Фильтр референциальных конфликтов в модели референциального выбора

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

Ключевые слова: референциальный выбор, референциальный конфликт, компьютерное моделирование, частичное обучение, машинное обучение, корпус MoRA 2015.

Referential conflict filter in a model of referential choice

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

My research focuses on the phenomenon of referential conflict which has not been taken into consideration in the predictive model of referential choice trained on WSJ MoRA corpus. Referential conflict is a situation in a discourse when there are two or more activated referents and the usage of reduced referential device leads to ambiguity. In this paper I describe the implementation of referential conflict filter in the computational model of referential choice by means of corpus analysis and semi-supervised learning. I show that this modification of a model can improve accuracy of prediction.

Key words: referential choice, referential conflict, computational modeling, semi-supervised learning, machine learning, MoRA 2015 corpus.

1. Введение

Явление референции широко изучается в компьютерной лингвистике. Организовываются различные соревнования алгоритмов моделирования референциального выбора ([Belz et al., 2010]), создаются системы разрешения кореферентности ([Haghighi and Klein, 2007], [Rahman and Ng, 2009]), модули разрешения анафоры и генерации рефренциальных выражений входят в состав систем порождения текстов на естественном языке ([Gasperin, 2006], [Gatt et al., 2016]). Однако, до сих пор остаются открытыми вопросы касательно природы референциальной неоднозначности, её влияния на качество автоматических систем генерации референциальных выражений и разрешения анафоры, а также учёта данного фактора в подобных моделях.

Объекты внеязыковой действительности, упоминаемые говорящим, называются **референтами**. Подобными объектами могут являться живые существа, предметы, абстрактные понятия и т.п. Языковое выражение, которым обозначен референт в конкретном случае, называется **референциальным выражением**. Референциальные выражения могут быть полными и редуцированными [Kibrik, 2011:37]. К первому типу относятся определенные и неопределенные дескрипции разной степени распространенности, а также имена собственные. Редуцированными референциальными выражениями считаются местоимения и нулевые выражения.

В моей работе в качестве теоретической базы выступает когнитивный многофакторный подход [Kibrik, 1996], согласно которому референциальный выбор зависит от степени активации референта в рабочей памяти говорящего. Так, главный закон референциального выбора звучит следующим образом: «Если референт обладает высоким уровнем активации в рабочей памяти говорящего, то используется редуцированное референциальное выражение. При низком уровне активации, напротив, употребляется полное референциальное средство» [Kibrik, 2011:363].

В рамках данного подхода существует ряд факторов, которые влияют на активацию референта и, следовательно, на референциальный выбор. Для измерения степени активации в данном подходе было введено понятие коэффициента активации. Каждый фактор имеет численное значение (тот вклад, который он вносит в активацию референта), и для получения счета активации референта суммируются значения всех факторов. Обычно коэффициенты активации лежат в пределах от 0 до 1. В зависимости от значения коэффициента активации будет использоваться либо полное референциальное выражение, либо редуцированное.

Материалом статьи является корпус WSJ MoRA 2015, который был создан и аннотирован исследовательской группой под руководством А. А. Кибрика специально для изучения референциального выбора. Прежнее название корпуса RefRhet [Кибрик и др., 2010]. Генеральная выборка, на которой происходит обучение модели референциального выбора, состояла из 2249 объектов – маркабул, взятых из корпуса WSJ MoRA 2015.

Целью моей работы является построение модели референциального выбора, которая имела бы в своей архитектуре фильтр референциальных конфликтов, позволяющий обрабатывать случаи с потенциальной референциальной неоднозначностью.

2. Референциальная неоднозначность

Неоднозначность является распространенным явлением в языке, которое может наблюдаться на всех его уровнях [Piantadosi et al., 2012], и референциальные выражения также могут быть неоднозначными. Ниже приведен пример, иллюстрирующий явление референциального конфликта.

(1) **Василий$_i$** зашел в комнату, в которой сидел пожилой **мужчина$_j$**. ??**Он$_{i/j}$** поинтересовался, который час.

Оба референта в данном примере имеют достаточно высокий уровень активации, и, следовательно, они оба могут выступать в качестве антецедента местоимения *он*. Однако, употребление местоимения может привести к нежелательной неоднозначности, или референциальному конфликту.

Явление, при котором невозможно однозначно соотнести анафор с тем или иным антецедентом, называется референциальным конфликтом (РК). В работе [Фёдорова, Успенская, 2011: 198] приведено следующее определение 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]». То есть РК возникает в том случае, когда в РП говорящего содержится более одного высоко активированного референта.

Проблемой референциальной неоднозначности занимаются и в области компьютерной лингвистики. В работе [McCoy and Strube, 1999] описывается опыт

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

создания системы генерации референциальных выражений. Авторы статьи выделяют четыре фактора, определяющих форму референциального выражения, и среди них указан фактор референциальной неоднозначности. Они считали референциальное выражение неоднозначным, если есть конкурирующий антецедент, совпадающий в числе и роде, употребленный в предшествующем предложении или слева от анафора в текущем предложении. Для того, чтобы выбрать форму референциального выражения для таких случаев, авторы статьи обратились к алгоритму разрешения анафоры [Strube, 1998]. Идея заключалась в следующем: когда мы хотим употребить анафорическое местоимение по отношению к референту E, но есть конкурирующий антецедент C, то мы обращаемся к алгоритму разрешения анафоры. Если он соотнесет местоимение с референтом E, то местоимение можно употребить в данном контексте; в другом случае необходимо употребить полную ИГ, чтобы предотвратить неоднозначность.

В статье [Yang et al., 2010] описан подход к автоматическому определению потенциальных случаев с референциальной неоднозначностью, а именно таких, которые могут по-разному интерпретироваться разными читателями. Авторы исследования описывают общие принципы, определяющие их модель: коллекция с суждениями носителей языка относительно случаев с неоднозначностью; эвристики, используемые для определения наиболее вероятного антецедента для неоднозначного анафора и модуль с машинным обучением для определения вероятности выбора того или иного антецедента. Статья [Sharma et al., 2016] также посвящена этой тематике, а именно выделению в тексте потенциально неоднозначных случаев. Авторы статьи рассматривают данную проблему как проблему классификации. В качестве обучающей выборки они используют корпус, в который включены тексты различных требований и документов, в которых есть примеры с референциальной неоднозначностью. Ввиду отсутствия достаточного количества обучающих примеров с неоднозначностью, и как следствие дисбаланса классов, в своей статье авторы решили применить обучение с частичным привлечением учителя (semi-supervised learning) [Kocaguneli, Cukic and Lu, 2013]. Наилучший результат классификации был достигнут с помощью алгоритма байесовской сети – полнота составила 95%.

С проблемой референциальной неоднозначности сталкиваются не только при моделировании референциального выбора, но и в задачах разрешения анафоры и выделения кореферентных цепочек. В работе [Toldova et al., 2016] рассматриваются различные ошибки систем распознавания кореферентных цепочек. Одной из проблем, которая влияла на качество подобных систем, была морфологическая неоднозначность в русском языке и возникающие из-за этого референциальные конфликты.

3. Признак «Наличие референциального конфликта»

Фильтр референциального конфликта будет представлен в моей модели признаком «Наличие референциального конфликта». Ситуация референциального конфликта подразумевает наличие в дискурсе двух или более высоко активированных референтов, поэтому в своем исследовании для поиска примеров с референциальным конфликтом я буду для каждого анафора искать ближайший ложный антецедент. С помощью системы подсчета, разработанной мной специально для этого исследования, я вычислю коэффициенты активации референтов, а затем все случаи, когда оба референта будут высоко активированы, я проверю по корпусу WSJMoRA 2015 на наличие в них возможного референциального конфликта. Полученные таким образом данные будут использоваться для разметки признака «Наличие референциального конфликта», который впоследствии будет добавлен в предсказательную модель референциального выбора.

3.1. Поиск ближайших ложных антецедентов

Чтобы предсказывать наличие потенциального референциального конфликта, необходимо знать о наличии в тексте для каждого анафора не только его истинного антецедента, но и ложного, такого что, если в тексте будет употреблено редуцированное референциальное выражение, то слушающий (или читатель) не сможет однозначно соотнести анафор с антецедентом или сделает это неправильно.

Для поиска ложных антецедентов в моем исследовании применяются конвенциональные референциальные средства (род и число). Для поиска используется расстояние в словах, так как оно является наиболее эффективным и простым для реализации способом. Для нахождения в тексте ближайшего ложного антецедента с помощью специального алгоритма, написанного на языке программирования Python, используются следующие признаки:

1) Согласование по роду и числу с анафором;
2) Ложный антецедент должен быть не кореферентен анафору, то есть они должны входить в разные референциальные цепочки;
3) Ложный антецедент находится в тексте до анафора и является ближайшим к нему референциальным выражением.

Данный алгоритм нашел ложных антецедентов для 2146 объектов из 2249 объектов в генеральной совокупности.

3.2.Система подсчета коэффициентов активации

Система подсчета коэффициентов активации необходима мне для выявления в текстах случаев с референциальной неоднозначностью. В своем исследовании я буду разрабатывать систему подсчета коэффициентов активации референтов с помощью алгоритма LASSO регрессии, которая широко применяется для отбора признаков [Tibshirani, 1996]. Процедура отбора признаков позволяет исключить из набора неинформативные признаки («шумовые признаки»), которые приводят к снижению точности.

В процессе работы алгоритма LASSO регрессии величина приписанных алгоритмом коэффициентов будет пропорциональна важности соответствующих переменных для классификации, а для переменных, которые дают наименьший вклад в устранение ошибки, коэффициенты станут нулевыми. Таким образом, более значимые признаки сохранят свои коэффициенты ненулевыми, а менее значимые – обнулятся. Стоит также отметить, что большие по модулю отрицательные значения коэффициентов тоже говорят о сильном влиянии. Данное свойство используется в моей работе для создания новой системы подсчета коэффициентов активации референтов.

В моем исследовании подсчет коэффициентов активации будет осуществляться не только для истинных антецедентов, но также и для ложных. Модель подсчета коэффициентов активации, разработанная с помощью метода LASSO регрессии, представлена в таблице ниже.

Признак	Значение	Вес
Одушевленность	Animate	0,245047
	Inanimate	0,0964
	Collective	0,067953
Синтаксическая роль антецедента	Subj	0,12918
	Dir_Obj	-0,010517
	Indir_Obj	-0,042649
	Obl	0,003323
	Attribute	-0,080366
	Possessor	-0,017485
	Specification	-0,025451
Линейное расстояние в клаузах	0	0,162889
	1	0,127714
	2	-0,094214
	3	-0,080336
	> 3	-0,116053

Линейное расстояние в предложениях	0 1 2 3 > 3	0,439952 0,12602 0,10889 0,08418 0,086778

Таблица 1. Числовые веса значений факторов.

Для того, чтобы выявить случаи с возможным референциальным конфликтом, по новой модели подсчета коэффициентов активации были вычислены коэффициенты активации референтов по признакам истинного антецедента и ложного. Затем отбирались те случаи, в которых оба коэффициента активации попадают в интервал от 0,7 до 0,977. Всего было выявлено 256 таких случаев.

После этого я анализировала все выделенные случаи по корпусу WSJ MoRA 2015 для того, чтобы выяснить, могут ли они являться примерами потенциального референциального конфликта. Всего в процессе анализа корпусных данных мной было обнаружено 88 примеров с потенциальным референциальным конфликтом. В соответствии с этими данными в качестве дополнительного признака в обучающую выборку был добавлен признак наличия/отсутствия референциального конфликта. То есть 88 пар анафор-антецедент в выборке получат отметку о наличии возможного референциального конфликта (значение признака «Наличие РК» равно 1), а в остальных 168 случаях (из тех 256, упомянутых ранее) будет отмечено отсутствие референциального конфликта (значение признака «Наличие РК» равно 0).

4. Моделирование референциального выбора

Моделирование референциального выбора разрабатывалось мной для двухклассовой задачи – выбор между полной именной группой и местоимением. Генеральная выборка состояла из 2249 объектов – маркабул, взятых из корпуса WSJ MoRA 2015. Для моделирования был выбран алгоритм машинного обучения градиентный бустинг (Gradient Boosting Classifier), так как он показывает наилучшее качество на этих данных. Моделирование референциального выбора проводилось с помощью пакета средств машинного обучения scikit-learn[3] на языке программирования Python. Для оценки качества работы алгоритма использовалась процедура кросс-валидации. Прогноз модели оценивался с помощью такой метрики, как аккуратность, которая является отношением

[3] http://scikit-learn.org

правильно предсказанных форм к общему числу предсказанных референциальных выражений.

Для оценки аккуратности предсказаний в качестве золотого стандарта выступают референциальные выражения, которые были употреблены в корпусе.
Всего для моделирования референциального выбора в настоящей работе используются следующие 25 признаков, автоматически извлекаемых из корпуса: одушевленность, число, род, лицо, фразовый типа анафора, фразовый тип антецедента, одушевленность антецедента, грамматическая роль анафора, грамматическая роль антецедента, тип дескрипции антецедента, тип местоимения антецедента, является ли антецедент группой, тип группы антецедента, наличие количественного числительного, тип имени собственного, тип атрибута, риторическое расстояние, расстояние от анафора до антецедента в словах, расстояние от анафора до антецедента в абзацах, расстояние от анафора до антецедента в предложениях, расстояние от анафора до антецедента в маркабулах, расстояние от анафора до антецедента в клаузах, длина маркабулы-антецедента в словах, количество маркабул до последнего упоминания в форме полной именной группы, порядковый номер маркабулы в цепочке.

Аккуратность классификации на данном наборе признаков составила 0.9053. Теперь посмотрим на качество модели с признаком наличия/отсутствия референциального конфликта, который был проверен по корпусу (см. Раздел 3.2.). Ввиду того, что размеченных данных не так много, я решила воспользоваться идеей из статьи [Sharma et al., 2016] и применить обучение с частичным привлечением учителя (частичное обучение) для своих данных по референциальной неоднозначности. Напомню, что частичное обучение – это разновидность обучения с учителем, которое использует небольшое количество размеченных данных и много неразмеченных. В своей работе я буду использовать метод частичного обучения, основанный на графах [Bengio, Delalleau, Le Roux, 2006: 193-216], который представлен в пакете машинного обучения scikit-learn моделью Label Spreading. Идея, лежащая в основе этого метода, заключается в следующем: создается граф, в узлах которого находятся размеченные или неразмеченные данные, и, если точки обладают похожими признаками, они размечаются одинаково. Следовательно, метки, которые нам известны, исходя из данных корпуса, как бы распространяются на остальные примеры из выборки на основании сходства.

Таким образом, с помощью частичного обучения я получу аннотацию неразмеченных объектов в выборке, чтобы затем добавить в модель признак «Наличие референциального конфликта», размеченный с использованием частичного обучения.

Для метода частичного обучения, который я использую в своем исследовании, существует ряд параметров, от настройки которых зависит качество приписывания меток объектам. Тип ядра – это один из основных параметров, который может принимать такие значения, как RBF-ядро (от radial basis functions) и kNN-ядро. На практике линейно разделимые классы встречаются не очень часто, поэтому, чтобы выборка стала линейно разделима, применяют так называемый Kernel Trick, то есть отображают исходное пространство признаков в какое-то новое, где классы можно разделить, и делается это как раз с помощью ядер. Точность классификации, в частности, зависит от выбора ядра. Для этих ядер также в свою очередь есть настраиваемые числовые параметры – коэффициенты.

Я провела частичное обучение с целевым признаком «Наличие референциального конфликта» в двух вариантах – с двумя типами ядер, при этом перебирая различные значения коэффициентов, от которых также зависит качество классификации, проводимой с использованием того или иного ядра. В результате было получено 5 вариантов разметки целевого признака с kNN-ядром и 80 с RBF.

Для того, чтобы понять, какой из этих 85 вариантов лучший, я внедряла каждый из этих вариантов в предсказательную модель референциального выбора наряду с остальными признаками из стандартного набора, состоящего из 25 признаков. Машинное обучение проводилось также с использованием алгоритма градиентного бустинга. Не все варианты вносили улучшения в модель, однако некоторые показывали результаты лучше, чем те, что были достигнуты на данный момент. Отмечу среди них один, показавший наибольшую точность классификации. Аккуратность классификации для двухклассовой задачи с использованием признака "Наличие РК", размеченным с помощью частичного обучения, составила 0.9116 (алгоритм градиентного бустинга).
Этот результат был получен со следующими параметрами алгоритма частичного обучения: RBF-ядро и значение коэффициента равно 44. С добавлением признака «Наличия РК» качество предсказания референциального выбора улучшилось на 0.63% по сравнению с аккуратностью, полученной на наборе из 25 признаков. Следовательно, новый фактор вносит положительный вклад в модель и повышает аккуратность классификации.

5. Вывод

Нельзя проигнорировать важность фактора «Наличие РК», который базировался на корпусном анализе и был размечен с помощью частичного обучения – именно добавление этого признака в стандартный набор позволило достичь максимального качества

классификации. Этот фактор оказался высокоинформативным и значительно улучшил качество предсказания формы референциальных выражений. Таким образом, я делаю вывод, что референциальная неоднозначность играет большую роль в моделировании референциального выбора, и она обязательно должна учитываться в предсказательной модели, например, как в моем исследовании, в виде отдельного признака. Я считаю, что аннотация данного признака в корпусе, несмотря не времязатратность и трудоёмкость, может быть довольно перспективной с точки зрения улучшения аккуратности предсказания референциального выбора.

6. Список использованной литературы

1) Belz A. et al. Generating referring expressions in context: The GREC task evaluation challenges //Empirical methods in natural language generation. – Springer, Berlin, Heidelberg. – 2010. – С. 294-327.
2) Gasperin C. Semi-supervised anaphora resolution in biomedical texts // Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06. – 2006. – С. 96–103.
3) Gatt A., Marín N., Portet F., Sánchez D. The role of graduality for referring expression generation in visual scenes //International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. – Springer, Cham. – 2016. – С. 191-203.
4) Haghighi A., Klein D. Unsupervised coreference resolution in a nonparametric bayesian model //Proceedings of the 45th annual meeting of the association of computational linguistics. – 2007. – С. 848-855.
5) Kibrik A. A. Anaphora in Russian narrative discourse: A cognitive calculative account //In: Barbara A. Fox (ed.), Studies in anaphora. Amsterdam: Benjamins. – 1996. – С. 255-304.
6) Kibrik A. A. Reference in discourse//Oxford University Press. – 2011.
7) Kocaguneli E., Cukic B., Lu H. Predicting more from less: Synergies of learning //Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), 2013 2nd International Workshop on. – IEEE. – 2013. – С. 42-48.
8) McCoy K., Strube M. Generating anaphoric expressions: Pronoun or definite description //Proceedings of the ACL Workshop on The Relation of Discourse/Dialogue Structure and Reference. – 1999. – С. 63-71.

9) Piantadosi S. T., Tily H., Gibson E. The communicative function of ambiguity in language //Cognition. – 2012. – Т. 122. – №. 3. – С. 280-291.
10) Rahman A., Ng V. Supervised models for coreference resolution //Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. – Association for Computational Linguistics. – 2009. – С. 968-977.
11) Sharma R., Sharma N., Biswas K. K. Machine Learning for Detecting Pronominal Anaphora Ambiguity in NL Requirements //Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science & Engineering (ACIT-CSII-BCD), 2016 4th Intl Conf on. – IEEE. – 2016. – С. 177-182.
12) Strube M. Never look back: An alternative to centering //Proceedings of the 17th international conference on Computational linguistics-Volume 2. – Association for Computational Linguistics. – 1998. – С. 1251-1257.
13) Tibshirani R. Regression shrinkage and selection via the lasso //Journal of the Royal Statistical Society. Series B (Methodological). – 1996. – С. 267-288.
14) Toldova, S., Azerkovich, I., Roytberg, A., Ladygina, A., and Vasilyeva, M. Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language. // Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016). – 2016. – С. 74–83.
15) Yang H. et al. A methodology for automatic identification of nocuous ambiguity //Proceedings of the 23rd International Conference on Computational Linguistics. – Association for Computational Linguistics. – 2010. – С. 1218-1226.
16) 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.
17) 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.