ОПТИМИЗАЦИЯ МОДЕЛИ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА, ОСНОВАННОЙ НА МАШИННОМ ОБУЧЕНИИ

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

Линник А. С. (skylinnik@gmail.com),
Добров Г. Б. (wslc@rambler.ru),
Худякова М. В. (mariya.kh@gmail.com)
МГУ им. М. В. Ломоносова

В статье обсуждаются различные способы оптимизации системы, моделирующей референциальный выбор (РВ) на основе аннотированного корпуса с использованием машинного обучения. Аннотационная схема, использовавшаяся в наших более ранних исследованиях, была улучшена и расширена. На следующем этапе был имплементирован более «дешевый» набор параметров с целью сокращения времени обработки и трудозатратности аннотации. Наши результаты свидетельствуют о том, что, несмотря на возможность исключения наиболее «дорогих» факторов при моделировании РВ, лучшая аккуратность предсказания достижима только при использовании максимального количества доступной информации. Жанровая принадлежность текстов была введена в систему в качестве одного из параметров и послужила повышению показателя аккуратности. И наконец, была запущена серия психолингвистических экспериментов по изучению категоричности выбора, совершаемого говорящими/пишущими. Первые полученные нами результаты оказались многообещающими: они показали, что в случаях, в которых системе не удается дать однозначное предсказание, согласно человеческой оценке, возможно с равной вероятность использование более одного референциального средства.

Ключевые слова: референциальный выбор, риторическая структура, референция, компьютерное моделирование, машинное обучение, аннотированный корпус

Введение

Целостность и состоятельность производимого дискурса (текста/речи) напрямую связана с повторным (многократным) упоминанием одних и тех же сущностей — референцией. В зависимости от того, какое референциальное средство использует говорящий — имя собственное, местоимение или дескриптивную именную группу (см. пример (1)), — может меняться как смысл высказывания в целом, так и степень доступности его смысла для адресата.

(1) *Mr. Akerson said MCI recorded "another solid cash positive quarter," its fourth in a row, but declined to comment on whether the company is considering a dividend or is planning any acquisition. "The current quarter," he said, "looks fine".*

Настоящее исследование посвящено изучению механизмов, используемых говорящим при выборе способа отсылки к референту, называемом *референциальным выбором* (РВ). Известно, что за этот процесс ответственна кратковременная память, в частности, *активация* того или иного референта в сознании говорящего (Kibrik 1996, 1999, 2011). Именно эта взаимосвязь лежит в основе главного принципа референциального выбора в модели, предложенной и разработанной А. А. Кибриком. Он заключается в том, что при низкой активации референта говорящий (или пишущий) использует наиболее лексически полные выражения, в то время как для высоко активированных референтов обычно используются более краткие формы. Данное исследование проводилось на материале английского языка (подробнее в разделе «Корпус RefRhet»), в котором основным редуцированным референциальным средством является местоимение (нули используются в крайне ограниченном количестве конструкций, а указательные местоимения чаще всего используются для ситуативной референции). Полными же референциальными средствами считаются имена собственные и определенные дескрипции.

Моделирование процесса референциального выбора на данный момент является одной из основных задач в области синтеза естественного языка. При построении модели РВ используются два варианта данной задачи: (а) выбор между редуцированным и полным референциальным средством, и более сложный трехчастный (б) выбор между местоимением, определенной дескрипцией и именем собственным. Кроме высокой точности предсказания средства отсылки к референту, необходимо наличие возможности имплементации его в общую модель порождения дискурса. В настоящем исследовании референциальный выбор моделируется при помощи машинного обучения на корпусе текстов большого объема.

Корпус RefRhet

Корпус RefRhet был создан на базе известного англоязычного корпуса RST Discourse Treebank (http://www.isi.edu/~marcu/discourse/Corpora.html), состоящего из тестов Wall Street Journal и получившего под руководством Д. Марку (Carlson et al. 2003) разметку в соответствии с Теорией Риторической структуры, далее — ТРС (Mann & Thompson,1987; Taboada & Mann, 2006). Результаты более ранних исследований (Fox, 1987; Hobbs, 1985; Kehler, 2002; Kibrik, 1996) указывают на наличие зависимости между дискурсивной структурой и референциальным выбором. Наиболее логично среди различных моделей семантико-дискурсивной структуры текста (например, Hobbs, 1985; Joshi, Prasad & Miltsakaki, 2006; Miltsakaki, Prasad, Joshi & Webber, 2004; Polanyi, 1985; Wolf & Gibson, 2003) иерархическую организацию дискурса описывает ТРС. Выбор RST Discourse Treebank обоснован наличием в нем аннотации иерархической

структуры, которая требует специальной подготовки аннотаторов и значительных затрат времени. В корпусе содержится 176 383 словоупотребления и 21 789 элементарных дискурсивных единиц (ЭДЕ). Риторическое расстояние, которое представляет собой длину пути между фрагментами текста по построенной риторической сети, считается важным фактором при референциальном выборе (Fox 1997, Kibrik 1996), поскольку позволяет учесть связь между фрагментами текста, далекими друг от друга по линейному расстоянию, но близкими по структуре изложения.

Корпус RefRhet был создан путем нанесения на RST Discourse Treebank референциальной разметки. Между всеми референциальными выражениями (markable, далее маркабула), устанавливаются отношения кореферентности — каждое непервое упоминание референта (анафор) связывается с предшествующим ему упоминанием (антецедентом). Разметка осуществлялась при помощи так называемой аннотационной схемы (подробно об исходной схеме разметки RefRhet см. Krasavina and Chiarcos 2007) — набора признаков (грамматическая роль, одушевленность др.), приписываемых каждой маркабуле, которые могут оказывать влияние на РВ, с использованием программного инструмента MMAX-2 (http://mmax2.sourceforge.net/). Каждый текст был размечен двумя независимыми аннтаторами. Мы применяли различные алгоритмы машинного обучения к данному корпусу, в результате чего была достигнута аккуратности предсказания референциальной формы до 90% для двухчастной задачи и около 80% — для трехчастной. Подробные отчеты содержатся в (Kibrik et al. 2010) и (Loukachevitch et al. 2011).

В целях оптимизации процесса аннотации было принято решение модифицировать аннотационную схему, что и было произведено к февралю 2012 года. Подкорпус RefRhet, названный RefRhet2, был аннотирован по новой схеме, которая была названа MoRA (Moscow Referential Annotation). В настоящей работе представлены результаты обработки RefRhet2 , количественные характеристики которого указаны в Таблице 1.

Табл. 1. Количественные характеристики корпуса RefRhet2

Параметр	*Количество в корпусе*
Количество слов	45 016
Количество ЭДЕ	5497
Количество маркабул (референциальных выражений)	11 461
Количество пар «анафор-антецедент»	3692
Количество референциальных цепочек (последовательностей упоминаний одного и того же референта)	1511

Методы машинного обучения

Для экспериментов были выбраны несколько алгоритмов машинного обучения, относящихся к разным типам: логические алгоритмы классификации

(деревья решений C4.5, алгоритм решающих правил JRip), результаты работы которых легко интерпретировать, а также логистическая регрессия, позволяющая получить оценки вероятности принадлежности каждому из классов и качество работы которой превосходит качество работы логических алгоритмов. Нами также были использованы композиции классификаторов: баггинг (bagging) и бустинг (boosting). Подробнее работа данных алгоритмов уже обсуждалась нами ранее (Loukachevitch et al. 2011). Критерием выбора наилучшего набора признаков и алгоритма является *аккуратность* — отношение правильно предсказанных типов референциальных выражений к их общему количеству.

Результаты обработки корпуса RefRhet2 приведены в Табл. 2.

Табл. 2. Результаты работы алгоритмов машинного обучения на корпусе RefRhet2

Алгоритм	**Аккуратность классификации для двухклассовой задачи**	**Аккуратность классификации для трехклассовой задачи**
Решающие правила	85.9%	75.7%
Деревья решений	87.1%	76.8%
Логистическая регрессия	87.9%	77.6%
Баггинг	88%	78.6%
Бустинг	88.7%	78.7%

Результаты работы некоторых алгоритмов (напр., лог. регрессии и деревьев решений) улучшились с изменением аннотационной схемы по сравнению с результатами, опубликованными в (Loukachevitch et al., 2011), однако по другим алгоритмам показатели оказались несущественно ниже, чем в прошлых исследованиях. Это связано с тем, что новая аннотационная схема включает в себя «сложные» случаи, такие как групповая аннотация и неопределенные дескрипции. Поскольку количество маркабул увеличилось почти втрое, работу систем машинного обучения можно оценить как стабильную.

Факторы, влияющие на референциальный выбор

Наш подход к моделированию референциального выбора основан на представлении о многофакторности этого процесса (Kibrik, 1996, 1999, 2011). Нашей задачей был не только поиск и изучение факторов, влияющих на РВ, но также изучение их индивидуального вклада в аккуратность предсказания РВ, с целью уменьшения их количества до необходимого минимума (чтобы снизить трудозатратность процесса аннотации). Полный набор факторов включает себя различные характеристики как анафора

и антецедента, так и самого референта, а также некоторые общие дискурсивные характеристики.

- Признаки референта: одушевленность, протагонизм (значимость референта в дискурсе), род и число;
- Признаки антецедента: входит ли в состав прямой речи, тип синтаксической группы, грамматическая роль, референциальная форма, длина антецедента в словах, количество антецедентов в цепочке от текущего места до полной именной группы
- Признаки анафора: первое/непервое упоминание в дискурсе, входит ли в состав прямой речи, тип синтаксической группы, грамматическая роль, число упоминаний референта в цепочке
- Расстояния между анафором и антецедентом: линейное расстояниев словах, линейное расстояние в клаузах, линейное расстояние в предложениях, расстояние в маркабулах, риторическое расстояние в ЭДЕ, расстояние в абзацах.

Список параметров периодически пополняется. Например, одно из последних дополнений к нему связано с исследованием влияния жанрового разнообразия корпуса на РВ. Этот вопрос неоднократно поднимался в предыдущих исследованиях РВ (Biber, Johansson, Leech, Conrad & Finegan, 1999: 235; Fox, 1987; Garrod, 2011; Longo & Todirasu, 2010; Toole, 1996; Strube & Wolters, 2000, De Clercq et al. 2011). Исследователи пришли к выводу, что жанр дискурса влияет на референцию. Несмотря на относительную однородность корпуса RefRhet2, в нем можно выделить три основных жанра — это наиболее часто встречающиеся в корпусе тексты (1) информационные заметки и отчеты (корпоративные и финансовые новости), (2) биографические очерки и (3) аналитические статьи и рецензии. Принадлежность к одному из трех жанров была размечена на подкорпусе и использовалась в качестве одного из параметров машинного обучения. Результатом стало повышение аккуратности около 0.5 %–1 % в зависимости от использованного алгоритма. Вклад этого параметра может показаться несущественным, однако, есть несколько причин включить его в полный набор при тренировке алгоритмов машинного обучения. Трудозатратность разметки жанровой принадлежности не высока, улучшение в аккуратности предсказания, которое она дает, сравнимо с вкладом других отдельно взятых факторов, и, что самое важное, это улучшение нельзя проигнорировать или объяснить прочими характеристиками тестового материала.

Некоторые из параметров, используемых нами при моделировании, делают процесс аннотации трудоемким и времязатратным — к примеру, риторическое расстояние. Наличие разметки риторической структуры в корпусе послужило причиной выбора его для данного исследования, кроме того, в наших предыдущих работах уже обсуждался вклад отдельных факторов в общее значение аккуратности (Loukachevitch et al., 2011). Однако так ли необходима разметка по ТРС для конечного результата? Исключение риторического расстояния из списка факторов дает ухудшение в качестве

предсказания всего на 0.6 % для двухклассовой и 0.1 % для трехклассовой задачи. Результат созвучен общепринятому мнению о том, что риторическое расстояние не влияет на выбор между именем собственным и определенной дескрипцией.

Чтобы оценить необходимость приложения дополнительных усилий на этапе аннотации, мы решили элиминировать из исходного набора параметров те, аннотация которых требует существенно больше усилий. Результаты работы системы (с использованием логистической регрессии) представлены в Таблице 5.

Табл. 3. Сравнение результатов машинного обучения с использование более и менее «дорогого» набора параметров

	Аккуратность для двухклассовой задачи	**Аккуратность для трехклассовой задачи**
Полный набор параметров	87.9 %	77.6 %
«Дешевый» набор параметров	85.2 %	75.5 %

Наши результаты свидетельствуют об ощутимой (хотя и не критической в целом) потере в аккуратности. Это подтверждает, что ни один из факторов не теряет своей значимости, если мы стремимся достичь наилучшего результата.

Деревья решений — взгляд изнутри

Один из наиболее интересных вопросов относится к задаче референциального выбора между местоимением, именем собственным и определенной дескрипцией. Изначально столь высокий показатель аккуратности был не вполне ожидаем, поскольку эксперименты по выявлению признаков, влияющих на различение говорящим имени собственного и определенной дескрипции, дали достаточно скромный результат (Linnik 2010). В настоящем исследовании мы предприняли попытку пойти «обратным путем» и разобрать детально результат работы одного из алгоритмов машинного обучения — деревьев решений. В Таблице 4 представлены некоторые из сгенерированных системой правил с высокой степенью вероятности предсказания.

Основная тенденция, которую можно вывести из этих правил, заключается в наличии эффекта *референциальной инерции* при выборе между вариантами полной именной группы. А именно, если референт упоминается с помощью полной ИГ, будь то дескрипция или имя собственное, при следующем упоминании с высокой вероятностью также будет использована полная и именная группа (см. Krahmer & Theune, 2002).

Табл. 4. Результат обратной обработки деревьев решений (SD — расстояние в предложениях)

Если	То	Количество правильно предсказанных форм	Процент правильных предсказаний из всех случаев, удовлетворяющих условиям из колонки «если»
Референциальная форма антецедента: не дескрипция SD=1 Одушевленность: collective	имя собственное	290	70%
Референциальная форма антецедента: не дескрипция SD>1 Число: единственное	имя собственное	522	86%
Референциальная форма антецедента: дескрипция без определителя Референциальная форма антецедента: не имя собственное	дескрипция	315	90%
Референциальная форма антецедента: дескрипция с определенным артиклем	дескрипция	360	78%
Референциальная форма антецедента: дескрипция с неопределенным артиклем	дескрипция	72	84%

О категоричности референциального выбора

Даже при использовании полного набора факторов наша система не достигает 100% аккуратности предсказания РВ. Новые параметры, которые добавлялись к исходному набору, также не давали существенного улучшения. С целью выяснить, насколько вообще категоричен выбор, осуществляемый говорящими, была запущена серия экспериментов, первые результаты которых будут опубликованы в ближайшее время (Худякова 2012).

В эксперименте испытуемым предлагалось ответить на вопросы к текстам, на которых ошибалась система машинного обучения (в частности, в первом эксперименте этой серии рассматривались случаи, в которых исходно в тексте было употреблено имя собственное, однако система предсказывала появление местоимения). Половине испытуемых были предложены тексты в их исходном виде; в текстах, предложенных второй половине, имена собственные были заменены на соответствующие местоимения. Оговоримся, что машинное

обучение может давать предсказания с различной степенью вероятности — от 0.5 до 1, что означает, что в одних случаях можно с большей уверенностью ожидать то или иное референциальное выражение, в других же — с меньшей.

Результаты эксперимента показали, что в 95% случаев (из тех текстов, которые не вызвали затруднений в исходном варианте) имя собственное и местоимение взаимозаменяемы, то есть, по всей видимости, равновероятны. Это позволяет сделать вывод, что референциальный выбор не категоричен, и этот вывод подтверждается работой нашей системы машинного обучения.

Заключение

Исследования по изучению факторов, влияющих на референциальный выбор, продолжаются. Мы стремились оптимизировать набор используемых в машинном обучении параметров и аннотационную схему без потери в аккуратности предсказания референциального средства. Нам удалось показать, что существует возможность использовать сокращенный список факторов за счет исключения некоторых параметров, сложных в аннотации. Хотя потеря в аккуратности не критична, очевидно, что лучший показатель достигается использованием наиболее полного набора факторов. Мы также продолжаем тестировать различные признаки референта и дискурса в целом — например, его жанра, — с целью улучшить данный показатель. Основная гипотеза, эксперименты по проверке которой уже запущены, заключается в том, что при среднем уровне активации референта референциальный выбор может быть некатегорическим. Полученные на данный момент результаты свидетельствуют о том, что в большинстве случаев, когда используемая нами система машинного обучения ошибается, замена референциального средства не создает говорящим проблем восприятия. Это свидетельствует о том, что в при среднем уровне активации референта выбор той или иной опции действительно может быть равновероятен.

В будущем мы намерены продолжить данную серию экспериментов. Кроме того, по мере того как объем корпуса RefRhet2 будет расти, планируется провести оценку согласия аннотаторов (Artstein & Poesio 2008), что, возможно, также даст нам некоторое представление о том, как влияет качество аннотации на наши результаты, а также о более сложных случаях неоднозначности референциального выбора.

Литература

1. *Artstein, R., & Poesio, M.* (2008), Inter-coder agreement for computational linguistics (surveyarticle). Computational Linguistics, no. 34(4), pp. 555–596.
2. *Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E.* (1999), Longman grammar of spoken and written English. Pearson Education limited, Harlow.
3. *Carlson, L., Marcu, D., & Okurowski, M. E.* (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt &

R. Smith (Eds.), Current directions in discourse and dialogue. Kluwer, Dordrecht, pp. 85–112
4. *Fox, B. A.* (1987), Discourse structure and anaphora in written and conversational English. Cambridge, Cambridge University Press.
5. *Garrod, S. C.* (2011), Referential processing in monologue and dialogue with and without access to real world referents. In E. Gibson & N. J. Pearlmutter (Eds.), The processing and acquisition of reference (pp. 273–294). MIT Press, Cambridge, MA.
6. *De Clercq O., Hoste V., & Hendrickx I.* (2011), Cross-domain dutch coreference resolution. In Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011), pp. 186–193.
7. *Hobbs, J. R.* (1985), On the Coherence and Structure of Discourse (Report No. CSLI-85–37). Center for the Study of Language and Information, Stanford University.
8. *Joshi, A. K., Prasad, R., & Miltsakaki, E.* (2006), Anaphora resolution: Centering theory approach. In K. Brown (Ed.), Encyclopedia of language and linguistics. Elsevier, Oxford, pp. 223–230.
9. *Kehler, A.* (2002). Coherence, reference, and the theory of grammar. CSLI Publications, Stanford.
10. *Kibrik A. A.* (1996) Anaphora in Russian narrative discourse: A cognitive calculative account In B. Fox (ed.) Studies in anaphora. Benjamins, Amsterdam, pp. 255–304.
11. *Kibrik A. A.* (1999) Reference and working memory: Cognitive inferences from discourse observation In van Hoek, K., Kibrik A. A. & L. Noordman (eds.) Discourse studies in cognitive linguistics. Benjamins, Amsterdam, pp. 29–52.
12. *Kibrik, A. A.* (2011). Reference in discourse. Oxford University Press, Oxford.
13. *Kibrik, A. A., Dobrov, G. B., Zalmanov, D. A., Linnik, A. S., & Loukachevitch N. V.* (2010) Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factor probabilistic process]. In A. E. Kibrik (Ed.), Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, Moscow region, pp. 173–181. Эта работа по-русски, неясно почему она дана в транслите.
14. *Khudyakova M. V.* (2012) Akkuratnost' modelirovanoja referencial'nogo vybora: ocenka chitateljami [Accuracy of referential choice modeling: reader evaluation]. To be presented at The Fifth International Conference on Cognitive Science, to be held in Kaliningrad, Russia. Кириллица!
15. *Krahmer, E., & Theune, M.* (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), Information Sharing: Reference and Presupposition in Language Generation and Interpretation , pp. 223–264. CSLI Publications, Stanford.
16. *Krasavina, O. N. & Chiarcos, Ch.* (2007), PoCoS — Potsdam Coreference Scheme. In Proceedings of the Conference of the Association for Computational Linguistics (LAW). Prague, Czech Republic, pp. 156–163.

17. *Longo, L.,& Todirascu, A.* (2010). Genre-based Reference Chains Identification for French. Investigationes Linguisticae, no. 21, pp. 57–75.
18. *Linnik A. S.* (2010) Linguistic support for computational analysis of the corpus of texts, annotated with respect to the referential theory. Graduation thesis at the Moscow State University, Moscow.
19. *Loukachevitch, N. V., Dobrov, G. B., Kibrik, A. A., Khudyakova, M. V., & Linnik, A. S.* (2011). Factors of referential choice: computational modeling. In A. E. Kibrik (Ed.), Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, Moscow region, pp. 518–528.
20. *Mann, W. C. & Thompson, S. A.* (1987), Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8(3), pp. 243–281.
21. *Miltsakaki, E., Prasad, R., Joshi, A. & Webber, B.* (2004), The Penn Discourse Treebank. In Proceedings of LREC 2004, Lisbon, Portugal, pp. 2237–2240.
22. *Strube, M., & Wolters, M.* (2000). A Probabilistic genre-independent model of pronominalization. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, pp. 18–25.
23. *Taboada, M., & Mann, W. C.* (2006), Rhetorical structure theory : Looking back and moving ahead. In Discourse Studies, no. 8, pp. 423–459.
24. *Toole, J.* (1996). The effect of genre on referential choice. In T. Fretheim & J. K. Gundel (Eds.) Reference and referent accessibility. Benjamins, Amsterdam, pp. 263–290.
25. *Wolf, F., & Gibson, E.* (2003), Representing discourse coherence: A corpus-based study. In Computational Linguistics, no. 31(2), pp. 249–287.