А. О. Шелманов, Ю. М. Кузнецова, В. А. Исаков, И. В. Смирнов, Открытое извлечение информации из текстов. Часть II. Извлечения семантических отношений с помощью машинного обучения без учителя, *Искусственный интеллект и принятие решений*, 2019, выпуск 2, 39–49

DOI: 10.14357/20718594190204

Открытое извлечение информации из текстов Часть II. Извлечения семантических отношений с помощью машинного обучения без учителя*

А. О. Шелманов, Д. А. Девяткин, В. А. Исаков, И. В. Смирнов

Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

Аннотация. Работа посвящена «открытому извлечению информации» из текстов на естественном языке (open information extraction). Описывается подход к решению задачи извлечения семантических отношений из текстов на основе машинного обучения без учителя. Подход основан на методах глубокой кластеризации (deep clustering), в которых алгоритм кластеризации интегрирован внутрь многослойного нейросетевого автокодировщика. Эта модель применяется для объединения в группы поверхностных связей (триплетов), которые можно интерпретировать как семантические отношения. Представлен метод для извлечения терминов и поверхностных связей на основе правил и статистических данных.

Ключевые слова: открытое извлечение информации, семантические отношения, машинное обучение без учителя, нейронные сети, автокодировщик.

DOI 10.14357/20718594190204

Введение

Данная работа посвящена «открытому извлечению информации» из текстов на естественном языке (open informa tion extraction) [1]. В это направление входят различные задачи извлечения сущностей и семантических отношений, в которых типы отношений и сущностей заранее не задаются. В такой постановке нельзя вручную создать грамматику для выделения закрытого множества отношений. Также невозможно в полной мере использовать и методы машинного обучения с учителем, которые зависят от размеченных корпусов текстов. Качество извлечения сущностей и отношений с помощью методов открытого извлечения информации закономерно ниже, чем с помощью традиционных методов, основанных на обучающих корпусах или традиционных грамматиках. Однако такие методы позволяют создавать системы, которые не привязаны к конкретной предметной области и позволяют обрабатывать большие объемы разнородных текстов. Это свойство особенно полезно для поисково-аналитических систем общего назначения, а также для систем автоматизированного пополнения баз знаний.

Большое количество работ в области открытого извлечения информации посвящено проблеме извлечения «поверхностных» связей или триплетов: пар сущностей и фрагмента текста, указывающего на наличие какого-либо семантического отношения между ними. Среди по-

*Работа выполнена при поддержке РФФИ (проекты №17-07-01477 «А», № 16-29-12937 «офи_м»).

✉ *Смирнов Иван Валентинович. E-mail: ivs@isa.ru*

добных систем можно назвать TextRunner [2], Woe [3], ReVerb [4], ArgLearner [5]. Несмотря на то, что эти системы могут извлекать из текста большое количество связей между объектами, они не позволяют в явном виде представить семантику отношений между объектами. Чтобы получить непосредственно отношения, необходимо группировать связи по смыслу.

Данная работа продолжает исследование, представленное в [6], основная цель которого состоит в том, чтобы получить средства для автоматизированного пополнения баз знаний для различных предметных областей. В статье описывается подход к решению задачи извлечения семантических отношений из текстов на основе машинного обучения без учителя. Подход основан на методах глубокой кластеризации (deep clustering), в которых алгоритм кластеризации интегрирован внутрь многослойного нейросетевого автокодировщика. Эта модель применяется для кластеризации поверхностных связей (триплетов) и получения на этой основе семантических отношений. Также представлен метод для извлечения терминов и поверхностных связей на основе правил и статистических данных.

Статья структурирована следующим образом. В Разделе 1 представлены методы и подходы для извлечения семантических отношений на основе машинного обучения без учителя, в Разделе 2 – разработанные методы для извлечения поверхностных связей и семантических отношений, в Разделе 3 – экспериментальные исследования разработанных методов.

1. Обзор методов извлечения семантических отношений на основе машинного обучения без учителя

В области извлечения из текстов семантических отношений с помощью методов машинного обучения без учителя можно выделить два направления (две задачи): определение шаблонов для извлечения бинарных семантических отношений между парой сущностей для каждого упоминания пары в каждом конкретном предложении и извлечение отношений между сущностями как элементов знаний на основе всех имеющихся текстовых данных. В первом случае между парой сущностей могут проявляться различные отношения, которые можно определить в каждом отдельном упоминании пары, тогда как во втором случае между парой сущностей обычно находится одно семантическое отношение. Второе направление можно назвать извлечением отношений между концептами, которые могут представлять собой элементы базы знаний. В этом случае важно само наличие отношения без его привязки к конкретному месту в тексте.

Первое направление предполагает кластеризацию лексико-синтаксических шаблонов (путей в синтаксическом дереве, расширенных лексикой) на основе того, какие сущности связаны с помощью этого шаблона. Основная гипотеза такого подхода заключается в том, что различные шаблоны со схожими по смыслу связываемыми сущностями с высокой вероятностью выражают одно и то же семантическое отношение. Эта гипотеза была предложена в работе [7]. Исследователи взяли несколько начальных синтаксических шаблонов, тип семантической связи которых известен. Затем из крупного неразмеченного корпуса извлекли новые шаблоны, рассчитали их близость к начальным в соответствии с предложенной метрикой на основе PMI и сформировали кластеры, соответствующие семантическим отношениям, отсекая мало похожие по порогу. Этот подход имеет две проблемы: трудоемкость расчета совместной встречаемости шаблонов и пар сущностей, а также высокое потребление памяти. В работе [8] предложено несколько способов решения этих проблем с помощью метода аппроксимации подсчета частот, метода снижения размерности (метод главных компонент), а также с помощью использования векторных представлений слов из модели word2vec [9].

В работе [10] для решения проблемы извлечения семантических отношений без привлечения учителя предлагается использовать подходы тематического моделирования, основанные на латентном размещении Дирихле (LDA) [11]. Авторы разработали несколько генеративных тематических моделей с разным набором признаков. В них место слов занимают триплеты, состоящие из пары связываемых сущностей и признаков контекста, а место тем занимают семантические отношения. Основной недостаток подхода на основе кластеризации лексико-синтаксических шаблонов заключается в том, что каждый шаблон оказывается, закреплен лишь за одним типом отношений. Однако мно-

гие шаблоны неоднозначны и могут выражать различные типы отношений в зависимости от контекста. Эта проблема решается в работе [12]. В ней используется два этапа кластеризации. На первом этапе пары сущностей, связанные одним лексико-синтаксическим шаблоном, кластеризуются по темам (спорт, развлечения и др.) с помощью LDA. При тематическом моделировании в этом случае синтаксические шаблоны выступают в роли документов («документ» содержит в себе все пары сущностей, связанных одним и тем же лексико-синтаксическим шаблоном).

На втором этапе в рамках полученных таким образом тематических кластеров при разных лексико-синтаксических шаблонах проводится еще одна кластеризация, в результате которой формируются семантические отношения, определенные тематикой и группой соответствующих им лексико-синтаксических шаблонов. На втором этапе используется агломеративная кластеризация, в которой объектами выступают тематические кластеры при разных лексико-синтаксических шаблонах, признаки которых сформированы из совокупности пар сущностей, входящих в исходные кластеры.

Еще одна модификация подхода, основанная на тематическом моделировании с помощью LDA, представлена в работе [13]. В ней исследователи совместили метод тематического моделирования для выделения семантических отношений с «внешними» ограничениями, заданными предикатами первого порядка. Был предложен алгоритм, позволяющий построить тематическую модель, которая нежестко удовлетворяет заданным ограничениям, содержащим некоторые «внешние» знания о предметной области. Для этого исследователи модифицировали модель First-Order Logic Latent Dirichlet Allocation, представленную в [14], а для настройки параметров применили представленный в этой же работе один из вариантов EM-алгоритма. В работе был также предложен способ для автоматического построения ограничений без учителя, основанный на анализе статистической совместной встречаемости признаков друг с другом. Исследователи показали, что применение логических ограничений в тематических моделях повышает качество извлечения отношений по сравнению с моделью LDA на связях.

В статье [15] для извлечения семантических отношений без учителя предлагается подход, основанный на минимизации ошибки восстановления входных данных (reconstruction error minimization). Модель представляет собой вариационный автокодировщик [16] с дискретными внутренними представлениями. В ней исходное признаковое представление преобразуется в параметры распределения, которое генерирует вектора, из которых затем восстанавливаются входные данные. Таким образом, сеть обучается генерировать такие параметры распределения, из которых затем можно восстановить исходную информацию. Авторы построили модель порождения семантического отношения при заданных аргументах и их контекстах (кодировщик) и модель порождения аргумента при заданном отношении и заданном другом аргументе (декодировщик). Параметры моделей обучаются совместно с помощью градиентного метода с адаптивным шагом. Обученный кодировщик используется для предсказания отношений между аргументами, а декодировщик является побочным результатом работы метода.

Второе направление, в котором из текстов извлекаются отношения между концептами, обычно предполагает кластеризацию пар сущностей по их контекстам. В работе [17] применяется иерархическая кластеризация пар сущностей, найденных в одном предложении. При этом в качестве признаков используются не сами сущности, а агрегированные статистики, построенные по совокупности их контекстов, состоящие из слов, находящихся между парой сущностей. В результате кластеризации группируются пары сущностей, между которыми имеется некоторое семантическое отношение. В работе [18] описывается метод извлечения отношений из текстов путем агломеративной кластеризации текстовых сообщений. В ней использовался корпус новостных текстов (рассматривались только первые 10 предложений), считалось, что каждый текст описывает какое-то одно отношение. При слиянии кластеров рассчитывались две оценки близости: по схожести лексики (без учета сущностей) и по схожести лексико-синтаксических шаблонов, при этом каждая из оценок имела свой порог. Авторы отмечают, что оценка близости по лексике необходима для того, чтобы отношения разных

типов, имеющие схожие шаблоны общего характера, не сливались в один кластер.

В работе [19] авторы извлекают шаблоны с помощью разметки на основе Wikipedia, а затем кластеризуют их на основе синтаксических признаков, извлеченных из предложений, в которых встречаются пары сущностей, а также поверхностных шаблонов, полученных с использованием поисковой машины в Интернете (в частности, исследователи использовали Google). Для получения данных, необходимых для построения поверхностных шаблонов, формируется поисковый запрос, используя глаголы и существительные из предложения, в котором встретилась соответствующая пара сущностей. Полученные из поисковой выдачи сниппеты используются для генерации поверхностных шаблонов вида «<последовательность слов1> аргумент1 <последовательность слов2> аргумент2 <последовательность слов3>». Для кластеризации был предложен метод, основанный на алгоритме k-средних, в котором при расчете расстояния используются и синтаксические признаки, и поверхностные шаблоны.

Важно также отметить ряд новых работ в области глубокой кластеризации (deep clustering) [20–24]. Этот подход совмещает в себе глубокую нейросетевую архитектуру типа автокодировщик (autoencoder) [25] и возможность получать кластеры непосредственно внутри сети без использования дополнительных алгоритмов кластеризации типа k-средних или аггломеративных методов. Глубокую кластеризацию активно начинают применять для анализа изображений. Однако на сегодняшний день не известны работы, в которых автокодировщики и глубокая кластеризация такого рода применялась бы для решения задач извлечения семантических отношений.

2. Разработанные методы и подходы

2.1. Метод извлечения терминов и поверхностных связей

Выделение концептов и поверхностных связей происходит с помощью следующего алгоритма:

1. Извлекаются все именные группы из синтаксических деревьев зависимостей:

- выполняется поиск вершины синтаксического поддерева (как правило, это глагол);
- происходит спуск по дереву до первого существительного.

2. Найденное существительное сохраняется вместе с потомками в качестве именной группы, если их части речи входят в следующий список: существительное, прилагательное, местоимение, числительное, имя собственное, союз, наречие, причастие. Генерация именной группы происходит для каждого существительного в поддереве.

3. Для каждой именной группы рассчитывается ее вес C-value согласно [26]. C-value учитывает длину именной группы, ее частоту, а также вложенность в другие именные группы:

$$\text{C-Value(a)} = \begin{cases} \log_2 |a| \times freq(a), & \text{если именная группа не вложена в другие} \\ \log_2 |a| \times freq(a) - \dfrac{1}{P(T_a)} \times \sum_{b \epsilon T_a} freq(b), \end{cases}$$

где a – именная группа; $|a|$ – количество слов в именной группе; $freq(a)$ – частота встречаемости a; T_a – именные группы, в которые входит a; $P(T_a)$ – количество именных групп, содержащих a.

4. В результате ранжирования списка именных групп по убыванию C-Value и, отсекая варианты с низким весом по порогу, формируется список многословных «терминов» корпуса.

5. Между терминами извлекаются поверхностные связи. Для этого выполняется спуск от глагола (вспомогательные глаголы игнорируются) до двух вершин терминов, которые назначаются аргументами триплета. При этом связь не устанавливается, если термины разделены пунктуацией или союзами. Триплеты также фильтруются дополнительным набором эвристик.

6. Глагол с двумя аргументами сохраняется в один триплет.

2.2. Метод извлечения семантических отношений на основе машинного обучения без учителя

Базовая модель для извлечения семантических отношений имеет архитектуру нейросетевого автокодировщика. Расширенная модель имеет также дополнительный кластеризующий слой. Автокодировщик состоит из кодировщика и декодировщика, которые совместно обучаются сжимать исходные вектора признаков в

векторные представления малой размерности и восстанавливать из этих представлений исходные данные с малой ошибкой. Таким образом, автокодировщик обучается выделять семантически важные компоненты, наиболее полезные для восстановления исходных данных.

В разработанном методе извлечения семантических отношений обучающими объектами для кодировщика являются признаковые описания выделенных триплетов, которые обозначают поверхностные связи. Обозначим за X – множество обучающих триплетов C_i, извлеченных из всего корпуса текстов: $X = \{C_i, i = 1..n\}$ (n – количество обучающих триплетов). Выход сети при обучении представляет собой множество восстановленных признаковых описаний триплетов $X' = \{C'_i, i = 1..n\}$. Пусть каждый триплет задается кортежем $C_i = \langle A_i^-, V_i, A_i^+ \rangle$, где A_i^-, A_i^+, V_i – признаковые описания первого и второго аргументов, а также глагола соответственно. Каждый аргумент A_i^-, A_i^+ задается кортежем $A_i^* = \langle E_{A_i^*}, P_{A_i^*}, \Lambda_{A_i^*} \rangle$, где $E_{A_i^*}$ – векторные представления лемм каждого отдельного слова, входящего в аргумент; $P_{A_i^*}$ – предлог главного слова в синтаксическом поддереве аргумента; $\Lambda_{A_i^*}$ – векторное представление леммы главного слова в синтаксическом поддереве аргумента. Признаковое описание глагола V_i представляет собой векторное представления его леммы. Пусть задано количество кластеров (семантических отношений) R – гиперпараметр модели.

Архитектура нейронной сети представлена на Рис. 1. Последовательности векторных представлений слов, входящих в аргументы, преобразуются однослойной одномерной сверточной сетью с выбором максимума. Применение слоя выбора максимума обусловлено тем, что аргументы часто состоят из нескольких словоупотреблений, лишь небольшое число из которых значимы при решении задачи. Полученные вектора конкатенируются с другими векторными представлениями признаков аргументов и глагола. Объединенный вектор проходит еще через три полносвязных слоя, в которых в качестве функции активации используется гиперболический тангенс, в результате чего формируется сжатое внутреннее векторное представление триплета z_i, которое используется далее в кластеризующем слое и декодировщике. Каждый последующий слой имеет меньше нейронов, чем предыдущий, поэтому размерность векторных представлений z_i существенно меньше размерности исходных векторов признаков триплетов.

Декодировщик пропускает векторное представление z_i через два полносвязных слоя, в которых также используется гиперболический тангенс в качестве функции активации. Каждый следующих слой имеет больше нейронов, чем предыдущий, таким образом, происходит декомпрессия внутреннего представления. После этого получившийся вектор направляется в кусочный выходной слой, состоящий из семи компонент: два сверточных транспонированных слоя для восстановления векторных представлений последовательностей слов двух аргументов, два полносвязных слоя с softmax активацией для восстановления предлогов аргументов, два слоя с tanh активацией для восстановления векторных представлений лемм главных слов аргументов, а также один слой с tanh активацией для восстановления леммы предикатного слова.

Рассмотрим функцию потерь, используемую в модели. Функция потерь для одного обучающего объекта в виде триплета C_i представляет собой взвешенную сумму трех элементов:

$$L = L_{rec} + \gamma_c L_c + \gamma_r L_2,$$

где L_{rec} – ошибка восстановления признакового описания триплета декодировщиком из внутреннего векторного представления; L_c – функция потерь на кластеризующем слое; L_2 – L2 регуляризация параметров нейронной сети; γ_c и γ_r – скалярные значения, являются гиперпараметрами модели и подбираются эмпирически.

Рассмотрим функцию L_{rec}, определяющую потерю восстановления исходных данных. Обозначим восстановленные векторные представления аргументов и глагола как $E_{A_i^{*\prime}}$, $P_{A_i^{*\prime}}$, $\Lambda_{A_i^{*\prime}}$, V_i'. Ошибка восстановления признаков триплета декодировщиком вычисляется как взвешенная сумма ошибки восстановления признаковых описаний на каждом фрагменте выходного кусочного слоя:

$$L_{rec} = \gamma_1 \left(L_{A_{lex}^+} + L_{A_{lex}^-} \right) + \gamma_2 \left(L_{A_{prep}^+} + L_{A_{prep}^-} \right) + \\ + \gamma_3 \left(L_{A_{lemma}^+} + L_{A_{lemma}^-} \right) + \gamma_4 L_V.$$

Здесь скалярные значения $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ также являются гиперпараметрами модели и подбираются эмпирически. Ошибки восстановления векторных представлений слов, входящих в аргументы $L_{A_{lex}^*}$, а также векторных представле-

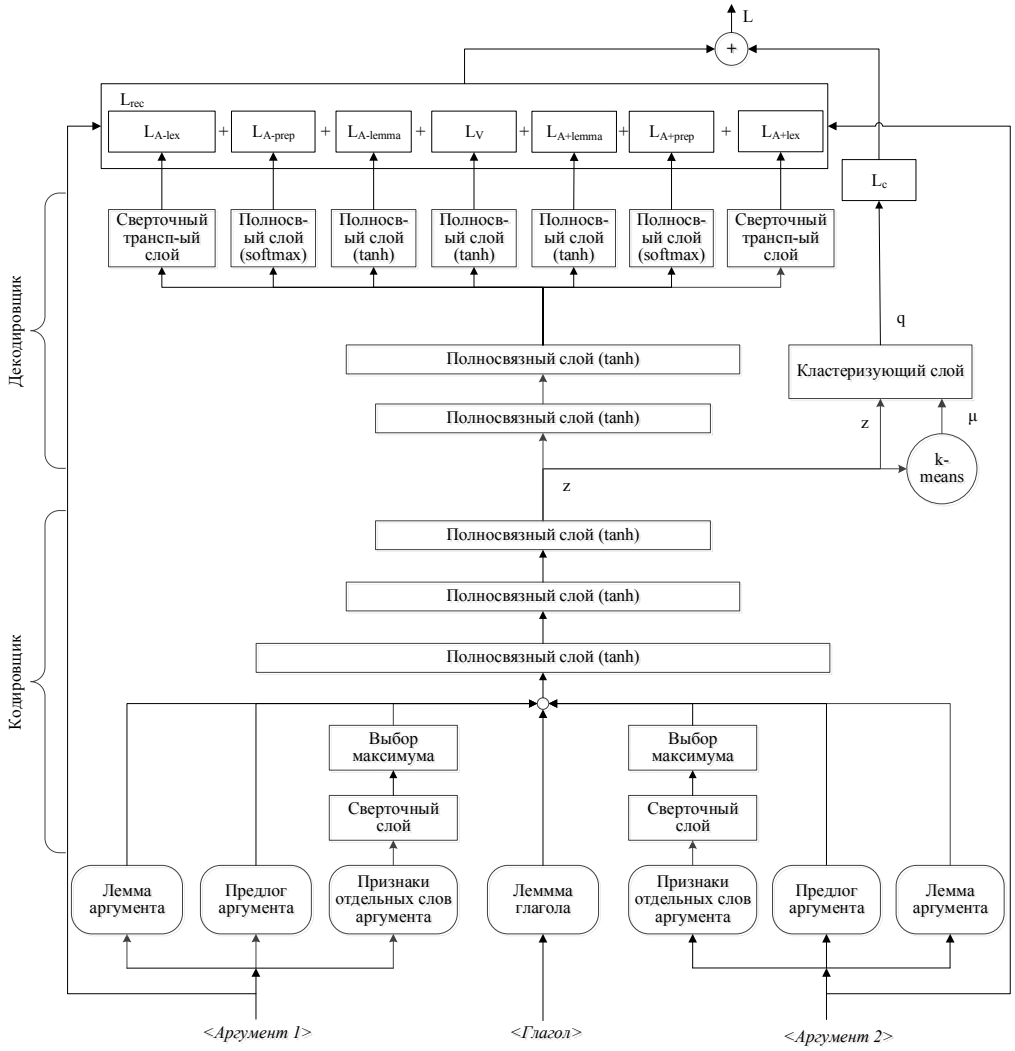

Рис. 1. Архитектура нейронной сети, применяемой в методе извлечения семантических отношений на основе машинного обучения без учителя

ний лемм главного слова аргумента и глагола $L_{A^+_{lemma}}$, L_V рассчитываются как среднеквадратичная ошибка восстановленных векторных представлений от исходных:

$$L_{A^*_{lex}} = \frac{1}{n}\sum_{i=1}^{n}\left(E_{A_i^{*\prime}} - E_{A_i^*}\right)^2,$$

$$L_{A^*_{lemma}} = \frac{1}{n}\sum_{i=1}^{n}\left(\Lambda_{A_i^{*\prime}} - \Lambda_{A_i^*}\right)^2,$$

$$L_V = \frac{1}{n}\sum_{i=1}^{n}(V_i' - V_i)^2.$$

Ошибка восстановления векторных представлений предлогов L_{feat} рассчитывается как бинарная кросс-энтропия:

$$L_{A^*_{prep}} = \sum_{i=1}^{n}\left(P_{A_i^*}\log P'_{A_i^*} + \left(1 - P_{A_i^*}\right) * \log\left(1 - P'_{A_i^*}\right)\right).$$

Процесс обучения сети проходит в два этапа. На первом этапе модель предобучается с использованием лишь одной функции потерь L_{rec} в течение нескольких итераций. На втором этапе к функции потерь добавляется ошибка кластеризации L_c. Дополнительное слагаемое в функции потерь необходимо, чтобы сжатые векторные представления обучающих объектов могли быть хорошо раскластеризованы. Глубокий автокодировщик выполняет эффективное сжатие информации так, чтобы из этого представления возможно было восстановить исходную информацию, однако подобные представления не обязательно позволяют эффективно группировать

исходные объекты. Так, например, метод главных компонент также осуществляет проекцию всех точек в исходном многомерном пространстве на подпространство меньшей размерности, однако такое проецирование не помогает группировке исходных объектов.

Для глубокой кластеризации был выбран подход, представленный в работе [20]. Ошибка кластеризации L_c представляет собой дивергенцию Кульбака-Лейблера между целевым высококонтрастным распределением P и распределением на выходе кластеризующего слоя Q:

$$L_c = KL(P||Q).$$

Распределение Q формируется следующим образом. С помощью распределения Стьюдента для каждого аргумента с внутренним представлением z_i формируется оценка вероятности, с которой объект принадлежит каждому кластеру (семантическому отношению) j:

$$q_{ij} = \frac{\left(1 + \left\|z_i - \mu_j\right\|^2\right)^{-1}}{\sum_j \left(1 + \left\|z_i - \mu_j\right\|^2\right)^{-1}}.$$

Здесь $\{\mu_j\}_{j=1}^R$ – центроиды кластеров, построенных на предыдущем шаге. Начальные значения $\{\mu_j\}_{j=1}^R$ формируются как центроиды кластеров, построенных на представлениях объектов z_i, некоторым стандартным алгоритмом кластеризации (в частности, использовался алгоритм k-means). Начальные представления объектов строятся с помощью предобученного автокодировщика без кластеризующего слоя. На последующих шагах параметры кластеризующего слоя $\{\mu_j\}_{j=1}^R$ преобразуются с помощью метода обратного распространения ошибки совместно с параметрами автокодировщика. Форма распределения P была предложена авторами работы [20] в соответствии с эвристиками, позволяющими повысить качество кластеризации:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})}.$$

После обучения модели с помощью нее можно проанализировать обучающий корпус текстов заново и построенные кластеры разметить смысловыми метками (семантическими отношениями). Это можно сделать, например, вручную, поскольку количество кластеров-отношений в данном случае будет невелико.

Общая схема извлечения отношений между сущностями с использованием обученной модели на основе глубокой кластеризации следующая:

- Извлечь из текста все триплеты: пары сущностей (аргументы) и глагол между ними.
- С помощь набора эвристик отфильтровать извлеченные триплеты (например, по частоте глагола или аргументов).
- С помощью кодировщика и кластеризующего слоя из вышеописанной модели назначить триплетам номера кластеров.
- Выбрать те объекты, которые лежат близко к центроидам кластеров (предсказанная моделью вероятность принадлежности к кластеру выше заданного порога).
- По номеру кластера назначить каждому триплету смысловую метку – семантическое отношение.

3. Экспериментальные исследования разработанных методов

В экспериментальных исследованиях метода извлечения отношений между сущностями обучение проводилось на русскоязычном корпусе текстов по информационным технологиям, а тестирование – на его размеченном тестовом подкорпусе. Исходный обучающий корпус содержит в себе более 40 тыс. текстов суммарным объемом более 29 млн токенов с учетом пунктуации.

Корпус был предобработан с помощью лингвистического анализатора, разработанного в ФИЦ ИУ РАН [27], который выполняет морфологический, синтаксический и семантический анализ текста. На основе корпуса было построено более 600 тыс. триплетов. Часть триплетов была отфильтрована по частоте входящего в него глагола. Для фильтрации также использовался ряд других незначительных эвристик. Для формирования векторных представлений слов использовалась модель RusVectores, построенная по корпусу русскоязычных новостей [28]. При тестировании учитывались только 10 отношений в тестовом подкорпусе.

Для оценки качества использовался подход, схожий с тем, что применялся в работах [29, 30] для оценки качества определения ролевых структур высказываний с помощью машинного обучения без учителя. В этих работах предлага-

ется оценивать метрики чистоты (purity), коллокации (collocation) и их среднего гармонического, используя тестовый размеченный корпус. Точность в этих работах рассчитывается как среднее значение максимального процента примеров, входящих в построенный кластер, которые имеют одну и ту же метку в тестовом корпусе. Полнота рассчитывается как среднее значение максимального процента примеров, имеющих одну и ту же метку в тестовом корпусе, входящих в один построенный кластер. Пусть N обозначает количество тестовых примеров, G_j – количество примеров, имеющих j-ую метку в тестовом корпусе, C_i – множество примеров, принадлежащих i-му кластеру. Тогда можно записать следующие формулы для расчета чистоты Pu, коллокации Co и F_1 меры:

$$\mathrm{P}u = \frac{1}{N}\sum_i \max_j |G_j \cap C_i|$$

$$Co = \frac{1}{N}\sum_j \max_i |G_j \cap C_i|$$

$$F_1 = \frac{2 \cdot Co \cdot Pu}{Co + Pu}$$

В отличие от работ [29, 30] ищется наилучшее сопоставление кластеров семантическим отношениям, при котором F_1-мера максимальна, но в котором один и тот же кластер нескольким меткам не назначается. Наилучшее распределение кластеров по семантическим отношениям можно найти, решив задачу о назначениях. Предложенная метрика позволяет более объективно оценить кластеризацию и всегда меньше, чем метрика, предложенная в работах [29, 30].

Разработанный метод сравнивался с двумя базовыми подходами: тривиальный и случайный. В тривиальном подходе всем триплетам назначается одно семантическое отношение. В случайном всем триплетам отношения назначаются случайным образом. Результаты экспериментальных исследований представлены в Табл. 1.

Табл. 1. Результаты экспериментальных исследований метода кластеризации триплетов

Подход	F_1, %
Тривиальный	10,7
Случайный	25,2
Глубокая кластеризация на векторных представлениях глаголов	46,3
Глубокая кластеризация на всех признаках	53,0

Полученные результаты демонстрируют, что метод глубокой кластеризации дает осмысленную группировку триплетов, которые могут рассматриваться как реализации одного и того же семантического отношения.

Заключение

В статье представлены методы открытого извлечения информации из текстов, основанные на машинном обучении без учителя. Показано, что метод на основе глубокой кластеризации позволяет сгруппировать связи в семантические отношения, которые соотносятся с экспертной разметкой. В дальнейшем признаки групп потенциально можно использовать для разметки новых текстов. Однако на данный момент метод имеет ряд ограничений. Для его эффективного использования необходима большая тематически однородная коллекция. Необходимы также дополнительные эвристики для фильтрации данных (в первую очередь по частоте слов, входящих в триплеты).

В дальнейшем планируется исследовать возможность применения других подходов глубокой кластеризации и современные способы аддитивной регуляризации. Кроме того, в настоящей работе не исчерпаны возможности по наращиванию признакового пространства для представления триплетов. В будущем также планируется улучшить качество извлечения поверхностных связей за счет разработки более качественных эвристик, что может также повысить качество выделения семантических отношений. Стоит отметить, что разработанный метод хорошо масштабируется, что открывает возможности для проведения экспериментальных исследований на корпусах большего размера.

Литература

1. Open information extraction from the web. / Michele Banko, Michael J Cafarella, Stephen Soderland et al. // IJCAI. — Vol. 7. — 2007. — P. 2670–2676.
2. Textrunner: open information extraction on the web / Alexander Yates, Michael Cafarella, Michele Banko et al. // Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations / Association for Computational Linguistics. — 2007. — P. 25–26.
3. Wu F., Weld D. S. Open information extraction using wikipedia // Proceedings of the 48th annual meeting of the as-

sociation for computational linguistics / Association for Computational Linguistics. — 2010. — P. 118–127.
4. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the conference on empirical methods in natural language processing / Association for Computational Linguistics. — 2011. — P. 1535–1545.
5. Open information extraction: The second generation. / Oren Etzioni, Anthony Fader, Janara Christensen et al. // IJCAI. — Vol. 11. — 2011. — P. 3–10.
6. Открытое извлечение информации из текстов. Часть 1. Постановка задачи и обзор методов / А.О. Шелманов, В.А. Исаков, М.А. Станкевич, И.В. Смирнов // Искусственный интеллект и принятие решений. — 2018. — № 2. — С. 47–61.
7. Lin D., Pantel P. Discovery of inference rules for question-answering // Natural Language Engineering. — 2001. — Vol. 7, no. 4. — P. 343–360.
8. Takase S., Okazaki N., Inui K. Fast and large-scale unsupervised relation extraction // Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. — 2015. — P. 96–105.
9. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — P. 3111–3119.
10. Structured relation discovery using generative models / Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum // Proceedings of the Conference on Empirical Methods in Natural Language Processing / Association for Computational Linguistics. — 2011. — P. 1456–1466.
11. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Vol. 3, no. Jan. — P. 993–1022.
12. Yao L., Riedel S., McCallum A. Unsupervised relation discovery with sense disambiguation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 / Association for Computational Linguistics. — 2012. — P. 712–720.
13. De Lacalle O. L., Lapata M. Unsupervised relation extraction with general domain knowledge // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 415–425.
14. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic / David Andrzejewski, Xiaojin Zhu, Mark Craven, Benjamin Recht // IJCAI Proceedings-International Joint Conference on Artificial Intelligence. — Vol. 22. — 2011. — P. 1171.
15. Marcheggiani D., Titov I. Discrete-state variational autoencoders for joint discovery and factorization of relations // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 231–244.
16. Kingma D. P., Welling M. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. — 2013.
17. Hasegawa T., Sekine S., Grishman R. Discovering relations among named entities from large corpora // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics / Association for Computational Linguistics. — 2004. — P. 415.
18. Shinyama Y., Sekine S. Preemptive information extraction using unrestricted relation discovery // Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics / Association for Computational Linguistics. — 2006. — P. 304–311.
19. Unsupervised relation extraction by mining wikipedia texts using information from the web / Yulan Yan, Naoaki Okazaki, Yutaka Matsuo et al. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 / Association for Computational Linguistics. — 2009. — P. 1021–1029.
20. Xie J., Girshick R., Farhadi A. Unsupervised deep embedding for clustering analysis // International conference on machine learning. — 2016. — P. 478–487.
21. Deep clustering with convolutional autoencoders / Xifeng Guo, Xinwang Liu, En Zhu, Jianping Yin // International Conference on Neural Information Processing / Springer. — 2017. — P. 373–382.
22. Tian K., Zhou S., Guan J. Deepcluster: A general clustering framework based on deep learning // Joint European Conference on Machine Learning and Knowledge Discovery in Databases / Springer. — 2017. — P. 809–825.
23. Learning deep representations for graph clustering. / Fei Tian, Bin Gao, Qing Cui et al. // AAAI. — 2014. — P. 1293–1299.
24. Auto-encoder based data clustering / Chunfeng Song, Feng Liu, Yongzhen Huang et al. // Iberoamerican Congress on Pattern Recognition / Springer. — 2013. — P. 117–124.
25. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks // science. — 2006. — Vol. 313, no. 5786. — P. 504–507.
26. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms:. the c-value/nc-value method // International journal on digital libraries. — 2000. — Vol. 3, no. 2. — P. 115–130.
27. Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов / И. В. Смирнов, А. О. Шелманов, Е. С. Кузнецова, И. В. Храмоин // Искусственный интеллект и принятие решений. — 2014. — № 1. — С. 11–24.
28. Kutuzov A., Kuzmenko E. Webvectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2016. — P. 155–161.
29. Lang J., Lapata M. Unsupervised semantic role induction via split-merge clustering // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies / Association for Computational Linguistics. — 2011. — P. 1117–1126.
30. Titov I., Khoddam E. Unsupervised induction of semantic roles within a reconstruction-error minimization framework // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — P. 1–10.

Open information extraction from texts Part II. Extraction of semantic relations using unsupervised machine learning

A.O. Shelmanov, D.A. Devyatkin, V.A. Isakov, I.V. Smirnov

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

Abstract. In this paper, we discuss open information extraction from natural language texts. We present the approach to extraction of semantic relations using unsupervised machine learning. The presented approach is based on deep clustering methods in which clusterization algorithm is integrated in multi-layer autoencoder neural network. This method allows to generalize surface relations (triplets) into semantic relations. This paper also provides the method of surface relation extraction.

Keywords: open information extraction, semantic relations, unsupervised machine learning, neural networks, autoencoder.

DOI 10.14357/20718594190204

References

1. Open information extraction from the web. / Michele Banko, Michael J Cafarella, Stephen Soderland et al. // IJCAI. — Vol. 7. — 2007. — P. 2670–2676.
2. Textrunner: open information extraction on the web / Alexander Yates, Michael Cafarella, Michele Banko et al. // Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations / Association for Computational Linguistics. — 2007. — P. 25–26.
3. Wu F., Weld D. S. Open information extraction using wikipedia // Proceedings of the 48th annual meeting of the association for computational linguistics / Association for Computational Linguistics. — 2010. — P. 118–127.
4. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the conference on empirical methods in natural language processing / Association for Computational Linguistics. — 2011. — P. 1535–1545.
5. Open information extraction: The second generation. / Oren Etzioni, Anthony Fader, Janara Christensen et al. // IJCAI. — Vol. 11. — 2011. — P. 3–10.
6. Shelmanov, A.O., V.A. Isakov, M.A. Stankevich and I.V. Smirnov. 2018. Otkrytoe izvlechenie informatsii iz tekstov chast' 1. Postanovka zadachi i obzor metodov. [Open information extraction from texts part 1. Problem statement and survey of methods]. Iskusstvennyj intellekt i prinyatie reshenij [Artificial intelligence and decision-making] 2:47-61.
7. Lin D., Pantel P. Discovery of inference rules for question-answering // Natural Language Engineering. — 2001. — Vol. 7, no. 4. — P. 343–360.
8. Takase S., Okazaki N., Inui K. Fast and large-scale unsupervised relation extraction // Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. — 2015. — P. 96–105.
9. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — P. 3111–3119.
10. Structured relation discovery using generative models / Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum // Proceedings of the Conference on Empirical Methods in Natural Language Processing / Association for Computational Linguistics. — 2011. — P. 1456–1466.
11. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. — 2003. — Vol. 3, no. Jan. — P. 993–1022.
12. Yao L., Riedel S., McCallum A. Unsupervised relation discovery with sense disambiguation // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 / Association for Computational Linguistics. — 2012. — P. 712–720.
13. De Lacalle O. L., Lapata M. Unsupervised relation extraction with general domain knowledge // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 415–425.
14. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic / David Andrzejewski, Xiaojin Zhu, Mark Craven, Benjamin Recht // IJCAI Proceedings-International Joint Conference on Artificial Intelligence. — Vol. 22. — 2011. — P. 1171.
15. Marcheggiani D., Titov I. Discrete-state variational autoencoders for joint discovery and factorization of relations // Transactions of the Association for Computational Linguistics. — 2016. — Vol. 4. — P. 231–244.
16. Kingma D. P., Welling M. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. — 2013.
17. Hasegawa T., Sekine S., Grishman R. Discovering relations among named entities from large corpora // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics / Association for Computational Linguistics. — 2004. — P. 415.
18. Shinyama Y., Sekine S. Preemptive information extraction using unrestricted relation discovery // Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Asso-

ciation of Computational Linguistics / Association for Computational Linguistics. — 2006. — P. 304–311.
19. Unsupervised relation extraction by mining wikipedia texts using information from the web / Yulan Yan, Naoaki Okazaki, Yutaka Matsuo et al. // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 / Association for Computational Linguistics. — 2009. — P. 1021–1029.
20. Xie J., Girshick R., Farhadi A. Unsupervised deep embedding for clustering analysis // International conference on machine learning. — 2016. — P. 478–487.
21. Deep clustering with convolutional autoencoders / Xifeng Guo, Xinwang Liu, En Zhu, Jianping Yin // International Conference on Neural Information Processing / Springer. — 2017. — P. 373–382.
22. Tian K., Zhou S., Guan J. Deepcluster: A general clustering framework based on deep learning // Joint European Conference on Machine Learning and Knowledge Discovery in Databases / Springer. — 2017. — P. 809–825.
23. Learning deep representations for graph clustering. / Fei Tian, Bin Gao, Qing Cui et al. // AAAI. — 2014. — P. 1293–1299.
24. Auto-encoder based data clustering / Chunfeng Song, Feng Liu, Yongzhen Huang et al. // Iberoamerican Congress on Pattern Recognition / Springer. — 2013. — P. 117–124.
25. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks // science. — 2006. — Vol. 313, no. 5786. — P. 504–507.
26. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms:. the c-value/nc-value method // International journal on digital libraries. — 2000. — Vol. 3, no. 2. — P. 115–130.
27. Smirnov, I.V., A.O. Shelmanov, E.S. Kuznetsova and I.V. Hramoin. 2014. Semantiko-sintaksicheskij analiz estestvennykh yazykov chast' II. Metod semantiko-sintaksicheskogo analiza tekstov. [Semantico-syntactic analysis of natural languages Part II. Method of semantic-syntactic analysis of texts]. Iskusstvennyj intellekt i prinyatie reshenij [Artificial intelligence and decision-making] 1:11-24.
28. Kutuzov A., Kuzmenko E. Webvectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2016. — P. 155–161.
29. Lang J., Lapata M. Unsupervised semantic role induction via split-merge clustering // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies / Association for Computational Linguistics. — 2011. — P. 1117–1126.

Titov I., Khoddam E. Unsupervised induction of semantic roles within a reconstruction-error minimization framework // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015. — P. 1–10.