Р.Е. Суворов, А.О. Шелманов, М.А. Каменская, И.В. Смирнов

Активное машинное обучение в задаче извлечения информации из научных текстов[1]

Аннотация. Рассматривается задача извлечения информации из текстов с применением методов машинного обучения. Традиционно для построения системы извлечения информации на основе машинного обучения требуется разметка достаточно больших корпусов текстов. Другой проблемой, возникающей при создании такого рода систем, является необходимость построения специального признакового пространства. Для решения первой проблемы предложены методы извлечения информации на основе активного машинного обучения. Для решения второй проблемы предложены методы генерации признакового пространства на основе результатов полного лингвистического анализа. Проведены экспериментальные исследования предложенных методов. Показано, что использование активного машинного обучения существенно сокращает трудоемкость создания системы извлечения информации, сохраняя при этом качество решения задачи.

Ключевые слова: извлечение информации из текстов, полный лингвистический анализ, активное машинное обучение, построение признакового пространства, обработка научных текстов.

Введение

На основе анализа первичных научных публикаций (статей, монографий, докладов) можно оценить результативность представленных в них исследований, выявить элементы новизны, выражающиеся, например, введением новых категорий. Таким образом, большую важность для такого анализа представляют содержащиеся в этих текстах формулировки результатов и определений новых понятий.

В данной работе решаются задачи извлечения результатов и новых понятий из научных текстов. Процесс извлечения информации включает в себя подзадачу выявления в тексте фрагментов (абзацев, предложений, подстрок и т.п.), потенциально содержащих релевантную информацию. Эта задача может быть решена с помощью классификации фрагментов текста. В нашей работе предлагаются методы извлечения определений и результатов, основанные на машинном обучении, в которых эти задачи рассматриваются как задачи классификации предложений.

Для минимизации трудозатрат на создание размеченной обучающей выборки текстов предлагается использовать методы активного обучения (от англ., active learning), а для автоматизации процесса построения признакового пространства предлагаются обобщенные, непривязанные к конкретным задачам подходы: на основе n-грамм лемм и частей речи, ролевых структур, синтаксических конструкций (словосочетаний), а также автоматически сгенерированных правил анализа контекста слов.

Суть активного обучения состоит в том, что обучаемая модель итеративно перестраивается в процессе разметки экспертом примеров, автоматически выбранных на основе анализа ответов текущей версии модели. Примеры для разметки выбираются в соответствии с некоторой заданной стратегией, например, с целью получения наибольшего потенциального прироста качества модели. Для этого обычно используются предсказания частично обученной модели на неразмеченных данных и степень уверенности модели. Такой подход к разметке обучающих выборок может быть особенно вы-

[1] Работа выполнена при поддержке РФФИ, проект №14-29-05023 «офи_м».

годным, когда доля положительных объектов среди всех объектов обучающей выборки очень мала, поскольку в активном обучении можно выбрать стратегию, которая позволит более эффективно находить положительные примеры. Именно эта ситуация возникает при решении задач извлечения определений и результатов из текстов научных публикаций.

В работе представлены новые стратегии активного обучения с учетом плотности примеров, проведена экспериментальная оценка и сравнение качества извлечения информации из текстов в рамках парадигм активного и традиционного обучения с учителем. Приводится сравнение нескольких методов машинного обучения и способов построения признакового пространства. Оценивается выгода от использования активного машинного обучения с точки зрения трудозатрат экспертов на разметку текстов.

1. Обзор смежных работ

Можно выделить две наиболее близкие к настоящей статье области исследований: извлечение информации из текстов в целом и анализ научных текстов. С учетом того, что за последнее время был опубликован ряд обзорных работ, посвященных, в том числе перечисленным тематикам [1, 2], в приведенном обзоре акцент сделан на наиболее релевантных и современных работах, опубликованных в предыдущие два-четыре года.

1.1. Подходы к снижению трудозатрат на построение систем извлечения информации из текстов

В области извлечения информации существует разрыв в подходах, используемых в академической среде, и в системах промышленного уровня [3]. Разработчики промышленных систем используют в первую очередь подходы на основе ручного построения правил. Несмотря на их очевидное достоинство в том, что подобные правила понятны человеку, проблема таких подходов состоит в больших трудозатратах на отладку и поддержание правил. В академической среде исследователи пытаются решить эту проблему с помощью методов машинного обучения. Однако, несмотря на очевидные успехи в этой области, классические подходы на основе обучения с учителем также сталкиваются с рядом проблем. При разработке систем извлечения информации из текстов часто возникает проблема отсутствия обучающей выборки или недостаточности ее размера, а ее построение или расширение может быть слишком трудоемко. В этом подразделе мы рассмотрим ряд подходов, которые призваны частично решить эту проблему.

В задаче извлечения информации из текстов для снижения трудозатрат чаще всего применяются: бутстрэппинг (bootstrapping), активное обучение, составление правил с подсказками, а также обучение без учителя или с минимальным привлечением учителя (distant supervision).

Бутстрэппинг – общий итерационный алгоритм построения классификатора, при котором на начальном этапе задается небольшое множество положительных примеров, а затем начинается чередование шагов уточнения классификатора и нахождения новых примеров. Бутстрэппинг применяется для расширения словарей и построения правил (лексико-синтаксических шаблонов) для извлечения информации [4], а также используется совместно со сложными классификаторами наподобие условных случайных полей и рекуррентных нейронных сетей [5].

Активное обучение и составление правил с подсказками являются интерактивными итерационными подходами, учитывающими обратную связь от эксперта в процессе построения классификатора. В активном обучении обратная связь от эксперта обычно заключается в выставлении эталонных меток для малого числа специально отобранных примеров. Существуют также работы, в которых вместо разметки примеров выполняется разметка признаков. Показано, что это значительно ускоряет и упрощает процесс разметки обучающей выборки и построения классификатора [6]. Составление правил с подсказками – вариант подхода с разметкой признаков, который применяется в классификаторах, основанных на правилах [7]. Актуальным направлением исследований в области активного обучения является использование внешних источников знаний, не требующих дополнительных ручных трудозатрат, например, векторных представлений слов [8] или онтологий для более адекватного выбора примеров для разметки [9].

Обучение без учителя или с минимальным привлечением его в большей степени относятся к областям открытого извлечения информации

(open information extraction), построения и расширения баз знаний [10]. Однако эти подходы не применимы в задачах извлечения информации, когда упоминания одних и тех же фактов встречаются редко. Такой задачей, например, является извлечение результатов исследований из научных статей.

Среди всех вышеперечисленных подходов, на наш взгляд, активное обучение является наиболее перспективным для использования, как в академических, так и в коммерческих системах. Оно может значительно экономить трудозатраты при решении многих задач, при этом позволяет достигать эффективного баланса между качеством моделей и затраченных на разметку ресурсов.

В данной работе предлагается ряд стратегий активного обучения, которые позволяют либо ускорить по сравнению с другими стратегиями рост качества модели в зависимости от количества размеченных примеров, либо ускорить разметку только заданных типов обучающих объектов (например, только положительных примеров).

1.2. Задачи извлечения информации из научных текстов

Обзор литературы показал, что в области анализа научных текстов решаются следующие основные задачи: извлечение метаданных [11], анализ структуры (дискурса) [12], извлечение определений, извлечение таблиц и графиков [13, 14], извлечение специализированной информации (например, извлечение результатов экспериментов особенно популярно в области биомедицины [15]).

Известно несколько работ, посвященных извлечению определений понятий из текстов. Часть из них направлена на извлечение определений из коллекций документов и использует общую статистику и связанность по цитированиям [16]. Мы не следуем этому подходу, так как для него необходимо наличие достаточно большой и связной коллекции документов по заданной тематике. Появляются работы, в которых предпринимаются попытки решать более общую задачу — выделение ключевых элементов статьи (задач, процессов/инструментов, материалов/корпусов/наборов данных) и отношений между ними [17]. Однако показатели качества решений этих задач весьма низки. Например, доля правильных ответов, показанных системами участников дорожки ScienceIE SemEval 2017, не превышает 0,3.

Задача извлечения определений в наиболее близкой нам постановке решается в [18-20]. В [20] для решения задачи используются вручную составленные лексико-синтаксические шаблоны и машинное обучение в зависимости от типа определения. На выбранных трех типах определений достигается F1-мера от 0,50 до 0,71. В [18] используются шаблоны без применения машинного обучения для извлечения определений из ненаучных текстов (например, Википедии). Достигаются показатели точности от 30 до 70% на различных англоязычных корпусах. Авторы делают заключение, что подавляющее число определений выделяется малым числом простых правил, однако для выделения остальных требуется составлять большое число сложных эвристик. В [19] применяются условные случайные поля и бутстрэппинг для классификации предложений, потенциально содержащих определения. В работах российских исследователей [21] извлечение определений основывается на составлении правил и лексико-синтаксических шаблонов.

2. Подходы и методы извлечения информации из научных публикаций на основе активного обучения

2.1. Методы машинного обучения, признаковые пространства

В данной работе задача извлечения определений и результатов из научных текстов решается как задача классификации предложений, содержащих формулировки дефиниций или полученных результатов. В ходе проведения исследования было выявлено, что в научных публикациях большая часть дефиниций занимает не более одного предложения. Наиболее значимая часть результатов также чаще всего представлена в виде одного предложения.

При создании систем извлечения информации часто используют узкоспециализированные признаки, пригодные лишь для решения частных задач. Целью же настоящей работы является разработка методов для ускоренного создания размеченных корпусов и обучения моделей, пригодных для решения широкого

круга задач, связанных с извлечением информации, а не только лишь с извлечением определений и научных результатов. Поэтому одним из этапов исследования была разработка обобщенных способов формирования признаковых описаний фрагментов текстов. В работе было рассмотрено два способа формирования признаковых описаний:

1. n-граммы лемм слов и частей речи;
2. n-граммы, дополненные словосочетаниями, выделенными в синтаксическом дереве предложения, признаками на основе ролевых структур предложения [22], а также правилами, в которых анализируется контекст леммы в некотором окне (контекстные правила).

Первый подход, основанный на n-граммах, использует лишь морфологическую информацию, извлеченную из текстов. Для него необходим только морфологический анализ, который в настоящей работе проводился с помощью библиотеки АОТ [23]. Это пространство признаков включает леммы и части речи слов в предложении, их биграммы и триграммы.

Во втором подходе используется также синтаксическая и семантическая информация. Синтаксический анализ выполнялся с помощью анализатора MaltParser [24], обученного на корпусе СинТагРус [25]. Определение ролевых структур выполнялось с помощью семантического анализатора, разработанного в ИСА ФИЦ ИУ РАН [26]. Признаки на основе словосочетаний формировались только из синтаксически связанных пар слов, в которых одно слово было глаголом или глагольной формой. Признаки на основе ролевых структур представляли собой пары вида: семантическая роль аргумента + лемма предикатного слова. Контекстные правила представляли собой пары и тройки вида (здесь => обозначает отношение «следует за» в некотором окне токенов): «лемма главного слова» => «морфологические признаки слова справа от главного слова»; «морфологические признаки слова слева от главного слова» => «лемма главного слова»; «морфологические признаки слова слева от главного слова» => «лемма главного слова» => «морфологические признаки слова справа от главного слова». При извлечении признаков каждое слово предложения по очереди назначается «главным», и для него генерируются вышеупомянутые правила, в ходе чего просматриваются левые и правые слова, которые находятся не дальше заданного предела в токенах от «главного» слова. Подобные правила призваны находить информативные шаблоны, менее зависимые от конкретных позиций слов, чем n-граммы.

Для применения методов машинного обучения был создан конвейер обработки данных, состоящий из двух компонент: компонент для предварительного отбора признаков (по дисперсии значений признака и по среднему весу признака в логистической регрессии с L1 регуляризацией) и классификатора. Использование предварительного отбора в некоторых случаях позволяет сократить более 99% незначимых признаков и ускорить обучение.

Непосредственно для классификации применялись следующие методы машинного обучения: логистическая регрессия с L2 регуляризацией, линейный SVM, случайный лес деревьев решений, метод градиентного бустинга (реализация из пакета LightGBM[2]), нейронная сеть (реализована на основе пакета Keras[3]), а также ансамбль из нескольких классификаторов.

Используемая в работе нейронная сеть представляет собой трехслойный перцептрон. Два первых слоя имеют ректификационную функцию активации, последний слой – сигмоидную. Количество нейронов на первом слое – 1/3 от количества признаков, на втором слое – 1/6. Сеть регуляризована с помощью методики выборочного обновления весов (dropout), в первых двух слоях используется локальная нормализация (batch normalization).

Ансамбль классификаторов реализует простое усреднение вероятностей, полученных от нескольких других классификаторов, включая: четыре классификатора на основе LightGBM, каждый из которых обучался на подмножестве признаков; один классификатор на основе LightGBM, обученный на полном наборе признаков; четыре нейронные сети, каждая из которых обучалась на случайном подмножестве признаков; одна нейронная сеть, обученная на всех признаках. Подобное ансамблирование призвано сделать итоговый классификатор более устойчивым к новым данным, поскольку каждая группа методов аппроксимирует целевую функцию по-разному, и таким образом может нивелировать ошибки остальных.

[2] https://github.com/Microsoft/LightGBM

[3] https://keras.io/

2.2. Стратегии активного обучения

В данной работе были реализованы и экспериментально оценены пять стратегий выбора примеров:

1. Случайная стратегия (*random*).
2. Выбор по наименьшей уверенности (*uncertainty sampling, US*).
3. Выбор по наименьшей уверенности с учетом плотности примеров (*adaptive density weighted uncertainty sampling, ADWUS*).
4. Выбор по наибольшей предсказываемой вероятности (*maximum probable error, MPERR*).
5. Выбор по наибольшей предсказываемой вероятности с учетом плотности (*ADWUS-MPERR*).

Случайная стратегия реализуется путем случайного выбора заданного числа примеров из множества всех неразмеченных примеров с одинаковой вероятностью для всех примеров. Эта стратегия, по сути, приблизительно соответствует ручной разметке без активного обучения в том смысле, что выбор следующего примера для разметки не является целенаправленным.

Стратегия выбора по наименьшей уверенности (*US*) ориентирована на разметку примеров, наиболее близких к разделяющей гиперповерхности. В случае с классификатором, предсказывающим вероятности (например, логистической регрессией), в качестве оценки перспективности примера используется близость вероятности к 0,5 (для бинарной классификации). Эта стратегия стремится уточнять положение гиперповерхности путем относительно небольших изменений. Однако, если начальное приближение гиперповерхности далеко от оптимального, сходимость может потребовать большого числа итераций. К тому же в этой стратегии не учитывается в явном виде совместное распределение объектов, поэтому выбор может падать на выбросы, правильная классификация которых бывает затруднительна даже для человека и бесполезна для обучения модели.

В настоящей работе предложена новая стратегия выбора по наименьшей уверенности с учетом плотности (*ADWUS*), которая призвана исправить недостатки простой стратегии выбора по неуверенности. В ней примеры ранжируются по убыванию рейтинга, который вычисляется как гармоническое среднее двух величин – оценки неуверенности и оценки центральности примера. Оценка центральности показывает, как много других примеров находятся близко к заданному. Максимальная центральность соответствует модам совместного распределения признаков, минимальная – выбросам. В качестве оценки центральности использовалось среднее значение косинусной меры сходства данного примера и его ближайших соседей. При этом учитывалось только заданное количество (10-100) наиболее близких соседей. Для поиска похожих использовался метод, реализованный в библиотеке Annoy[4]. Необходимо заметить, что существуют более корректные способы оценки центральности, например, PageRank и другие. Однако процедура их вычисления является значительно более трудоемкой. К тому же на практике даже такая приблизительная оценка центральности дает удовлетворительные результаты. После разметки примера оценка его центральности становится равной 0, а оценки его соседей понижаются пропорционально их близости к размеченному примеру. Таким образом, стратегия помечает хорошо изученные области пространства признаков и обучающей выборки, чтобы не возвращаться к ним снова.

Стратегия выбора по наибольшей предсказываемой вероятности (*MPERR*) ставит целью разметку тех неразмеченных примеров, которые классификатор относит к положительным с наибольшей уверенностью. При данном подходе вся имеющаяся выборка (включающая и размеченные, и неразмеченные примеры) делится на две части: положительные и отрицательные примеры. При этом положительными считаются только те, которые эксперт пометил как положительные. Отрицательными примерами считаются все остальные. Классификатор обучается на составленной таким образом выборке, после чего строятся предсказания для неразмеченных примеров. Эксперту предъявляются примеры, получившие наибольшую вероятность принадлежности к положительному классу.

Стратегия выбора по наибольшей предсказываемой вероятности с учетом плотности (*ADWUS-MPERR*) представляет собой комбинацию двух последних стратегий, при этом вместо оценки неуверенности используется предсказываемая вероятность отнесения к положительному классу.

[4] https://github.com/spotify/annoy

3. Экспериментальные исследования методов извлечения определений терминов из научных текстов

3.1. Корпус для обучения и тестирования

В ходе работы был создан корпус текстов, в котором были размечены определения понятий. В корпус вошли статьи журналов из перечня ВАК, а также труды российских конференций. Под понятием подразумевалось используемое или нововведенное научное знание, выраженное с помощью специальных языковых средств, таких как: «X – это…», «Под X подразумевается что», «Под X понимается что», «X называется чем» и др.

Полный объем корпуса составил примерно 65 тыс. предложений. Было выделено более 1200 фрагментов, соответствующих определениям. Для проведения экспериментов использовался подкорпус, который был ранее применен в работе [27]. Он был дополнительно верифицирован экспертом – удалены ошибочно выделенные фрагменты, добавлены пропущенные, дополнительно рассмотрены сложные случаи. Подкорпус состоит из 36 документов, совокупным объемом более 180 000 токенов. В нем размечено 439 определений. После разделения корпуса на предложения было сформировано 408 положительных примеров (присутствие определения в предложении) и 10 256 отрицательных примеров. Как видно из приведенной выше статистики, распределение классов в сформированной выборке сильно скошено.

3.2. Оценка качества методов извлечения определений без активного обучения

Цель данного эксперимента – оценить наилучшее возможное качество извлечения целевой информации с применением машинного обучения и ограниченной обучающей выборки.

В этом эксперименте весь корпус несколько раз случайным образом разбивался на две непересекающиеся части: обучающую и тестовую выборки, оценки усреднялись по всем разбиениям. При этом разбиение выполнялось не по отдельным примерам, а по документам. Такой подход к разбиению на обучающие и тестовые данные необходим для того, чтобы оценки качества не были завышены из-за переобучения классификатора под лексику конкретных документов. Были использованы случайные разбиения вместо стандартного подхода, реализованного в перекрестной проверке, поскольку количество документов было мало. Они имели различающийся размер и содержали разное количество положительных примеров, из-за чего разбиения при перекрестной проверке могли быть несбалансированными. Проводилось 50 итераций обучения–тестирования на разных разбиениях для каждого классификатора и каждого признакового пространства. Такой подход к оценке качества извлечения информации должен давать более объективные результаты в случае малых выборок.

В Табл. 1 приведены результаты оценки основных метрик качества для первого эксперимента – точность, полнота, F1-мера, площадь под кривой количества ошибок первого и второго рода (ROC AUC).

В проведенных экспериментах наилучшее значение F1-меры составляет 57,6%, что является достаточно высоким показателем, учитывая сильный дисбаланс между положительными и отрицательными примерами в корпусе, использование лишь обобщенных (неспециализированных) наборов признаков, а также малый размер обучающей выборки и малое количество положительных примеров. Эти особенности выборки частично объясняют также высокую дисперсию показателей качества даже для линейных моделей.

Полученные результаты свидетельствуют, что применение глубокого лингвистического анализа в настоящей задаче оправдано. Качество классификации по основной метрике F1 на втором наборе признаков имеет значительный прирост по сравнению с простыми n-граммами лемм и частей речи. Большую значимость во втором наборе показывают семантические признаки, в которых учитываются ролевые структуры предложения. Это происходит благодаря тому, что при построении ролевых структур учитывается большое количество лингвистической информации. В результате семантического анализа эта информация получает удобное представление в виде ролевых структур, которые могут быть использованы как обобщенные признаки в задачах извлечения информации из текстов. Контекстные правила также являются значимыми признаками при построении многих классификаторов, однако их вклад оказался значительно меньше, чем вклад семантических признаков.

Табл.1. Сравнение качества извлечения информации с помощью различных методов машинного обучения и методов построения признакового пространства без использования активного обучения

Признаки	Классификатор	Точность, %	Полнота, %	F_1, %	ROC AUC, %
n-граммы	Лог. рег.	62,8 ± 11,5	36,0 ± 7,8	45,3 ± 8,3	88,9 ± 2,6
	Линейный SVM	66,8 ± 11,5	35,2 ± 8,3	45,6 ± 9,0	-
	Random Forest	38,6 ± 8,6	30,4 ± 8,3	33,0 ± 6,6	88,2 ± 2,3
	LightGBM	56,4 ± 9,8	**53,9 ± 8,9**	54,8 ± 8,5	92,9 ± 1,9
	3-layer MLP	46,9 ± 10,3	49,0 ± 8,0	47,3 ± 7,7	85,9 ± 3,5
	Ансамбль	66,2 ± 10,7	50,0 ± 8,3	56,8 ± 8,7	93,6 ± 1,8
n-граммы + сем. и синт. призн. + конт. правила	Лог. рег.	64,6 ± 8,6	38,7 ± 8,3	48,0 ± 8,1	89,6 ± 3,1
	Линейный SVM	67,9 ± 9,5	37,8 ± 8,1	48,2 ± 8,5	-
	Random Forest	50,9 ± 11,1	33,2 ± 9,3	39,3 ± 8,4	90,4 ± 2,4
	LightGBM	59,3 ± 9,2	52,6 ± 9,1	55,4 ± 8,2	93,5 ± 1,7
	3-layer MLP	51,4 ± 9,9	51,7 ± 8,5	50,9 ± 7,4	87,2 ± 3,5
	Ансамбль	**68,7 ± 8,7**	49,9 ± 6,6	**57,6 ± 6,4**	**93,8 ± 2,1**

Стоит также отметить, что во всех экспериментах высокое качество предсказуемо показывают методы на основе ансамблей классификаторов (градиентный бустинг, композиция) и нейронные сети. Композиция из нейронных сетей и моделей на основе градиентного бустинга существенно опережает единичные классификаторы на втором наборе признаков, однако требует больших ресурсов для обучения, поэтому ее применение в активном обучении затруднено.

3.3. Оценка качества методов извлечения определений, основанных на активном обучении

Цель данного эксперимента – оценить и сравнить различные стратегии выбора примеров при активном обучении как с точки зрения построения наиболее качественного классификатора, так и с точки зрения экономии трудозатрат, на создание размеченного корпуса удостоверенного качества.

В эксперименте использовалась логистическая регрессия и пространство признаков на основе n-грамм, синтаксиса, семантики и контекстных правил. Эксперимент с активным обучением проводился в режиме симуляции без участия человека. На каждой итерации автоматически размечалось 10 наиболее перспективных примеров, и вычислялись точность, полнота и F1-мера на обучающей и на тестовой выборках, а также процент размеченных положительных примеров. В качестве тестовой выборки использовалось 30% всех имеющихся примеров. Результирующие показатели были вычислены путем усреднения по десяти разбиениям (для оценки различных стратегий выбора использовались одни и те же разбиения). Как и в экспериментах без активного обучения разбиения проводились по документам.

На Рис. 1 приведены графики роста F1-меры в зависимости от номера итерации на обучающей выборке, а на Рис. 3 – на тестовой.

Из Рис. 1 можно сделать вывод о том, что данные в выбранном пространстве признаков достаточно хорошо разделяются линейным классификатором. Все стратегии, кроме случайной, одинаково хорошо справляются с исследованием обучающей выборки.

Графики на Рис. 2 показывают, какая часть имеющихся положительных примеров оказывается размеченной за количество итераций, отложенное по горизонтальной оси. Вполне предсказуемо, что стратегия выбора по наибольшей ожидаемой вероятности (простая и с учетом плотности) исследует область положительных примеров быстрее остальных стратегий. Меньше всего положительных примеров размечается при использовании случайной стратегии.

Из Рис. 3 можно сделать вывод о том, что максимальное качество достигается при использовании всех стратегий примерно одинаково (кроме случайной стратегии). Стабилизация происходит чуть раньше, чем на обучающей выборке (около 80-90 итераций).

Учет плотности распределения примеров в целом незначительно ускоряет достижение максимальных сбалансированных оценок каче-

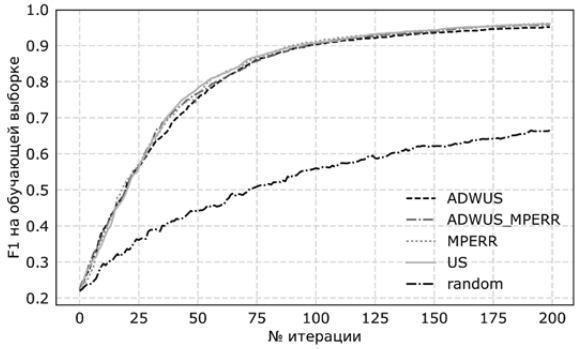

Рис. 1. Графики роста F_1-меры на обучающей выборке в зависимости от номера итерации

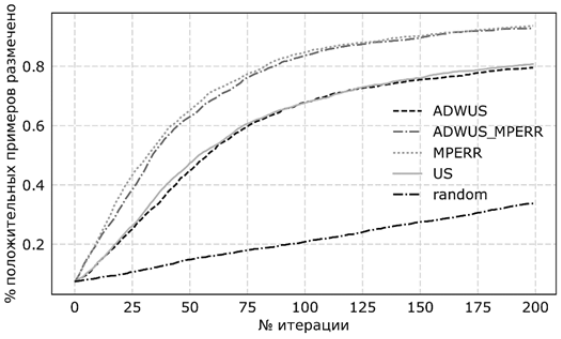

Рис. 2. Графики роста доли размеченных положительных примеров в зависимости от номера итерации

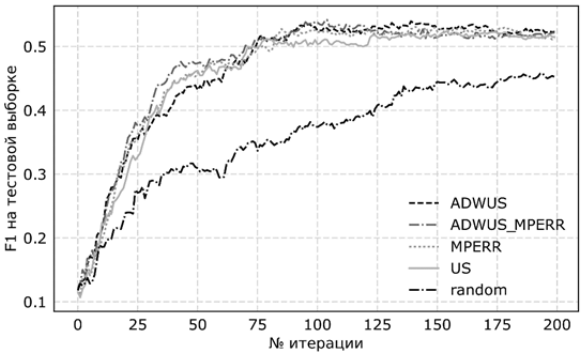

Рис. 3. Графики роста F_1-меры на тестовой выборке в зависимости от номера итерации

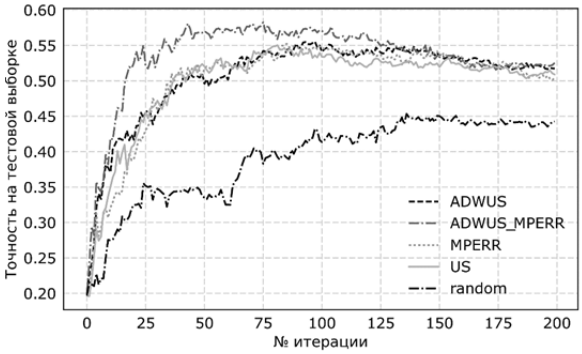

Рис. 4. Графики роста точности на тестовой выборке в зависимости от итерации

ства (F1-меры) на обучающей выборке. Однако профиль роста точности на тестовой выборке (Рис. 4) и скорость исследования области положительных примеров (Рис. 2) у всех стратегий разные. Это может говорить о том, что в разных конечных задачах целесообразно применять разные стратегии активного обучения (Рис. 5). Например, стратегия ADWUS-MPERR наиболее эффективно исследует область положительных примеров тестовой выборки и может лучше подходить для разметки корпусов со скошенным распределением классов.

Суммарную трудоемкость построения классификатора с применением активного обучения можно оценить как процент размеченных примеров от всего количества объектов в выборке, при котором достигается примерно такое же качество, как и при обучении на полной выборке. В нашем эксперименте максимальное качество достигается на 100-й итерации ($F1 \approx 51\%$), что соответствует разметке 1000 примеров, которые составляют в среднем 14% от общего количества примеров в обучающей выборке. Таким образом, для достижения максимального качества классификации требуется разметка корпуса в 7 раз

Рис. 5. Доля выборки, разметка которой необходима для достижения $F_1 = 51\%$, при применении разных стратегий активного обучения

меньшего по сравнению с исходным. При этом учет плотности существенно улучшает работу базовых стратегий (US и MPERR).

4. Экспериментальные исследования методов активного обучения для извлечения описаний научных результатов из текстов статей

Для исследования методов извлечения описаний научных результатов из текстов статей вручную был размечен корпус из 207 документов. Разметка результатов проводилась путем

Табл.2. Результаты оценки качества извлечения результатов до применения активного обучения к корпусу и после

Корпус	Классификатор	Точность, %	Полнота, %	F_1, %	ROC AUC, %
До активного обучения	Лог. рег.	35,8 ± 8,2	10,9 ± 3,0	16,6 ± 4,2	75,1 ± 3,3
	LightGBM	29,4 ± 6,2	18,4 ± 4,3	22,3 ± 4,4	79,8 ± 2,2
После активного обучения	Лог. рег.	**77,9** ± 6,0	39,4 ± 4,8	52,0 ± 3,9	88,5 ± 2,3
	LightGBM	62,4 ± 5,1	**50,1** ± 5,8	**55,2** ± 3,4	**89,2** ± 1,7

выделения предложений, в которых говорится о получении определенного результата: о разработке модели или метода, об их программной реализации, о проведении экспериментов и т.п. Результаты работ других авторов или работ, предшествующих текущему исследованию, описанные в разделах «Обзор» или «Обсуждение результатов», не выделялись. Суммарный объем корпуса составил около 2 млн словоупотреблений, из них к результатам было отнесено около 16 тыс. словоупотреблений (<1%). После преобразования корпуса в обучающую/тестовую выборку путем деления на предложения было получено 721 положительных и 56 369 отрицательных примеров.

Во время проведения предварительных экспериментов было установлено, что размеченный корпус содержит большое количество шума (пропуски и ошибочные положительные примеры). Для корректировки корпуса было предложено применить активное обучение. При этом основной целью являлось наиболее быстрое нахождение всех положительных примеров, а не скорейший рост качества классификации. Такая постановка задачи извлечения информации актуальна в случаях, когда необходимо извлечь все факты из фиксированного корпуса документов (например, при составлении аналитического обзора), а обобщающая способность классификатора менее важна.

Методика оценки целесообразности применения активного обучения для улучшения корпусов текстов следующая. Сначала на имеющемся корпусе оценивается качество извлечения информации без активного обучения. Затем проводится переразметка корпуса с активным обучением. При этом изначально размеченными примерами считаются только размеченные положительные примеры. Исходя из предыдущих результатов, в настоящем эксперименте было сочтено целесообразным применить стратегию на основе выбора по наибольшей вероятности с учетом плотности (ADWUS-MPERR). После переразметки корпуса с помощью активного обучения количество положительных примеров возросло до 1110 (более чем на 50%).

Для оценки эффективности применения активного обучения в соответствии с методикой была проведена оценка качества классификации на исходном корпусе и на переразмеченном. Для генерации признаков был использован метод, учитывающий различные n-граммы, синтаксические словосочетания, ролевые структуры, а также контекстные правила. В Табл. 2 приведены оценки качества для двух методов машинного обучения: логистической регрессии и градиентного бустинга (LightGBM).

Полученные результаты показывают, что после переразметки корпуса с помощью активного обучения качество классификации значительно возросло. Это произошло из-за того, что, с одной стороны, увеличилось количество обучающих примеров, а с другой стороны уменьшилось количество ошибок в тестовой выборке – тех примеров, положительный класс которых при тестировании был корректно определен классификатором, но которые ошибочно были отнесены экспертом при разметке к отрицательным примерам.

Заключение

В статье представлены подходы к извлечению информации из текстов, основанные на активном обучении, а также способы формирования признаковых описаний фрагментов текстов. Экспериментально показано, что использование полной лингвистической информации (морфологии, синтаксиса и семантики) дает в среднем более высокие показатели качества, чем n-граммы лемм и

частей речи. Анализ информативности признаков показал, что признаки, построенные на основе результатов глубокого лингвистического анализа, а именно на основе ролевых структур, а также контекстные правила, являются важными для классификации.

Предложены две новые стратегии активного обучения, осуществляющие выбор примеров для разметки с учетом их плотности в признаковом пространстве. Проведено экспериментальное исследование нескольких различных стратегий, в результате которого было показано:

• все стратегии, кроме случайной, требуют разметки примерно одинакового количества примеров для достижения максимально возможного качества;

• стратегии, использующие выбор по неуверенности, быстрее повышают качество модели в начале и середине разметки;

• стратегия выбора по наибольшей ожидаемой вероятности находит положительные примеры быстрее остальных стратегий, и поэтому она подходит для тех случаев, когда распределение классов в выборке скошено, а важность положительных примеров гораздо выше отрицательных;

• предложенная стратегия выбора по наибольшей ожидаемой вероятности с учетом плотности более эффективно исследует область положительных примеров.

На задачах извлечения определений и результатов из научных публикаций продемонстрировано преимущество использования активного обучения при разметке обучающих корпусов. Показано, что применение активного обучения может сократить до семи раз объем разметки обучающего корпуса, необходимого для построения модели, имеющей схожее качество, по сравнению с обучением на полной выборке. Показано, что переразметка корпуса с использованием активного обучения и предложенных в работе стратегий значительно повысила полноту корпуса и качество построенных на нем моделей машинного обучения.

Основные направления дальнейших исследований включают разработку интегрированных программных средств для автоматизации процесса построения специализированных систем извлечения информации, применение предложенных методов к другим предметным областям, а также исследование более сложных алгоритмов генерации признаковых пространств по графовому представлению текстов.

Литература

1. Chiticariu L., Li Y., Reiss F. Transparent machine learning for information extraction // The materials of the tutorial of the Conference on Empirical Methods in Natural Language Processing. — 2015.
2. Piskorski J., Yangarber R. Information extraction: past, present and future // Multi-source, multilingual information extraction and summarization. — 2013. — P. 23–49.
3. Chiticariu L., Li Y., Reiss F. R. Rule-based information extraction is dead! Long live rule-based information extraction systems! // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 827–832.
4. Gupta S., Manning C. D. SPIED: Stanford pattern-based information extraction and diagnostics // Association for Computational Linguistics (ACL) Workshop on Interactive Language Learning, Visualization, and Interfaces. — 2014.
5. Espinosa K. J., Batista-Navarro R., Ananiadou S. Learning to recognise named entities in tweets by exploiting weakly labelled data // Proceedings of the 2nd Workshop on Noisy User-generated Text (W-NUT). — 2016.
6. Learning from human-generated lists / Kwang-Sung Jun, Jerry Zhu, Burr Settles, Timothy Rogers // International Conference on Machine Learning. — 2013. — P. 181–189.
7. IKE - an interactive tool for knowledge extraction / Bhavana Dalvi, Sumithra Bhakthavatsalam, Chris Clark et al. // Proceedings of the 5th Workshop on Automated Knowledge Base Construction. — 2016. — P. 12–17.
8. The benefits of word embeddings features for active learning in clinical information extraction / Mahnoosh Kholghi, Lance De Vine, Laurianne Sitbon et al. // Proceedings of the Australasian Language Technology Association Workshop. — 2016. — P. 25–34.
9. External knowledge and query strategies in active learning: a study in clinical information extraction / Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, Anthony Nguyen // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. — 2015. — P. 143–152.
10. Augenstein I., Maynard D., Ciravegna F. Relation Extraction from the Web Using Distant Supervision // Knowledge Engineering and Knowledge Management: 19th International Conference (EKAW 2014). — 2014. — P. 26–41.
11. OCR++: A robust framework for information extraction from scholarly articles / Mayank Singh, Barnopriyo Barua, Priyank Palod et al. // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — 2016. — P. 3390–3400.
12. Liu H. Automatic argumentative-zoning using word2vec // arXiv preprint arXiv:1703.10152. — 2017.
13. Figureseer: Parsing result-figures in research papers / Noah Siegel, Zachary Horvitz, Roie Levin et al. // European Conference on Computer Vision. — 2016. — P. 664–680.
14. Clark C. A., Divvala S. K. Looking beyond text: Extracting figures, tables and captions from computer science papers // AAAI Workshop: Scholarly Big Data. — 2015.

15. Lever J., Jones S. J. M. VERSE: Event and relation extraction in the BioNLP 2016 shared task // Proceedings of the 4th BioNLP Shared Task Workshop. — 2016.
16. Which techniques does your application use?: An information extraction framework for scientific articles / Soham Dan, Sanyam Agarwal, Mayank Singh et al. // arXiv preprint arXiv:1608.06386. — 2016.
17. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications / Isabelle Augenstein, Mrinal Kanti Das, Sebastian Riedel et al. // Proceedings of the 11th International Workshop on Semantic Evaluation. — 2017.
18. Kovar V., Mociarikova M., Rychly P. Finding definitions in large corpora with sketch engine // Proceedings of the Tenth International Conference on Language Resources and Evaluation. — 2016.
19. DEFEXT: A semi supervised definition extraction tool / Luis Espinosa-Anke, Roberto Carlini, Horacio Saggion, Francesco Ronzano // GLOBALEX 2016: Lexicographic Resources for Human Language Technology Workshop. — 2016.
20. Del Gaudio R. Automatic extraction of definitions : Ph.D. thesis / R. Del Gaudio. — 2014. — University of Lisbon.
21. Bolshakova E. I., Efremova N. E. A heuristic strategy for extracting terms from scientific texts // International Conference on Analysis of Images, Social Networks and Texts. — 2015. — P. 297–307.
22. Смирнов И. В., Шелманов А. О., Кузнецова Е. С. , Храмоин И. В. Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов // Искусственный интеллект и принятие решений. — № 1. — С. 11–24.
23. Сокирко А. В. Морфологические модули на сайте www.aot.ru // Труды международной конференции «Диалог 2004». — 2004. — P. 559–565.
24. MaltParser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95–135.
25. Апресян Ю. Д. , Богуславский И. М. , Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка.— 2005. — С. 193–214.
26. Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2014". — No. 13. — 2014. — P. 607–620.
27. Шелманов А.О. , Каменская М.А. , Ананьева М.И. , Смирнов И.В. Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. — 2016. — № 4. — С. 47–61.

Суворов Роман Евгеньевич. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил Рыбинский государственный авиационный технический университет им. П.А. Соловьева в 2012 году. Количество печатных работ: 23. Область научных интересов: машинное обучение, интеллектуальный анализ данных, анализ текстов на естественном языке, извлечение информации, распределенные вычисления. E-mail: rsuvorov@isa.ru

Шелманов Артем Олегович. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил Национальный исследовательский ядерный университет «МИФИ» в 2011 году. Кандидат технических наук. Количество печатных работ: 28. Область научных интересов: искусственный интеллект, компьютерная лингвистика, машинное обучение, информационно-аналитические системы. E-mail: shelmanov@isa.ru

Каменская Маргарита Александровна. Инженер-исследователь ИСА ФИЦ ИУ РАН. Окончила Российский университет дружбы народов в 2014 году. Количество печатных работ: 7. Область научных интересов: компьютерная лингвистика, обработка естественного языка, автоматический анализ текстов, разрешение референции. E-mail: mak@isa.ru

Смирнов Иван Валентинович. Заведующий лабораторией «Компьютерная лингвистика и интеллектуальный анализ информации», доцент ИСА ФИЦ ИУ РАН. Окончил Российский университет дружбы народов в 2003 году. Кандидат физико-математических наук. Количество печатных работ: 66. Область научных интересов: обработка естественного языка, интеллектуальный анализ информации. E-mail: ivs@isa.ru

Information Extraction from Scientific Texts Using Active Machine Learning

R.E. Suvorov, A.O. Shelmanov, M.A. Kamenskaya, I.V. Smirnov

The paper addresses the task of information extraction from natural language texts using machine learning methods. For creating an information extraction system based on machine learning, usually, large annotated text corpora are required. Another problem that arises during the development of such systems is feature engineering. To solve the first problem, we propose methods of information extraction based on active machine learning techniques. To solve the second problem, we investigate methods for generating the feature space based on the results of the deep linguistic analysis. Experimental studies of the proposed methods showed that active learning significantly reduces the amount of labor required for creating an information extraction system, while maintaining the quality of the trained models. Using the

results of deep linguistic analysis for generating feature space improves the quality of models for information extraction.
Keywords: information extraction, deep linguistic analysis, active machine learning, multipurpose feature engineering, scientific texts processing.

References

1. Chiticariu L., Li Y., Reiss F. Transparent machine learning for information extraction // The materials of the tutorial of the Conference on Empirical Methods in Natural Language Processing. — 2015.
2. Piskorski J., Yangarber R. Information extraction: past, present and future // Multi-source, multilingual information extraction and summarization. — 2013. — P. 23–49.
3. Chiticariu L., Li Y., Reiss F. R. Rule-based information extraction is dead! Long live rule-based information extraction systems! // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — P. 827–832.
4. Gupta S., Manning C. D. SPIED: Stanford pattern-based information extraction and diagnostics // Association for Computational Linguistics (ACL) Workshop on Interactive Language Learning, Visualization, and Interfaces. — 2014.
5. Espinosa K. J., Batista-Navarro R., Ananiadou S. Learning to recognise named entities in tweets by exploiting weakly labelled data // Proceedings of the 2nd Workshop on Noisy User-generated Text (W-NUT). — 2016.
6. Learning from human-generated lists / Kwang-Sung Jun, Jerry Zhu, Burr Settles, Timothy Rogers // International Conference on Machine Learning. — 2013. — P. 181–189.
7. IKE - an interactive tool for knowledge extraction / Bhavana Dalvi, Sumithra Bhakthavatsalam, Chris Clark et al. // Proceedings of the 5th Workshop on Automated Knowledge Base Construction. — 2016. — P. 12–17.
8. The benefits of word embeddings features for active learning in clinical information extraction / Mahnoosh Kholghi, Lance De Vine, Laurianne Sitbon et al. // Proceedings of the Australasian Language Technology Association Workshop. — 2016. — P. 25–34.
9. External knowledge and query strategies in active learning: a study in clinical information extraction / Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, Anthony Nguyen // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. — 2015. — P. 143–152.
10. Augenstein I., Maynard D., Ciravegna F. Relation Extraction from the Web Using Distant Supervision // Knowledge Engineering and Knowledge Management: 19th International Conference (EKAW 2014). — 2014. — P. 26–41.
11. OCR++: A robust framework for information extraction from scholarly articles / Mayank Singh, Barnopriyo Barua, Priyank Palod et al. // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — 2016. — P. 3390–3400.
12. Liu H. Automatic argumentative-zoning using word2vec // arXiv preprint arXiv:1703.10152. — 2017.
13. Figureseer: Parsing result-figures in research papers / Noah Siegel, Zachary Horvitz, Roie Levin et al. // European Conference on Computer Vision. — 2016. — P. 664–680.
14. Clark C. A., Divvala S. K. Looking beyond text: Extracting figures, tables and captions from computer science papers // AAAI Workshop: Scholarly Big Data. — 2015.
15. Lever J., Jones S. J. M. VERSE: Event and relation extraction in the BioNLP 2016 shared task // Proceedings of the 4th BioNLP Shared Task Workshop. — 2016.
16. Which techniques does your application use?: An information extraction framework for scientific articles / Soham Dan, Sanyam Agarwal, Mayank Singh et al. // arXiv preprint arXiv:1608.06386. — 2016.
17. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications / Isabelle Augenstein, Mrinal Kanti Das, Sebastian Riedel et al. // Proceedings of the 11th International Workshop on Semantic Evaluation. — 2017.
18. Kovar V., Mociarikova M., Rychly P. Finding definitions in large corpora with sketch engine // Proceedings of the Tenth International Conference on Language Resources and Evaluation. — 2016.
19. DEFEXT: A semi supervised definition extraction tool / Luis Espinosa-Anke, Roberto Carlini, Horacio Saggion, Francesco Ronzano // GLOBALEX 2016: Lexicographic Resources for Human Language Technology Workshop. — 2016.
20. Del Gaudio R. Automatic extraction of definitions : Ph.D. thesis / R. Del Gaudio. — 2014. — University of Lisbon.
21. Bolshakova E. I., Efremova N. E. A heuristic strategy for extracting terms from scientific texts // International Conference on Analysis of Images, Social Networks and Texts. — 2015. — P. 297–307.
22. Smirnov, I.V., A.O. Shelmanov, E.S. Kuznetsova, I.V. Khramoin. 2014. Semantiko-sintaksicheskiy analiz estestvennykh yazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov [Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts]. Iskusstvennyy intellekt i prinyatie resheniy [Artificial intelligence and decision making]. 1: 11–24.
23. Sokirko, A.V. 2004. Morfologicheskie moduli na sayte www.aot.ru [Morphological modules on the site www.avt.ru]. Trudy mezhdunarodnoy konferentsii "Dialog 2004" [Proceedings of the International Conference "Dialogue-2004"]. 559–565.
24. MaltParser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95–135.
25. Apresyan, Yu. D., I.M. Boguslavskiy, B.L. Iomdin, L.L. Iomdin, A.V. Sannikov, V.Z. Sannikov ... and L.L. Tsinman. 2005. Sintaksicheski i semanticheski annotirovannyy korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy [Syntactically and semantically annotated corpus of the Russian language: current state and prospects]. Natsional'nyy korpus russkogo yazyka, 2003-2005 [National corpus of Russian Language, 2003-2005]. 193–214.
26. Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual

Technologies. Papers from the Annual International Conference "Dialogue 2014". — No. 13. — 2014. — P. 607–620.
27. Shelmanov, A.O., M.A. Kamenskaya, M.I. Anan'eva, I.V. Smirnov. 2016. Semantiko-sintaksicheskiyanaliz tekstov v zadachakh voprosno-otvetnogo poiska i izvlecheniya opredeleniy [Semantic-syntactic analysis for question-answering and definition extraction]. Iskusstvennyy intellekt i prinyatie resheniy [Artificial intelligence and decision-making].4: 47-61.

Suvorov Roman Evgen'evich. Fellow of ISA FRC CSC RAS, Moscow, pr. 60-letiya Oktyabrya, 9. Graduated from Rybinsk State Aviation Technical University in 2012. The number of publications: 22. Research interests: machine learning, data mining, natural language texts analysis, information extraction, distributed computing. E-mail: rsuvorov@isa.ru

Shelmanov Artem Olegovich. Fellow of ISA FRC CSC RAS. PhD. Graduated from National Research Nuclear University «MEPhI» in 2011. The number of publications: 20. Research interests: artificial intelligence, computer linguistics, machine learning, search and analytical systems. E-mail: shelmanov@isa.ru

Kamenskaya Margarita Aleksandrovna. Research engineer of ISA FRC CSC RAS, Moscow, pr. 60-letiya Oktyabrya, 9. Graduated from RUDN University in 2014. The number of publications: 7. Research interests: computer linguistics, natural language processing, automatic text analysis, reference resolution. E-mail: mak@isa.ru

Smirnov Ivan Valentinovich. Head of laboratory "Computational linguistics and intelligent data analysis" of ISA FRC CSC RAS, Moscow, pr. 60-letiya Oktyabrya, 9. Graduated from RUDN University in 2003. PhD, associate professor. The number of publications: 52. Research interests: natural language processing, intellectual analysis of information. E-mail: ivs@isa.ru