А.О. Шелманов, М.А. Каменская, М.И. Ананьева, И.В. Смирнов

Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений[1]

Аннотация. В работе исследуется вклад семантического и семантико-синтаксического видов анализа в эффективность решения прикладных задач обработки текстов: вопросно-ответного поиска и извлечения определений из научных публикаций. Представлены методы решения этих задач, использующие помимо морфологических и синтаксических структур также семантические структуры текстов. Проведено экспериментальное исследование методов, а также сравнение на рассматриваемых задачах двух подходов к выполнению синтаксического и семантического анализа: раздельного последовательного синтаксического и семантического анализа и совмещенного семантико-синтаксического анализа.

Ключевые слова: семантический анализ, семантико-синтаксический анализ, вопросно-ответный поиск, извлечение информации из текстов, извлечение определений.

Введение

Решение многих задач обработки текстов, например, таких как вопросно-ответный поиск и извлечение информации, естественным образом может быть основано на анализе эксплицитного представления синтаксической и семантической структур текста. Однако многие системы используют методы, которые непосредственно не оперируют этими структурами и опираются в основном на результаты поверхностного лингвистического и статистического анализа. Широкому применению методов синтаксического и семантического анализа мешают связанные с ними ограничения. В частности, по сравнению, например, с методами морфологического анализа или стемминга, они требуют гораздо больших вычислительных ресурсов, их качество долгое время считалось недостаточным для применения во многих прикладных задачах, кроме того, создание глубоких лингвистических анализаторов и их настройка на предметную область весьма трудоемки. Однако успехи современных исследований в области синтаксического и семантического анализа смогли в значительной степени нивелировать эти ограничения, что позволило использовать эти виды анализа при решении прикладных задач в ряде информационно-поисковых и информационно-аналитических систем [1, 2].

В этой работе представлены методы решения двух прикладных задач: вопросно-ответного поиска и извлечения определений из текстов научных публикаций, которые используют результаты семантического и семантико-синтаксического анализа.

В качестве основной модели семантики используется реляционно-ситуационная модель [3], которая опирается на теорию коммуникативной грамматики Г.А. Золотовой [4]. Используемые в настоящей работе методы семантического и семантико-синтаксического анализа строят ролевые структуры и семантические сети предложений. В ролевой структуре семантика предложения представляется в виде совокупности предикатных слов, их аргументов и семантических ролей [5, 6]. Под предикатным словом подразумевается лексема или синтакси-

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

ческая конструкция, которая обозначает в тексте ситуацию и обладает набором семантических ролей. Такими словами являются глаголы, отглагольные существительные, причастия и некоторые другие части речи. Семантические аргументы – это синтаксические конструкции, которые обозначают в тексте участников ситуации, заданной предикатным словом. Семантическая роль – часть семантики предиката, отражающая общие свойства его аргумента. Это значение синтаксической конструкции (аргумента) при предикатном слове. Помимо ролей в реляционно-ситуационной модели строится семантическая сеть. Она описывает выявленные в тексте отношения между концептами в некоторой предметной области. Применяемые в настоящей работе методы семантического анализа устанавливают семантические отношения только между синтаксемами, являющимися семантическими аргументами предикатных слов.

Данная статья является продолжением работ [7-9], в которых представлено подробное описание методов семантического и семантико-синтаксического анализа. Ее основное назначение заключается в экспериментальном исследовании методов решения прикладных задач, использующих семантические структуры, а также в сравнении на этих задачах двух подходов к выполнению синтаксического и семантического видов анализа: раздельного последовательного синтаксического и семантического анализа, совмещенного семантико-синтаксического анализа.

В разделе 1 проведен краткий обзор подходов к решению задач вопросно-ответного поиска и извлечения определений. В разделе 2 представлены методы, использующие результаты семантического и семантико-синтаксического анализа. В разделе 3 описаны эксперименты, в которых исследуются вышеупомянутые методы.

1. Обзор методов вопросно-ответного поиска и извлечения определений из научных текстов

1.1. Методы вопросно-ответного поиска

Задача вопросно-ответного поиска имеет множество постановок. Ответ на вопрос может искаться в структурированной базе данных [10, 11], извлекаться из текста документа коллекции [12] или генерироваться с применением логического вывода [13]. В настоящей работе постановка задачи вопросно-ответного поиска наиболее близка ко второму варианту: ответ должен извлекаться из неструктурированных текстовых документов.

Большое количество систем, решающих рассматриваемую задачу, опирается только на лексические шаблоны и синтаксическую структуру текста. Реализованные в них методы сравнивают синтаксические поддеревья запроса пользователя и тексты документов из поискового массива (например, [14, 15]). Обобщение этого подхода представлено в работе [16], где отдельные деревья предложений объединяются по различным семантическим и кореферентным связям в «чащи разбора» для целых параграфов, что дает преимущество при их сопоставлении с запросом. Отметим также систему, представленную в работе [17]. Она принимала участие в соревнованиях по вопросно-ответному поиску РОМИП-2006 [18]. Для ранжирования ответов эта система помимо синтаксических шаблонов использует семантические правила и учитывает семантические классы объектов в тексте.

Существует также ряд систем, в которых для вопросно-ответного поиска используются ролевые структуры предложений. Отметим некоторые из них. В работе [19] авторы продемонстрировали повышение качества метода вопросно-ответного поиска с применением ролевой структуры предложений по сравнению с подходом, основанным лишь на сопоставлении синтаксических поддеревьев. В работе [20] рассматриваются вопросы индексации ролевых структур для осуществления на этой основе вопросно-ответного поиска. В [21] сравнивается система вопросно-ответного поиска, использующая методы построения ролевых структур высказываний, и система, основанная на извлечении именованных сущностей. Было показано, что использование семантических ролей позволяет значительно улучшить качество ответов на категориальные вопросы общего характера. В отличие от методов, применяемых в вышеупомянутых работах, метод вопросно-ответного поиска, представленный в настоящей работе, помимо синтаксических деревьев и семантических ролей использует также семантические отношения между концептами, которые позволяют некоторой степени абстрагироваться от ролевых структур и проводить более общее сопоставление вопросов и текстов из поискового массива.

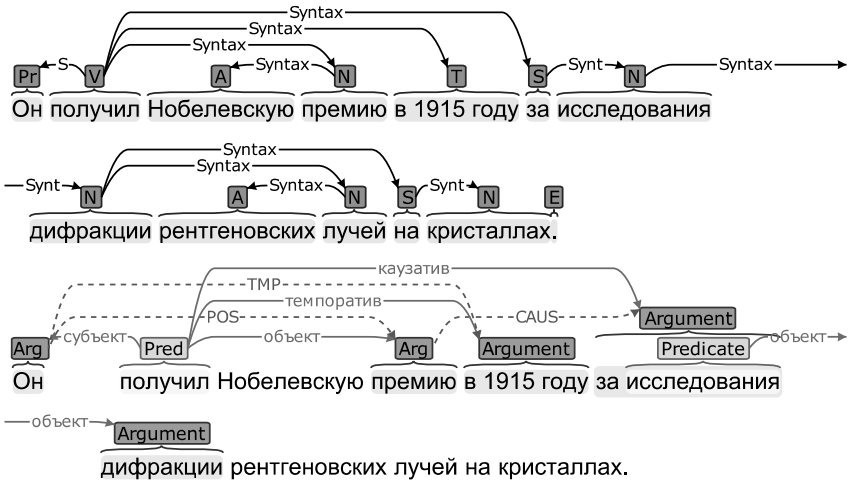


Рис. 1. Пример синтаксической и семантической структуры предложения

1.2. Методы извлечения определений из научных текстов

Наиболее близким подходом к методу извлечения определений, представленному в настоящей статье, является подход, предложенный в работах [22-25]. В них описывается система, применяющая лексико-синтаксические шаблоны для извлечения определений и терминов различного типа из текстов научных публикаций. В настоящей работе было предложено помимо лексической, морфологической и синтаксической информации при построении фреймов использовать также семантические роли. Использование информации о семантических ролях позволяет, с одной стороны, упростить создание фреймов, а с другой - одним фреймом покрыть больше конструкций, задающих определения терминов.

2. Методы решения прикладных задач обработки текстов на основе семантического и семантико-синтаксического анализа

Представленные в настоящей работе методы решения прикладных задач на вход получают морфологические, синтаксические и семантические структуры предложений, построенные либо с помощью раздельного выполнения синтаксического и семантического анализа, либо с помощью совмещенного метода семантико-синтаксического анализа. Пример подобных структур представлен на Рис. 1. Вверху находится синтаксическое дерево предложения, внизу – семантическая структура. Семантические аргументы помечены токенами «Arg» и «Argument», предикатные слова - токенами «Pred» и «Predicate». В семантической структуре на сплошных стрелках указаны семантические роли, пунктирные стрелки обозначают семантические отношения в семантической сети.

2.1. Метод ранжирования ответов для вопросно-ответного поиска на основе результатов семантического и семантико-синтаксического анализа

У многих поисковых систем в Интернете модуль вопросно-ответного поиска либо отсутствует, либо обладает ограниченными возможностями (например, отвечает только на вопросы «Где?», «Когда?»), поэтому ответы на вопросы пользователей, заданные на естественном языке, часто оказываются в глубине поисковой выдачи. Этот недостаток решают некоторые метапоисковые системы, главной особенностью которых является то, что они могут не иметь собственного поискового индекса и формируют свою выдачу на основе результатов других поисковиков. Метапоисковые системы делают запросы сразу к нескольким поисковым машинам, получают от каждой из них сниппеты (фрагменты текста, которые выводятся рядом со ссылкой в поисковой выдаче) из топа ответов некоторой глубины, агрегируют все сниппеты в единую выдачу и с помощью методов интеллектуальной обработки

ранжируют ее так, чтобы наиболее релевантная информация находилась в ее начале. В системе Exactus (http://www.exactus.ru) имеется модуль метапоиска, который позволяет находить непосредственные ответы на категориальные вопросы общего характера (т.е. вопросы, в которых вначале присутствуют вопросительные слова: «кто», «что», «где», «когда», «куда» и др.), заданные на естественном языке. В этом разделе описывается метод ранжирования сниппетов для вопросно-ответного поиска, реализованный в этой системе. В методе помимо морфологической и синтаксической структуры текста используются также семантические роли и отношения.

Общий алгоритм ранжирования

Пусть задан вопросительный запрос и имеется список сниппетов, полученных от разных поисковых систем. Для простоты описания алгоритма положим, что запрос пользователя состоит из одного предложения. Алгоритм легко может быть обобщен на случай, когда запрос состоит из нескольких предложений. Чтобы отсортировать список, для каждого сниппета f (fragment) необходимо вычислить оценку его релевантности запросу r^f. Рассмотрим общий алгоритм вычисления этой оценки.

Сниппет разбивается на предложения, для каждого из которых вычисляется две оценки релевантности предложению запроса: лексико-семантическая оценка (ЛС-оценка предложения r_{ls}^s), учитывающая как лексику, так и семантическую структуру предложений; лексическая оценка (Л-оценка предложения, r_l^s), учитывающая только совпадение лексики предложений сниппета и запроса. Отметим, что при сопоставлении лексики стоп-слова и вопросительные слова запроса, такие как «что», «где» и др., не учитываются.

Затем определяется оценка соответствия лексики сниппета запросу по всем предложениям – Л-оценка сниппета r_l^f. Она вычисляется как отношение мощности пересечения множества лемм слов сниппета L_f и множества лемм слов запроса L_q к мощности множества лемм слов запроса:

$$r_l^f = \frac{|L_f \cap L_q|}{|L_q|}. \quad (1)$$

С помощью линейной свертки максимальной ЛС-оценки предложения сниппета и Л-оценки сниппета вычисляется ЛС-оценка сниппета:

$$r_{ls}^f = \alpha_{ls} * \max_i\left(r_{ls}^{s_i}\right) + \alpha_l * r_l^f. \quad (2)$$

Добавление r_l^f в формулу релевантности призвано повысить релевантность тех сниппетов, в которых присутствует больше слов из запроса вне зависимости от того, как они распределены по предложениям, при прочих равных оценках. На практике коэффициент при лексико-семантической оценке должен быть значительно больше, чем коэффициент при лексической оценке $\alpha_{ls} > \alpha_l$, Например, в экспериментах использовались эмпирически подобранные коэффициенты $\alpha_{ls} = 0{,}9;\ \alpha_l = 0{,}1$.

Для каждого сниппета вычисляется также полная Л-оценка r_{fl}^f, которая представляет собой сумму всех Л-оценок предложений сниппета:

$$r_{fl}^f = \sum_i r_l^{s_i}. \quad (3)$$

После того как вычислены полные Л-оценки и ЛС-оценки всех сниппетов, рассчитывается финальная оценка релевантности каждого сниппета r^f. Она представляет собой линейную свертку ЛС-оценки сниппета и полной Л-оценки сниппета, нормированной на максимальную полную оценку среди всех сниппетов:

$$r^f = \beta_{ls} * r_{ls}^f + \beta_l * \frac{r_{fl}^f}{\max_j r_{fl}^{f_j}}. \quad (4)$$

На практике нормированная полная Л-оценка сниппета должна мало влиять на релевантность сниппета вопросительному запросу. Она добавляется в формулу релевантности для того, чтобы при прочих равных оценках повысить релевантность сниппетов с большим количеством лексики, совпавшей с лексикой запроса, и большим размером. Более крупные сниппеты с большей вероятностью содержат ответ, нежели короткие сниппеты, что может помочь при поиске, если оценки, опирающиеся на семантику, малы. Кроме того, крупные сниппеты, как правило, содержат более развернутый ответ. В экспериментах использовались эмпирически подобранные коэффициенты $\beta_{ls} = 0{,}95, \beta_l = 0{,}05$.

Алгоритм определения лексико-семантической оценки релевантности предложения сниппета запросу (ЛС-оценки предложения r_{ls}^{s})

Лексико-семантическая оценка $r_{ls}^{s} \in [0,1]$ релевантности предложения сниппета вопросительному запросу вычисляется как линейная свертка трех компонент:

- лексической оценки (Л-оценка) предложения r_{l}^{s} (lexis) – характеризует близость запроса и предложения сниппета по лексике;
- оценки семантических ролей (СР-оценка) r_{sr}^{s} (semantic roles) – характеризует близость запроса и предложения сниппета по предикатно-аргументным структурам (ПА-структурам), используя семантические роли аргументов;
- оценки семантических отношений (СО-оценка) r_{sn}^{s} (semantic net) – характеризует близость запроса и предложения сниппета по семантическим отношениям семантической сети.

Оценка

$$r_{ls}^{s} = \gamma_{l} * r_{l}^{s} + \gamma_{sr} * r_{sr}^{s} + \gamma_{sn} * r_{sn}^{s}, \qquad (5)$$

где $\gamma_{l} \in [0,1]$, $\gamma_{sr} \in [0,1]$, $\gamma_{sn} \in [0,1]$ – параметры алгоритма такие, что $\gamma_{l} + \gamma_{sr} + \gamma_{sn} = 1$.

Лексическая оценка релевантности вычисляется как отношение мощности пересечения множества лемм запроса L_q со множеством лемм предложения сниппета L_s к мощности множества лемм запроса. Эти множества не включают в себя леммы стоп-слов: предлогов, союзов, пунктуации и др., кроме того, в них не входят вопросительные слова запроса. Если предложение сниппета является вопросительным, то релевантность занижается путем умножения Л-оценки на понижающий коэффициент $\delta_{l} \in (0,1)$, который является параметром алгоритма. Это необходимо, чтобы уменьшить значимость вопросительных предложений по сравнению с утвердительными, поскольку в сниппетах поисковиков часто вместо ответа на вопрос находится сам вопрос. Кроме того, ответ в виде утвердительного предложения является более предпочтительным для пользователя. На практике релевантность вопросительных предложений необходимо занижать примерно в два раза:

$$r_{l}^{s} = \delta_{l} \frac{|L_q \cap L_s|}{|L_q|}, \qquad (6)$$

$$\delta_{l} = \begin{cases} 1, \text{ если s – утвердительное предл.} \\ \delta_{l}^{q} \in (0,1), \text{ если s – вопросительное предл.} \end{cases}$$

Рассмотрим метод расчета СР-оценки. Пусть в запросе определено P_q ПА-структур. СР-оценка релевантности предложения сниппета запросу r_{sr}^{s} представляет собой взвешенную нормализованную сумму оценок близости отдельных ПА-структур запроса и предложения сниппета $r_{sr}^{p_i}$ $(i = 1, \dots P_q)$. Пусть в предложении сниппета присутствует P_s ПА-структур. Каждая структура запроса p_i $(i = 1, \dots P_q)$ сопоставляется со всеми ПА-структурами предложения сниппета p_j $(j = 1, \dots P_s)$, и для каждой такой пары вычисляется оценка сходства $r_{sr}^{p}(p_i, p_j)$. Для заданной ПА-структуры p_i запроса выбирается структура предложения сниппета с максимальной оценкой сходства, и эта оценка становится оценкой $r_{sr}^{p_i}$:

$$r_{sr}^{p_i} = \max_{j} r_{sr}^{p}(p_i, p_j). \qquad (7)$$

Каждая оценка сходства пары ПА-структур $r_{sr}^{p}(p_i, p_j)$ в свою очередь является суммой оценок сходства семантических аргументов $r_{sr}^{pa}(a^q, a^s)$ в ПА-структуре запроса и структуре предложения сниппета. Пусть в ПА-структуре запроса находится K аргументов: $p^q = \{a_1^q, a_2^q, \dots, a_K^q\}$, а в ПА-структуре предложения сниппета – M аргументов $p^s = \{a_1^s, a_2^s, \dots, a_M^s\}$. Каждый семантический аргумент из ПА-структуры запроса, которому назначена семантическая роль $a_k^q \in p^q$, $(k = \overline{1,K})$, сопоставляется с каждым семантическим аргументов ПА-структуры предложения сниппета $a_m^s \in p^s$, $(m = \overline{1,M})$, в результате чего вычисляется оценка сходства аргументов $r_{sr}^{pa}(a_k^q, a_m^s)$, $(k = \overline{1,K}, m = \overline{1,M})$.

Представим семантический аргумент в виде четверки:

$$a = < role, pred, syn, neg >, \qquad (8)$$

где $role$ – семантическая роль; $pred$ – лемма предикатного слова аргумента или идентификатор словарной статьи предикатного слова; syn – синтаксическое поддерево аргумента; neg – флаг, указывающий на наличие в тексте признаков отрицания аргумента, например, в

отрывке «*бумагу не изготавливают из нефти*» аргумент «бумагу» имеет признаки отрицания.

Пусть имеется семантический аргумент запроса $a^q =< role^q, pred^q, syn^q, neg^q >$ и семантический аргумент предложения сниппета $a^s =< role^s, pred^s, syn^s, neg^s >$. Рассмотрим алгоритм вычисления оценки сходства семантических аргументов $r_{sr}^{pa}(a^q, a^s)$.

1) Зададим начальное значение $r_{sr}^{pa}(a^q, a^s) = 0$.

2) Если $role^s \neq role^q$, то оценка $r_{sr}^{pa}(a^q, a^s) = 0$, выход. Иначе перейти к следующему шагу.

3) Если синтаксической вершиной syn^s является вопросительное слово, то оценка $r_{sr}^{pa}(a^q, a^s) = 0$, выход. Иначе перейти к следующему шагу.

4) Проверить совпадение по предикатному слову. Если $pred^s = pred^q$ (совпадают леммы предикатных слов или идентификаторы словарных статей предикатных слов), то модифицировать $r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_p$. Вес $w_p > 0$. Перейти к следующему шагу.

5) Если синтаксической вершиной syn^q является вопросительное слово, то перейти к шагу 8. Иначе перейти к следующему шагу.

6) Сравнить синтаксические поддеревья аргумента запроса syn^q и аргумента сниппета syn^s. Оценка сходства синтаксических поддеревьев $sim(syn^s, syn^q)$ вычисляется путем сопоставления лемм слов с учетом уровня слов в поддереве аргумента запроса и аргумента сниппета. Модифицировать $r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_a * sim(syn^s, syn^q)$. Вес $w_a > 0$. Перейти к следующему шагу.

7) Если есть совпадения по предикатным словам и синтаксическим поддеревьям, то модифицировать $r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_f$. Вес $w_f > 0$. Выход.

8) Если нет совпадения по предикатным словам, то модифицировать $r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_{q1}$, вес $w_{q1} > 0$, выход. Иначе перейти к следующему шагу.

9) Если $neg^s \neq neg^q$, выход. Иначе перейти к следующему шагу.

10) Модифицировать $r_{sr}^{pa}(a^q, a^s) := r_{sr}^{pa}(a^q, a^s) + w_{q2}$, вес $w_{q2} > 0$ может быть модифицирован набором эвристик, например, если аргумент сниппета представлен местоимением, то вес w_{q2} снижается. На этом шаге также выделяется ответ – синтаксическое поддерево аргумента сниппета заданной максимальной глубины. Выход.

11) Конец.

Величины w_p, w_a, w_f, w_{q1}, w_{q2} являются параметрами алгоритма, их можно варьировать в соответствии с эвристическими предположениями о значимости вклада компонентов в оценку релевантности или, например, настраивать с помощью методов оптимизации на заданной размеченной выборке.

Среди всех оценок пар сходства семантических аргументов для заданного аргумента запроса и заданной ПА-структуры предложения сниппета находится максимальная, которая, в конечном счете, участвует в расчете оценки релевантности всей ПА-структуры:

$$r_{sr}^{p}(p_i, p_j) = \sum_{k=1}^{K_i} \max_m r_{sr}^{ra}(a_k^q, a_m^s). \qquad (9)$$

Оценка сходства между ПА-структурами $r_{sr}^{p}(p_i, p_j)$ может быть больше единицы. При этом максимальная оценка для разных по размеру и признакам ПА-структур может отличаться. Структуры, в которых присутствует больше семантических аргументов, которым назначены роли, потенциально могут получить большую оценку, чем те, в которых семантических аргументов меньше. Это соответствует представлению о том, что наличие сходства крупных ПА-структур вносит более крупный вклад в оценку релевантности запроса сниппету, чем наличие сходства мелких ПА-структур. Кроме того, структуры запроса, в которых присутствует аргумент-вопросительное слово, также считаются более важными при оценке релевантности по семантическим аргументам.

В итоге СР-оценка запроса и предложения сниппета определяется следующим образом:

$$r_{sr} = \delta_{sr} \frac{\sum_{i=1}^{P_q} \max_j r_{sr}^{p}(p_i, p_j)}{Z_{sr}}, \qquad (10)$$

где Z_{sr} – коэффициент нормализации, который необходим для того, чтобы конечная оценка релевантности по семантическим ролям не выходила за рамки отрезка [0,1].

Z_{sr} вычисляется как сумма всех максимально возможных оценок сходства ПА-структур

запроса вне зависимости от ПА-структур предложения сниппета. Коэффициент $\delta_{sr} \in [0,1)$ понижает СР-оценку релевантности в соответствии с набором эвристик. Эти эвристики учитывают различную лингвистическую информацию о предложении, например, тип предложения: утвердительное, предложение-условие («*Если бы произошел финансовый кризис в 2012 году…*»), вопрос. Релевантность утвердительного предложения должна быть выше, чем предложения-условия или вопросительного предложения.

СО-оценка r_{sn}^{s} рассчитывается по схожему принципу, однако семантические отношения не привязаны к предикатным словам и ПА-структурам, поэтому СО-оценка вычисляется как нормализованная сумма оценок сходства семантического отношения в запросе с семантическим отношением в предложении сниппета $r_{sn}^{sr}(r^q, r^s)$.

Пусть в предложении запроса имеется R_q отношений, в предложении сниппета – R_s отношений. Представим отношение в виде тройки:

$$r =< type, syn_{left}, syn_{right} >, \qquad (11)$$

где $type$ – тип семантического отношения; syn_{left} – синтаксическое поддерево, в его вершине находится слово, из которого выходит семантическое отношение; syn_{right} – синтаксическое поддерево, в его вершине находится слово, в которое ведет семантическое отношение. Каждое отношение запроса $r^q =< type^q, syn_{left}^q, syn_{right}^q >$ попарно сравнивается с отношением предложения сниппета $r^s =< type^s, syn_{left}^s, syn_{right}^s >$. Рассмотрим алгоритм сравнения отношений.

1) Зададим начальное значение $r_{sn}^{sr}(r^q, r^s) = 0$.

2) Если $type^q \neq type^s$, то оценка $r_{sn}^{sr}(r^q, r^s) = 0$, выход. Иначе перейти к следующему шагу.

3) Если синтаксической вершиной syn_{left}^s или syn_{right}^s является вопросительное слово, то оценка $r_{sn}^{sr}(r^q, r^s) = 0$. Иначе перейти к следующему шагу.

4) Если синтаксической вершиной syn_{left}^q или syn_{right}^q является вопросительное слово, то перейти к шагу 8. Иначе перейти к следующему шагу.

5) Оценить сходство синтаксических поддеревьев $sim(syn_{left}^q, syn_{left}^s) \in [0,1]$. Модифицировать $r_{sn}^{sr}(r^q, r^s) := r_{sn}^{sr}(r^q, r^s) + w_{sl} * sim(syn_{left}^q, syn_{left}^s)$. Вес $w_{sl} > 0$. Перейти к следующему шагу.

6) Оценить сходство синтаксических поддеревьев $sim\left(syn_{right}^q, syn_{right}^s\right) \in [0,1]$. Модифицировать $r_{sn}^{sr}(r^q, r^s) := r_{sn}^{sr}(r^q, r^s) + w_{sr} * sim(syn_{right}^q, syn_{right}^s)$. Вес $w_{sr} > 0$. Перейти к следующему шагу.

7) Модифицировать $r_{sn}^{sr}(r^q, r^s) := \delta_{sn} * r_{sn}^{sr}(r^q, r^s)$, $\delta_{sn} > 1$. Коэффициент δ_{sn} необходим для того, чтобы повысить различие между оценками сходства при высоких и при низких суммарных оценках сходства синтаксических поддеревьев. Выход.

8) Без ограничения общности положим, что вопросительное слово является вершиной syn_{left}^q. Модифицировать $r_{sn}^{sr}(r^q, r^s) := r_{sn}^{sr}(r^q, r^s) + w_{rq}$. Перейти к следующему шагу.

9) Оценить сходство синтаксических поддеревьев $sim\left(syn_{right}^q, syn_{right}^s\right) \in [0,1]$. Модифицировать $r_{sn}^{sr}(r^q, r^s) := r_{sn}^{sr}(r^q, r^s) + w_{lq} * sim\left(syn_{right}^q, syn_{right}^s\right)$. Перейти к следующему шагу.

10) Модифицировать $r_{sn}^{sr}(r^q, r^s) := \delta_{sn} * r_{sn}^{sr}(r^q, r^s)$. Выход.

11) Конец.

Среди всех оценок пар сходства семантических отношений для заданного семантического отношения запроса находится максимальная, которая, в конечном счете, участвует в расчете СО-оценки r_{sn}:

$$r_{sn} = \frac{\sum_{i=1}^{R_q} \max_j r_{sn}^{sr}\left(r_i^q, r_j^s\right)}{Z_{sn}}, \qquad (12)$$

где Z_{sn} – коэффициент нормализации, который необходим для того, чтобы конечная оценка релевантности по семантическим отношениям не выходила за рамки отрезка [0,1].

Z_{sn} вычисляется как сумма всех максимально возможных оценок сходства семантических отношений предложения запроса вне зависимости от отношений в сниппете. Значения w_{sl}, w_{sr}, w_{rq}, w_{lq}, δ_{sn} являются параметрами алгоритма.

2.2. Метод извлечения определений из текстов научных публикаций на основе анализа семантико-синтаксической структуры предложений

Для извлечения определений из текстов научных публикаций был предложен метод, основанный на сравнении лексико-синтаксической и семантической структуры предложения со списком фреймов (шаблонов). Построение фреймов проводилось вручную. Для этого экспертами в области лингвистики предварительно был исследован корпус научно-технических текстов, в результате чего были выделены конструкции, которые часто указывают на присутствие в предложении определения термина. В результате обобщения этих конструкций были сформированы фреймы, позволяющие как находить предложение, в котором дается определение, так и выделить определяемый термин. Во фреймах содержится ряд правил и шаблонов, учитывающих разнородную информацию, извлеченную из текста с помощью автоматического лингвистического анализа. На основе данных, полученных в результате нескольких итераций экспериментов на частично размеченном корпусе для разработки, было проведено уточнение фреймов. Основной особенностью рассматриваемого в настоящей статье подхода к извлечению определений является то, что во фреймах помимо лексики, морфологических признаков слов и синтаксических деревьев учитываются также семантические ролевые структуры. В Табл. 1 приведено несколько примеров частей этих фреймов, в которых учитываются семантические роли.

Основным преимуществом использования семантики, а именно ролевых структур, является то, что на их основе можно формировать более простые и более общие правила извлечения информации, которые абстрагируются от специфичных синтаксических деревьев и форм предикатных слов. Например, в каждом из нижеперечисленных определений для извлечения термина во фрейме можно указать правило, по которому будет искаться синтаксическая группа, в вершине которой находится семантический аргумент с ролью «делибератив»:

- «Атрибут (attribute) может быть формально определен как функция, отображающая набор сущностей или набор связей в набор значений или Декартово произведение наборов значений».
- «Аксиоматический метод традиционно определяется как такой способ дедуктивного построения научной теории, когда ее основу составляют лишь некоторые, принятые без доказательств положения – аксиомы, а все остальные положения теории выводятся из них путем рассуждений, корректных относительно принимаемой этой теорией логики».

Табл. 1. Примеры фреймов, в которых учитываются семантические роли

Фрейм	Примеры
ЧР(Сущ.) && Сем.роль(эстиматив) + Л(«называться»)	Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая обобщенный категориальный смысл и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков.
Сем.роль(делибератив) + ПС(«определять») + Л(«как»)	Аксиоматический метод традиционно определяется как такой способ дедуктивного построения научной теории, когда ее основу составляют лишь некоторые, принятые без доказательств положения – аксиомы, а все остальные положения теории выводятся из них путем рассуждений, корректных относительно принимаемой этой теорией логики.
Л(«под») + Сем.роль(эстиматив) + Л(«быть»)&&Время(буд.) + (Л(«понимать») || Л(«пониматься») || Л(«подразумевать») || Л(«подразумеваться»))	Под алгеброй мы будем понимать линейное пространство A (например, над полем действительных чисел R), снабженное ассоциативной операцией умножения и единицей…
Обозначения: «Л» – лемма; «Сем.роль» – семантическая роль; «ЧР» – часть речи; «ПС» – идентификатор предикатного слова; && – логическое и; || – логическое или.	

- «Информацию определим как формы эксплицитного и отчужденного от человека представления его знаний, предназначенные для передачи, непосредственного сенсорного восприятия и понимания их другими людьми».
- «… Ч. Тилли, который определяет государство как «социальную организацию, отличную от домашних хозяйств и родовых групп, обладающую правом на применение насилия и очевидным приоритетом в определенных аспектах над другими организациями на значительной территории».

В каждом из примеров меняются синтаксические функции группы определяемого термина, позиция относительно предикатного слова и форма предикатного слова, однако семантическая роль остается неизменной. Поэтому эти особенности уже не нужно учитывать при построении правила для извлечения определяемого термина.

Рассмотрим общий алгоритм извлечения определений и определяемых терминов из текстов научных публикаций. В первую очередь в тексте выделяются предложения, которые содержат ключевые слова и токены, которые, в свою очередь, сопоставлены с фреймами для извлечения терминов. Предложения, содержащие ключевые слова и токены, проходят ряд общих для всех фреймов простых проверок (таких как проверка на тип предложения: вопросительное / утвердительное). Подобная предварительная фильтрация с помощью простых правил и ключевых слов позволяет с небольшими вычислительными затратами отсеять большое количество нерелевантных предложений, не содержащих определений терминов.

Далее каждое отобранное предложение сопоставляется с соответствующими ключевым словам фреймами. Каждый из фреймов независимо от других осуществляет ряд более сложных проверок и, либо отвергает предложение как несоответствующее заложенным во фрейме правилам, либо извлекает один или несколько терминов. После этого происходит постобработка терминов с помощью набора эвристик, которые находят и отсеивают часто встречающиеся ошибочные случаи, общие для всех фреймов. Например, такими случаями являются подписи к рисункам или таблицам, мелкие математические обозначения. В результате работы алгоритма на выходе формируется список терминов, которым дано определение, и предложения, в которых встречаются эти термины. Предложения, в которых встречаются извлеченные термины, помечаются как определения.

На основе метода извлечения определений реализована функция полнотекстового поиска по определениям в поисково-аналитической системе Exactus Expert (http://expert.exactus.ru/) Извлеченные определения и термины помечаются в поисковом индексе отдельными тегами. Поисковый механизм позволяет задавать ограничение зоны поиска по таким тегам. Благодаря этому пользователи могут осуществлять поиск не по полным текстам статей, а по терминам, для которых есть определение. Это позволяет пользователю быстро отыскать непосредственно определение неизвестного термина, что является весьма полезным при эксплоративном поиске.

3. Экспериментальное исследование методов

В экспериментальных исследованиях сравнивались два подхода к выполнению синтаксического и семантического видов анализа: раздельный последовательный синтаксический и семантический анализ, совмещенный семантико-синтаксического анализ. Применяемый в работе синтаксический анализатор основан на MaltParser [26], способ обучения которого описан в работе [7]. Используемый метод семантического анализа строит ролевые структуры высказываний на основе семантического словаря [7, 8]. Для семантико-синтаксического анализа используется метод, описанный в [9]. В нем сначала выполняется синтаксический анализ, затем полученное дерево корректируется на основе информации, полученной от семантического анализатора, и на исправленном дереве заново проводится семантический анализ. За счет этого повышается качество как синтаксического, так и семантического анализа.

3.1. Экспериментальное исследование метода ранжирования ответов для вопросно-ответного поиска

Для экспериментальной проверки метода ранжирования сниппетов для вопросно-ответного поиска был создан размеченный корпус вопросительных запросов и ответов. Двое разметчиков сформировали список вопросительных запросов на русском языке. Вопросы являются

категориальными и содержат вопросительные слова «кто» (и др. формы «кем», «кому» и т.д.), «что» (и др. формы «чем», «чему» и т.д.), «кто такой», «что такое», «зачем», «когда», «где», «сколько», «куда», «откуда», «почему». Всего было сформулировано 195 запросов. Они были заданы четырем поисковым системам в Интернете: Yandex, Google, Bing и Yahoo. От каждой системы был получен топ ответов глубиной 20. Каждый ответ состоит из поискового сниппета и его заголовка. Таким образом, для каждого заданного вопроса был получен список, состоящий из около 50-70 ответов без повторений: сниппетов и их заголовков.

Для каждого запроса разметчики оценили каждый ответ четырех поисковых систем и определили, является ли он ответом на поставленный вопрос. Оценивался и сниппет, и его заголовок. Ответ поисковой системы считался ответом на поставленный вопрос если:

- сниппет или его заголовок содержат прямой ответ на вопрос;

Пример:

Вопрос: «Из чего делают сахар?»

Ответ:

Заголовок: «Из чего делают сахар ? - Большой ...»

Сниппет: «Сахар (или сахароза) делают в основном из *сока сахарного тростника или сахарной свеклы.*»

- если информация в заголовке соотносится с заданным вопросом, а в сниппете содержится краткий правильный ответ;

Пример:

Вопрос: «Кто написал «что делать»?»

Ответ:

Заголовок: «Ответы@Mail.Ru: кто написал произведение что делать?»

Сниппет: «*Николай Гаврилович Чернышевский.* Нравится Пожаловаться. 3 ответа. опорполп Мастер (2258) 4 года назад. Чернышевский. Нравится Пожаловаться.",»

- если в заголовке или сниппете отсутствует прямой подробный ответ на вопрос, но содержится корректная информация, которую человек интуитивно может интерпретировать как ответ.

Пример:

Вопрос: «Чем стричь газоны?»

Ответ:

Заголовок: «Стрижка газона - ее особенности»

Сниппет: «Стрижка газона по видам. Для партерных газонов можно использовать готовый рулонный газон, который необходимо подстригать на высоту 5 см. — это в первые 2 года. *Газона хорошего у нас пока нет, еще только строимся, но траву стрижем. Самое первое, что подарили мне дети после покупки деревенского дома — газонокосилку.*»

Качество автоматического ранжирования для вопросно-ответного поиска оценивалось по точности p, которая вычислялась как отношение количества вопросов, ответы на которые система дала на глубине не больше чем $d > 0$, ко всему количеству запросов. Точность оценивалась при $d = 1$, $d = 2$, $d = 3$, $d = 4$. Наиболее важным показателем в вопросно-ответном поиске считается точность при $d = 1$ [27, 28] – это оценка того, как часто ответ на заданный вопрос находится в первом сниппете или заголовке выдачи метапоисковой системы. Например, это важно для задачи извлечения непосредственно ответа на вопрос.

Сравнивались четыре системы ранжирования (Табл. 2):

- система, которая ранжирует ответы случайным образом (среднее по 5 прогонам). Обозначим ее как «Случ. ранж.»;

Табл. 2. Точность ранжирования сниппетов для вопросно-ответного поиска, %

Максимальная глубина, *d*	**Системы ранжирования**			
	Сем. ранж. Сем. ан.	Сем. ранж. Сем.-син. ан.	Лексич. ранж.	Случ. ранж.
1	53,9	**58,0**	46,1	26,0
2	73,1	**75,1**	64,8	40,1
3	79,3	**79,8**	71,5	52,8
4	82,9	**83,4**	75,6	62,3

- система, которая ранжирует ответы на основе лишь лексического критерия – все веса и коэффициенты при семантических оценках – нули. Обозначим ее как «Лексич. ранж.»;
- система, которая использует описанный выше алгоритм вычисления релевантности с учетом семантической информации, полученной от анализатора, в котором синтаксический и семантический анализ выполняются раздельно. Обозначим ее как «Сем. ранж. Сем. ан.»;
- система, которая использует описанный выше алгоритм вычисления релевантности с учетом семантической информации, полученной от семантико-синтаксического анализатора. Обозначим ее как «Сем. ранж. Сем.-син. ан.».

Для оценки качества вопросно-ответного поиска часто используют также метрику mean reciprocal rank (MRR), которая вычисляется следующим образом:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{r_i}, \qquad (13)$$

где r_i – ранг первого правильного ответа на i-й вопрос в выдаче; Q – общее количество вопросов.

Эта метрика учитывает правильные ответы системы на небольшой глубине выдачи, которые не оказались первыми, но при этом штрафует их, снижая вклад в итоговую оценку качества. Таким образом, она является более мягкой оценкой по отношению к точности при $d = 1$, но более жесткой по отношению к точности при $d > 1$.

Обычно максимальный ранг ограничен. В проведенных экспериментах использовалась метрика MRR, которая применялась на соревнованиях TREC по вопросно-ответному поиску; в ней максимальный ранг ограничивается 5 [12]. Сравнивались три способа ранжирования: «Лексич. ранж.», «Сем. ранж. Сем. ан.», «Сем. ранж. Сем.-син. ан.». Оценки MRR для них представлены в Табл. 3.

Табл. 3. Оценки MRR систем ранжирования сниппетов для вопросно-ответного поиска

Системы ранжирования	$\boldsymbol{MRR \cdot 100}$
Лексич. ранж.	59,0
Сем. ранж. Сем. ан.	67,2
Сем. ранж. Сем.-син. ан.	**69,6**

Результаты показывают следующее:

1) Разработанный критерий вносит большой вклад в качество ранжирования ответов поисковых систем: точность на 30-35% выше по сравнению со случайным ранжированием.

2) Учет семантической структуры предложения при оценке релевантности ответов в вопросно-ответном поиске позволяет повысить точность на 10-12% и значение MRR на более чем 8% по сравнению с лексическим критерием ранжирования, а также позволяет извлекать из сниппета сам ответ на вопрос. В качестве иллюстрации на Рис. 2 представлен скриншот веб-интерфейса модуля метапоиска системы Exactus.ru, в котором реализован описанный метод. В этом примере ответы на вопросы в сниппетах выделены курсивом.

3) Качество ранжирования при использовании семантико-синтаксического анализатора существенно выше, чем при использовании анализатора, в котором синтаксический и семантический анализ выполняются раздельно – точность выше на 4% при $d = 1$, существенно выше и значение MRR. Повышение качества построения синтаксических деревьев зависимостей и определения ролевых структур высказываний за счет использования метода семантико-синтаксического анализа позволяет более точно проводить анализ как вопросительных запросов, так и поисковых сниппетов. На Рис. 3 представлен пример, в котором семантико-синтаксический анализатор исправил синтаксическое дерево в сниппете ответа на вопрос «*откуда в Россию привезли картофель?*». За счет этого были установлены семантическая роль «директив» и семантическое отношение «TRA» , которые помогли поднять ответ на вопрос на первое место в выдаче.

3.2. Экспериментальное исследование метода извлечения определений из текстов научных публикаций

Для экспериментальной проверки разработанного метода был размечен корпус научных публикаций. В состав корпуса вошли статьи журналов из перечня ВАК, доклады на российских научных конференциях и авторефераты диссертаций. Размер размеченного корпуса составляет более 110 000 токенов, в корпусе выделено более 380 определений. Рассчитывались нестрогие метрики: точность p, полнота r и

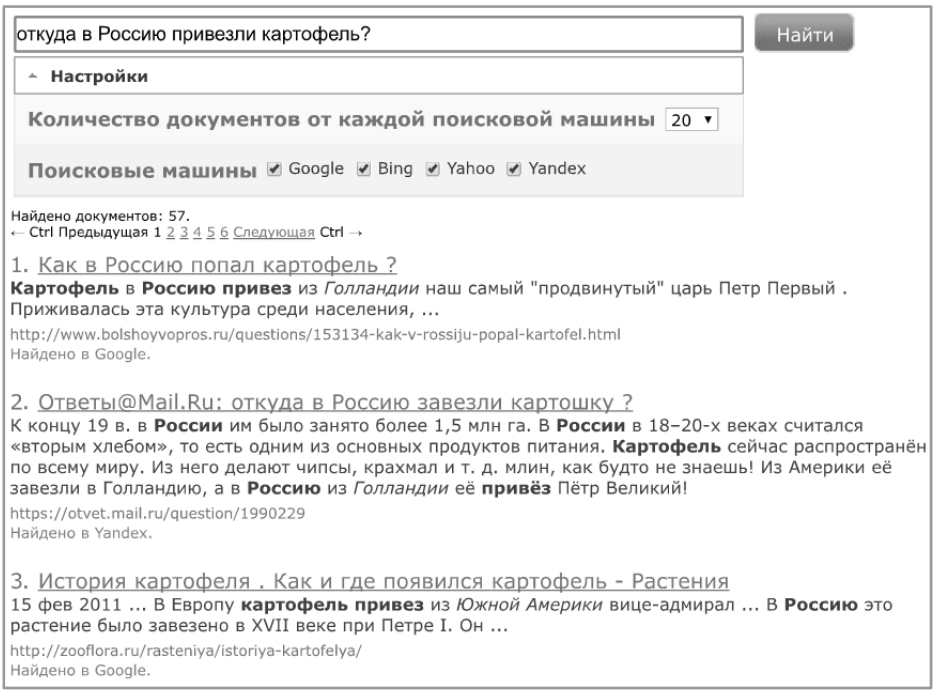


Рис.2. Результат работы модуля метапоиска системы Exactus

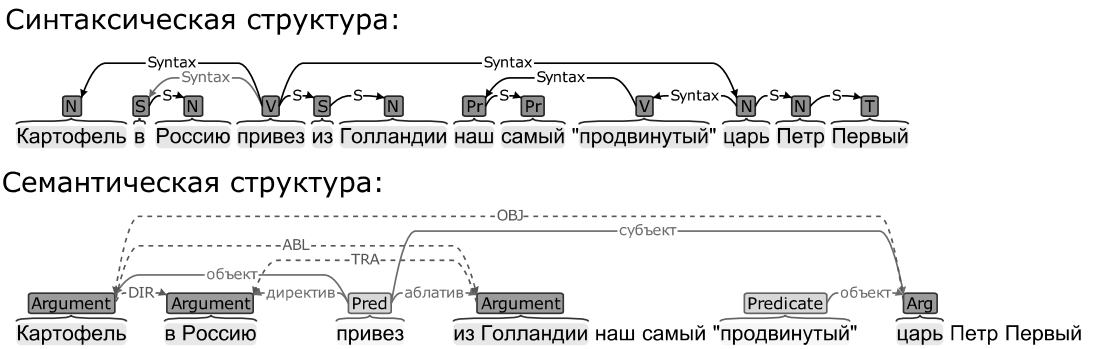


Рис.3. Пример предложения в сниппете ответа на вопрос «откуда в Россию привезли картофель?», в котором семантико-синтаксический анализатор исправил синтаксическое дерево

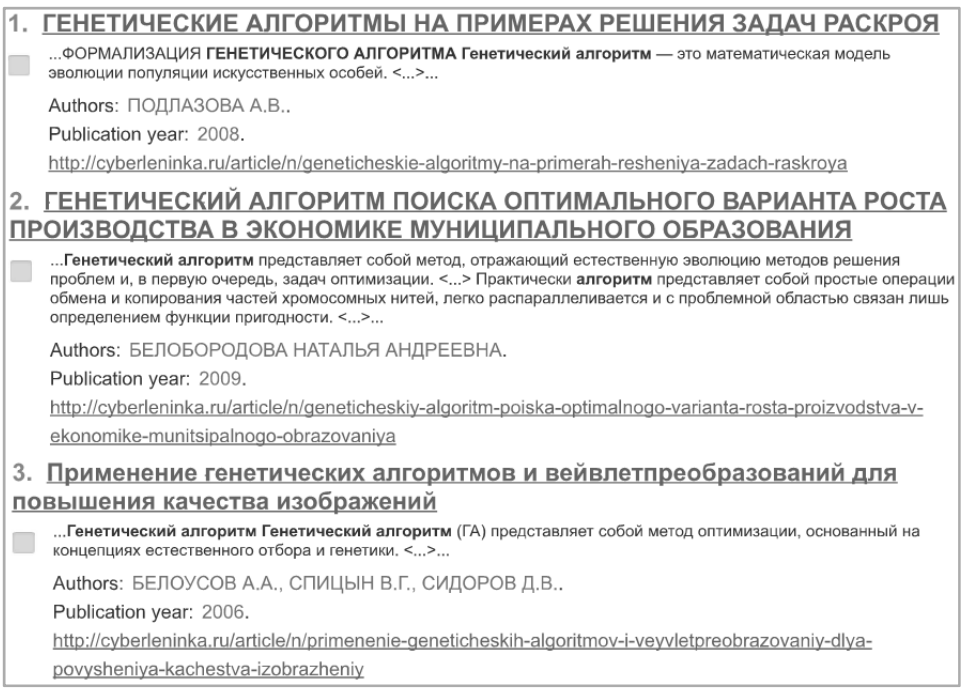
мера F_1, в которых пересечение отрезка текста в проверочном корпусе с отрезком, полученным автоматически от анализатора, даже без точного совпадения краев отрезков, считалось правильным ответом. Сравнивались результаты трех систем для извлечения определений:

- Система, в которой реализованы фреймы, использующие семантические роли, полученные от анализатора, в котором синтаксический и семантический анализ выполняются раздельно. Обозначим эту систему как «Сем».
- Система, в которой реализованы фреймы, использующие семантические роли, полученные от семантико-синтаксического анализатора. Обозначим эту систему как «Сем.-син.».
- Система, в которой отсутствуют фреймы, использующие семантические роли. Обозначим эту систему как «Без сем.».

Результаты экспериментов представлены в Табл. 4.

Табл. 4. Оценки качества метода извлечения определений из текстов научных публикаций

Система	***p*,%**	***r*,%**	***F*$_1$,%**
Сем	80,6	67,4	73,4
Сем.-син.	**80,7**	**67,6**	**73,6**
Без сем.	76,7	52,5	62,3


1. **ГЕНЕТИЧЕСКИЕ АЛГОРИТМЫ НА ПРИМЕРАХ РЕШЕНИЯ ЗАДАЧ РАСКРОЯ**

...ФОРМАЛИЗАЦИЯ **ГЕНЕТИЧЕСКОГО АЛГОРИТМА Генетический алгоритм** — это математическая модель эволюции популяции искусственных особей. <...>...

Authors: ПОДЛАЗОВА А.В..

Publication year: 2008.

http://cyberleninka.ru/article/n/geneticheskie-algoritmy-na-primerah-resheniya-zadach-raskroya

2. **ГЕНЕТИЧЕСКИЙ АЛГОРИТМ ПОИСКА ОПТИМАЛЬНОГО ВАРИАНТА РОСТА ПРОИЗВОДСТВА В ЭКОНОМИКЕ МУНИЦИПАЛЬНОГО ОБРАЗОВАНИЯ**

...**Генетический алгоритм** представляет собой метод, отражающий естественную эволюцию методов решения проблем и, в первую очередь, задач оптимизации. <...> Практически **алгоритм** представляет собой простые операции обмена и копирования частей хромосомных нитей, легко распараллеливается и с проблемной областью связан лишь определением функции пригодности. <...>...

Authors: БЕЛОБОРОДОВА НАТАЛЬЯ АНДРЕЕВНА.

Publication year: 2009.

http://cyberleninka.ru/article/n/geneticheskiy-algoritm-poiska-optimalnogo-varianta-rosta-proizvodstva-v-ekonomike-munitsipalnogo-obrazovaniya

3. **Применение генетических алгоритмов и вейвлетпреобразований для повышения качества изображений**

...**Генетический алгоритм Генетический алгоритм** (ГА) представляет собой метод оптимизации, основанный на концепциях естественного отбора и генетики. <...>...

Authors: БЕЛОУСОВ А.А., СПИЦЫН В.Г., СИДОРОВ Д.В..

Publication year: 2006.

http://cyberleninka.ru/article/n/primenenie-geneticheskih-algoritmov-i-veyvletpreobrazovaniy-dlya-povysheniya-kachestva-izobrazheniy

Рис.4. Поисковая выдача, полученная в результате поиска по определениям по запросу «{генетический алгоритм}»

Полученные результаты показывают, что фреймы, использующие семантические роли, обрабатывают существенное количество случаев определения терминов в научных публикациях (более 15% по полноте), тем самым вносят значительный вклад в решение поставленной задачи. Использование системы семантико-синтаксического анализа не дало значимого прироста качества по сравнению с использованием системы, в которой синтаксический и семантический анализ выполняются раздельно. Причина этого заключается в том, что наибольшую роль в решении этой задачи играют лексико-морфологические правила, не учитывающие семантику и синтаксис предложений, следовательно, прирост полноты определения ролевых структур не сильно отражается на конечном результате. Примерно половина выделенных определений подпадает под шаблоны вида «*определяемый термин – [это] определение*».

Анализ ошибок выявил, что значительное число не извлеченных терминов требуют для их нахождения разрешения кореферентных связей как на уровне одного предложения, так и на уровне всего текста. Другая значительная часть ненайденных терминов, размеченных в проверочном корпусе, соответствует отдельным редким паттернам. На точность в значительной степени негативно влияют ошибки фрейма, где определяемый термин и его определение разделены тире или дефисом. Фрейм, который бы соответствовал лишь определениям в таких случаях, выделить довольно трудно из-за большого количества различных лингвистических конструкций, использующих тире или дефис, а также ошибок их употребления в текстах на естественном языке.

В целом полученные результаты иллюстрируют применимость разработанного анализатора, в котором реализованы правила, использующие семантические роли, полученные в результате семантико-синтаксического анализа, для решения задачи извлечения определений и определяемых терминов в прикладных приложениях. На Рис. 4 представлен пример работы функции полнотекстового поиска по определениям в системе Exactus Expert, в которой внедрен предложенный метод.

Заключение

В статье представлены методы решения двух прикладных задач: вопросно-ответного поиска и извлечения определений из научных публикаций, в которых используются результаты семантического и семантико-синтаксического анализа.

В методе ранжирования сниппетов для вопросно-ответного поиска в метапоисковой системе наряду с лексикой учитываются семантические роли и отношения. Экспериментально показано, что семантическая информация вносит большой вклад в точность работы алгоритма ранжирования и позволяет извлекать из сниппетов непосредственно ответы на вопросы. Кроме этого, результаты экспериментальных исследований свидетельствуют о значительном преимуществе использования семантико-синтаксического анализатора при решении этой задачи по сравнению с применением анализатора, в котором синтаксический и семантический анализ выполняются раздельно.

В методе извлечения определений и определяемых терминов из текстов научных публикаций реализованы фреймы, учитывающие семантические роли. Экспериментально показана эффективность разработанного метода и значимость вклада этих фреймов в решение поставленной задачи. Преимущество использования результатов семантического анализа заключается в том, что учет семантических ролей во фреймах упрощает построение правил для извлечения определений и терминов.

В дальнейших исследованиях планируется применить к задачам вопросно-ответного поиска и извлечения определений подходы, основанные на активном машинном обучении.

Литература

1. Осипов Г. С., Смирнов И. В., Тихомиров И. А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. — 2008. — № 2. — С. 3–10.
2. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности / И.А. Тихомиров, И.В. Смирнов, И.В. Соченков и др. // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Б.: БГТУ. — Т. 4. — 2012. — С. 100–108.
3. Osipov G. Methods for extracting semantic types of natural language statements from texts // 10th IEEE International Symposium on Intelligent Control. — Monterey, California, USA, 1995.
4. Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка // Институт русского языка РАН им. В. В. Виноградова. — 2004.
5. Gildea D., Jurafsky D. Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.
6. Тестелец Я. Г. Введение в общий синтаксис. — М.: Изд-во РГГУ, 2001.
7. 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. — № 1. — С. 11–24.
8. 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". — N. 13. — 2014. — P. 607–620.
9. Осипов Г. С., Шелманов А. О. Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой международной конференции "Системный анализ и информационные технологии" (САИТ). — Т. 1. — 2015. — С. 229–240.
10. Semantic analysis and question answering: a system under development / Igor Boguslavsky, Vyacheslav Dikonov, Leonid Iomdin et al. // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2015". — 2015. — P. 62–79.
11. Yao X., Van Durme B. Information extraction over structured data: Question answering with Freebase // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. — Association for Computational Linguistics, 2014. — P. 956–966.
12. Voorhees E. M. et al. The TREC-8 question answering track report // TREC. — Vol. 99. — 1999. — P. 77–82.
13. Building Watson: An overview of the DeepQA project / David Ferrucci, Eric Brown, Jennifer Chu-Carroll et al. // AI magazine. — 2010. — Vol. 31, N. 3. — P. 59–79.
14. ILQUA–an IE-driven question answering system / Min Wu, Michelle Duan, Samira Shaikh et al. // TREC. — 2005.
15. Shen D., Klakow D. Exploring correlation of dependency relation paths for answer extraction // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics / Association for Computational Linguistics. — 2006. — P. 889–896.
16. Matching sets of parse trees for answering multi-sentence questions / Boris Galitsky, Dmitry Ilvovsky, Sergey Kuznetsov, Fedor Strok // Proceedings of Recent Advances in Natural Language Processing (RANLP 2013). — 2013. — P. 285–293.
17. Огарок А. Стокона на РОМИП-2006 // Труды РОМИП'2006. — 2006. — С. 86–91.
18. Некрестьянов И., Некрестьянова М. РОМИП'2006: отчет организаторов // Труды РОМИП'2006. — 2006. — С. 7–29.
19. Shen D., Lapata M. Using semantic roles to improve question answering. // EMNLP-CoNLL. — 2007. — P. 12–21.
20. Pizzato L. A., Mollá D. Indexing on semantic roles for question answering // Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering / Association for Computational Linguistics. — 2008. — P. 74–81.
21. Combining semantic information in question answering systems / Paloma Moreda, Hector Llorens, Estela Saquete, Manuel Palomar // Information Processing & Management. — 2011. — Vol. 47, N. 6. — P. 870–885.
22. Васильева Н. Э. Шаблоны употреблений терминов и их использование при автоматической обработке

научно-технических текстов // Труды международной конференции "Диалог 2004". — 2004. — P. 96–101.
23. Большакова Е. И., Носков А. А. Программные средства анализа текстов на основе лексико-синтаксических шаблонов языка LSPL // Программные системы и инструменты: Тематический сборник. — 2010. — № 11. — С. 61–73.
24. Терминологический анализ текста на основе лексико-синтаксических шаблонов / Н. Э. Ефремова, Е. И. Большакова, А. А. Носков, В. Ю. Антонов // Труды международной конференции "Диалог 2010". — Т. 9. — М.: Изд-во РГГУ, 2010. — С. 124–129.
25. Большакова Е. И. Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития // Программные системы и инструменты: Тематический сборник. — 2014. — № 15. — С. 15–26.
26. MaltParser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, N. 2. — P. 95–135.
27. Learning to rank for robust question answering / Arvind Agarwal, Hema Raghavan, Karthik Subbian et al. // Proceedings of the 21st ACM international conference on Information and knowledge management. — 2012. — P. 833–842.
28. Dang H. T., Kelly D., Lin J. J. Overview of the TREC 2007 question answering track // Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007. — 2007.

Шелманов Артем Олегович. Младший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил Национальный исследовательский ядерный университет «МИФИ» в 2011 году. Автор 16 печатных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, информационно-аналитические системы, машинное обучение. E-mail: shelmanov@isa.ru

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

Ананьева Маргарита Игоревна. Инженер-исследователь ИСА ФИЦ ИУ РАН. Окончила Московский государственный лингвистический университет в 2013 году. Автор одной печатной работы. Область научных интересов: компьютерная лингвистика, анализ дискурса. E-mail: ananyeva@isa.ru

Смирнов Иван Валентинович. Заведующий лабораторией «Компьютерная лингвистика и интеллектуальный анализ информации» ИСА ФИЦ ИУ РАН. Окончил Российский университет дружбы народов в 2003 году. Кандидат физико-математических наук, доцент. Автор 48 печатных работ. Область научных интересов: обработка естественного языка, интеллектуальный анализ информации. E-mail: ivs@isa.ru

Semantic-syntactic analysis for question answering and definition extraction

A.O. Shelmanov, M.A. Kamenskaya, M.I. Ananyeva, I.V. Smirnov

Abstract. We research the contribution of semantic-syntactic analysis to the effectiveness of solving applied text processing tasks: question-answering and definition extraction from scientific publications. The paper presents methods for solving these tasks that in addition to morphological and syntactic structure use also semantic structure of texts. We conducted the experimental evaluation of these methods and experimental comparison of two approaches to syntactic and semantic analysis: separate parsing and join semantic-syntactic parsing.

Keywords: semantic parsing, joint semantic-syntactic parsing, question-answering, information extraction, definition extraction.

Artem O. Shelmanov. Junior Fellow in the Institute for Systems Analysis of the Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. Graduated from MIFI National Nuclear Research University in 2011. Author of 16 scientific papers. Research interests: artificial intelligence, computational linguistics, information analysis system, machine learning. E-mail: shelmanov@isa.ru

Margarita A. Kamenskaya. Research engineer in the Institute for Systems Analysis of the Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. Graduated from Peoples' Friendship University of Russia in 2014. Author of 5 scientific papers. Research interests: computational linguistics, information analysis system, data mining. E-mail: mak@isa.ru

Margarita I. Ananyeva. Research engineer in the Institute for Systems Analysis of the Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. Graduated from Moscow State Linguistic University in 2013. Author of 5 scientific papers. Research interests: methods for linguistic analysis, information technology, discourse analysis. E-mail: ananyeva@isa.ru

Ivan V. Smirnov. Head of laboratory “Computational linguistics and intelligent data analysis” in the Institute for Systems Analysis of the Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. Graduated from Peoples' Friendship University of Russia in 2003. Candidate of Physical and Mathematical Sciences, associate professor. Author of 48 scientific papers. Research interests: natural language processing, intelligent data analysis. E-mail: ivs@isa.ru