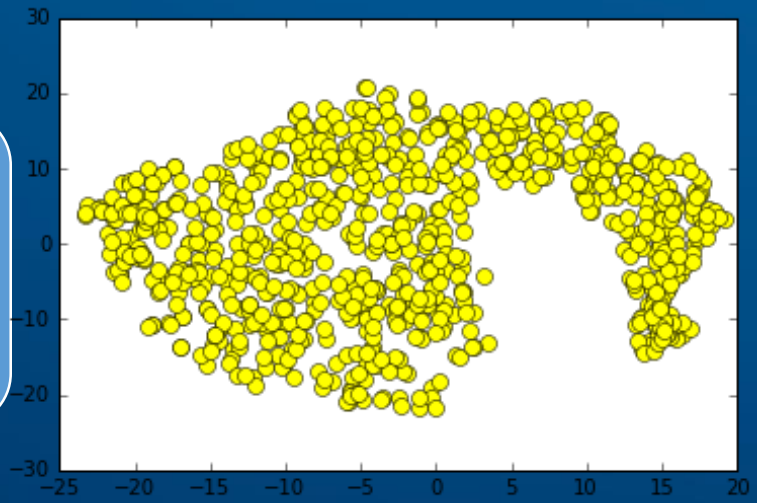
Крайне несбалансированная выборка (в реальности анафорических пар существует намного меньше, чем неанафорических). 10 % верных анафор.

2684 текста для обучения с портала mvd.ru, 262804 объектов-пар на обучении.

1000 текстов для теста, около 165000 объектов-пар.

Рамка в 3 предложения, иначе – еще более несбалансированная выборка.

Выборка линейно неразделима, а при визуализации методом tNSE кажется, что верные анафоры располагаются в пространстве почти хаотично.



4 Признаки

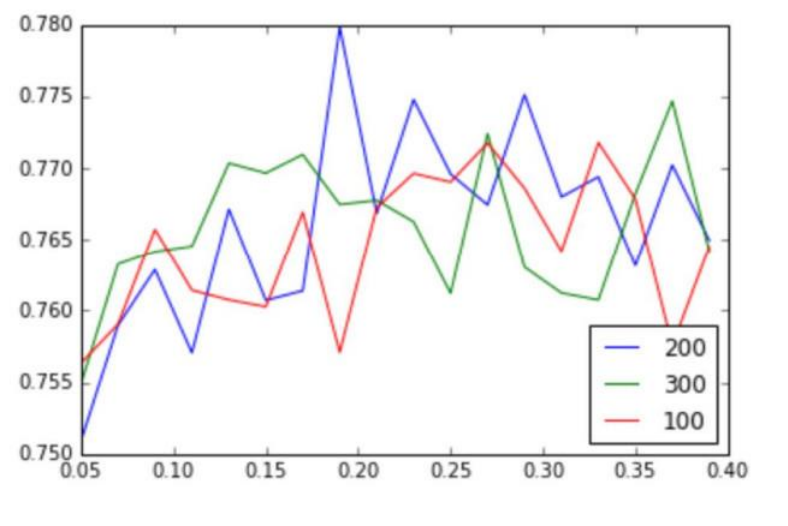
Описание группы признаков	Источники признаков	Кол-во признаков
1. Бинарные категориальные признаки, основанные на относительно сложных правилах, например, «сущность, обозначенная антецедентом, является человеком», «анафора и антецедент согласуются по лицу, числу и падежу».	правила, mystem	7
2. Небинарные категориальные признаки, основанные на простых грамматических характеристиках анафоры и антецедента (часть речи, число, род, падеж, одушевленность, тип анафорического местоимения, синтаксические отношения)	mystem, Syntaxnet	82
3. Вещественные признаки, полученные на основании векторов Word2Vec для синтаксических контекстов. Сюда вошли всевозможные расстояния между векторами контекстов антецедента и анафоры, а также расстояния между собственными векторами анафоры и антецедента до средних векторов их контекстов.	Word2Vec, SyntaxNet	28
4. Вещественные признаки, полученные в результате подсчета различных линейных (т. е. не синтаксических) расстояний, например, расстояния в словах, в предложениях, в существительных, в глаголах между анафорой и антецедентом.	правила, mystem	13
5. Транспонированные векторы, полученные на основе tf-idf векторизации морфологических и синтаксических тэгов, взятых для анафоры и антецедента по трем направлениям синтаксического контекста: дочерние узлы синтаксического дерева, родительский узел, а также узлы, находящиеся на том же уровне, что анафора или антецедент.	SyntaxNet	110

5 Алгоритм

В основе получившейся модели лежат 2 алгоритма машинного обучения:

1. **Balance Cascade** для преобразования выборки с целью балансировки классов (после применения алгоритма доля верных анафорических пар стала составлять 25%).
2. **Extra Trees** в качестве финального классификатора со следующими параметрами:
   - Количество деревьев: 200
   - Доля признаков: 0.23
   - Метод взвешивания: balanced

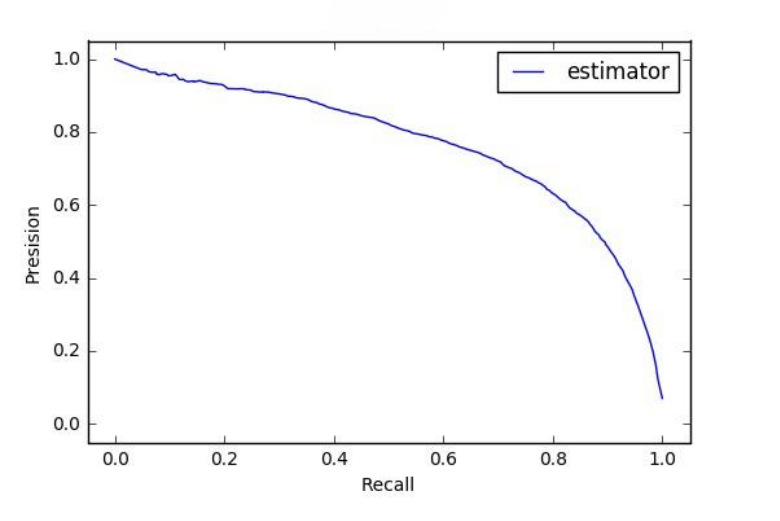
График подбора параметров по сетке (**по горизонтали** – фракция признаков, **по вертикали** – f-мера, **цветом** – количество деревьев)



6 Результаты

F-мера=0.71

Precision	Recall	Threshold
0.657	0.779	0.280
0.660	0.775	0.285



Пример работы программы:

Кировский районный суд Казани признал активиста «Молодой гвардии Единой России» Эдуарда Салахутдинова и ЕГО подельника Тимура Ильина виновными в вымогательстве денег у владельцев подпольных игорных заведений. Об этом сообщает «Интерфакс». Суд счел установленным, что Салахутдинов и Ильин вымогали у предпринимателей деньги, угрожая передать компрометирующую информацию в правоохранительные органы и СМИ. В общей сложности, по версии следствия, ОНИ вымогали более трех миллионов рублей.