**Реферат статьи**

***Yasheng Wang, Yang Zhang, Bing Liu***

**Sentiment Lexicon Expansion Based on Neural PU Learning, Double Dictionary Lookup, and Polarity AssociationSentiment Lexicon Expansion Based on Neural PU Learning, Double Dictionary Lookup, and Polarity Association**

Рассмотрев посты в социальной сети Weibo, Yasheng Wang, Yang Zhang и Bing Liu выяснили, что многие китайские оценочные слова не включены в словари, потому что эти слова слишком новые или употребляются в письменной речи более “низкого стиля”. Авторы поставили перед собой задачу расширить sentiment словарь китайского языка за счёт машинного обучения. Это задача PU Learning.

Что такое PU (Positive/Unlabeled) Learning? PU Learning - это задача где есть выборка, часть объектов из которой размечены как принадлежащие к одному классу (P), а все остальные - не размечены вообще (U). При этом подразумевается, что среди неразмеченных объектов есть как принадлежащие к этому классу, так и не принадлежащие к нему. Задача PU Learning состоит в том, чтобы построить такой классификатор, который разделял бы тестовую выборку на объекты, принадлежащие к заданному классу и не принадлежащие к нему. Отличие от бинарной классификации состоит в том, что мы сначала размечаем часть наиболее “надёжных” примеров в тестовой подвыборке и на них обучаем классификатор.

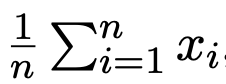
В отличие от предыдущих подходов к решению задачи PU авторы использовали в качестве классификатора не SVM (Support Vector Machine, метод опорных векторов), а Augmented Multilayer Peceptron. Авторы решили выделить как отдельный класс не Positive, а Unlabeled и потом постепенно исключать Positive из Unlabeled. Этот метод авторы назвали SE-AMP-DL (*Spy-based Elimination of* P *class instances using AMP and Double dictionary Lookup*).

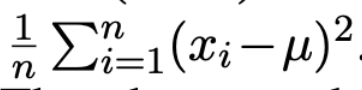
Обычно оценочную лексику ищут либо с применением словарей либо с применением корпусов. Применяя словари, мы ищем антонимы и синонимы оценочных слов, и похожие на них слова по расстоянию в Word-Net. Если использовать корпус, важно обращать внимание на союзы (“красивый и красный” - оба прилагательных положительные, “красивый но красный” - положительно только красивый). Одна из важных проблем, возникающих при корпусном подходе - слова, которые несут разную оценку в разных контекстах.

Традиционным методом для решения задачи PU Learning является Spy SVM. Его примерная схема действий такова: случайным образом выбираются 10 процентов Positive слов и помещаются в Unlabeled массив. Это слова-шпионы. После этого запускается SVM, который тренируется доставать все остальные Positive слова из массива. Всем словам приписывается вероятность оценочности, а потом определяется порог попадания в оценочные, который позволяет идентифицировать слова-шпионы.

Авторы предлагают свой подход к решению данной-задачи – SE-AMP (Spy-based Elimination of P class instances using AMP). Основные его отличия от S-SVM заключаются в следующем: 1) вместо SVM используется AMP 2) после каждой эпохи обучения из Unlabeled-массива (объединённого с массивом слов-шпионов) удаляются слова, распознанные как принадежащие к классу P. Также авторы предлагают объединить свой метод с двойным поиском по словарю (DL, Double Lookup), таким образом получая метод SE-AMP-DL (SE-AMP and Double Lookup). Метод двойного поиска по словарю заключается в том, что для каждого слова из словаря достаётся его определение. Затем определение классифицируется при помощи ранее обученного классификатора предложений. Если определение было классицифицировано как положительное, то далее классификатор применяется к каждому слову в определении. Если их определения классифицируются как положительное, то слово, на которое мы смотрели в начале признаётся положительным.

Использование AMP вместо SVM помогает значительно улучшить результаты использования алгоритма. Архитектура сети содержит три слоя. Первый принимает в себя вектор слова и даёт им пятидесяти размерный выход. Второй принимает выход первого и сокращает количество размерностей до двух. Нейроны и в первом и втором слое активируются функцией RELU (все веса меньше нуля автоматически становятся нулями). Выход второго слоя конкатенирует с векторами, зависящими от частей речи (для существительного - [1, 0, 0, 0, 0], для глагола - [0, 1, 0, 0, 0], и т.д. для прилагательных, наречий и прочих), и получается вектор с семью размерностями. Функция активации последнего слоя - сигмоида. Важно сначала сократить размерность вектора с двухсот до двух, и только после этого добавлять информацию о частях речи, чтобы их использование было максимально эффективно. SE-AMP удаляет из Unlabeled (которые маркируются +1) все Positive (-1) и получается модель, которая классифицирует всех Unlabeled и всех “шпионов”. Чаще всего модель делают так, что она считает, что есть определённый процент Unlabeled слов, которые надо классифицировать как Positive. Авторы же используют порог, который определяется с помощью гауссианы. Параметры гауссианы получают методом максимального правдоподобия:

μ = 

σ2 = 

Порог (θ) равен μ + σ. Он достаточно жёсткий и только веса тех слов, которые абсолютно точно Positive, могут его преодолеть.

Далее следует двойной этап работы со словарём. Каждое слово, по весам проходящее выше порога, находят в словаре и потом находят все Positive слова из его определения. Если такие слова есть, для каждого находится определение в словаре и выбираются все Positive слова из этого определения. Если такие слова есть, изначальное слово тоже оценочное. Проверка словарём должна быть двойная, потому что некоторые оценочные слова на самом деле оценочные не во всех контекстах. Если мы предполагаем, что в словарном объяснении оценочных слов будут другие Positive слова, надо проверить, что другие слова на самом деле находятся в списке Positive.

Чтобы понять, отрицательная или положительная оценка, используется структура Китайского языка. Оценка слова связана с оценками иероглифов, которые его составляют. Каждому иероглифу приписывается вектор - доля Positive слов с этим иероглифом, которые имеют положительную оценку, и доля Positive слов с этим иероглифом, которые имеют отрицательную оценку. Наивный байесовский классификатор учится распределять слова согласно векторам их иероглифов. Новым иероглифам приписывается вектор (0.5; 0.5).

Изначально у авторов был словарь Dalian University of Technology, Information Retrieve Lab, в котором 27466 Positive слов. Корпусом служил Weibo (китайский Twitter), где примерно 4.4 миллиона постов. Stanford Chinese word segmenter собирает части речи всех слов и дробит предложения на части. Wordvec находит словам двухсотмерные эмбеддинги. Случайным образом были выбраны 200 000 сообщений. В сообщениях было примерно 55000 значимых слов, из них 5000 - Positive. Для валидации случайным образом были отобраны 300 Positive слов и 700 Unlabeled (предполагалось, что они не Positive). Носители языка проверили деление и выделили 200 Positive слов, из них 80 с положительной оценкой и 120 с отрицательной. 500 Positive слов были выбраны “шпионами”.

Были опробованы 7 методик, среди них были и просто SVM, и просто двойной поиск по словарю и поиск слов определённой полярности (PMI с позитивным порогом вхождения для положительных определений и негативным для отрицательных). Лучше всего по точности, полноте и F-мере оказались методики, использующие AMP, хотя довольно близких результатов удалось добиться и с помощью PMI. SE-AMP через много итераций позволяет значительно улучшить результаты, SE-AMP-DL улучшает ещё, но совсем немного, потому что многих слов нету в словаре и это сокращает список Positive. Определение полярности работает совсем плохо.

“A typical PU learning algorithm works by first identifying a small set of reliable N class examples (RN) from the unlabeled set U and then running a supervised learning method (e.g., SVM) iteratively to add more and more data to the RN set to finally build a classifier (Liu, 2011)”