Домашнее задание по компьютерной лингвистике: обзор статьи

Никишина Ирина
группа МКЛ171

Contextual Spelling Correction Using Latent Semantic Analysis

Michael P. Jones
and James H. Martin

1 Постановка задачи

В статье описываются возможности использования Латентно-семантического анализа (ЛСА) для контекстного исправления опечаток. Авторы убеждены, что данная задача является достаточно сложной и нетривиальной, так как слово, написанное неправильно, может быть не просто набором символов, а словом, существующим в языке, но не употребляемым в другом контексте. Кроме того, опечаткой или ошибкой также можно считать и само слово, употребленное в неверном контексте. Важно отметить, что исследователи зачастую не различают данные типы ошибок и относят их к «контекстным ошибкам».

Авторы утверждают, что на момент написания статьи подавляющее большинство линвистических систем выделяет только ошибки ии опечатки, и не учитывает слова, неверно употребленные в том или ином контексте.

Для решения данной задачи авторы предлагают использовать латентно-семантический анализ (ЛСА). Традиционно ЛСА используется в информационном поиске, однако в последнее время его применяют также для индексирования материалов конференций, для картирования интеллектуальных активов (Expert Locator) и др. Таким образом, в работе рассматриваются возможности данного метода для исправления «опечаток», а также сравнивается эффективность применения LSA с эффективностью метода, основанного на Байесовском классификаторе.

2 Описание метода

Метод, используемый в данном исследовании – латентно-семантический анализ – это способ обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами, в них встречающимися, сопоставляющий некоторые факторы (тематики) всем документам и терминам. Данный метод позволяет частично разрешить полисемию и синонимию в текстам, что достигается путем уменьшения размерности пространства, в котором располагаются вектора термов и документов.

Корпус текстов представляется в форме матрицы, в которой строки соответсвуют термам (m термов), а столбцы - документам (n документов). В каждой ячейчке на пересечении строки и столбца располагаются веса, учитывающие частоты использования каждого терма в каждом документе и участие терма во всех документах (TF-IDF). Затем к данной матрице M порядка $m \times n$ применяют сингулярное разложение (SVD) – разложение вида $M = U\Sigma V^*$ где Σ – матрица размера $m \times n$ с неотрицательными элементами, у которой элементы, лежащие на главной диагонали – это сингулярные числа (а все элементы, не лежащие на главной диагонали, являются нулевыми), а матрицы U (порядка $m \times m$ и V (порядка $n \times n$ – это две унитарные матрицы, состоящие из левых и правых сингулярных векторов соответственно (а V^* – это сопряжённо-транспонированная матрица к V). Это позволяет приближать заданную матрицу M некоторой другой матрицей M_k меньшего ранга k.

Для данной задачи текст разбивается на открывки, в которых присутсвуют слова, употребленные в неверном контексте. Сходство между вектором отрывка и вектора данного слова может быть использована для предсказания наиболее вероятного слова для данного контекста.

Материалами исследования послужили независимые подкорпуса, составленные на базе Брауновского корпуса (Kucera and Francis, 1967). Корпус был разделен на предложения, которые были случайно распределены между тренировочными данными (80%) и тестовыми (20%). Для

построения латентно-семантического пространства тренировочного корпуса были использованы только те предложения, в которых содержатся неоднозначные слова. Подобным же образом были отобраны предложения для тестового подкорпуса.

Перед тем как строить лексико-семантическое пространство каждое предложение (в матрице ЛСА оно является «документом») проходит следующие этапы предобработки:

- *уменьшение контекста* – в среднем, длина предложения составляет 28 слов, тогда как при уменьшении контекста авторы используют окно размером 7 (или до границ предложения).Данный этап значительно (почти в половину) сокращает время обработки данных, однако эффективность использования данного этапа предобработки незначительна.

- *стемминг* – процесс нахождения основы (в данном случа морфологического корня) слова для заданного исходного слова с целью распознавать одно и то же слово в любой его форме. В данном иследовании для стемминга использовался алгоритм Портера (Porter, 1980)

- *выделение биграм* – биграмы составляются из слов, оставшихся после первого этапа. Они используются в ЛСА матрице в качестве дополнительных термов.

- *подсчет весов для термов* – состоит из локального и глобального весов. Локальный вес – логарифм от произведения количества употреблений слова в предложении на «близость» слова к «неоднозначному». Глобальный вес в данном исследовании – логарифмическая энтропия (Lochbaum and Streeter, 1989).

Термы, содержащиеся в корпусе только один раз, не использовались.

Затем создается латентно-семантическое пространство, которое можно использовать для предсказания «неоднозначного» слова в предложении. Авторы рассматривали каждое предложение как «предложение с неизвестным»: необходимо было предсказать то самое «неоднозначное» слово. Каждый из вариантов по очереди вставляется на место «неизвестного» слова, затем производятся те же преобразования, что и во время тренировки алгоритма и строится латентно-семантическое пространство и ЛСА-вектор, используя термы предложения и биграмы со словами-кандитатами. Искомое слово определяется косинусным расстоянием между вектором слова-кандидата и ЛСА-вектором.

3 Результаты

Для оценки результатов были использованы 18 паронимов английского языка, взятых из Golding (1995; 1996), 7 из них – паронимы одной части речи («raise», «rise»), 11 – разных частей речи («its», «it's»). Особое внимание авторы уделяют результатам именно «неоднозначным» словам одной части речи, отмечая тот факт, что, как и предполагалось, результаты для оставшихся 11 пар слов оказались недостаточно высокими. Полученные результаты также сопоставлялись с результатами системы Tribayes (Golding and Schabes, 1996), использовавших триграмы.

В качестве baseline-метода используется система без учета контекста, выбирающее наиболее частотное «неоднозначное» слово. Система, использующая латентно-семантический анализ, в среднем увеличила результативность на 14% (а у слов одной части речи даже на 16%).

В статье авторы сравнивают свои результаты с результатами Tribayes, несмотря на то, что они неодинаково делили корпус в своих исследованиях, а также не реализовывали их методы для своих данных. Различия в baseline-методе обосновываются разным распределением предложений корпуса между тренировочной и тестовой выборкой, о чем свидетельствую разные частоты употребления слов в тренировочной и тестовой выборке. В итоге авторы статьи сравнивают различие в эффективности ЛСА и Tribayes по сравнению с собственными baseline-методами каждого исследования.

Сопоставив результаты, авторы делают вывод, что использование латентно-семантического анализа в среднем позволяет получит более высокий результат для слов одной части речи, тогда как Tribayes демонстрирует гораздо более высокие результаты чем LSA-метод для слов разных частей речи.

4 Обсуждение результатов

Выделим основные вопросы, возникшие у авторов статьи:

- параметры «неоднозначных» слов (паронимов) могут определяться отдельно для каждого сета (позволило увеличить эффективность для пары «amount»/«number» на 6%)
- контест «неоднозначных» слов часто включает в себя словоформы «of» и «the» (исключение данных форм позволяет увеличить эффективность на 13%)

Хотелось бы отметить следующие пункты касательно данной статьи:

+ в статье рассматривается (и подверждается) сама возможность использования латентно-семантического анализа для контексного исправления ошибок

- использование случайного распределения предложений корпуса между тренировочной и тестовой выборкой в данной статье и статье Golding и Schabes не позволяет более явно, точно и очевидно сравнить эффективность LSA и Tribayes

- неочевидно соотношение количества «неоднозначных» слов в тренировочной и тестовой выборках

- ограниченная применимость метода (только для слов одной части речи)

Статьи, упоминаемые авторами

1. Andrew R. Golding. 1995. A Bayesian hybrid method for context-sensitive spelling correction. In Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA.
2. Andrew R. Golding and Yves Schabes. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Clara, CA, June. Association for Computational Linguistics.
3. Henry KuSera and W. Nelson Francis. 1967. Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.
4. Peter W. Foltz. 1995. Improving human-proceedings interaction: Indexing the CHI index.In Human Factors in Computing Systems: CHI'95 Conference Companion, pages 101-102. Associations for Computing Machinery (ACM), May.
5. Karen E. Lochbaum and Lynn A. Streeter. 1989. Comparing and combining the effectiveness of Latent Semantic Indexing and the ordinary vector space model for information retrieval. Information Processing CJ Management, 25(6):665-676.
6. Susan W. McRoy. 1992. Using multiple knowledge sources for word sense disambiguation. Computational Linguistics, 18(1):1-30, March.
7. M. F. Porter. 1980. An algorithm for suffix stripping. Program, 14(3):130-137, July.