

Обзор статьи «Resolving Ambiguities in Biomedical Text With Unsupervised Clustering Approaches»

ПетрГУ, 508 группа, Ярышкина Е.А.

11 ноября 2014 г.

В статье [1] изучаются уже существующие методы кластеризации без учителя и их эффективность для решения лексической многозначности при обработке текстов по биомедицине. Решение проблем лексической многозначности в данной области включает в себя не только традиционные задачи присвоения ранее определенных смысловых значений для терминов, но так же и обнаружения новых значений для них, ещё не включённых в данную онтологию.

Авторы описали методологию метода решения лексической многозначности без учителя, учитываемые лексические признаки и наборы экспериментальных данных. В качестве оценки эффективности алгоритмов кластеризации текста была предложена F-мера.

Подход для решения поставленной задачи – это разделение контекстов (фрагментов текста), содержащих определенное целевое слово на кластеры, где каждый кластер представляет собой различные значения целевого слова. Каждый кластер состоит из близких по значению контекстов. Задача решается в предположении, что используемое целевое слово в аналогичном контексте будет иметь один и тот же или очень похожий смысл.

Процесс кластеризации продолжается до тех пор, пока не будет найдено предварительно заданное число кластеров. В данной статье выбор шести кластеров основан на том факте, что это больше, чем максимальное число возможных значений любого английского слова, наблюдаемое среди данных (большинство слов имеют два-три значения). Нормализация текста не выполняется.

Данные в этом исследовании состоят из ряда контекстов, которые включают данное целевое слово, где у каждого целевого слова вручную отмечено – какое значение из словаря было использовано в этом контексте. Контекст – это единственный источник информации о целевом слове. Цель исследования – преобразовать контекст в контекстные вектора первого и второго порядка [2]. Контекстные вектора содержат следующие «лексические свойства»: биграммы, совместную встречаемость и совместную встречаемость целевого слова. Биграммами являются как двухсловные словосочетания, так и любые два слова, расположенные рядом в некотором тексте. Для лингвистических исследований могут быть полезны только упорядоченные наборы биграмм [3].

Экспериментальные данные – это набор NLM WSD [4] (NLM – национальная библиотека медицины США), в котором значения слов взяты из UMLS (единая система медицинской терминологии). UMLS имеет три базы знаний:

- Метатезаурус включает все термины из контролируемых словарей (SNOMED-CT, ICD и другие) и понятия, которые представляют собой кластера из терминов, описывающих один и тот же смысл.
- Семантическая сеть распределяет понятия на 134 категории и показывает отношения между ними. SPECIALIST-лексикон содержит семантическую информацию для терминов Метатезауруса.
- Medline – главная библиографическая база данных NLM, которая включает приблизительно 13 миллионов ссылок на журнальные статьи в области науки о жизни с уклоном в биомедицинскую область.

Авторы успешно проверили по три конфигурации существующих методов (PB – Pedersen and Bruce [5], SC – Schütze [6]) и оценили эффективность использования SVD (сингулярное разложение матриц). Методы PB основаны на контекстных векторах первого порядка – признаки одновременного присутствия целевого слова или биграммы. Рассчитывается среднее расстояние между кластерами или применяется метод бисекций. PB методы подходят для работы с довольно большими наборами данных. Методы SC основаны на представлениях второго порядка – матрицы признаков одновременного присутствия или биграммы, где каждая строка и столбец – вектор признаков первого порядка данного слова. Так же

рассчитывается среднее расстояние между кластерами или применяется метод бисекций. SC методы подходят для обработки небольших наборов данных.

Метод SC2 (признаки одновременного присутствия второго порядка, среднее расстояние между элементами кластера в пространстве подобия) с применением и без SVD показал лучшие результаты: всего 56 сравниваемых экземпляров, в 47 случаях метод SC2 показал наилучшие результаты, в 7 случаях результаты незначительно отличаются от других проверяемых методов.

Все эксперименты, указанные в исследовании, выполнялись с помощью пакета SenseClusters [7]. В ходе исследования было проведено два эксперимента для разных наборов данных. Маленький тренировочный набор – это набор NLM WSD, который включает 5000 экземпляров для 50 часто встречаемых неоднозначных терминов из Метатезауруса UMLS. Каждый неоднозначный термин имеет по 100 экземпляров с указанным вручную значением. У 21 термина максимальное число экземпляров находится в пределах от 45 до 79 экземпляров. У 29 терминов число экземпляров от 80 до 100 для конкретного значения. Стоит отметить, что каждый термин имеет категорию «ни одно из вышеупомянутых», которая охватывает все оставшиеся значения, не соответствующие доступным в UMLS. Большой тренировочный набор является реконструкцией «1999 Medline», который был разработан Weeber [8]. Были определены все формы из набора NLM WSD и сопоставлены с тезисами «1999 Medline». Для создания тренировочного набора экземпляров использовались только те тезисы из «1999 Medline», которым было найдено соответствие в наборе NLM WSD.

Использование целиком текста аннотации статьи в качестве контекста приводит к лучшим результатам, чем использование отдельных предложений. С одной стороны, большой объем контекста, представленный аннотацией, дает богатую коллекцию признаков, с другой стороны, в коллекции WSD представлено небольшое число контекстов.

Список литературы

- [1] Savova G. Resolving ambiguities in biomedical text with unsupervised clustering approaches. University of Minnesota Supercomputing Institute Research Report, 2005.

- [2] Епрев А. С. Применение контекстных векторов в классификации текстовых документов. 2010. <http://jre.cplire.ru/iso/oct10/1/text.html>.
- [3] Аверин А.Н. Разработка сервиса поиска биграмм // Труды международной конференции «Корпусная лингвистика–2006». СПб: С.Петербург. ун-та, 2006.
- [4] UMLS Terminology Services (UTS). <http://umlsks.nlm.nih.gov/kss/servlet/Turbine/template>.
- [5] Pedersen T., Bruce R. Distinguishing word senses in untagged text. Proc.EMNLP.Providence, RI, 1997.
- [6] Schutze H. Automatic Word Sense Discrimination. Computational Linguistics, vol. 24, number 1., 1998.
- [7] SenseClusters. <http://senseclusters.sourceforge.net>.
- [8] Weeber M., Mork J., Aronson A. Developing a test collection for biomedical word sense disambiguation. Proc. AMIA., 2001.

Ярышкина Екатерина Александровна
 Студентка
 Математический факультет
 Петрозаводский государственный университет
 пр-кт Ленина, 33, Петрозаводск, Республика Карелия
 +7 (8142) 71-10-78
kate.rysh@gmail.com

Yaryshkina Ekaterina Alexandrovna
 Student
 Faculty of Mathematics
 Petrozavodsk State University
 Prospect Lenina, 33, Petrozavodsk, Republic of Karelia
 +7 (8142) 71-10-78
kate.rysh@gmail.com