Семенова Наталья, МКЛ-171

Домашнее задание: Компьютерная лингвистика

Обзор статьи

Contextual Spelling Correction Using Latent Semantic Analysis

by Michael P. Jones and James H. Martin

В данной статье рассматривается подход к идентификации слов, неправильно использованных в предложении, основанный на Латентном семантическом анализе (LSA). Традиционные spell-checkers помечают слова с опечатками, однако, если слово перепутано с другим, похожим на него словом, как в случае *quite/quiet*, а также в ситуации с омофонами, значения которых часто смешивают, например, *affect/effect*, то такое слово будут считаться корректным. Проблема может быть решена посредством исследования контекста, окружающего конкретное слово. Слова, в которых часто ошибаются по одной из вышеперечисленных причин (опечатки и контекстные орфографические ошибки), авторы называют *confusion sets*. Авторы выделяют среди *confusion sets* два типа: слова, относящиеся к одной части речи, и слова, принадлежащие разным частям речи.

В статье авторы используют LSA в качестве метода правки контекстных орфографических ошибок для заданной коллекции из *confusion sets*. Первоначально этот метод возник как модель информационного поиска, однако он показал себя и в других задачах. Целью LSA является выявить “доказательства” (слова), которые показывают и открывают лежащую в основе текста семантику. Однако, поскольку в языке существуют полисеманты и синонимы, они создают “шум”. LSA призван отделить шум от данных для первого представления текста в высокоразмерном пространстве и затем сократить размерность с помощью сингулярного разложения только до самых важных пространств.

Авторы выбрали LSA, поскольку хотели протестировать его эффективность в предсказании слов на заданных предложениях и сравнить его с байесовским классификатором (система Tribayes).

Авторы взяли те же 18 *confusion sets (слова, которые путают, будем называть их “проблемными словами”)*, что и в системе Tribayes, описанной у Golding and Schabes (1996), а также в качестве материалов - корпус Brown, чтобы потом с ними же и сравниться. Для обучающей и тестовой выборки, авторы данной статьи приняли пропорцию 80% к 20%, также, как у Golding and Schabes.

Рассмотрим, как авторы применили выбранный **экспериментальный метод**. Подготовка данных для обучения происходила следующим образом. Корпус данных был создан на основе брауновского корпуса (Brown corpus), который был распарсен на отдельные предложения, которые случайным образом распределили в обучающую и тестовую выборки. Для построения LSA-пространства для каждого *confusion sets* извлекли из обучающей выборки только те предложения, которые содержали слова из этих наборов. Аналогичные действия были произведены с тестовой выборкой. Опишем кратко этот метод. Коллекция текстов представлена в формате матрицы A, где строки – термы, а столбцы – документы. Каждое значение отдельной ячейки основано на TF-IDF. Затем матрицу раскладывают с помощью сингулярного разложения на три матрицы: T – ортогональная матрица векторов термов, D’ – ортогональная матрица векторов документов, S – диагональная матрица ранка r, значения на диагонали которой называются сингулярными значениями матрицы, они расположены в порядке убывания. Если линейно перемножить эти матрицы TSD’, то получится матрица, достаточно точно приближенная к исходной матрице A. Идея состоит в том, чтобы оставить в матрице S только k наибольших сингулярных значений (первые k значений, а матрица S превратится в матрицу S’ ранка k), а в матрицах T и D’ – первые k столбцов и первые k строк, соответственно, и таким образом сократить наименее важные сингулярные значения, которые скорее всего являются шумом. Произведение получившихся матриц будет наилучшим приближением к исходной матрице A. Полученная в результате матрица определяет пространство, которое предсказывает частоту, с которой каждый терм в этом пространстве будет появляться в конкретном документе или текстовом сегменте на большой выборке из семантически близких

текстов. Новые текстовые проходы можно проектировать на пространство посредством подсчета среднего веса векторов термов, которые согласовываются со словами в новых текстах. В задаче коррекции контекстных орфографических ошибок можно генерировать вектор, который является репрезентацией для каждого текстового прохода, в каждом из которых появляется слово, которое часто путают с другим (confusion word). Подобие вектора тестового прохода и векторов слова может быть использовано для предсказывания наиболее подходящего слова, данного в контексте или тексте, в котором это слово появится.

Обучение состоит из обработки обучающих предложений и создания LSA-пространства из них. LSA требует, чтобы корпус был разбит на документы. Каждое обучающее предложение – это столбец в матрице LSA. Перед этим предложения проходят следующие трансформации: сокращение контекста, стемминг, разбиение на биграммы и присвоение весов каждому терму.

Сокращение контекста – шаг, на котором происходит усечение предложения до размера слова из *confusion sets* плюс по 7 слов с каждой стороны от него либо до конца предложения. Изначальная средняя длина предложений в корпусе Brown – 28 слов, получается, что авторы сократили контекст в 2 раза. В действительности, такой шаг совсем немного повлиял на прогнозирование LSA-модели.

Стемминг – это усечение слов до их морфологических корней с целью того, чтобы снять различные морфологические варианты одного и того же слова. Авторы попробовали разные алгоритмы стемминга, все они улучшали предсказание LSA, но в конечном варианте остановились на алгоритме Портера (1980).

Авторы разбили на биграммы те слова, которые не были удалены на шаге сокращения контекста. Биграммы формируются между всеми смежными парами слов. Биграммы рассматриваются как дополнительные термы в процессе создания LSA-пространства (занимают строку в матрице LSA).

Присвоение весов термам является попыткой увеличения важности определенных термов в высокоразмерном пространстве. Для каждого терма в каждом предложении присваиваются локальный и глобальный вес. Первый – это комбинация посчитанных строк для определенного терма в предложении и близость терма к проблемному слову (confusion word). От локального веса каждого слова берется логарифм по основанию 2. Глобальный вес присваивается каждому терму, как попытка измерить его предсказательную мощность в корпусе в целом. Авторы обнаружили, что энтропия показывает лучший результат, как глобальная мера. Кроме того, слова, которые не появлялись больше, чем в 1 предложении обучающей выборки, удалялись.

Тестирование предсказательной точности модели LSA происходило следующим образом: выбирается предложение из тестовой выборки, и место проблемного слова рассматривается как неизвестное слово, которое нужно предсказать. По одному слова из *confusion set* вставлялись на это место, и после этого предложение подвергали тем же трансформациям, что и перед обучающей выборкой, которые мы описали выше. Затем вставленное слово удалялось из предложения (но не его биграммы), поскольку их наличие оказывало влияние на сравнение, которое происходило позже. Вектор в LSA-пространстве создавался на основании результирующего терма. Сравнивая схожесть каждого вектора тестового предложения с каждым вектором проблемного слова из LSA-пространства, авторы определили предсказанные наиболее подходящие для предложения слова. Вектор подобия оценивается косинусной мерой между двумя векторами. После этого определяется пара из векторов предложения и проблемного слова с наибольшей косинусной мерой, соответствующее проблемное слово выбирается, как наиболее подходящее слово для тестового предложения. Предсказанное слово сравнили с корректным, таким образом получили правильность предсказания.

Авторы сравнивали свои результаты с полученными у Golding and Schabes, поэтому из 18 *confusion sets* у них было 7 наборов слов одной части речи и 11 наборов слов из разных частей речи. Golding and Schabes показали, что использование модели на триграммах эффективно работает для проблемных слов из разных частей речи, поэтому авторы рассматриваемой статьи сфокусировались на тех 7 наборах, содержащих слова одной части речи (для каждого набора). Остальные 11 авторы

также сравнивают с результатами Tribayes, однако делают пометку о том, что предполагали, что они не будут особо хороши. Авторы утверждают, что считают свою систему дополнительной к подобной Tribayes системе.

Авторы представляют результаты своей работы в терминах базовой предсказывающей системы (Baseline Prediction System - BPS), которая игнорирует контекст, содержащийся в тестовом предложении и всегда предсказывает проблемное слово, наиболее часто встречающееся в обучающей выборке. Авторы представляют результаты в виде таблиц. В таблице 2 представлены результаты работы LSA в задаче коррекции контекстных орфографических ошибок в сравнении с работой BPS и Tribayes. Во всех случаях, кроме сета *{amount, number},* LSA показала себя лучше, чем BPS. В среднем LSA сработала на 14% лучше на всех сетах и на 16% лучше на сетах, состоящих из слов одной части речи. Стоит отметить, что перед тем, как сравнивать LSA и Tribayes, авторы уточняют, что данное сравнение является в некотором роде косвенным, поскольку деление корпуса для данного исследования происходило не также, как в работе Golding and Schabes. В этом же состоит причина различий baseline predictor для каждой системы: корпус Brown случайным образом делился на обучающую (80%) и тестовую (20%) выборки. Однако baseline predictor в данном исследовании и у Golding and Schabes основан на одном и том же методе, поэтому можно сравнить полученные результаты, чтобы получить представление о распределении предложений, содержащих наиболее частотные слова для каждого *confusion set.*

В связи с проблемой, описанной выше, авторы приняли решение сравнивать результаты работы каждой системы с baseline score для каждой из них. Авторы получили, что LSA работает в среднем немного лучше, чем Tribayes, для тех *confusion sets,* которые содержат слова одной части речи. В то время, как Tribayes явно превосходит LSA на словах, относящихся к разным частям речи. Таким образом, LSA работает лучше, чем байесовский компонент Tribayes, но, так как он не включает в себя информацию о частях речи, он не способен работать также хорошо, как триграммы Tribayes. Тем не менее, авторы считают LSA альтернативой байесовскому классификатору для предсказаний среди слов одной части речи.

Авторы попробовали сделать настройку алгоритма для различных сетов на примере одного из них (*{amount, number}*), потому что до этого они использовали единообразную настройку для всех *confusion sets.* Посредством удаления “плохого” локального контекста (те слова, которые встречались в обучающей выборке чаще с одним из пары слов в *confusion set,* чем увеличивали вес этого слова) для этих слов в обучающей и тестовой выборке, авторы запретили его появление и на шаге предсказания, то есть этот контекст просто не рассматривался и не портил картину. Так авторы получили 32% прибавку к производительности LSA, и он показал результат на 13% лучше, чем baseline predictor.

Несмотря на неплохие результаты авторы уточняют, что все это ничего не говорит о том, хорошо ли работает техника на неотредактированном тексте. Процедура тестирования предполагает, что проблемное слово вообще не было поставлено автором, либо люди путают эти слова в 50% случаях. В заключении авторы указывают, что предсказательная точность LSA равна 91%. Они также отмечают, что на практике в 95% случаев люди используют правильное слово в рассмотренных предложениях, получается, что LSA внес 4% ошибок в процесс выбора слова.

На мой взгляд, данная статья заслуживает внимания, поскольку в ней подробно и наглядно показана работа лексико-семантического анализа в задаче spell-checking, хотя и на сегодняшний день ее можно считать несколько устаревшей, поскольку этот метод сейчас используют во многих исследованиях, иногда даже неожиданных, таких как изучение и моделирование когнитивного развития детей. Тем не менее, для меня эта статья оказалась полезной, она помогла мне в точности понять, как работает LSA и как можно применить его для задач компьютерной лингвистики.