Word Sense Inspector – induced word senses exploration toolkit

Zhikov V. B. (s02180164@stud.cs.msu.ru), Arefyev N. V. (nick.arefyev@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

We present a system that allows exploration of word senses induced from unlabeled text corpora for the purposes of linguistic research. The system performs word sense induction by clustering lexical substitutes generated by a neural language model for an ambiguous word. It provides convenient user interface to study clusters of the word usages, often corresponding to different senses of this word. The system allows exploring word senses expressed in a given corpus, as well as the difference between word senses in two corpora. This may be useful for studying lexical semantic change over time or between domains.

Key words: Word Sense Induction, Lexical Semantic Change Detection, Lexical Substitution, Neural Language Models

Word Sense Inspector – инструмент исследования автоматически обнаруженных значений слов

Жиков В. Б. (s02180164@stud.cs.msu.ru), Арефьев Н. В. (nick.arefyev@gmail.com)

МГУ им. М.В. Ломоносова, Москва, Россия

Мы представляем систему, позволяющую исследовать значения слов, выявленные автоматически на основе неразмеченных текстовых коллекций, для проведения лингвистических исследований. Система обнаруживает значения слов с помощью кластеризации лексических подстановок, сгенерированных нейронной языковой моделью для многозначного слова. Предоставляется удобный пользовательский интерфейс для изучения выявленных кластеров словоупотреблений, обычно соответствующих разным значениям многозначного слова. Система позволяет исследовать значения слова, встретившиеся в некотором корпусе, а также разницу в значениях слова, представленных в двух корпусах. Система может использоваться для изучения изменения значений слов со временем, а также разницы в значениях слов в двух предметных областях.

1 Введение

В данной работе мы представляем программную систему, предназначенную для помощи в проведении лингвистических исследований. Функционал системы позволяет исследователю увидеть, на какие семантические кластеры можно разделить употребления тех или иных слов в корпусе, а также оценить, как сильно различаются контексты употребления слов в двух разных корпусах.

2 Обзор литературы

В публикациях [1] и [2] для решения задачи выявления значений слов (Word Sense Induction) использовался подход на основе кластеризации bag-of-words векторов лексических подстановок, сгенерированных нейросетевыми языковыми моделями. Общая идея метода состоит в следующем: для некоторого контекста употребления интересующего нас слова в целевом корпусе это целевое слово заменяется специальным токеном

mask (или последовательностью из нескольких таких токенов). В таком виде контекст подается на вход модели, которая предсказывает возможные замены для токена *mask* (или последовательности таких токенов) в этом контексте. Наиболее вероятные (по предсказанию модели) замены будем называть контекстными (или лексическими) подстановками.

В реализации системы мы применяем данный подход, используя маскированную языковую модель XLM-R [3]. Это мультиязычная языковая модель, способная работать со 100 различными языками. Это определило наш выбор модели, т.к. хотелось дать системе возможность работать с разными языками. Примеры, на которых показывается вывод системы, взяты из датасета RUSSE'2018 Word Sense Induction [5] для русского языка.

3 Описание системы

Описываемая система имеет два режима работы. В первом режиме на вход системе подается текстовый корпус и набор целевых слов. В качестве опционального параметра могут быть переданы также данные разметки вхождений целевых слов в корпус по значениям для сравнения предсказаний системы с этой разметкой. Выводом системы является набор гистограмм, демонстрирующих результаты кластеризации примеров, а также описание кластеров, включающее в себя наиболее характерные для кластера подстановки и несколько конкретных примеров использования слова (контекст + подстановки), которые система отнесла к данном кластеру.

Во втором режиме работы система получает вместе с набором ключевых слов не один, а два корпуса. В таком случае, система кластеризует примеры из двух корпусов по одному набору кластеров и продемонстрирует, как отличается распределение примеров использования слова по кластерам в зависимости от корпуса. Для каждого кластера будут отдельно приводиться значимые подстановки и примеры из одного и из другого корпуса. При задании порогов k и n система способна выделять слова, имеющие специфические значения в одном из двух корпусов. Значение будет считаться специфическим для определенного корпуса, если соответствующий ему кластер будет содержать не менее чем n примеров из этого корпуса и не более чем k примеров из другого. k и n являются гиперпараметрами, задающими чувствительность системы

Цикл работы системы состоит из следующих этапов:

- **Загрузка данных**
  На этом этапе происходит загрузка корпуса, поиск в нем вхождений целевых слов и формирование примеров типа [контекст + целевое слово + позиции символов целевого слова в контексте].

- **Загрузка лексических подстановок**
  Поскольку генерация подстановок с использованием нейронных языковых моделей весьма ресурсоемка (требуется сервер с графическим ускорителем GPU), в систему загружаются заранее сгенерированные подстановки. В результате генерируются наборы подстановок, характеризующие конкретные примеры.

- **Обработка подстановок**
  На этом этапе сгенерированные наборы подстановок подвергаются очистке от мусора. Удаляются специальные символы, пунктуация, словосочетания и бессмысленные части слов. Кроме того, подстановки лемматизируются.

- **Векторизация**
  Следующий этап - получение по списку подстановок bag of words векторов для каждого примера вхождения каждого из целевых слов.

- **Кластеризация**
  Полученные векторы кластеризуются. Количество кластеров может быть задано вручную или выбрано при помощи метрики silhouette.

Для генерации подстановок система использует языковую модель XLM-R [3]. В качестве алгоритма кластеризации был выбран алгоритм иерархической кластеризации с average linkage. Система реализована на языке Python версии 3.7.3. Веб-интерфейс создавался при помощи библиотеки Flask.

Для каждого кластера, на которые разбились примеры слова, система выводит характерные подстановки, упорядочивая их тремя разными способами:

- **Top P** - подстановки с максимальной условной вероятностью $P(s) = Count^s_{cluster}/Size_{cluster}$ где $Count_s$ - количество примеров в данном кластере, в которых встретилась подстановка, $Size_{cluster}$ - количество примеров в кластере

- **Top PMI** - подстановки с наибольшим значением PMI (Pointwise mutual information) $PMI(s) = \frac{Count^s_{cluster}/Size_{cluster}}{Count^s/Size}$, где $Count^s$ - общее количество примеров, куда вошла подстановка s, $Size$ - общее количество примеров в датасете

- **Top P (sorted by PMI)** - подстановки, входящие в top-100 по вероятности, отсортированные в соответствии с PMI

Для каждого из способов сортировки выводится top-15 подстановок.

4 Примеры работы системы

В качестве примеров работы системы рассмотрим несколько слов из размеченного для задачи WSI датасета RUSSE'2018 bts-rnc (train) [5]. В качестве оценки качества используется метрика Adjusted Rand Index (ARI) [6]. На тренировочной подвыборке этого датасета система достигает ARI 0.59 (при фиксированном количестве кластеров, равном трем). На рис. 1 изображены значения этой метрики для слов датасета. При выборе числа кластеров по метрике silhouette система дает значение ARI 0.5 при том же наборе гиперпараметров.

Во втором режиме работы - для задачи обнаружения изменения значений слов - на варианте русскоязычного датасета RuMacro [4] система демонстрирует точность 0.68 (при правильном подборе порогов n, k). На датасетах, опубликованных в контексте соревнования SemEval'2020, система показала следующие результаты по точности: 0.708 для немецкого языка, 0.649 для английского, 0.375 для латыни и 0.742 для шведского.

Слово *лира*, ARI=0.95 На рис. 2(а) видно, что два первых кластера соответствуют разным значениям *лира*. Судя по таблице 2, в первых кластер попали примеры использования слова *лира* в значении «музыкальный инструмент», а во второй - в значении «валюта». В третий кластер попал лишь один пример для первого значения. Т.к. целевое количество кластеров было фиксированным, кластер не мог остаться пустым, поэтому, в него попал наиболее выбивающийся из общего контекста пример.

На рис. 1(б) приводится гистограмма расстояний между векторами подстановок. Синим цветом показано распределение расстояний между векторами примеров для одного и того же значения, оранжевым цветом - для примеров, соответствующих разным значениям. Можно видеть явное различие двух распределений.

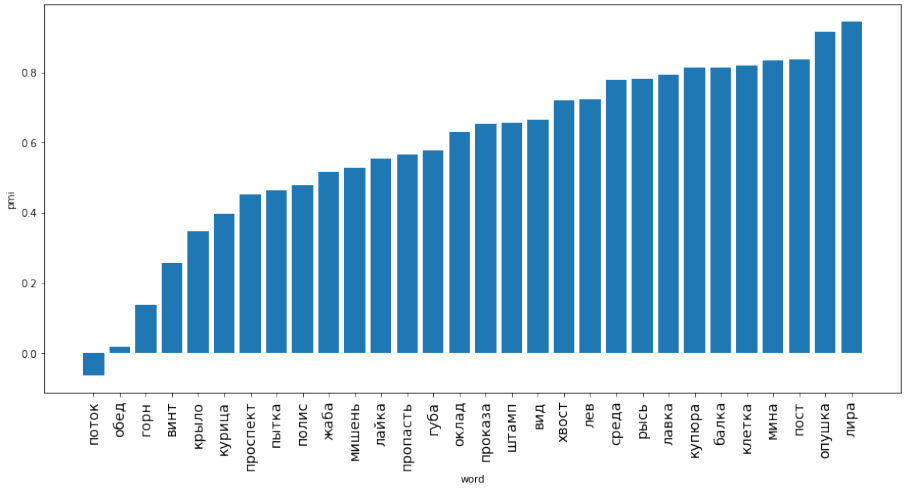
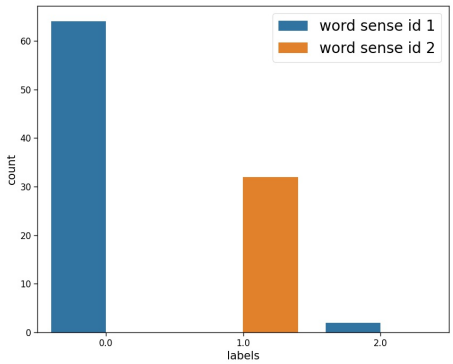
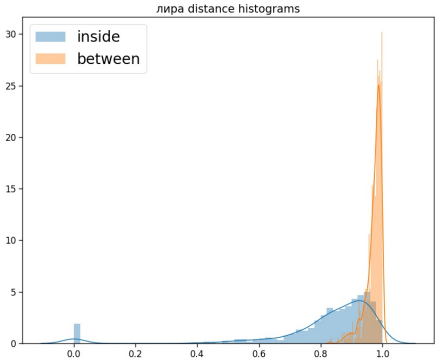


Рис. 1: ARI по словам датасета



(a) *лира* - распределение векторов, соответствующих примерам разных значений, по кластерам, предсказанным системой



(b) *лира* - гистограмма расстояний между векторами - внутри одного значения и между разными значениями

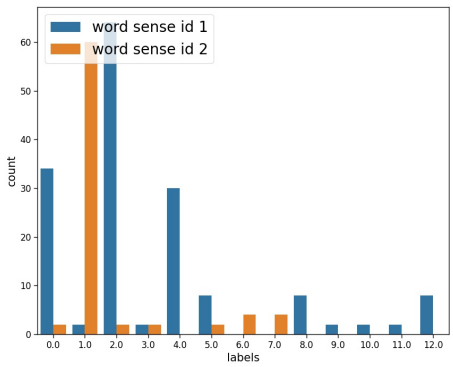
Рис. 2: Анализ слова *лира*

Таблица 1: Наиболее частотные подстановки для кластеров слова *лира*

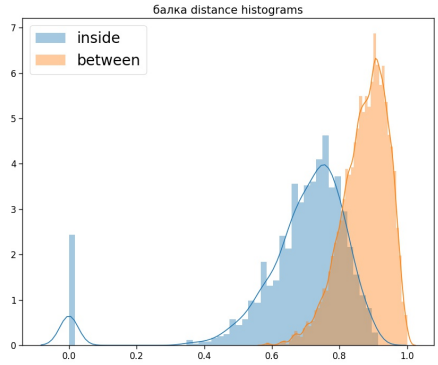
кластер	top P	top PMI	top P (sorted by PMI)
кластер 0	стих 0.59, песня 0.56, муза 0.56, птица 0.53, мелодия 0.53	пика 0.43, грусть 0.43, современность 0.43, корабль 0.43, солдат 0.43	нота P=0., PMI=0.43,груша P=0., PMI=0.43,горка P=0., PMI=0.43,стрела P=0., PMI=0.43,гитара P=0., PMI=0.43,поэт P=0., PMI=0.43
кластер 1	доллар 1.00, фунт 1.00, франк 1.00, копия 1.00, крона 0.8	us 1.12, контракт 1.12, ус 1.12, рэп 1.12, сант 1.12, сдача 1.12	ус P=0.50, PMI=1.12, вилл P=0.50, PMI=1.12, евро P=0.50, PMI=1.12, депозит P=0.50, PMI=1.12, франк P=1.00, PMI=1.12

Слово *балка*, ARI=0.25

В данном примере количество кластеров было выбрано по silhouette score. Как можно видеть на рис. 3(a), кластеров выделилось гораздо больше, чем выделено значений у этого слова.



(a) *балка* - распределение векторов, соответствующих примерам разных значений, по кластерам, предсказанным системой



(b) балка - гистограмма расстояний между векторами - внутри одного значения и между разными значениями

Рис. 3: Анализ слова *балка*

Рассмотрим наиболее объемные кластеры - 0, 1, 2 и 4.

Таблица 2: Наиболее частотные подстановки для кластеров слова *балка*

кластер	top P	top PMI	top P (sorted by PMI)
кластер 0	труба 1.00, коробка 1.00, опора 1.00, доска 0.94, панель 0.88	ракета 1.88, деревушка 1.88, копия 1.88, метка 1.88, железка 1.77	железка P=0.44, PMI=1.77, строчка P=0.44, PMI=1.77, корочка P=0.38, PMI=1.75, броня P=0.33, PMI=1.73, проводка P=0.55, PMI=1.62
кластер 1	лес 0.96, скал 0.93, река 0.90, холм 0.90, гора 0.90	исток 1.34, пешка 1.34, пешком 1.34, тайга 1.34, равнина 1.34	тайга P=0.35, PMI=1.34, равнина P=0.41, PMI=1.34, дупло P=0.35, PMI=1.34, карьера P=0.35, PMI=1.34, пустота P=0.38, PMI=1.34
кластер 2	камень 1.00, блок 1.00, ступень 0.96, доска 0.96, коробка 0.93	лиана 1.28, скат 1.28, стенание 1.28, листовой 1.28, витраж 1.28	сруб P=0.45, PMI=1.15, металлический P=0.42, PMI=1.08, крепление P=0.75, PMI=1.06, контур P=0.57, PMI=1.04, дрова P=0.78, PMI=1.04
кластер 4	крыша 1.00, опора 1.00, стена 1.00, труба 1.00, блок 0.93	материал 2.07, деталь 2.07, дверинять 2.07, паз 2.07, конструкция 1.98	деталь P=0.46, PMI=2.07, конструкция P=0.73, PMI=1.98, основание P=0.53, PMI=1.95, платформа P=0.40, PMI=1.91

На первый взгляд, разделение на кластеры не столь очевидно. Наиболее вероятные подстановки для кластеров 0 и 4 сильно пересекаются. Одно и то же значение разбилось на несколько кластеров. Однако если посмотреть на примеры контекстов, выяснится, что в нулевой кластер входят преимущественно примеры, связанные с балками в конструкции самолета: *...разрушения основной титановой балки. В практике крушений самолетов..., ...«реактивный момент», когда задний винт сносит балку... и т.п.* В четвертом же кластере встречаются примеры, относящиеся к железнодорожному транспорту: *...По вагонам, по колесным парам, надрессорным балкам..., ...Цельнометаллическое тело, усиленное передними и задними поперечными балками...* В обоих

кластерах, однако, встречаются примеры и из других предметных областей, что показывает, что векторы подстановок не всегда в состоянии в достаточной мере разделить вхождения по семантике.

5 Заключение

В данной работе была представлена программная система для исследования автоматически обнаруженных значений слов. Данная система применима для исследования типичных значений некоторых слов в языке, а также различий в значениях слов, возникающих с течением времени или наблюдающихся в разных предметных областях. Применение этой системы может быть интересно исследователям в области лексикографии. В перспективе планируется реализовать заменяемость отдельных компонент системы, в том числе алгоритма кластеризации, алгоритмов решения задачи WSI и т.п.

References

Список литературы

[1] Asaf Amrami and Yoav Goldberg. Towards better substitution-based word sense induction. *ArXiv*, abs/1905.12598, 2019.

[2] Nikolay Arefyev, Boris Sheludko, and Alexander Panchenko. Combining lexical substitutes in neural word sense induction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 62–70, Varna, Bulgaria, 09 2019. INCOMA Ltd.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[4] Elizaveta Kuzmenko and Andrey Kutuzov. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. In *Quantitative approaches to the russian language*, pages 105–122. Routledge, 2017.

[5] Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 547–210, Moscow, Russia, 2018. RSUH.

[6] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.