Референциальный выбор как многофакторный вероятностный процесс[1]

Referential choice as a multi-factor probabilistic process

Кибрик А. А. (aakibrik@gmail.com), Институт языкознания РАН

Добров Г. Б. (wslc@rambler.ru), МГУ им. М. В. Ломоносова

Залманов Д. А. (dm.zalmanov@gmail.com)

Линник А. С. (skylinnik@gmail.com), **Лукашевич Н. В.** (louk@mail.cir.ru), МГУ им. М. В. Ломоносова

Один из важнейших процессов, участвующих в порождении дискурса — референциальный выбор, то есть выбор языкового выражения при упоминании лица или объекта. Референциальный выбор зависит от большого числа одновременно действующих дискурсивных факторов. В докладе предлагается модель, основанная на методах машинного обучения и описывающая референциальный выбор в аннотированном корпусе английских текстов.

1. Вводные замечания

При порождении дискурса говорящие/пишущие постоянно сталкиваются с необходимостью упоминать те или иные лица или объекты, то есть осуществлять референцию. В естественном дискурсе примерно каждое третье слово так или иначе связано с референцией. Способность адекватно использовать референциальные выражения — одна из наиболее важных языковых способностей. В данной работе предпринимается попытка смоделировать референциальные процессы, происходящие при порождении дискурса.

2. Референциальный выбор

Среди типов референции наиболее частотным типом является конкретная определенная референция. В естественном дискурсе постоянно встречаются повторные и многократные упоминания конкретных определенных референтов. В таких случаях говорящий может использовать не только полные референциальные выражения (имена нарицательные, имена собственные, именные группы с модификаторми), но и редуцированные референциальные выражения, в первую очередь анафорические местоимения. Выбор референциального выражения говорящим для конкретного референта — это **референциальный выбор**.

Рассмотрим отрывок естественного письменного дискурса, иллюстрирующий референциальный выбор.

(86) Tandy said consumer electronics sales at its Radio Shack stores have been slow, partly because a lack of hot, new products. Radio Shack continues to be lackluster, said Dennis Telzrow, analyst with Eppler, Guerin Turner in Dallas. He said Tandy has done a decent job increasing sales by manufacturing computers for others and expanding sales of its Grid Systems Corp. subsidiary, which sells computers to bigger businesses, but it 's not enough to offset the problems at Radio Shack. Sales at Radio Shack stores open more than a year grew only 2 % in the

[1] Данное исследование поддержано грантом № 09-06-00 390 Российского фонда фундаментальных исследований.

Dialog'2010.indb 173 11.05.2010 16:57:14

quarter from a year earlier, he said. As a result, Mr. Telzrow said he cut his fiscal 1990 per-share earnings estimate for Tandy to $ 4,05 from $ 4,2.

В этом отрывке статьи из Wall Street Journal повторно упоминаются три референта: компания Tandy Corp. (пять раз), магазины Radio Shack (четыре раза) и лицо по имени Dennis Telzrow (шесть раз). Из пяти упоминаний Tandy три осуществляются при помощи полных ИГ (конкретно, при помощи имени собственного), а два — при помощи местоимения *it*. Лицо Dennis Telzrow упоминается дважды посредством полных ИГ и четырежды посредством местоимения *he*.

От чего зависит референциальный выбор? Этому вопросу посвящена огромная литература, обозреть которую не представляется возможным; но см., например, Givón 1983, Fox 1987, Chafe 1994, Arnold 2008. Существует целый ряд работ, в которых авторы пытались выделить тот или иной единичный фактор как объясняющий референциальный выбор. Практика, однако, показала, что никакой единичный фактор не в состоянии описать все случаи референциального выбора. В работах Kibrik 1996, 1999 была предложена следующая модель. Референциальный выбор непосредственно зависит от статуса референта в когнитивной системе говорящего в данный момент. Когнитивный компонент, отвечающий за референциальный выбор — рабочая память. Если референт высоко активирован в рабочей памяти говорящего, то используется редуцированное референциальное средство. Напротив, если референт характеризуется низкой степенью активации, то используется полная ИГ. Имеются также промежуточные уровни активации, при которых возможно использование и полных, и редуцированных референциальных средств.

Уровень активации референта, в свою очередь, зависит от ряда характеристик референта и дискурсивного контекста. Эти характеристики именуются факторами активации. К числу важнейших факторов активации относятся: одушевленность референта; значимость референта в дискурсе (протагонизм); расстояние до предшествующего упоминания референта (антецедента), измеряемое в клаузах; наличие/отсутствие дискурсивной границы (например, границы абзаца в письменном тексте) между данной точкой и антецедентом; роль антецедента в своей клаузе (подлежащее, дополнение…) и некоторые другие.

Каждый фактор активации принимает в конкретном случае конкретное значение, которое вносит определенный вклад в интегральный коэффициент активации (КА). КА, таким образом, представляет собой равнодействующую всех факторов и непосредственно определяет референциальный выбор. Референциальный выбор зависит и от некоторых других компонентов, в частности от фильтра референциального конфликта (неоднозначности); подробнее см. указанные выше работы.

3. Арифметическая и нейросетевая модели референциального выбора

Общий подход, описанный в предыдущем разделе, был ранее реализован в виде двух математических моделей. Первая из этих моделей, которую можно назвать арифметической, была описана в работах Kibrik 1996, 1999 применительно к русскому и английскому нарративному дискурсу, соответственно. В этой модели КА варьировал примерно в диапазоне от 0 до 1, а числовые веса значений факторов были подобраны соответственно. Эти веса представляли собой десятичные дроби (от −0,4 до 0,7), а их взаимодействие было смоделировано как простое сложение. Диапазон КА был разделен на несколько интервалов. Так, согласно модели английской референции, при КА менее 0,3 непременно используется полная ИГ, при КА более 1,0 — непременно местоимение, и есть три промежуточных интервала, в которых и полные ИГ, и местоимения возможны, но обладают разной степенью предпочтительности.

В работах Grüning and Kibrik 2003, 2005 была предпринята попытка построить более математически адекватную модель, в которой вклад отдельных факторов в референциальный выбор определялся бы автоматически, а взаимодействие между факторами могло бы быть нелинейным. Эта модель была основана на методе нейросетей — одном из широко известных алгоритмов машинного обучения. Эта работа показала, что алгоритмы машинного обучения, подбирающие оптимальный набор параметров, влияющих на конечных выбор, в принципе пригодны для моделирования такого многофакторного процесса, как референциальный выбор.

Следует отметить, что в исследовании по нейросетевому моделированию референциального выбора общий когнитивный подход к референциальному выбору был существенно редуцирован. В отличие от арифметической модели, в которой имеется интегрирующий когнитивный компонент (коэффициент активации), в нейросетевой модели набор значений факторов непосредственно отображается на референциальный выбор. Это является недостатком данной модели, поэтому в работе Grüning and Kibrik 2005 была поставлена задача восстановить когнитивную адекватность в дальнейших исследованиях, основанных на методах машинного обучения.

Во всех упомянутых работах исследовались небольшие образцы дискурса, насчитывающие 100–150 референциальных выражений. Моделирование при помощи алгоритмов машинного обучения требует значительно больших объемов данных. В связи с этим была поставлена задача создания большого референциального корпуса.

4. Корпус для исследований по моделированию референциального выбора

Согласно результатам работ Kibrik 1996, 1999, в число факторов активации входит фактор **риторического расстояния** от текущей точки дискурса до антецедента. Понятие риторического расстояния (RhetD) основано на теории риторической структуры (Mann and Thompson 1988). Эта теория описывает иерархическую смысловую организацию дискурса. Каждая элементарная дискурсивная единица (чаще всего — клауза) является узлом риторической сети, терминальные узлы объединяются в группы по близости, а между узлами (как терминальными, так и внутренне сложными) устанавливаются отношения — симметричные (такие как последовательность или конъюнкция) или асимметричные (такие как причина, условие, уступка и т. д.). Как было показано в работе Fox 1987, близость к антецеденту по иерархической структуре дискурса не менее важна для референции, чем линейная близость. На основе этой идеи в Kibrik 1996 было предложено измерение RhetD, то есть расстояние от клаузы, в которой происходит референциальный выбор, до клаузы антецедента, подсчитанное в числе шагов по риторической сети. Эмпирические исследования показали, что этот фактор является наиболее мощным фактором референциального выбора. См. Kibrik and Krasavina 2005 о некоторых дискуссионных вопросах, связанных с подсчетом RhetD.

Фактор RhetD является одновременно наиболее «дорогим», поскольку для идентификации его значений для каждого референциального выражения необходима полная разметка риторической структуры дискурса — трудо- и времяемкое дело. В связи с этим в качестве корпуса для исследования референциального выбора крайне желательно было использовать корпус, в котором разметка по риторической структуре уже произведена. На время начала данного проекта (середина 2000-х гг.) такой корпус имелся ровно один: англоязычный корпус RST Discourse Treebank, созданный коллективом под руководством Д. Марку (http://www.isi.edu/~marcu/discourse/Corpora.html), см. Carlson et al. 2003.

Этот корпус включает 385 статей экономической или политической тематики из газеты Wall Street Journal, в этих статьях содержатся 176 383 словоупотреблений и 21 789 элементарных дискурсивных единиц. Пример фрагмента текста из корпуса был приведен выше (пример (1), текст №1374). На рис. 1 показан пример риторического графа, соответствующего фрагменту другого текста (№1315). В этом фрагменте можно видеть как симметричные отношения (Contrast, List), так и большое число асимметричных.

Тексты, входящие в RST Discourse Treebank, составили основу нового корпуса, в котором была осуществлена аннотация по целому ряду потенциальных факторов активации.

5. Референциальная аннотация

Референциальная разметка была произведена при помощи программы MMAX-2, специально созданной группой немецких компьютерных лингвистов для этих целей; см. http://mmax2.sourceforge.net/. Разметка в MMAX-2 осуществляется при помощи так называемой аннотационной схемы, включающей набор размечаемых параметров (факторов). Эта схема была составлена О. Н. Красавиной и К. Чиаркосом (Krasavina and Chiarcos 2007). Она основана на языке XML и устроена по принципу stand-off annotation (файлы с аннотацией отдельны от самих текстов). Схема включает три последовательных компонента:

- выбор аннотируемых элементов
- разметка анафорических связей между ними
- разметка дополнительных признаков аннотируемых элементов.

Аннотируемый, или маркируемый, элемент (*markable*), далее **маркабула**, **—** это составляющая текста, которая может быть референциальным выражением. В качестве маркабул выступают либо

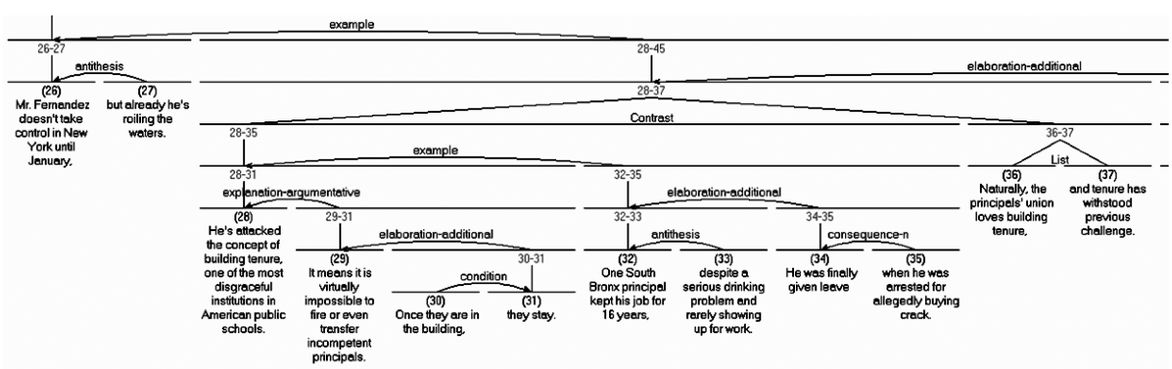

Рис. 1. Фрагмент риторического графа из корпуса RST Discourse Treebank

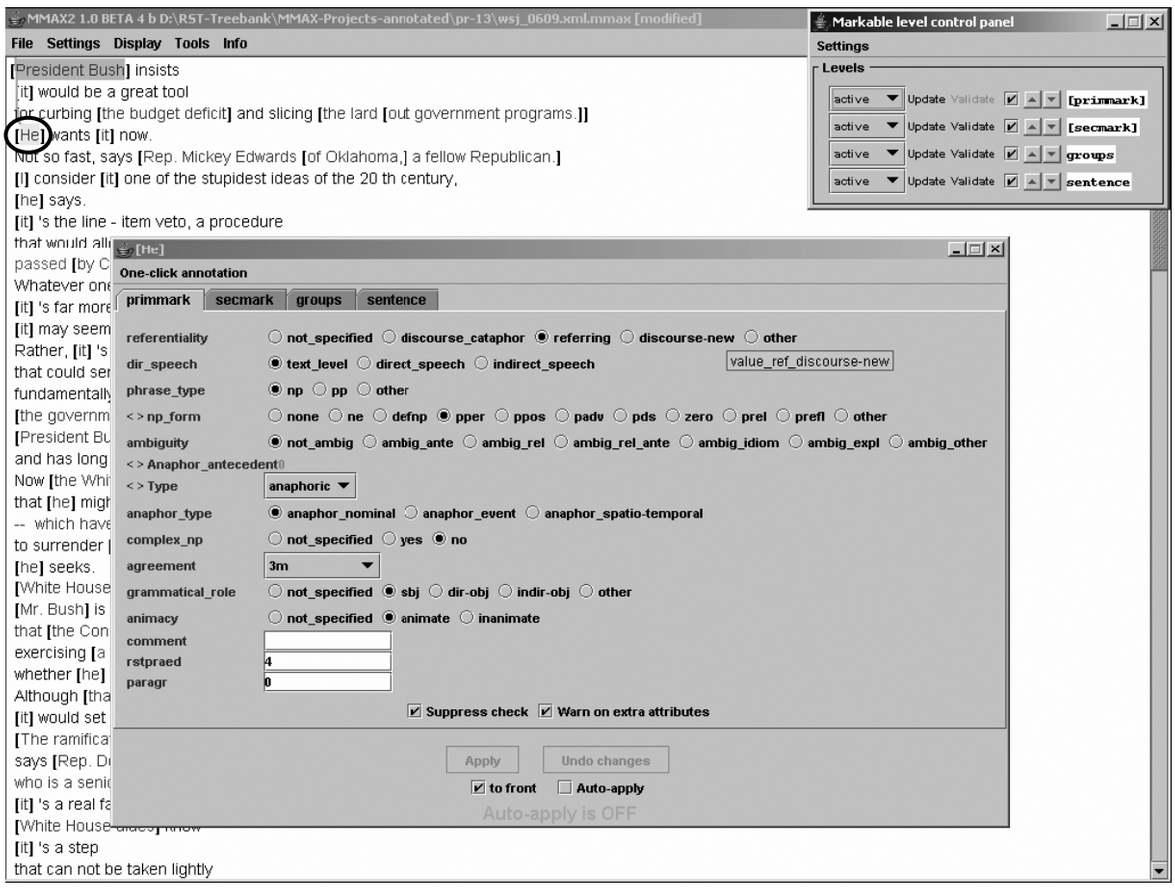

Рис. 2. Окна программы MMAX-2

именные, либо предложные группы. На рис. 2 можно видеть текстовое окно программы MMAX-2, в котором маркабулы обозначены при помощи квадратных скобок (текст №0609 корпуса). В рамках данной аннотационной схемы выделяется два типа маркабул:

- основные маркабулы, которые могут использоваться в анафорической функции; к ним принадлежат:
  - определенные, указательные и посессивные ИГ
  - имена собственные
  - личные и указательные местоимения
- второстепенные маркабулы, которые соответствуют дискурсивно-новым референтам и не могут употребляться в анафорической функции, однако могут являться антецедентами анафорических выражений; к ним относятся:
  - неопределенные ИГ (с неопределенным артиклем и без артикля)
  - элементы, которые не были классифицированы как основные маркабулы, но являются антецедентами основных маркабул.

Между выделенными маркабулами устанавливаются **отношения кореферентности**. Отношение кореферентности соединяет каждое непервое упоминание *n* некоторого референта с предшествующим ему упоминанием *n–1*, то есть антецедентом. Так, от местоимения [he] (выделено на рис. 2 при помощи овала) проходит линия к антецеденту — именной группе [President Bush].

Наконец, последний компонент схемы предполагает разметку **признаков** каждой маркабулы. Основные и второстепенные маркабулы имеют разные, но пересекающиеся наборы признаков. На рис. 2 в центре можно видеть рабочее окно программы MMAX-2, в котором показаны значения признаков для вышеупомянутой маркабулы [he]. Признаки соответствуют потенциальным факторам активации, таким как грамматическая роль или одушевленность; см. ниже.

Аннотация корпуса осуществлялась в основном вручную студентами-практикантами ОТиПЛ МГУ, однако перед этим тексты прошли автоматическую обработку. Так, автоматически выделялись маркабулы, которые легко обнаружить по формальным признакам: местоимения; практически все имена собственные; именные группы с артиклями. (Эта работа в значительной мере была проведена студенткой ОТиПЛ МГУ А. Антоновой при помощи частеречного парсера компании Cognitive Technologies; см. Антонова 2004.) На уровне анафорических связей автоматическая обработка была минимальной: связи устанавливались только для тех пар «анафор —

антецедент», где анафором являлось личное или притяжательное местоимение. Также автоматически присваивались значения некоторых признаков там, где это можно было сделать, опираясь на одно лишь лексическое наполнение маркабул. Частично автоматизированным был также процесс проверки конечных версий разметки.

В настоящее время аннотирование корпуса по референции осуществлено примерно на 64 %. Описанные ниже результаты основаны на 247 текстах, содержащих около 110 000 словоупотреблений. В этих текстах размечено 26 024 маркабул, в том числе 7097 имен собственных, 8560 определенных дескрипций, 1797 личных местоимений 3 лица. (К числу других маркабул относятся местоимения 1/2 лица, притяжательные местоимения, указательные местоимения, неопределенные ИГ и некоторые другие категории.) В данном сегменте корпуса на настоящий момент получено 3502 надежных пар «анафор — антецедент». Входящие в них анафоры разбиваются на классы так: имена собственные — 1541 (44 %), определенные дескрипции — 976 (28 %), местоимения — 985 (28 %).

Две версии корпуса — риторическая и референциальная — образуют в совокупности единый продукт, который мы называем «корпус RefRhet». Некоторые предварительные сведения о раннем варианте этого корпуса были представлены в работе Красавина 2006.

6. Факторы референциального выбора

Из референциальной и риторической разметок, а также из структуры текста как такового для каждого референциального выражения извлекаются значения следующих потенциальных факторов активации (в скобках указаны названия признаков, принятые в аннотационной схеме):

Признаки референта:
- первое/непервое упоминание в дискурсе (referentiality)
- одушевленность (animacy)
- протагонизм

Признаки антецедента:
- Входит ли в состав прямой речи (dir_speech)
- Тип синтаксической группы (phrase_type)
- Грамматическая роль (gramm_role)
- Референциальная форма (np_form, def_np_form)

Признаки анафора:
- Входит ли в состав прямой речи (dir_speech)
- Тип синтаксической группы (phrase_type)
- Грамматическая роль (gramm_role)

Расстояния между анафором и антецедентом:
- Расстояние в словах
- Расстояние в маркабулах между анафором и антецедентом
- Линейное расстояние в клаузах
- Риторическое расстояние в элементарных дискурсивных единицах

7. Методы машинного обучения при моделировании референциального выбора

Моделирование референциального выбора было осуществлено при помощи система Weka (http://www.cs.waikato.ac.nz/ml/weka/, см. Hall et al. 2009), в которой реализовано множество алгоритмов машинного обучения и автоматизирована оценка их качества. Для экспериментов были выбраны несколько алгоритмов машинного обучения, относящихся к разным типам: логические алгоритмы классификации и логистическая регрессия. При выборе алгоритмов деревьев решений и решающих правил (логических алгоритмов) мы руководствовались интерпретируемостью получаемых ими описаний классов в виде понятных человеку конструкций. В качестве таких классификаторов мы выбрали алгоритм деревьев решений C4.5 и алгоритм решающих правил JRip. Выбор логистической регрессии обусловлен двумя соображениями: во-первых, качество работы этого алгоритма несколько превосходит качество работы логических, а во-вторых, логистическая регрессия позволяет получить оценки вероятности принадлежности каждому из классов (см. ниже).

Для контроля качества использовалась процедура скользящего контроля:

1. Обучающее множество делится на 10 частей.
2. Затем классификатор строит решающую функцию по 9 частям из 10.
3. Построенная решающая функция тестируется на оставшейся десятой части.

Процедура повторяется для всех возможных разбиений, а результаты потом усредняются. Критерием выбора наилучшего набора признаков и алгоритма является аккуратность — отношение правильно предсказанных типов референциальных выражений к их общему количеству. Ср. недавнюю работу Greenbacker and McCoy 2009, в которой реализуется сходный подход.

8. Основные результаты

Пары анафор–антецедент, где анафор представлен полной ИГ (именем собственным или определенной дескрипцией), составляют 72 % в данной вы-

борке. Соответственно, классификатор будет иметь смысл, если будет показывать качество выше 72 %. При использовании всех вышеперечисленных признаков с помощью логистической регрессии удалось достичь уровня аккуратности предсказания 86,8 %. Логические алгоритмы показали качество чуть более 85 %. Приведем в качестве примера несколько правил, порожденных алгоритмом JRip:

- *(грамматическая роль антецедента= подлежащее) И (Риторическое расстояние ≤ 1,5) И (Расстояние в словах ≤ 7) => местоимение*
- *(Расстояние в словах ≤ 20) И (грамматическая роль антецедента = подлежащее) И (2-я модель протагонизма ≤ 0,117647) И (одушевленный) => местоимение*
- *(одушевленный) И (Расстояние в маркабулах ≥ 2) И (Расстояние в словах ≤ 11) => местоимение*

Кроме того, было осуществлено моделирование троичного референциального выбора: определенная дескрипция vs. имя собственное vs. местоимение. Этот выбор предсказать оказалось значительное труднее, показатели аккуратности снизились следующим образом: логистическая регрессия показала результат 76 %, логические алгоритмы — 74 %. Такое снижение уровня аккуратности не является удивительным, поскольку набор факторов активации изначально ориентирован на базовый референциальный выбор между полной и редуцированной ИГ. Для того, чтобы достаточно хорошо предсказывать различие между именем собственным и определенной дескрипцией, нужны дополнительные факторы. Следует отметить, что характер этих дополнительных факторов требует дальнейших исследований. В современных работах по референции вопрос о том, как говорящие выбирают между разновидностями полных ИГ, является малоизученным, см. Линник 2009.

9. Вероятностный характер референциального выбора

Как было отмечено выше, существует значительное количество референциальных выражений, которые выбираются не детерминированным, а вероятностным образом. В работе Kibrik 1999 данная проблема была подвергнута детальному анализу. При помощи многоэтапной экспериментальной процедуры, включающей опрос значительного числа носителей английского языка, каждому фактическому референциальному выражению было поставлено в соответствие «потенциальное референциальное выражение» — оценка того, какие референциальные средства в принципе могли бы быть использованы в данной точке дискурса. Имеется пять типов таких потенциальных референциальных выражений. В арифметической модели референциального выбора каждому из этих типов соответствует определенный интервал значений коэффициента активации.

В таблице 1 показаны 6 возможных соответствий между пятью потенциальным и двумя фактическими типами референциальных выражений, а также количественные данные по небольшому набору данных, на котором была основана работа Kibrik 1999.

Как экстраполировать эти результаты на корпус RefRhet? Учитывая, что статьи Wall Street Journal представляют собой хорошо отредактированные тексты, можно предположить, что категории (2) и (4) не вызывают проблем, и в таких случаях использованы оптимальные референциальные средства — полные ИГ и местоимения, соответственно. Однако остается категория (3) — случаи, в которых местоимение и полная ИГ могут быть использованы с равной вероятностью. В таких случаях предсказать фактический референциальный выбор практически невозможно. Даже идеальный алгоритм должен предсказывать референциальный выбор с некоторым количеством ошибок. Можно предположить, что целевой достижимый уровень аккуратности вряд ли превысит 90 %. Напомним, что на данный момент алгоритмы машинного обучения предсказывают референциальный выбор с уровнем аккуратности около 85 %, то есть отклоняются от фактического референциального выбора в 15 % случаев. Если вышеприведенные рассуждения верны, то достигнутый результат весьма близок к «потолку», в принципе возможному для такого вероятностного процесса, как референциальный выбор. Вопрос, однако, в том, совпадают ли те случаи, когда алгоритмы дают отличный от фактов прогноз, с теми

Таблица 1. Потенциальные и фактические референциальные выражения (по Kibrik 1999).

Потенциальные референциальные выражения	(1) Только полная ИГ (19 %)	(2) Полная ИГ, ?местоимение (21 %)	(3) Местоимение или полная ИГ (28 %)*	(4) Местоимение, ?полная ИГ (23 %)	(5) Только местоимение (9 %)
Фактические референциальные выражения	Полная ИГ (49 %)			Местоимение (51 %)	

* В том числе 9 % фактических полных ИГ и 19 % фактических местоимений.

случаями, когда предсказание действительно невозможно. Ответ на этот вопрос требует дальнейших исследований.

Один из алгоритмов машинного обучения, а именно логистическая регрессия, обладает свойствами, которые позволяют смоделировать вероятностный характер референциального выбора. Для каждого референциального выражения логистическая регрессия выдает количественное значение, которое можно считать оценкой вероятности использования местоимения в данной точке дискурса. Эта оценка является некоторым аналогом коэффициента активации — интегрального показателя, представляющего собой равнодействующую всех одновременно действующих факторов активации (см. раздел 3).

10. Заключительные замечания

В данной статье описана модель референциального выбора в английском тексте. Работа основана на корпусе RefRhet, включающем несколько сотен текстов, размеченных по риторической структуре и по целому ряду потенциальных факоторов референциального выбора (свойства референта, свойства антецедента, расстояние до антецедента и др.) Был использован ряд алгоритмов машинного обучения, способных предсказывать референциальный выбор (в первую очередь, выбор между полной и редуцированной ИГ) на основе имеющихся признаков. Алгоритмы продемонстрировали аккуратность предсказания референциального выбора в районе 85 %.

Таким образом, референциальный выбор зависит от множества одновременно действующих факторов и в этом смысле представляет собой **многофакторный** процесс. Кроме того, референциальный выбор является **вероятностным** процессом: существует достаточно большое число случаев, в которых говорящий (пишущий) может использовать более чем одну референциальную опцию. Учитывая это обстоятельство, очевидно, что референциальный выбор не может быть предсказан с аккуратностью 100 %. Полученные на данный момент результаты предсказания довольно близки к теоретически возможному максимуму аккуратности.

Описанный в данной работе теоретический и методологический подход может быть применен к широкому кругу других языковых процессов, которые являются многофакторными и вероятностными — например, к выбору порядка слов, к выбору грамматического оформления предиката, к выбору просодических характеристик и т. д.

Литература

1. *Антонова А.* 2004. Алгоритм определения антецедентов анафорических местоимений и автоматическая референциальная разметка корпуса газетных статей Wall Street Journal. Курсовая работа. МГУ им. М. В. Ломоносова.
2. *Красавина О. Н.* 2006. Корпусно-ориентированное исследование референции (принципы аннотации и анализ данных). Дисс. ... кандидата филологических наук. М.: МГУ им. М. В. Ломоносова.
3. *Линник А. С.* 2009. Выбор именной группы в качестве референциального средства в зависимости от риторического расстояния до антецедента и уровня активации референта. Курсовая работа. МГУ им. М. В. Ломоносова.
4. *Arnold, Jennifer.* 2008. Reference Production: Production-internal and addressee-oriented processes // Language and Cognitive Processes 23,4: 495–527.
5. *Carlson, Lynn D., Daniel Marcu and Mary Ellen Okurowski.* 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory // Jan van Kuppevelt and Ronnie Smith (eds.) Current directions in discourse and dialogue. Dordrecht: Kluwer, 85–112.
6. *Chafe, W. L.* 1994. Discourse, consciousness, and time. Chicago: University of Chicago Press.
7. *Fox, B.* 1987. Discourse structure and anaphora in written and conversational English. Cambridge: Cambridge University Press.
8. *Givón, T.* 1983. Topic continuity in discourse: An introduction // T. Givón (Ed.), Topic continuity in discourse: A quantitative cross-language study. Amsterdam: Benjamins, 1–42.
9. *Greenbacker, Charles F., and Kathleen F. McCoy.* 2009. Feature selection for reference generation as informed by psycholinguistic research // Production of referring expressions (PRE-CogSci) 2009: Bridging the gap between computational and empirical approaches to reference. Proceedings of the conference, Amsterdam, July 29, 2009.
10. *Grüning, André, and Andrej A. Kibrik.* 2003. A neural network approach to referential choice // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог-2003. М.: Наука, 260–266.

11. *Grüning, André, and Andrej A. Kibrik.* 2005. Modeling referential choice in discourse: A cognitive calculative approach and a Neural Networks approach // António Branco, Tony McEnery and Ruslan Mitkov (eds.). Anaphora processing: Linguistic, cognitive and computational modelling. Amsterdam: Benjamins, 163–198.
12. *Hall, M., F. Eibe, G. Holmes, B. Pfahringer, P. Reutemann, Ian H. Witten*. Witten. 2009. The WEKA data mining software: An update. SIGKDD Explorations, Volume 11, Issue 1.
13. *Kibrik, A. A.* 1996. Anaphora in Russian narrative discourse: A cognitive calculative account // B. Fox (ed.) Studies in anaphora. Amsterdam: Benjamins, 255–304.
14. *Kibrik, A. A.* 1999. Reference and working memory: Cognitive inferences from discourse observation // Discourse studies in cognitive linguistics. Ed. by K. van Hoek, A. A. Kibrik and L. Noordman. Amsterdam: Benjamins, 29–52.
15. *Kibrik, A. A., and Olga N. Krasavina.* 2005. A corpus study of referential choice: The role of rhetorical structure // Диалог. Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог'2005. Ред. И. М. Кобозева, А. С. Нариньяни, В. П. Селегей. М.: Наука, 561–569.
16. *Krasavina, Olga, and Christian Chiarcos,* Ch. 2007. PoCoS — Potsdam Coreference Scheme // Proceedings of the Linguistic Annotation Workshop (LAW). June 28–29, 2007, Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics, 156–163.
17. *Mann, William C., and Sandra A. Thompson.* 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation // Text 8(3): 243–281.