Применение встроенных методов отбора признаков для оптимизации модели референциального выбора

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

Ключевые слова: референциальный выбор, референциальное выражение, компьютерное моделирование, отбор признаков, машинное обучение, корпус MoRA 2015.

Embedded feature selection methods for optimizing a model of referential choice

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

Referential choice, that is the choice of referential expression, depends on various interconnected factors. This paper presents a study on optimization of the computational model of referential choice by means of embedded feature selection methods: logistic regression and random forest. The elimination of a number of features in the modeling studies showed that the accuracy of classification did not fall down. The analysis of the data demonstrates that not all of the factors are equally significant, and, thus, uninformative features can be easily removed from the model.

Key words: referential choice, referential expression, computational modeling, feature selection, machine learning, MoRA 2015 corpus.

1. Введение

В процессе порождения дискурса говорящий или пишущий постоянно упоминает определенные объекты внеязыковой действительности, называемые референтами. В настоящей работе исследуется механизм референциального выбора, то есть то, как говорящий выбирает языковое выражение, которым будет закодирован референт. Выделяются два основных типа референциальных выражений: полные и редуцированные. К полым референциальным выражениям относят имена собственные и дескрипции, а к редуцированным – местоимения и нулевые выражения.

Теоретической базой данной работы является 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. На степень активации, в свою очередь, влияет большое число взаимосвязанных факторов, каждый из которых может быть связан как с самим референтом, так и с контекстом высказывания. Менее активированные в сознании говорящего референты выражаются более полными номинациями, в противном случае говорящий использует редуцированные номинации. Таким образом, типичным средством обозначения высоко активированных референтов являются личные местоимения и референциальные нули.

Референциальный выбор является одним из вопросов, изучаемых в рамках такой области лингвистики, как генерация текстов на естественном языке (*NLG – natural language generation*). Генерация референциальных выражений (*REG* или *GRE - Referring Expressions Generation или Generating Referring Expressions*) представляет собой неотъемлемую часть программ порождения текстов ([Dale, 1992], [van Deemter, 2002]), программ автоматического реферирования [Nenkova, 2008], а также программ машинного перевода [Jurafsky, 2000: 864]. В модели референциального выбора, о которой идет речь в данной работе, стоит задача выбора между полным и редуцированным референциальным выражением. Целью настоящего исследования является оптимизация системы моделирования референциального выражения, а именно, выбор оптимального набора признаков, при котором точность алгоритма, обученного на этом подмножестве признаков, будет максимальной.

2. Обзор предыдущих исследований

Исследования по отбору признаков в модели референциального выбора уже проводились для описанного в данной работе материала (см. Раздел 3). В статье [Loukachevitch et al., 2011] изучается важность конкретных признаков для трехклассовой задачи референциального выбора – выбор между полной именной группой, дескрипцией и местоимением. Аккуратность предсказания при использовании всех факторов составила 80.7%. Затем авторы статьи исключали из модели некоторые признаки или группы признаков (различные метрики расстояний). При этом любые модификации системы давали снижение качества классификации. Например, исключение только одного признака «Протагонизм» вело за собой ухудшение качества на 0.7%. Таким образом, авторы статьи заключают, что большинство признаков важны для модели и не могут быть исключены из нее.

Оптимизация модели референциального выбора также описана в работе [Кибрик, 2012]. В ней сравнивалось качество модели, в которую были включены все признаки, и модели, из которой исключили признаки, требующие больших усилий при аннотации (например, аннотация иерархической структуры дискурса). Результаты моделирования показали, что использование «дешевого» набора параметров влечет за собой снижение аккуратности на 2.7% (для двухклассовой задачи). Авторы статьи заключают, что «ни один из факторов не теряет своей значимости, если мы стремимся достичь наилучшего результата» [Кибрик, 2012: 242]. То есть, утверждается, что наилучшего качества классификации можно добиться только при использовании максимально полного набора факторов.

В своем исследовании я намереваюсь опровергнуть выводы, сделанные в данных статьях, и доказать, что количество признаков, использующихся в модели, можно сократить без ущерба для качества классификации.

3. Корпус MoRA 2015

Корпус WSJ MoRA 2015, используемый в настоящей работе, создавался специально для исследований референциального выбора и состоит из статей журнала Wall Street Journal. Аннотация корпуса представляет собой референциальную разметку и разметку риторической структуры текстов, которая была взята из корпуса RST Discourse Treebank, размеченного в рамках теории риторической структуры [Carlson, Marcu, Okurowski, 2003]. Выбор корпуса RST Discourse Treebank в качестве основы объясняется тем, что он уже содержит аннотацию иерархической структуры дискурса, а создание новой разметки риторической структуры является трудоемким и времязатратным процессом.

Программа MMAX Annotation Tool [Müller & Strube, 2006] применялась для разметки корпуса, а сама схема разметки называется MoRA (Moscow Reference Annotation scheme, 2011). Единицей разметки корпуса является маркабула, эквивалентная референциальному выражению.

В таблице 1 приведены основные количественные характеристики корпуса WSJ MoRA 2015.

Признак	**Количество в корпусе**
Тексты	64
Маркабулы	6294
Пары анафор-антецедент	1852
Предложения	976
Слова	23952
ЭДЕ[1]	2928
Референциальные цепочки	866

Таблица 1. WSJ MoRA 2015 Corpus: количественные характеристики.

4. Признаки, используемые при моделировании референциального выбора

Из разметки корпуса автоматически извлекается целый ряд признаков, которые влияют на выбор той или иной формы референциального выражения. Всего в модели

[1] Элементарная дискурсивная единица. Данное понятие подробно описывается в [Кибрик, Подлесская, 2009]. ЭДЕ автоматически извлекались из корпуса RST Discourse Treebank.

выделяется 25 признаков. Подробнее признаки описаны в работах [Loukachevitch et al., 2011] и [Кибрик, 2012].

Всё множество признаков можно поделить на несколько групп: признаки референта (одушевленность, число, род, лицо), признаки анафора (грамматическая роль, фразовый тип – беспредложная/предложная группа), признаки антецедента (грамматическая роль, одушевленность, наличие атрибута, наличие количественного числительного, тип дескрипции и т.д.), а также различные метрики расстояний между анафорой и антецедентом (линейное расстояние в клаузах, словах, маркабулах, предложениях, абзацах, риторическое расстояние и т.п.).

5. Отбор признаков

Отбор признаков позволяет исключить из набора неинформативные признаки («шумовые признаки»), которые приводят к снижению точности. Помимо этого, отбор признаков перед обучением модели необходимо производить по ряду других немаловажных причин. Во-первых, он позволяет уменьшить размерность пространства признаков, а также сокращается время обучения модели. Во-вторых, уменьшается вероятность переобучения - отсутствие избыточных данных исключает возможность обобщений на основании случайных закономерностей в обучающих данных.

Методы отбора признаков делятся на три группы: методы-фильтры (filters), методы-обертки (wrappers) и встроенные методы (embedded) [Ladha & Deepa, 2011]. Методы-фильтры применяются на этапе предобработки, до запуска алгоритма обучения, поэтому они оценивают признаки только на основе информации, полученной из обучающей выборки. Методы-обертки основаны на том, что классификатор запускается на конкретных подмножествах обучающей выборки, а затем выбирается подмножество наиболее информативных для обучения признаков.

Преимуществом встроенных методов, которые используются в данной работе является то, что они позволяют не отделять отбор признаков и обучение классификатора. Данный тип методов также обладает рядом других преимуществ: они хорошо приспособлены к конкретной модели; не требуется выделять специальное подмножество для тестирования, как в предыдущих методах, и, как следствие из этого, меньше риск переобучения.

В качестве алгоритмов отбора признаков в настоящей работе будут использоваться LASSO регрессия и случайный лес (Random Forest).

5.1. Моделирование референциального выбора без отбора признаков

Моделирование референциального выбора разрабатывалось автором данной статьи 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 было выбрано несколько алгоритмов машинного обучения разного типа: логистическая регрессия (Logistic Regression), градиентный бустинг (Gradient Boosting Classifier), метод опорных векторов (SVM) и случайный лес (Random Forest). Для оценки качества работы алгоритма использовалась процедура кросс-валидации. Прогноз модели оценивался с помощью такой метрики, как аккуратность, которая является отношением правильно предсказанных форм к общему числу предсказанных референциальных выражений. Корректно

предсказанными формами считались те, что совпадали с эталонными (исходными), представленными в корпусе WSJ MoRA 2015.

Результаты работы алгоритмов машинного обучения приведены в таблице 2:

Алгоритм	**Аккуратность классификации для двухклассовой задачи**
Логистическая регрессия	0.8893
Градиентный бустинг	0.9053
Метод опорных векторов	0.89
Случайный лес	0.8856

Таблица 2. Аккуратность классификации для двухклассовой задачи (полный набор признаков)

5.2. Отбор признаков методом LASSO регрессии

LASSO регрессия широко применяется для отбора признаков [Tibshirani, 1996]. В процессе работы алгоритма 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.

Ниже приведена таблица с ранжированными коэффициентами важности признаков, приписанными алгоритмом LASSO регрессии.

Признак	**Коэффициент**
Род	-0.071699
Фразовый тип анафора	-0.067430
Риторическое расстояние	-0.065404
Расстояние от анафора до антецедента в абзацах	-0.021806
Длина маркабулы-антецедента в словах	-0.012545
Одушевленность	-0.001942
Расстояние от анафора до антецедента в словах	-0.001399
Одушевленность антецедента	-0.001197
Расстояние от анафора до антецедента в предложениях	-0.000445
Лицо	0
Количество маркабул до последнего упоминания в форме полной именной группы	0
Порядковый номер маркабулы в цепочке	0
Фразовый тип антецедента	0
Число	0
Тип дескрипции антецедента	0
Тип местоимения антецедента	0
Тип группы антецедента	0
Является ли антецедент группой	0
Наличие количественного числительного	0.000686

Расстояние от анафора до антецедента в маркабулах	0.001040
Тип имени собственного	0.005545
Тип атрибута	0.012806
Расстояние от анафора до антецедента в клаузах	0.013151
Грамматическая роль анафора	0.017538
Грамматическая роль антецедента	0.023104

Таблица 3. Коэффициенты важности признаков согласно алгоритму LASSO регрессии

Таким образом, 9 признаков получили нулевые коэффициенты важности, следовательно, они будут удалены из модели, чтобы выяснить, дает ли это улучшение качества. В таблице 4 приведены результаты моделирования с помощью алгоритмов, описанных ранее, но уже с исключенными признаками:

Алгоритм	**Аккуратность классификации для двухклассовой задачи**
Логистическая регрессия	0.89
Градиентный бустинг	0.9065
Метод опорных векторов	0.8881
Случайный лес	0.8887

Таблица 4. Аккуратность классификации для двухклассовой задачи (без признаков, исключенных методом LASSO регрессии)

Результаты работы алгоритмов изменились в незначительной степени (на ±0,003), то есть исключение 9 признаков из модели никак не ухудшает ее качества. Можно заключить, что признаки лица, количества маркабул до последнего упоминания в форме полной именной группы, порядкового номера маркабулы в цепочке, фразового типа антецедента, числа, типа дескрипции антецедента, типа местоимения антецедента, а также бинарный признак «Является ли антецедент группой» являются мало информативными для модели.

5.3. Отбор признаков методом случайного леса

Случайные леса, используемые для классификации объектов, могут естественным образом быть использованы для оценки важности признаков [Breiman, 2001]. Для этого мы обучаем алгоритм на некоторой тренировочной выборке и во время построения модели для каждого элемента этой выборки считаем out-of-bag ошибку (усредненная оценка алгоритмов на тех данных, на которых они не обучались). Затем для каждого объекта такая ошибка усредняется по всему случайному лесу. Чтобы оценить важность конкретного признака, его значения перемешиваются для всех объектов обучающей выборки и out-of-bag ошибка считается снова. Важность параметра подсчитывается путем усреднения по всем деревьям разности показателей out-of-bag ошибок до и после перемешивания значений. Параметры выборки, которые дают бóльшие значения, считаются более важными для тренировочного набора.

В таблице 5 приведены коэффициенты важности признаков, подсчитанные с помощью алгоритма случайного леса:

Признак	**Коэффициент**
Расстояние от анафора до антецедента в предложениях	0.197447
Расстояние от анафора до антецедента в словах	0.171997
Расстояние от анафора до антецедента в клаузах	0.122139
Расстояние от анафора до антецедента в абзацах	0.109338
Риторическое расстояние	0.087335
Грамматическая роль анафора	0.054079
Грамматическая роль антецедента	0.050881
Расстояние от анафора до антецедента в маркабулах	0.048543
Фразовый тип анафора	0.039568
Род	0.029817
Одушевленность антецедента	0.025816
Одушевленность	0.025146
Количество маркабул до последнего упоминания в форме полной именной группы	0.007250
Длина маркабулы-антецедента в словах	0.005625
Порядковый номер маркабулы в цепочке	0.005392
Тип местоимения антецедента	0.003206
Тип имени собственного	0.003111
Фразовый тип антецедента	0.002831
Тип атрибута	0.002465
Тип дескрипции антецедента	0.002463
Наличие количественного числительного	0.002102
Число	0.001670
Лицо	0.001345
Тип группы антецедента	0.000435
Является ли антецедент группой	0.000000

Таблица 5. Коэффициенты важности признаков согласно алгоритму случайного леса

В случае с алгоритмом случайного леса отобрать признаки не так просто, как для LASSO регрессии, где неинформативные признаки обнуляются. Здесь я элиминировала из модели признаки, коэффициент важности которых ниже 0.02, так как если удалять признаки с коэффициентом больше этого порога, то качество классификации начинает снижаться. Таким образом, были элиминированы такие признаки, как: количество маркабул до последнего упоминания в форме полной именной группы, длина маркабулы-антецедента в словах, порядковый номер маркабулы в цепочке, тип местоимения антецедента, тип имени собственного, фразовый тип антецедента, тип атрибута, тип дескрипции антецедента, наличие количественного числительного, число, лицо, тип группы антецедента и является ли антецедент группой.

В таблице 6 приведены значения аккуратности для модели, в которую не входят 13 признаков, перечисленных выше:

Алгоритм	**Аккуратность классификации для двухклассовой задачи**
Логистическая регрессия	0.8868
Градиентный бустинг	0.9008
Метод опорных векторов	0.8926
Случайный лес	0.883

Таблица 6. Аккуратность классификации для двухклассовой задачи (без признаков, исключенных методом случайного леса)

Как показывают данные, представленные в таблице, даже удаление 13 признаков из модели не ухудшает ее качества по сравнению с моделью, где использовался полный набор признаков, что говорит о том, что данные признаки были мало информативны и легко могут быть элиминированы из модели.

6. Заключение

В данном исследовании изучались факторы, влияющие на референциальный выбор. Была поставлена задача оптимизировать систему моделирования референциального выбора путем сокращения набора признаков, используемых в обучении. Для уменьшения количества признаков использовались встроенные методы отбора признаков – случайный лес и LASSO регрессия.

В данной работе удалось показать, что существует возможность использовать элиминированный набор признаков без ущерба для качества модели. Качество работы классификаторов после работы каждого из методов отбора показало, что различий в их эффективности нет. Однако стоит отметить, что метод случайного леса позволил исключить из модели 13 признаков, в то время как в ходе отбора методом LASSO регрессии было элиминировано 9 признаков. При этом, все 9 признаков, которым алгоритм LASSO регрессии приписал нулевые коэффициенты, входят в множество неинформативных признаков, выделенных с помощью случайного леса. Таким образом, можно предположить, что эти признаки являются наименее важными для модели.

Результаты исследований, описанных во втором разделе настоящей работы, противоположны результату, который был получен в ходе настоящей работы. Возможно, это объясняется тем, что в работах предыдущих лет признаки, которые были удалены из модели, выбирались эмпирически, то есть для этого не использовались никакие специальные алгоритмы.

В будущем планируется продолжить данную серию исследований с подробным вычислением параметров моделей отбора признаков для выяснения наиболее полного набора информативных признаков.

Библиография

1) Breiman L. Random forests //Machine learning. – 2001. – Т. 45. – №. 1. – С. 5-32.
2) 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.
3) Dale R. Generating referring expressions: Constructing descriptions in a domain of objects and processes// Bradford Books. – 1992.
4) Jurafsky D., Martin J. H. Speech & language processing// International Edition. – 2000.
5) Kibrik A. A. Reference in discourse//Oxford University Press. – 2011.
6) Ladha L., Deepa T. Feature selection methods and algorithms //International journal on computer science and engineering. – 2011. – Т. 3. – №. 5. – С. 1787-1797.
7) 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. – 2011.- vol. 10, pp. 458–467.
8) Müller C., Strube M. Multi-level annotation of linguistic data with MMAX2 //Corpus technology and language pedagogy: New resources, new tools, new methods. – 2006. – Т. 3. – С. 197-214.
9) Nenkova A. Entity-driven Rewrite for Multi-Document Summarization //Third International Joint Conference on Natural Language Processing. – 2008. – С. 118.
10) Tibshirani R. Regression shrinkage and selection via the lasso //Journal of the Royal Statistical Society. Series B (Methodological). – 1996. – С. 267-288.
11) Van Deemter K. Generating referring expressions: Boolean extensions of the incremental algorithm //Computational Linguistics. – 2002. – Т. 28. – №. 1. – С. 37-52.
12) Кибрик А. А., Подлесская В. И. Рассказы о сновидениях: корпусное исследование устного русского дискурса. — М.: ЯСК, 2009. — С. 736.
13) Кибрик, А. А., Линник, А. С., Добров, Г. Б., Худякова, М. В. Оптимизация модели референциального выбора, основанной на машинном обучении // Компьютерная лингвистика и интеллектуальные технологии. По материалам конференции Диалог-2012. – 2012. - vol. 11, pp. 237–246.