АКТИВАЦИЯ РЕФЕРЕНТОВ И ВЕРОЯТНОСТНАЯ ОЦЕНКА РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА (ИССЛЕДОВАНИЕ АНГЛОЯЗЫЧНЫХ ГАЗЕТНЫХ ТЕКСТОВ)

Кудрявцева А. С. (angelina_ku@mail.ru)
МГУ им. М.В. Ломоносова, Москва, Россия

Ключевые слова: референциальный выбор, референциальное выражение, компьютерное моделирование, некатегорический референциальный выбор, корпус WSJ MoRA 2015.

REFERENT ACTIVATION AND PROBABILISTIC EVALUATION OF REFERENTIAL CHOICE: A STUDY OF ENGLISH NEWSPAPER TEXTS

Kudriavtceva A. S. (angelina_ku@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

Machine learning algorithms modeling referential choice between a pronoun and a full noun phrase cannot predict referential choice with 100% accuracy because referential choice is not always fully categorical. To answer the question of whether it is possible to increase accuracy I will examine the errors made by the algorithm of logistic regression. The comparison of probabilities of referential devices with referent's activation cost makes it possible to say that a moderate level of activation correlates with a moderate degree of certainty in the prediction of a pronoun, and that the majority of the incorrectly predicted instances are characterised by the intermediate level of activation.

Key words: referential choice, referential expression, computational modeling, non-categorical referential choice, WSJ MoRA 2015 corpus.

Введение

В устной или письменной речи мы постоянно упоминаем определенные объекты внеязыковой действительности, которые называются референтами. Говорящий для каждого конкретного упоминания выбирает референциальное выражение, то есть то языковое средство, которым он будет закодирован. Данный процесс называется референциальным выбором. Существует два крупных типа референциальных выражений - редуцированные и полные. К редуцированным относят местоимения и нулевые выражения. Полными референциальными выражениями считаются имена собственные и дескрипции. В настоящем исследовании рассматривается референциальный выбор только между полной ИГ и анафорическим местоимением.

В данной работе используется многофакторный количественный подход к референции (Kibrik, 2011), который основан на представлении о том, что референциальный выбор зависит от степени активации референта в рабочей памяти говорящего. Степень активации, в свою очередь, связана с целым рядом факторов, которые определяются свойствами референта, анафора или антецедента и структурой дискурса. Для измерения уровня активации в рамках данной теории вводится такое понятие, как коэффициент активации. Чем больше значение коэффициента активации, тем больше вероятность употребления редуцированного референциального выражения. Референциальный выбор - это не всегда полностью детерминированный и категорический выбор (Kibrik, 1999; Belz & Varges, 2007; van Deemter et al., 2012). Существуют такие позиции, где используются только полные именные группы, а есть такие, где используются только местоимения. Есть промежуточные случаи, где могут употребляться и полные референциальные выражения, и редуцированные. Предполагается, что классифицирующие алгоритмы машинного обучения с меньшей точностью предсказывают форму референциального выражения именно ввиду того, что отклонения возникают именно в тех случаях, где сами говорящие допускают референциальную альтернативу.

Методы машинного обучения лишены когнитивного компонента – они не способны моделировать реальные когнитивные процессы порождения речи. В данной работе предлагается ввести данный компонент с помощью вероятностных характеристик, которые приписываются алгоритмом логистической регрессии. Таким образом, целью моего исследования является проверка гипотезы о том, что активация референтов и вероятностная оценка референциального выбора связаны.

1. Когнитивный многофакторный подход

Коэффициент активации, используемый в когнитивном многофакторном подходе для измерения степени активации, зависит от набора факторов, различающихся по языкам. Вес каждого из факторов суммируется для получения коэффициента активации, и в зависимости от его значения используется редуцированное или полное референциальное средство. Факторы активации, влияющие на выбор референциального выражения, могут быть связаны как с самим референтом, так и с контекстом высказывания. Среди факторов дискурсивной структуры, влияющих на референциальный выбор, центральное место принадлежит различным типам расстояний от анафора до антецедента: линейному расстоянию в клаузах, предложениях, абзацах, а также риторическому расстоянию. Поясню, что риторическим расстоянием называется расстояние от анафора до антецедента в иерархической структуре дискурса, представленной в рамках теории риторической структуры (Mann & Thompson, 1988).

Ещё одним важным компонентом данной когнитивной модели являются фильтры. Иногда в дискурсе могут возникать ситуации референциального конфликта (РК): «Будем

называть РК такую ситуацию, при которой в пределах текущего дискурсивного фрагмента адресат может отнести использованное говорящим РедуцРС[1] к нескольким референтам, активированным в его РП[2]» (Фёдорова, Успенская, 2011: 198). Фильтр референциальных конфликтов запрещает употребление местоимения при наличии высокого коэффициента активации у более чем одного референта и, тем самым, разрешает возникающие неоднозначности. Данный фильтр никак не влияет на коэффициент активации референта, и поэтому он представляет собой отдельный компонент когнитивной многофакторной модели.

Описанная когнитивная модель референциального выбора (Kibrik, 2011) представлена на рисунке 1:

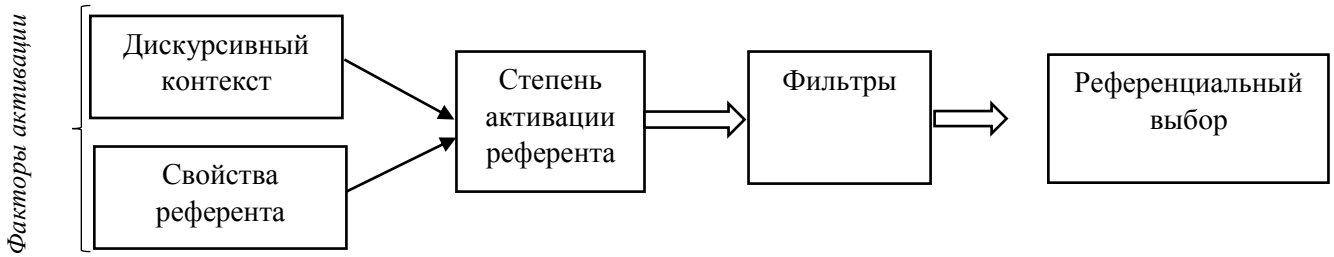

Рисунок 1. Когнитивная многофакторная модель референциального выбора.

2. Корпус WSJ MoRA 2015

В настоящей работе использовался корпус WSJ MoRA 2015 (прежнее название RefRhet), созданный и аннотированный группой исследователей под руководством А.А. Кибрика специально для исследований референциального выбора (Kibrik, 2011). В его основе лежит корпус RST Discourse Treebank, состоящий из англоязычных статей газеты The Wall Street Journal, размеченный в рамках теории риторической структуры (Carlson, Marcu, Okurowski, 2003). Большинство газетных статей The Wall Street Journal представляют собой новостные сообщения экономической или политической направленности.

Данный корпус является базой для целого ряда работ, посвященных изучению референции (Kibrik & Krasavina, 2005; Krasavina, 2006). Схема разметки корпуса MoRA использовалась в экспериментах машинного обучения по моделированию референциального выбора – (Кибрик и др., 2010); (Loukachevitch et al., 2011).

Для моделирования референциального выбора используются значения признаков из разметки корпуса WSJ MoRA 2015. Среди факторов, относящихся к референту, используются одушевленность, род и протагонизм. Для анафора и антецедента выделяют следующие признаки: фразовый тип (именная или предложная группа), грамматическая роль, референциальная форма (местоимение, дескрипция или имя собственное), лицо и число. Также используются такие расстояния между анафором и антецедентом, как расстояние в словах, маркабулах, клаузах, предложениях, абзацах и риторическое расстояние.

3. Модель подсчёта коэффициентов активации референтов

Когнитивная многофакторная модель, описанная ранее, обладает математическим компонентом, который позволяет вычислять коэффициент активации каждого референта. Утверждается, что каждый фактор активации характеризуется числовым весом, который отражает его вклад в общий коэффициент активации.

[1] РедуцРС – редуцированное референциальное средство. Сокращение из работы (Фёдорова, Успенская, 2011).
[2] РП – рабочая память. Сокращение из работы (Фёдорова, Успенская, 2011).

В работе (Kibrik, 1999) представлена система подсчета коэффициента активации на основании значений комплекса таких признаков, как риторическое, линейное расстояние, расстояние в абзацах, протагонизм референта, одушевленность, синтаксическая роль антецедента и т.д. В данном исследовании будет использоваться не весь набор признаков, а лишь те, что вносят наибольший вклад в коэффициент активации. Таким образом, система, представленная в (Kibrik, 1999), была модифицирована в соответствии с типом дискурса, используемым в данной работе, – газетные тексты The Wall Street Journal. Система подсчета, представленная в (Kibrik, 1999), и данная её модификация подбирались методом проб и ошибок, до тех пор, пока подобранные числовые веса не стали объяснять весь имеющийся материал. Модель подсчета коэффициентов активации, разработанная для настоящего исследования, отражена в таблице 1.

Признак	**Значение**	**Вес**
Одушевленность	LinD ≤ 2	0
	LinD ≥ 3: Animate Inanimate Collective[3]	 0.2 0.1 0
Синтаксическая роль антецедента	RhD> 3.5	0
	RhD ≤ 3.5: Subj Dir_Obj, Indir_Obj, Obl Attribute, Possessor Specification[4]	 0.3 0.2 0.1 0
Линейное расстояние между анафором и антецедентом в клаузах (LinD)	0 1 2 3 > 3	0.1 0 -0.1 -0.2 -0.3
Риторическое расстояние между анафором и антецедентом (RhD)	0; 1; 1.5 2; 2.5; 3 3.5 ≥ 4	0.6 0.5 0.4 0
Расстояние между анафором и антецедентом в абзацах	0 1 > 1	0 -0.2 -0.4

Таблица 1. Числовые веса значений факторов

Применив модель подсчета к материалу исследования, я установила следующее соответствие между потенциальными референциальными средствами и коэффициентами активации:

Референциальное выражение	Только полная ИГ	Полная ИГ ?местоимение	Полная ИГ или местоимение	Местоимение ?полная ИГ	Только местоимение
Коэффициент активации	**≤ 0,4**	**0,5**	**0,6** **0,7**	**0,8**	**0,9** **1**

Таблица 2. Соответствие между коэффициентом активации и типом референциального выражения

[3] К коллективным референтам (*collective*) относят организации, коллективы людей, государства и т.п.

[4] Спецификации в корпусе WSJ MoRA 2015 могут быть двух типов: постпозитивное уточнение (*with four lawyers, **all former assistant U. S. attorneys with extensive trial experience***) и предикат (*George Bush is **the president of America***).

Данные, представленные в таблице 2, демонстрируют, что коэффициенты активации, попадающие в интервал от 0,5 до 0,8, можно назвать промежуточными коэффициентами активации, так как они характеризуют случаи некатегорического референциального выбора.

4. Моделирование референциального выбора

В настоящей работе для моделирования референциального выбора использовалась программа WEKA (Frank E. et al., 2010), а именно, был применен алгоритм логистической регрессии, позволяющий видеть вероятностные характеристики тех форм, среди которых проводился выбор.

Моделирование референциального выбора было осуществлено для двухклассовой задачи – выбор между местоимением и полной именной группой. Обучающая выборка, подаваемая на вход программе WEKA, состоит из 2249 примеров, взятых из корпуса WSJ MoRA 2015. Однако не все из них использовались для построения модели – чтобы избежать проблемы переобучения вся генеральная совокупность была разделена на обучающую выборку (1124 примера) и тестовую (1125 примеров). Эффективность работы алгоритма оценивается с помощью такой метрики, как аккуратность, которая представляет собой отношение правильно предсказанных форм к общему числу предсказанных референциальных выражений. Правильно предсказанными формами считались те, что совпадали с эталонными, представленными в корпусе WSJ MoRA 2015. По данным нашего моделирования аккуратность работы алгоритма составила 86.6%, среди всех форм, предсказанных алгоритмом, 150 были предсказаны неверно.

Все отклонения алгоритма могут условно делиться на три категории. Первая включает в себя те примеры, где вероятность отклоняющегося варианта оказалась значительно выше эталонного варианта – вероятность предсказанной формы находится в интервале от 0,9 до 1. Второй тип случаев, который встречается среди отклоняющихся предсказаний алгоритма, - это такие, где референциальные выражения могут быть выбраны примерно с равными вероятностями – вероятность предсказанного варианта попадает в интервал от 0,5 до 0,55. Для исследования эти примеры представляют наибольший интерес, так как именно в этой категории доля случаев некатегорического референциального выбора максимальная среди всех групп. К третьей категории были отнесены все остальные неверные предсказания – те, в которых разница между вероятностями эталонной и предсказанной формы варьируется от 0,1 до 0,8. В таблице 3 представлены несколько примеров для всех типов отклонений:

Тип отклонения	**Эталонная форма**	**Предсказанная форма**	**Вероятность эталонной формы**	**Вероятность предсказанной формы**
I	Pronoun	Full Noun Phrase	0	1
II	Pronoun	Full Noun Phrase	0,46	0,54
III	Full Noun Phrase	Pronoun	0,283	0,717

Таблица 3. Примеры отклонений алгоритма логистической регрессии

5. Анализ отклонений при моделировании референциального выбора

Для всех 150 примеров отклонений вручную были подсчитаны коэффициенты активации, используя адаптированную систему подсчета, представленную в разделе 3 настоящей работы. В данном исследовании для проверки корректности построенной модели будут

также учитываться и случаи правильно предсказанных референциальных форм – всего 150 случаев.

Для того, чтобы определить, существует ли корреляция между вероятностной характеристикой анафорического местоимения и уровнем активации референта, для всех четырех случаев – правильных предсказаний и трех типов отклонений, описанных в разделе 4, - был подсчитан коэффициент корреляции Спирмена, выявляющий степень линейной связи между случайными величинами. Чем ближе коэффициент к |1|, тем теснее линейная связь. Значения коэффициентов, полученные для каждого случая, показаны в таблице 4:

Случай		**Коэффициент корреляции**
Правильно предсказанные референциальные выражения		0,765191
Референциальные выражения, отклоняющиеся от эталонных	Вероятность предсказанного выражения равна (или примерно равна) 1	0,305355
	Вероятности почти равны	0,157204
	Все остальные случаи	0,13152

Таблица 4. Корреляция между коэффициентами активации референтов и вероятностной характеристикой анафорических местоимений

Наиболее сильная зависимость между коэффициентом активации референта и вероятностной характеристикой наблюдается, как этого и следовало ожидать, в случае правильных предсказаний. Таким образом, каждому уровню активации можно привести в соответствие вероятностный интервал, характеризующий анафорическое местоимение, что позволит в дальнейших исследованиях интерпретировать одну величину с помощью другой. Данное соответствие было выявлено на основании диаграмм с размахом, построенных с помощью программы Microsoft Office Excel, и представлено в таблице 5:

Вероятностная характеристика	[0;0,004]	[0,006;0,27]	[0,023;0,59]	[0,16;0,67]	[0,51;0,82]	[0,77;0,87]	[0,86;0,95]
Коэффициент активации	**≤ 0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**

Таблица 5. Соответствие между вероятностной характеристикой референта и коэффициентом активации

Все типы отклонений характеризуются низким показателем зависимости. При этом умеренная корреляция (0,305355) наблюдается в случае, когда отклоняющееся референциальное выражение имело вероятность близкую или равную единице. В двух последних типах отклонений наблюдается слабая корреляция, что опровергает предположение о зависимости вероятностной характеристики и коэффициента активации в случае отклоняющихся предсказаний. Также было подсчитано, что 73% всех отклонений имели промежуточный коэффициент активации, то есть попадающий в интервал от 0,5 до 0,8. Следовательно, некатегорическая природа референциального выбора является одной из основных причин отклонений при моделировании референциальных выражений.

6. Выводы

В ходе работы выяснилось, что в случае правильных предсказаний алгоритма между вероятностной характеристикой и коэффициентом активации наблюдается сильная корреляция, что позволило установить соответствие между значениями этих двух признаков. Отсутствие корреляционной зависимости для отклоняющихся предсказаний показывает, что коэффициент активации и вероятностная характеристика в данных случаях не взаимосвязаны. Исследование также показало, что большинство случаев отклоняющихся предсказаний являются примерами некатегорического референциального выбора.

Библиография

Belz A., Varges S. Generation of repeated references to discourse entities//Proceedings of the Eleventh European Workshop on Natural Language Generation. – Association for Computational Linguistics. – 2007. – С. 9-16.

Carlson L., Marcu D., Okurowski M. E. Current Directions in Discourse and Dialogue, chapter Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory //IEEE Intelligent Systems. – 2003.

Frank E. et al. Weka-a machine learning workbench for data mining //Data Mining and Knowledge Discovery Handbook. – Springer US. – 2010. – С. 1269-1277.

Kibrik A. A. Reference and working memory: Cognitive inferences from discourse observation // In: Karen van Hoek, Andrej A. Kibrik, and Leo Noordman (eds.), Discourse Studies in Cognitive Linguistics. Proceedings of the 5th International Cognitive Linguistics Conference. Amsterdam: John Benjamins. – 1999. – С.29-52

Kibrik A. A. Reference in discourse//Oxford University Press. – 2011.

Kibrik A. A., Krasavina O. A corpus study of referential choice: The role of rhetorical structure //Proceedings of DIALOG'05. – 2005.

Krasavina Olga. Korpusno-orientirovannoe issledovanie referencii (principy annotacii i analiz dannyx) [A corpus-oriented study of reference (principles of annotation and data analysis)] // Ph.D. thesis. Dept. of Theoretical and Applied Linguistics, Moscow State University. – 2006

Loukachevitch N. V., Dobrov G. B., Kibrik A. A., Khudyakova M. V., and Linnik A. S., "Factors of referential choice: Computational modeling," //А.Е.Кибрик и др. (ред.) Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции Диалог, vol. 10, М.: РГГУ. – 2011. – С. 458–467.

Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization //Text-Interdisciplinary Journal for the Study of Discourse. – 1988. – Т. 8. – №. 3. – С. 243-281.

Van Deemter, K., Gatt, A., van Gompel, R. P., Krahmer, E. Toward a computational psycholinguistics of reference production //Topics in cognitive science. – 2012. – Т. 4. – №. 2. – С. 166-183.

Кибрик, А. А., Добров, Г. Б., Залманов, Д. А., Линник, А. С., Лукашевич, Н. В. Референциальный выбор как многофакторный вероятностный процесс// По материалам международной конференции Диалог. – 2010. – С.173-180.

Фёдорова О. В., Успенская А. М. Экспериментальный анализ дискурса: референциальный выбор в ситуации потенциального референциального конфликта (экспериментальное исследование на материале русского языка)//Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции Диалог (Бекасово, 25–29 мая 2011 г.), Вып. 10 (17), РГГУ Москва. – 2011. – С. 196–206.