РАЗРЕШЕНИЕ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ ПРИ ПОМОЩИ ЧАСТИЧНОГО ОБУЧЕНИЯ

Тесленко Д. А.[1] (teslenkoden@gmail.com),
Усталов Д. А.[2] (dmitry@informatik.uni-mannheim.de)

[1] Уральский федеральный университет, Екатеринбург, Россия
[2] Университет Мангейма, Мангейм, Германия

Abstract. In this paper, we present a semi-supervised approach for word sense disambiguation. We use a relatively small amount of manually annotated contexts to automatically expand the training set over non-annotated contexts. The expanded training set is then used for training a classifier that discriminates meanings of the target word. Our experiments on the RUSSE'2018 shared task indicate that our approach, although its simplicity, shown the 4th best result on the "bts-rnc" dataset.

Keywords: word sense disambiguation, semi-supervised learning, bootstrapping, lexical semantics.

1. Введение

Явление лексической многозначности заключается в том, что слово в тексте может быть употреблено в различных значениях [3]. Разрешение многозначности — открытая задача искусственного интеллекта, которая сложна не только для машин, но и для людей. Методы разрешения лексической многозначности (англ. *word sense disambiguation*) применяются в информационном поиске [11], машинном переводе, при построении вопросно-ответных систем, и т. д. В общем виде, задача формулируется следующим образом: для заданного слова w и его контекста C определить в каком из известных значений M слово w употреблено в контексте C. Данная работа посвящена применению частичного обучения с учителем для решения этой задачи.

Подход к разрешению многозначности на основе частичного обучения с учителем первоначально описан Д. Яровским в 1995 году [9]. В данном подходе предложено использовать небольшое количество вручную размеченных «посевных» коллокаций для *бутстреппинга* — автоматического расширения обучающей выборки по увеличению критерия уверенности системы в правильности такого расширения. Недостатком такого подхода является невозможность определить причину истинного или ложного ответа системы для целевого слова в контексте, а также ограниченность набора признаков в рамках допущения «одно значение на коллокацию».

Система *IMS* (сокр. англ. *It Makes Sense*) для английского языка основана на обучении с учителем: задача разрешения многозначности сводится к задаче классификации на один из нескольких классов [10]. Основным недостатком системы *IMS* является требование доступности большого количества семантически-размеченных контекстов для построения классификатора. Разметка этих данных является трудной задачей.

В данной работе мы предлагаем совместить эти два подхода. Как и в системе *IMS*, мы используем постановку задачи разрешения лексической многозначности в виде машинного обучения с учителем. С целью снижения трудозатрат при подготовке размеченных данных, мы стараемся расширить обучающую выборку новыми данными, близкими к уже размеченным «посевным» данным в пространстве признаков. Таким образом, мы представляем новый метод разрешения многозначности с учителем, осуществляющий расширение обучающей выборки.

2. Метод разрешения лексической многозначности при помощи частичного обучения

Предлагаемый подход основан на допущении о том, что похожие контексты соответствуют одному и тому же *независимому* значению слова. Например, слово «программа» в контексте «на уроке информатики мы написали программу» имеет такое же значение, как и в контексте «учитель информатики показал, как правильно написать программу». В другом контексте значение слова отличается: «профессиональные артисты балета, уникальные постановки, световое и звуковое оформление ставит новую программу в один ряд с лучшими цирковыми представлениями» из-за разных контекстов.

2.1. Общая схема подхода

На входе предлагаемый подход (рис. 1) получает размеченные S и неразмеченные контексты N целевого слова w, причём $|N| \gg |S|$. Под *разметкой* подразумевается ручное выставление каждому контексту $C \in S$ в соответствие некоторого значения из множества значений M данного слова.

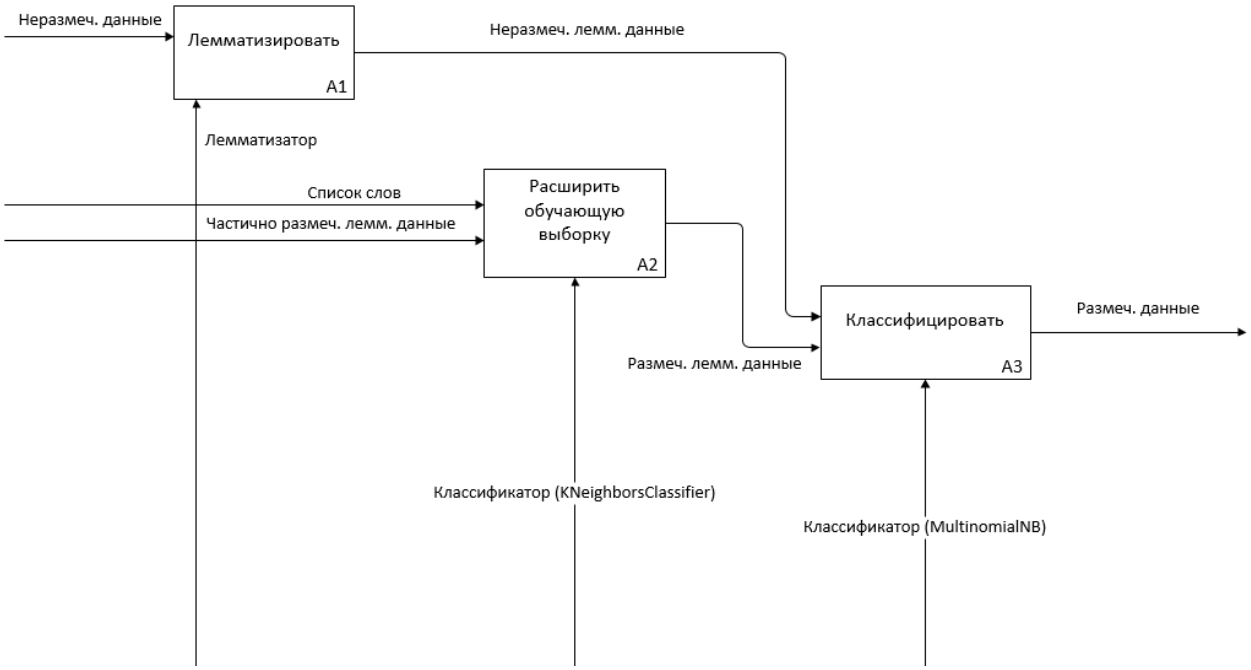

Рис. 1. Общая схема подхода

Подход состоит из двух этапов. Сначала производится построение расширенной обучающей выборки F, включающей все размеченные контексты из S и автоматически разрешённые контексты из N (раздел 2.2). Затем на основе F обучается классификатор, используемый для предсказания значения слова. Во всех случаях в качестве признаков используются вхождения слов и коллокации в контекстах; измерение значений признаков осуществляется при помощи меры tf–idf. Для каждого слова подход выполняется отдельным и независимым образом.

2.2. Автоматическое расширение обучающей выборки

Автоматическое расширение обучающей выборки основано на постепенном построении классификатора значений целевого слова и приведено в виде псевдокода в Алгоритме 1. Классификатор P осуществляет оценку распределения вероятности принадлежности контекста $C \in N$ слова w одному из значений $m \in M$ на основании обучающей выборки S, что будет обозначаться как $P(C \to m \mid S)$. Обученный классификатор запускается на неразмеченных данных и оценивает вероятность принадлежности каждого контекста к

одному из значений M. Если эта вероятность оказывается выше некоторого порога $0 < t \leq 1$, то контексту присваивается метка соответствующего значения и он добавляется в обучающую выборку; иначе контекст пропускается. Затем классификатор обучается заново на расширенной выборке и вновь оценивает вероятность принадлежности контекстов. Процедура повторяется до тех пор, пока не останется необработанных контекстов в N, которые поддаются обработке. На полученной расширенной обучающей выборке F обучается новый классификатор, используемый для предсказания значения слова w.

Алгоритм 1. Частичное обучение с учителем

Вход: слово w, множество значений M слова w, обучающая выборка S дляw, неразмеченная выборка Nдля w, классификатор Pдля w, пороговое значение t.

Выход: расширенная обучающая выборка F.

$F \leftarrow S$
while $N \neq \emptyset$ **do**
    $\Delta_F \leftarrow \emptyset$
    for each $C \in N$ **do**
        $p \leftarrow \max_{m \in M} P(C \rightarrow m \mid F)$
        if $p > t$ **then**
            $m^* \leftarrow \arg\max_{m \in M} P(C \rightarrow m \mid F)$
            $\Delta_F \leftarrow \Delta_F \cup \{C \rightarrow m^*\}$
            $N \leftarrow N \setminus \{C\}$
        end if
    end for
    $F \leftarrow F \cup \Delta_F$
end while
return F

3. Оценка эффективности

Оценка эффективности предложенного подхода производится по методологии и материалам дорожки по разрешению лексической многозначности RUSSE'2018 [4]. В данном случае предполагается, что предложенный подход осуществляет кластеризацию контекстов так, чтобы каждый кластер соответствует отдельному значению слова. Число значений и число кластеров заранее неизвестно. В качестве меры качества используется скорректированный коэффициент Рэнда (англ. *adjusted Rand index*, сокр. *ARI*). В качестве золотого стандарта используется набор данных "bts-rnc".

3.1. Экспериментальная установка

Контексты в дорожке RUSSE'2018 разделены лексически: словники обучающей и тестовой выборок различаются. Это делает невозможным применение модели, построенной только на обучающей выборке, к словам из тестовой выборки. В связи с этим необходимо осуществлять разметку контекстов каждого целевого слова. Для каждого слова из набора данных "bts-rnc" были случайно выбраны контексты из Национального корпуса русского языка (НКРЯ) [12]. Контексты были размечены вручную в соответствии с инвентарём значений RuWordNet [2]. Поскольку в процессе разметки оказалось, что RuWordNet покрывает не все значения слов, представленные в НКРЯ, то были введены некоторые дополнительные значения некоторых слов. В процессе предварительной обработки контекстов из НКРЯ удалялись служебные части речи, а также контексты, в которых содержится меньше трех слов после удаления служебных частей речи.

Полученный размеченный набор данных представляет контексты 51 слова; всего 4 449 контекстов: на одно значение в среднем приходится от 20 до 30 контекстов. Всего неразмеченных контекстов: 86 791, из них включено в расширенную выборку: 63 444.

Подбор параметров проводился следующим образом. Появления слов и коллокации преобразуются в разреженную векторно-пространственную модель общепринятым способом: один признак — одно измерение. Производится взвешивание при помощи меры tf–idf. Используется реализация векторно-пространственных представлений и методов машинного обучения из библиотеки scikit-learn [5]. При расширении обучающей выборки используется классификатор на основе ближайших соседей (*KNeighborsClassifier* в scikit-learn), при предсказании значений — мультиномиальный наивный байесовский классификатором (*MultinomialNB*). Выбор классификаторов производился в ходе эксперимента.

3.2. Результаты

В табл. 1 представлены результаты сравнения предложенного подхода с другими подходами в дорожке RUSSE'2018 [4] на итоговой закрытой выборке (private). Видно, что представленный подход занял четвёртое место в рейтинге и показал значение ARI выше, чем эталонный метод AdaGram [1], обученный на корпусе текстов НКРЯ [12], и система Watasense [8], основанная на автоматически построенном инвентаре значений слов [7] и векторных представлениях значений слов [6].

С целью изучения целесообразности расширения обучающей выборки, мы построили классификатор только на основе исходных размеченных данных. Такой подход показал низкий результат и занял тринадцатое место, что показывает и подтверждает целесообразность использования представленного в данной работе подхода к разрешению лексической многозначности.

Таблица 1. Сравнение предложенного подхода на основе частичного обучения с аналогичными методами по набору данных "bts-rnc" дорожки RUSSE'2018

Место	**Идентификатор**	**ARI**
1	jamsic	0,3384
2	Pavel	0,2818
3	joystick	0,2579
4	**Частичное обучение (раздел 2)**	**0,2576**
...	…	…
10	**AdaGram [1]**	**0,2132**
...	...	...
13	**Обучение с учителем без расширения (раздел 2 без 2.2)**	**0,0548**
14	Watasense (Dense) [8]	0,0469

3.3. Анализ результатов

Анализ результатов выявил классы ошибок, связанные с гранулярностью значений (раздел 3.3.1), некорректной лемматизацией (раздел 3.3.2), а также человеческим фактором (раздел 3.3.3).

3.3.1. У слова могут быть близкие по контексту значения. Например, у слова «обед» есть различные значения, близкие по смыслу друг у другу. Согласно толковому словарю Ушакова, у этого слова три значения [13]:

1. Прием пищи, обычно приуроченный к середине дня.
2. Самая пища, приготовленная для этой еды.
3. Время этого приема пищи; то же, что полдень (*разг.*).

Все значения связаны с приёмом пищи. Алгоритму трудно их различить, поскольку нарушается допущение о независимости значений.

3.3.2. Слово может быть некорректно приведено в начальную форму. Слово «гвоздика» при лемматизации заменялось на слово «гвоздик», что приводило к неверному выбору классификатора для слова.

3.3.3. Ошибки разметки. Поскольку контексты из НКРЯ выбирались случайным образом, для некоторых значений не нашлось примеров в полученной выборке. Например, такая проблема возникла со словами «тюрьма» и «карьер».

4. Заключение

Представленный подход к разрешению лексической многозначности на основе обучения с учителем позволяет использовать небольшую размеченную выборку для решения данной задачи. Результаты экспериментов на материалах дорожки RUSSE'2018 показывает перспективность представленного подхода. В свою очередь, результаты можно улучшить путём отбора дополнительных информативных признаков и автоматизации процесса формирования обучающей выборки, что является направлениями дальнейших исследований. Насколько нам известно, нами представлена первая система разрешения лексической многозначности в дорожке RUSSE'2018, использующая обучение с учителем: остальные системы основаны на обучении без учителя [4]. Исходный код представленной системы на языке программирования Python и результаты экспериментов доступны в репозитории на GitHub под открытой лицензией: https://github.com/Pushkinue/KR-WSD.

Благодарности. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а.

Список литературы

1. Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D.P. (2016), Breaking Sticks and Ambiguities with Adaptive Skip-gram, Journal of Machine Learning Research, Vol. 51, pp. 130–138.
2. Loukachevitch N.V., Lashevich G., Gerasimova A.A., Ivanov V.V., Dobrov B.V. (2016), Creating Russian WordNet by Conversion, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", Moscow, Russia, pp. 405–415.
3. Navigli R. (2012), A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches, Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012), pp. 115–129, Špindlerův Mlýn, Czech Republic.
4. Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N. (2018), RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", Moscow, Russia.
5. Pedregosa F. et al. (2011), Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
6. Pelevina M., Arefyev N., Biemann C., Panchenko A. (2016), Making Sense of Word Embeddings, Proceedings of the 1st Workshop on Representation Learning for NLP (Rep4NLP), pp. 174–183, Berlin, Germany.
7. Ustalov D., Panchenko A., Biemann C. (2017), Watset: Automatic Induction of Synsets from a Graph of Synonyms, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1579–1590, Vancouver, Canada.

8. Ustalov D., Teslenko D., Panchenko A. et al. (2018), An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pp. 1018–1022, Miyazaki, Japan.
9. Yarowsky D. (1995), Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995), pp. 189–196, Cambridge, MA, USA.
10. Zhong Z., Ng H.T. (2010), It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free Text, Proceedings of the ACL 2010 System Demonstrations (ACLDemos '10), pp. 78–83, Uppsala, Sweden.
11. Лукашевич Н.В. (2011), Тезаурусы в задачах информационного поиска [Tezaurusy v zadachah informacionnogo poiska], Изд-во МГУ, Москва.
12. Плунгян В.А. (2009), Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы [Nacional'nyj korpus russkogo jazyka: 2006—2008. Novye rezul'taty i perspektivy], Нестор-История, Санкт-Петербург.
13. Ушаков Д.Н. (2001), «Толковый словарь русского языка» в 3 т. на основе 4-томного издания 1948 г. [«Tolkovyj slovar' russkogo jazyka» v 3 t. na osnove 4-tomnogo izdanija 1948 g.], «Вече», «Си ЭТС», Москва.