Труды Кольского научного центра РАН. Информационные технологии. Вып. 12. 2021. Т. 12, № 5. С. 22–34.
Transactions of the Kola Science Centre. Information technologies. Series 12. 2021. Vol. 12, no. 5. P. 22–34.

Научная статья
УДК 004.853
DOI: 10.37614/2307-5252.2021.5.12.002

АУГМЕНТАЦИЯ ОБУЧАЮЩЕГО НАБОРА ПРИ ОБУЧЕНИИ НЕЙРОСЕТЕВОЙ ЯЗЫКОВОЙ МОДЕЛИ ДЛЯ НАПОЛНЕНИЯ ОНТОЛОГИИ*

Павел Андреевич Ломов [1]✉, Марина Леонидовна Малоземова [2]
[1, 2] Институт информатики и математического моделирования ФИЦ КНЦ РАН, Апатиты, Россия
[1]lomov@iimm.ru✉, https://orcid.org/0000-0002-0924-0188
[2]malozemova@iimm.ru, https://orcid.org/0000-0002-4358-2683

Аннотация

Данная работа является продолжением исследования, ориентированного на решение задачи наполнения онтологии с помощью обучения на автоматически формируемом обучающем наборе и последующего применения нейросетевой языковой модели для анализа текстов с целью обнаружения в них новых понятий для добавления в онтологию. Статья посвящена проблеме автоматического увеличения размера обучающего набора путем аугментации входящих в него образцов. Наряду с этим рассматривается решение проблемы уточнения найденных понятий (корректировка их границ в предложениях), которые были найдены при автоматическом создании обучающего набора. Представлен краткий обзор существующих подходов к аугментации текстовых данных, а также подходов к извлечению вложенных именованных сущностей (nested NER). Предложена процедура уточнения границ обнаруженных понятий обучающего набора и его аугментации для последующего обучения и применения нейросетевой языковой модели с целью выявления новых понятий онтологии в текстах предметной области. Рассмотрены результаты экспериментальной оценки обученной модели на аугментированном наборе и основные направления дальнейшего исследования.

Ключевые слова:

аугментация данных, нейронная сеть, наполнение онтологий

Финансирование

Работа выполнена в рамках выполнения гос. задания по теме НИР № 0226-2019-0036. При поддержке Российского фонда фундаментальных исследований, проект № 20-07-00754 А.

Для цитирования: Ломов П. А., Малоземова М. Л. Аугментация обучающего набора при обучении нейросетевой языковой модели для наполнения онтологии // Труды Кольского научного центра РАН. Информационные технологии. Вып. 12. 2021. Т. 12, № 5. С. 22–34. http://dx/doi.org/10.37614/2307-5252.2021.5.12.002.

Original article

TRAINING SET AUGMENTATION IN TRAINING NEURAL-NETWORK LANGUAGE MODEL FOR ONTOLOGY POPULATION

Pavel A. Lomov [1]✉, Marina L. Malozemova [2]
[1, 2, 3] Institute for Informatics and Mathematical Modeling Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1] lomov@iimm.ru✉, https://orcid.org/0000-0002-0924-0188
[2] malozemova@iimm.ru, https://orcid.org/0000-0002-4358-2683

Abstract

This paper is a continuation of the research focused on solving the problem of ontology population using training on an automatically generated training set and the subsequent use of a neural-network language model for analyzing texts in order to discover new concepts to add to the ontology. The article is devoted to the text data augmentation - increasing the size of the training set by modification of its samples. Along with this, a solution to the problem of clarifying concepts (i.e. adjusting their boundaries in sentences), which were found during the automatic formation of the training set, is considered. A brief overview of existing approaches to text data augmentation, as well as approaches to extracting so-called nested named entities (nested NER), is presented. A procedure is proposed for clarifying the boundaries of the discovered concepts of the training set and its augmentation for subsequent training a neural-network language model in order to identify new concepts of ontology in the domain texts. The results of the experimental evaluation of the trained model and the main directions of further research are considered.

Keywords:

data augmentation, neural network, ontology population

Funding

The article was supported by the federal budget to carry out the state task of the FRC KSC RAS No. 0226-2019-0036. The study was funded by RFBR, project number 20-07-00754 A.

For citation: Lomov P. A., Malozemova M. L. Training set augmentation in training neural-network language model for ontology population // Transactions of the Kola Science Centre. Information technologies. Series 12. 2021. Vol. 12, no. 5. P. 22–34. http://dx/doi.org/10.37614/2307-5252.2021.5.12.002.

Введение

Данная работа является продолжением исследования [1], направленного на автоматическую генерацию обучающего набора на основе анализа текстов предметной области и его использования для обучения нейросетевой модели, ориентированной на решение одной из подзадач обучения онтологий – задачи наполнения онтологии. Упоминаемая проблема обучения онтологий заключается в анализе естественно-языковых текстов с последующем извлечением из них концептов и отношений, а также логических выражений (аксиом) с последующим формированием онтологии [2]. Наполнение онтологии предполагает добавление в существующую онтологию новых экземпляров для заданных в ней классов без изменения структуры онтологии.

В предыдущей работе была предложена технология, предполагающая анализ онтологии для формирования списка ее понятий, сбор и анализ текстов, относящихся к предметной области онтологии, с формированием обучающего набора размеченных предложений. Далее данный набор применялся для обучения нейросетевой языковой модели, ориентированной на решение задачи извлечения именованных сущностей (NER). Модель впоследствии применялась для извлечения из текстов новых понятий – кандидатов на добавление в онтологию.

Ввиду того, что в основе предложенной технологии лежит обучение с учителем, необходимо обеспечить достаточно большой объем обучающего набора для успешного обучения. Одним из способов его увеличения является аугментация данных, которая предполагает автоматическое создание новых образцов путем некоторого изменения имеющихся. Это позволяет в некоторой степени повысить эффективность обучения и результативность модели.

В данной работе рассматривается проблема аугментации сгенерированного набора для повышения полноты и точности получаемой на его основе языковой модели в отношении обнаружения в текстах возможных новых элементов онтологии – классов и экземпляров. При этом важно обеспечить

соответствие метки ассоциированному с ней образцу (текстовому предложению), который подвергся изменению в результате аугментации. Так как в состав метки в данном случае входит извлекаемое понятие, представленное в виде своих границ (индексов первого и последнего токенов в предложении), то особую важность представляет правильное определение этих границ для правильного определения модифицируемой части предложения. Данная проблема напоминает проблему извлечения вложенных именованных сущностей (Nested Named Entity Recognition, Nested NER [3]), однако имеет некоторую специфику, обусловленную извлечением понятий для обучения онтологий. Таким образом, в данной работе предлагается после этапа генерации набора выполнять этап, на котором производится уточнение найденных понятий и последующее формирование дополнительных вариаций предложений, в которых они встречаются.

1. Обзор существующих подходов к аугментации данных

Аугментация текстовых данных обучающих наборов заключается в изменении содержания их образцов (обычно текстов или предложений) так, чтобы не был утрачен их смысл. При этом присвоенные образцам метки, как правило, остаются без изменения.

Среди общих видов техник аугментации можно выделить следующие:

- замена слов в предложении [4];
- перестановка слов в предложении или предложений внутри текстов [5];
- «зашумление», то есть добавление незначительных ошибок в слова, предложения или тексты (изменение регистра, знаков препинания и т.п.);
- генерация новых предложений или текстов на основе изменения структуры исходных [6].

Весьма распространенной практикой в последнее время стало использование предобученных на большем объеме текстов моделей с BERT (Bidirectional Encoder Representations from Transformers) [7] архитектурой. Ключевой особенностью данной архитектуры является возможность рассмотрения в процессе обучения отдельного слова в контексте окружающих его слов. Предобучение BERT-моделей на большом объеме текстов позволяет сформировать контекстуализированные векторные представления слов (contextualized word embeddings) для некоторого естественного языка, которые могут быть использованы в дальнейшем для решения различных NLP-задач, а также для выполнения аугментации.

Так, в работе [8] предлагается алгоритм коррекции слов с ошибками, основанный на использовании маскированной языковой модели (masked language model) на основе BERT. В предложенном алгоритме данная модель используется для представления вариантов замены маскированных ошибочных слов. Аугментация в данном случае выполняется путем конкатенации исходного предложения и его варианта, содержащего маскирующие токены вместо слов с ошибками. По словам авторов, такая аугментация позволяет получать варианты для замены, состоящие из большего или меньшего числа токенов, чем заменяемое слово, а также «отвлечь» модель от ошибочного слова при генерации вариантов его замены.

В работе [9] предлагается метод контекстной аугментации размеченных предложений с помощью условной BERT-модели. Данная модель является результатом настройки исходной BERT-модели с помощью набора данных, дополнительно включающего метки (позитивная/негативная). Это позволяет модели при аугментации предлагать замены маскированным токенам с учетом метки и тем самым обеспечивать правильность получаемых образцов.

В работе [10] предлагается метод аугментации с применением так называемой filtered-BERT модели для решения задачи деидентификации защищенной информации о здоровье (protected health information, PHI) в документах для вторичного использования. Filtered-BERT предсказывает маскированное слово, предоставляя несколько вариантов, и далее производит их фильтрацию путем сравнения косинусного расстояния между fastText-векторами слов-вариантов и заменяемого слова. В итоге аугментированные предложения формируются со словами, прошедшими через данный фильтр.

Упомянутая проблема уточнения границ сущности в предложении похожа на проблему извлечения вложенных сущностей (nested NER). В ранних работах, посвященных ее решению, используются подходы, основанные на правилах. Так, в работе [11], посвященной распознаванию биомедицинских сущностей предлагаются два таких подхода: подход на основе правил постобработки (post-processing) и подход на основе скрытой марковской модели (Hidden Markov Model, HMM). Подход на основе правил постобработки предполагает использование специально разработанных на основе корпуса GENIA паттернов, которые позволяют распознать наиболее длинные имена сущностей на основе более простых (вложенных). Подход на основе HMM, в свою очередь, предполагает использование двух предварительно обученных моделей: первая модель распознает короткие сущности, а вторая – используется для последующего расширения этих распознанных коротких сущностей в длинные.

В работе [3] представлен специализированный парсер для распознавания вложенных именованных сущностей. Данный парсер обучается на предложениях, представленных в виде синтаксических деревьев (parse tree), которые содержат информацию о составляющих (токенах) каждой именованной сущности – «родителя» и «прародителя», а также их части речи.

В недавних работах чаще всего используются подходы с применением нейросетевых моделей для распознавания вложенных сущностей. Например, в работе [12] для решения данной задачи предлагается простая нейросетевая модель. Она позволяет выделить и классифицировать все возможные фрагменты входной последовательности, в которых упоминается потенциальная вложенная сущность. Затем в этих выделенных областях с помощью слоя LSTM обнаруживаются сами сущности.

В следующей работе [13] предлагается нейронная модель для идентификации вложенных сущностей путем наложения друг на друга так называемых «плоских» слоев NER. «Плоский» слой используется для распознавания «плоских» сущностей – противоположность вложенных сущностей. Данный слой, в свою очередь, состоит из слоя LSTM, который захватывает двунаправленное контекстное представление последовательности, и слоя CRF, предсказывающего последовательность меток – теги BIO для этого представления. Количество плоских слоев зависит от уровня вложенности сущности (например, New York – 1 уровень, New York University – 2 уровень).

Процесс обнаружения сущностей прекращается, если текущий плоский слой NER не выявляет никаких сущностей.

В работе [14] предлагается операция регрессии для обнаружения вложенных именованных сущностей в предложении. Для ее выполнения предложение сначала преобразуется с помощью глубокой нейросети в рекуррентные карты признаков (recurrent feature maps), т.е. в абстрактные представления, фиксирующие семантические зависимости между словами. Каждая карта признаков определяет возможные границы сущности. Далее из этих карт признаков генерируются рамки (bounding boxes), которые представляют собой абстрактные представления именованных сущностей. Каждая рамка включает информацию о положении сущности (начальная позиция и длина) и категории класса. В процессе обучения операция регрессии предсказывает значение смещения начальной позиции и значение смещения текущей рамки относительно истинной рамки, соответствующей истинной именованной сущности. Предсказанные смещения позволяют корректно «сдвинуть» рамку, тем самым точно идентифицируя сущность.

В работе [15] предлагается итеративный алгоритм двунаправленного распознавания вложенных именованных сущностей. Он предполагает обучение двух нейросетевых моделей на одном наборе данных для идентификации именованных сущностей в двух направлениях: от общего к конкретному (снаружи внутрь) и от конкретного к общему (изнутри наружу). Каждое слово входной последовательности представляется в виде конкатенации трех векторов: контекстное представление символьной языковой модели, статическое векторное представление (word embedding) и multi-hot вектор закодированных предсказаний для данного слова из предыдущих итераций. На каждой итерации модель генерирует новые прогнозы на основе исходной последовательности слов и ранее сделанных прогнозов. Данный процесс завершается, когда новые сущности больше не выявляются. На выходе прогнозы обеих моделей фильтруются посредством выбранного критерия отбора (например, объединение результатов, пересечение результатов и др.) с последующим формированием окончательного набора обнаруженных сущностей.

В работе [16] предлагается метод декодирования, который итеративно распознает сущности по принципу от самых внешних к внутренним («outside-to-inside» способ). Он позволяет выявить в диапазоне каждой обнаруженной сущности внутренние вложенные сущности, используя алгоритм Витерби [17].

Предлагаемая в данной работе процедура аугментации также предполагает применение предобученной BERT-модели, ориентированной на решение задачи маскированного языкового моделирования (masked language modeling, MLM) [7]. Основное отличие состоит в подготовке маскированного предложения с учетом вероятного переопределения границ сущности, производимого в рамках ее уточнения.

2. Предлагаемая процедура уточнения понятий и аугментации размеченных предложений

Аугментация предполагает некоторое изменение исходного предложения. Однако в данном случае предполагается последующее извлечение сущности для ее добавления в онтологию, поэтому модификация предложения не должна ее

затронуть. Поэтому особую актуальность приобретает правильность определения границ сущности в предложении, то есть определения упорядоченного множества составляющих ее токенов.

Данное обстоятельство заставляет пересмотреть подход к выявлению понятий при формировании обучающего набора, предложенный в предыдущей работе [1]. Он предполагал поиск в текстах предложений, содержащих имена экземпляров классов уже существующих в исходной онтологии. Это позволяло свести задачу ее наполнения к задаче извлечения именованных сущностей (NER).

Однако анализ полученных таким образом размеченных наборов предложений выявил некоторые особенности, которые необходимо учитывать в контексте извлечения именно онтологических концептов. Например, при поиске предложений в корпусе новостных текстов, включающих экземпляр класса «Country» с именем «Russia», помимо предложений, содержащих соответствующий токен, как название страны, были найдены также предложения, включающие комбинации этого токена с другими: «government of Russia», «president of Russia», «company in Russia». Данные комбинации токенов могут быть проинтерпретированы и как классы онтологии («government», «president»), и как экземпляры классов («government of Russia», «president of Russia»). При этом они могут неявно определять некоторое отношение (например, company «is-located-in» Russia) к исходному классу «Country» и/или его экземпляру «Russia», которое также можно добавить в онтологию. Таким образом, идентификация онтологической сущности в предложениях при формировании обучающего набора, в отличие от просто именованных сущностей в задаче NER, имеет смысл осуществлять с учетом некоторого онтологического контекста, описывающего роль исходной сущности в онтологии.

Учет онтологического контекста при формировании обучающего набора, а также возможность его включения в состав меток размеченных предложений, предполагается рассмотреть подробнее в продолжении исследования. В рамках же данной работы для корректной аугментации необходимым является уточнение границ сущностей в предложениях, найденных в текстах предметной области на первом этапе. Для этого был введен дополнительный шаг, предполагающий определение положения исходной сущности в дереве зависимостей (dependency tree) предложения и включения в ее состав токенов, непосредственно связанных с ней синтаксическими отношениями.

Например, предложение «Международный аэропорт Шереметьево является крупнейшим в России.» имеет следующее синтаксическое дерево, как показано на рис. 1.

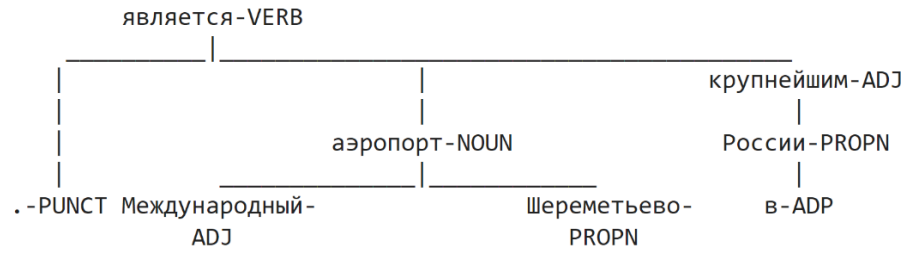

Рис. 1. Схема дерева зависимостей предложения

Согласно данному дереву, сущность онтологии «аэропорт» связана с токенами «Международный» и «Шереметьево». Следовательно, уточненная сущность будет «Международный Аэропорт Шереметьево».

После уточнения сущностей производится аугментация исходного предложения следующим образом:

1. На основе анализа дерева синтаксических зависимостей предложения среди токенов сущности определяется главный токен сущности (он расположен на более высоком уровне дерева, чем дочерние токены).
2. Выявляем ближайшие токены, расположенные слева и справа от токенов сущности и не являющиеся прилагательными, стоп-словами или предлогами. По дереву зависимостей определяем зависимые от них токены. Таким образом, получаем группы ближайших левых и правых токенов.
3. В полученных на 2-ом шаге группах отмечаем главные и дочерние токены как кандидаты на замену.
4. Формируем все комбинации индексов токенов, отмеченных на замену.
5. С использованием сформированных на предыдущем шаге комбинаций индексов создаем модификации исходного предложения, в которых разные комбинации токенов заменены токеном-маской.
6. С помощью предобученной BERT-модели получаем варианты токенов для замены и подставляем их вместо токенов-масок в модификации исходного предложения.
7. Для полученных аугментированных предложений производим переопределение границ онтологической сущности, поскольку замена токенов приводит к их изменению.

Например, рис. 1 показывает, что главный токен понятия «Международный аэропорт Шереметьево» – это «аэропорт». На следующем шаге мы получаем только группу ближайших правых токенов [«является», «крупнейшим»], а группа ближайших левых токенов является пустой, поскольку группа токенов сущности расположена в начале предложения. Эти два токена из группы ближайших правых токенов являются кандидатами на замену. Далее, имея сформированных комбинации индексов этих токенов, мы создаем модификации исходного предложения, где токены «является» и «крупнейшим» заменены токеном-маской:

Международный аэропорт Шереметьево {mask} крупнейшим в России.

Международный аэропорт Шереметьево является {mask} в России.

Используя BERT-модель, мы получили 16 возможных токенов для замены токенов-масок. В итоге, применив некоторые из предложенных токенов, получаем следующие модифицированные предложения:

Международный аэропорт Шереметьево был крупнейшим в России. Международный аэропорт Шереметьево является крупнейшим в России. Международный аэропорт Шереметьево является единственным в России.

Международный аэропорт Шереметьево является старейшим в России.

Для получения дерева синтаксических зависимостей была использована русскоязычная модель из фреймворка spaCy [18]. В качестве языковой модели для решения задачи MLM и предложения токенов для подстановки использовалась русскоязычная модель RuBERT из проекта deeppavlov [19].

Для того, чтобы BERT-модель предлагала на замену токены, релевантные предметной области предложения, необходимо избегать замены в предложении

слишком большого числа токенов. По этой причине на 4-ом шаге при генерации комбинаций используются параметры, определяющих одновременное максимально количество заменяемых и удаляемых токенов в комбинации. Им были присвоены значения 4 и 2 соответственно. Наряду с этим, во избежание частых ошибок согласования предлагаемого моделью токена и его окружения (например, несогласование по роду, числу или падежу), не рассматривались комбинации, в которых маскируемые токены располагались друг за другом.

Указанные параметры позволяют варьировать грамматическую и смысловую правильность формируемых предложений и объем получаемого в результате аргументированного набора. Чем больше комбинаций будет рассматриваться, тем больше вариантов будет сгенерировано для каждого исходного предложения. Однако при этом возрастает вероятность наличия в них грамматических и смысловых ошибок.

После завершения представленной процедуры аугментации сформированный набор может быть использован для обучения языковой модели.

3. Оценка эффективности предложенной процедуры аугментации

Для оценки эффекта предлагаемых процедур уточнения понятий и аугментации был проведен эксперимент, в рамках которого были обучены и оценены три модели. Первая модель была получена без применения предложенных процедур аугментации и уточнения понятий, вторая – с применением уточнения понятий, третья – с применением уточнения понятий и аугментации.

Для формирования начального обучающего набора использовался корпус новостных русскоязычных текстов интернет-издания Lenta.ru [20], который содержит около 800 тысяч новостных текстов различной тематики (политика, экономика, спорт и т.д.). В качестве набора понятий онтологии был вручную сформирован исходный список понятий, характерный для новостных текстов (например, «компания», «полиция», «акция» и др.).

В результате анализа текстового корпуса с применением исходного списка понятий был сформирован обучающий набор, включающий около 550 тысяч размеченных образцов – предложений с метками «понятие и ее категория». На данном наборе была обучена первая модель.

Далее к сформированному обучающему набору были применены предложенные процедуры уточнения и аугментации. В результате были получены еще два набора: набор с уточненными понятиями и аугментированный набор с уточненными понятиями. Размер последнего вырос с 550 тысяч до 2800 тысяч образцов. После этого на данных наборах было обучено еще две модели.

Проверка их качества выполнялась на тестовом наборе. Его формирование осуществлялось аналогично формированию обучающего, но при этом использовалась другая часть текстов новостного корпуса. Объем тестового набора составил 300 тысяч образцов.

Оценка качества производилась в рамках следующих экспериментов:

Эксперимент 1. Обнаружение «известных» моделям понятий тестового набора, т.е. тех понятий, которые присутствовали в обучающем наборе:

- модель, обученная без уточнения и аугментации: точность = 0.002, полнота = 0.024;

- модель, обученная с уточнением понятий: точность = 0.056, полнота = 0.694;
- модель, обученная с уточнением и аугментацией: точность = 0.055, полнота = 0.593.

Эксперимент 2. Обнаружение «неизвестных» моделям понятий тестового набора, т.е. понятий, не присутствовавших в обучающем наборе:

- модель, обученная без уточнения и аугментации: точность = 0.0, полнота = 0.01;
- модель, обученная с уточнением понятий: точность = 0.455, полнота = 0.424;
- модель, обученная с уточнением и аугментацией: точность = 0.435, полнота = 0.321.

Отдельно также была произведена экспертная оценка корректности обнаруженных моделями понятий, которых не было в исходном списке (и, соответственно, в тестовом и обучающем наборах). Таким образом, оценивалась доля тех понятий, которые могут быть использованы для наполнения онтологии.

Эксперимент 3. Обнаружение понятий, не представленных в исходном списке понятий:

- модель, обученная без уточнения и аугментации: всего новых понятий – 39, доля корректных понятий – 0,6;
- модель, обученная с уточнением понятий: всего новых понятий – 3566, доля корректных понятий – 0,82;
- модель, обученная с уточнением и аугментацией: всего новых понятий – 3254, доля корректных понятий – 0,85.

Наиболее показательными в отношении оценки эффективности использования полученных моделей для обучения онтологий является второй и третий эксперименты. В них оценивается способность моделей находить новые понятия на основе контекстов, в которых встречались понятия обучающего набора.

Результаты экспериментов показали, что основной вклад в увеличение эффективности привносит процедура уточнения понятий. Вероятно, это вызвано тем, что понятия из исходного списка, представляющие в экспериментах понятия наполняемой онтологии, дополняются при уточнении связанными с ними токенами из предложений анализируемых текстов. Это приводит к тому, что данные токены рассматриваются моделью при обучении как части понятия, а не его контекста. Например, уточнение понятия «аэропорт» до «Международный аэропорт Шереметьево», позволяет модели рассматривать контекст «... является крупнейшим в России», который с большей вероятностью может ассоциироваться с другими понятиями, чем контекст «Международный ... Шереметьево является крупнейшим в России», полученный без уточнения понятия. Таким образом, уточнение понятий позволяет скорректировать их контекст употребления, что положительно сказывается на способности модели находить новые понятия.

Применение аугментации также позволило немного повысить эффективность в отношении точности обнаружения новых понятий в третьем эксперименте. Однако этого удалось достичь после обучения модели на аугментированном наборе размером 2800 тысяч образцов. Такое обучение заняло в 5 раз больше времени, чем обучение на неаугментированном наборе.

Заключение

Обучение онтологий на сегодняшний день продолжает оставаться актуальной проблемой при разработке современных информационных систем, ориентированных на представление и оперирование знаниями предметной области. Использование существующих технологий NLP и машинного обучения имеет большой потенциал в отношении автоматизации связанных с этим задач – от сбора, предобработки и анализа естественно-языковых текстов до формирования начальной структуры понятий онтологии и ее последующего усложнения.

В данной работе рассмотрено расширение предложенной ранее технологии, ориентированной на решение задачи наполнения существующей онтологии новыми экземплярами классов, которые извлекаются из текстов предметной области с помощью обученной нейросетевой языковой модели. В качестве дополнительных шагов предложено при формировании обучающего набора производить уточнение понятий. Это предполагает анализ предложений, содержащих исходное понятие онтологии, и расширение его границ путем включения в него некоторых дополнительных токенов, связанных с ним. Это, с одной стороны, позволило представить понятие онтологии в том виде, в котором оно встречалось в анализируемых текстах, а с другой – скорректировать его контекст, распознавать который обучается модель.

Другим предложенным шагом стало выполнение аугментации предложений, которое предполагало замену некоторых токенов, не входящих в состав уточненного понятия, на вариант, предложенный предобученной языковой моделью-трансформером, способной представлять контекстуализированные векторные представления слов.

В результате экспериментов было установлено, что уточнение понятий положительно сказывается на обнаружении новых понятий в тех контекстах, которые были представлены в предложениях обучающего набора. В дальнейшем планируется само уточнение понятий производить с учетом онтологического контекста, который описывает положение исходного понятия в онтологии. Это позволит проинтерпретировать токены, обнаруженные в ходе уточнения, как еще одно возможное понятие онтологии, которое также следует представить в обучающем наборе. Последнее обстоятельство может потребовать повторения этапа анализа текстов для включения предложений с такими понятиями в обучающий набор.

Кроме того, планируется рассмотреть возможность включения онтологического контекста в метку, что позволит при использовании обученной модели не только обнаруживать понятия, но и указывать их возможное положение в онтологии (подкласс класса, экземпляр класса, носитель свойства, значение свойства и т.д.).

В отношении развития предложенной процедуры аугментации планируется рассмотреть возможность генерации новых образцов (текстовых предложений) путем изменения структуры исходных, то есть заменой/добавлением/удалением их частей (наборов синтаксически связанных токенов). Помимо проблемы определения изменяемой части и генерации заменяющей части, актуальной станет проблема проверки семантической корректности результата. Однако выполнение такой аугментации позволит

улучшить качество обученной модели при использовании аугментированного набора меньшего размера, чем при использовании текущего подхода.

Примечания

[*] Адаптированный перевод статьи: Lomov P.A. Data Augmentation in Training Neural-Network Language Model for Ontology Population / P.A. Lomov, M.L. Malozemova, M.G. Shishaev // Data Science and Intelligent Systems: Lecture Notes in Networks and Systems / ed. R. Silhavy. – Cham: Springer International Publishing, 2021. – pp. 669-679

Список литературы

1. Lomov P., Malozemova M., Shishaev M. Training and application of neural-network language model for ontology population // Software engineering perspectives in intelligent systems / под ред. R. Silhavy, P. Silhavy, Z. Prokopova. Cham: Springer International Publishing, 2020. Т. 1295. С. 919–926.
2. Wong W., Liu W., Bennamoun M. Ontology Learning from Text: A Look Back and into the Future // ACM Comput. Surv. - CSUR. 2011. Т. 44. С. 1–36.
3. Finkel J. R., Manning C. D. Nested named entity recognition // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 EMNLP '09. USA: Association for Computational Linguistics, 2009. С. 141–150.
4. Wang W. Y., Yang D. That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. С. 2557–2563.
5. Luque F. M. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis // ArXiv190911241 Cs. 2019.
6. Coulombe C. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs // 2018. С. 33.
7. Devlin J. и др. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // ArXiv181004805 Cs. 2018.
8. Sun Y., Jiang H. Contextual Text Denoising with Masked Language Model // Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong, China: Association for Computational Linguistics, 2019. С. 286–290.
9. Wu X. и др. Conditional BERT Contextual Augmentation // Computational Science – ICCS 2019 Lecture Notes in Computer Science. / под ред. J. M. F. Rodrigues и др. Cham: Springer International Publishing, 2019. С. 84–95.
10. Kang M., Lee K., Lee Y. Filtered BERT: Similarity Filter-Based Augmentation with Bidirectional Transfer Learning for Protected Health Information Prediction in Clinical Documents // Appl. Sci. 2021. Т. 11. С. 3668.
11. Zhang J. и др. Enhancing HMM-based biomedical named entity recognition by studying special phenomena // J. Biomed. Inform. 2004. Т. 37. № 6. С. 411–422.
12. Sohrab M. G., Miwa M. Deep Exhaustive Model for Nested Named Entity Recognition // Proceedings of the 2018 Conference on Empirical Methods in Natural

Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. C. 2843–2849.
13. Ju M., Miwa M., Ananiadou S. A Neural Layered Model for Nested Named Entity Recognition // Proceedings of NAACL-HLT 2018. , 2018. C. 1446–1459.
14. Chen Y. и др. A Boundary Regression Model for Nested Named Entity Recognition // ArXiv201114330 Cs. 2020.
15. Dadas S., Protasiewicz J. A Bidirectional Iterative Algorithm for Nested Named Entity Recognition // IEEE Access. 2020. T. 8. C. 135091–135102.
16. Shibuya T., Hovy E. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding // Trans. Assoc. Comput. Linguist. 2020. T. 8. C. 605–620.
17. Huang Z. и др. Iterative viterbi A* algorithm for K-best sequential decoding // 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2012. C. 611–619.
18. Russian spaCy Models Documentation [Электронный ресурс]. URL: https://spacy.io/models/ru#ru_core_news_sm
19. Pre-trained embeddings – DeepPavlov 0.15.0 documentation [Электронный ресурс]. URL: http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html#bert
20. News dataset from Lenta.Ru [Электронный ресурс]. URL: https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta

References

1. Lomov P., Malozemova M., Shishaev M. Training and Application of Neural-Network Language Model for Ontology Population. In: Silhavy R., Silhavy P., Prokopova Z. (eds) Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing, 2020, Vol. 1295, Springer, Cham, pp. 919–926.
2. Wong W., Liu W., Bennamoun M. Ontology Learning from Text: A Look Back and into the Future. ACM Comput. Surv, CSUR, 2011, Vol. 44, pp. 1–36.
3. Finkel J. R., Manning C. D. Nested named entity recognition. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1 – Vol. 1 EMNLP '09. USA: Association for Computational Linguistics, 2009, pp. 141–150.
4. Wang W. Y., Yang D. That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2557–2563.
5. Luque F. M. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis. ArXiv190911241 Cs. 2019.
6. Coulombe C. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs. 2018. pp. 33.
7. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs. 2018.
8. Sun Y., Jiang H. Contextual Text Denoising with Masked Language Model. Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong, China: Association for Computational Linguistics, 2019, pp. 286–290.

9. Wu X. et al. Conditional BERT Contextual Augmentation. Computational Science – ICCS 2019 Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 84–95.
10. Kang M., Lee K., Lee Y. Filtered BERT: Similarity Filter-Based Augmentation with Bidirectional Transfer Learning for Protected Health Information Prediction in Clinical Documents. Appl. Sci. 2021, Vol. 11, pp. 3668.
11. Zhang J. et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J. Biomed. Inform. 2004, Vol. 37, No 6, pp. 411–422.
12. Sohrab M. G., Miwa M. Deep Exhaustive Model for Nested Named Entity Recognition. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2843–2849.
13. Ju M., Miwa M., Ananiadou S. A Neural Layered Model for Nested Named Entity Recognition. Proceedings of NAACL-HLT 2018, 2018, pp. 1446–1459.
14. Chen Y. et al. A Boundary Regression Model for Nested Named Entity Recognition. ArXiv201114330 Cs, 2020.
15. Dadas S., Protasiewicz J. A Bidirectional Iterative Algorithm for Nested Named Entity Recognition. IEEE Access, 2020, Vol. 8, pp. 135091–135102.
16. Shibuya T., Hovy E. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding. Trans. Assoc. Comput. Linguist, 2020. Vol. 8, pp. 605–620.
17. Huang Z. et al. Iterative viterbi A* algorithm for K-best sequential decoding. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2012, pp. 611–619.
18. Russian spaCy Models Documentation. Available at: https://spacy.io/models/ru#ru_core_news_sm
19. Pre-trained embeddings – DeepPavlov 0.15.0 documentation. Available at: http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html#bert
20. News dataset from Lenta.Ru. Available at: https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta

Сведения об авторах
П. А. Ломов — кандидат технических наук, старший научный сотрудник ИИММ КНЦ РАН;
М. Л. Малоземова — инженер-исследователь ИИММ КНЦ РАН.

Information about the authors
P. A. Lomov — Candidate of Science (Tech.), Senior Research Fellow of the Institute for Informatics and Mathematical Modeling Kola Science Centre of the Russian Academy of Sciences;
M. L. Malozemova — research engineer of the Institute for Informatics and Mathematical Modeling Kola Science Centre of the Russian Academy of Sciences.

Статья поступила в редакцию 15.11.2021; одобрена после рецензирования 20.11.2021; принята к публикации 08.12.2021.
The article was submitted 15.11.2021; approved after reviewing 20.11.2021; accepted for publication 08.12.2021.