УДК 371.263+004.82
ББК 74.00

ОЦЕНКА ТРУДНОСТИ И СЛОЖНОСТИ УЧЕБНЫХ ЗАДАЧ НА ОСНОВЕ СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

Наумов И. С.[1], Выхованец В. С.[2]
(*ФГБУН Институт проблем управления РАН, Москва*)

Предложено решение задачи автоматической оценки сложности и трудности учебных задач на основе синтаксического анализа текста, выявления предикативной структуры его предложений и построения на этой основе семантической сети. Для подсчета объема знаний в семантической сети разработан математический аппарат, базирующийся на определения семантических расстояний между понятиями-словами. Показано, что объем знаний, содержащихся в семантической сети, является мерой на множестве семантических сетей, а введенное расстояние превращает это множество в метрическое пространство.

Ключевые слова: процесс обучения, педагогические измерения, трудность и сложность задач, синтаксический анализ, семантические сети, семантическое расстояние.

1. Введение

В последнее время большое число исследований направлено на автоматизацию и интеллектуализацию различных сфер деятельности. Однако в автоматизации обучения на сегодняшний день существует много нерешенных проблем. Такое положение дел связано с тем, что обучение представляет собой

[1] *Игорь Савельевич Наумов, инженер* (*naigsa@gmail.com*).
[2] *Валерий Святославович Выхованец, доктор технических наук, доцент* (*valery@vykhovanets.ru*).

сложный интеллектуальный процесс, плохо поддающийся формализации [3].

Заметной тенденцией в автоматизации обучения является разработка методов, методик и средств, реализующих индивидуальный подход к процессу обучения. К актуальным проблемам в этой области можно отнести проверку и оценку знаний, составление программ обучения, индивидуальный подбор учебных задач и др. Автоматический подбор учебных задач является одной из самых сложных проблем, так как связан с оценкой субъективной трудности и объективной сложности учебной задачи.

В процессе приобретения знаний условно можно выделить источник знаний – описание некоторой предметной области, или предмет обучения, и приемник знаний – субъект обучения. Знания предметной области признаются эталонными знаниями, которые остаются неизменными в течение всего процесса обучения, а знания обучающегося являются изменяемыми. Цель обучения – максимально полно передать знания предметной области обучаемому субъекту.

Основным средством проверки знаний является решение учебных задач. Учебную задачу можно рассматривать как содержащую знания предметной области, которые приобретает обучающийся в результате правильного решения задачи. Сравнение знаний, заключенных в учебной задаче, с соответствующими знаниями предметной области и текущими знаниями позволяет оценить сложность и трудность учебной задачи.

Однако в большинстве известных методик оценка сложности и трудности задачи осуществляется по косвенным критериям, например, по числу шагов решения, по времени решения, по вероятности решения и т.п. [28, 29].

Известен также подход, при котором сложность текста определяется по соотношению конкретных и абстрактных слов, по проценту новых слов, по длине предложений, по сложности логической структуры текста, а трудность оценивается после чтения текста на основе экспериментальной проверки понимания текста путем постановки вопросов к тексту и анализа ответов на эти вопросы [20].

Сегодня в области оценки сложности текстов популярны методы, основанные на анализе статистических закономерностей. Среди них можно выделить работы посвященные определению уровня читаемости текста [32, 33]. Недостаток таких методов заключается в том, что они не учитывают содержательные взаимосвязи слов, которые являются ключевыми для «понимания» смысла текста.

В итоге известные методы оценки сложности и трудности текстов и учебных задач не позволяют строить и использовать эффективные индивидуальные программы обучения. Поэтому проблема оценки сложности и трудности задач остается актуальной до сих пор.

В настоящей работе рассмотрен метод оценки сложности и трудности учебной задачи, основанный на автоматическом синтаксическом анализе текстов. В результате синтаксического анализа учебного материала строится семантическая сеть предметной области, которая представляет собой набор вершин (понятий предметной области) и набор связей (отношений между понятиями). Для построения семантической сети обучающегося строится объединенная семантическая сеть задач, которые правильно решил обучающийся. В этом случае оценка сложности и трудности новой учебной задачи осуществляется путем сравнения семантических сетей предметной области, семантической сети задачи и семантической сети обучающегося.

2. Сложность и трудность

Понятия «сложность» и «трудность» в научной литературе зачастую используют как синонимы. Между тем эти понятия имеют различное содержание и в рассматриваемой задаче играют ключевую роль. Обычно под сложностью понимают объективную оценку, а под трудностью – субъективную [18, с. 86].

Понятия «сложность» и «трудность» характеризуют связь между субъектом, решающим задачу, и объектом – учебной задачей. При этом трудность и сложность задачи зависят от различных объективных и субъективных факторов. Например, к объективным факторам относятся предмет задачи, требования

задачи, условия задачи, а к субъективным – способности и подготовку субъекта, его мотивация, психическое и физическое состояние и др. [4, с. 62].

В итоге имеем, что каждая задача может иметь две независимые оценки: сложность и трудность. Под сложностью задачи будем понимать объективную характеристику, которая определяется объемом предметных знаний, достаточных для ее решения. В свою очередь под трудностью задачи будем понимать субъективную характеристику, которая может быть получена путем сопоставления знаний, достаточных для решения задачи, со знаниями, имеющимися у обучающегося.

Следует заметить, что трудность задачи не может быть больше сложности: в процессе обучения сложность задачи является достижимым пределом трудности. Количественные оценки сложности и трудности задачи непосредственно связаны с процессами извлечения и сравнения знаний.

3. Знания и данные

В самом общем виде знание определяется как проверенный на практике результат отражения объективной действительности, представленный в сознании субъекта в виде понятий и суждений, утвержденных некоторой последовательностью умозаключений [10, ст. «Знание»]. С гносеологической точки зрения знание определяется как субъективно обоснованное убеждение [7, с. 12]. Тем самым признается субъективная (внутренняя) природа знания. В объективном смысле знание рассматривается как представленный во внешней форме результат субъективного познания, признаваемый объективно истинным в некоторый исторический момент [25, ст. «Знание»]. Считается, что идеальность знаний является адекватным следствием тех свойств внешнего мира, которую они отражают.

Знания формируются в результате целенаправленного педагогического процесса, самообразования и жизненного опыта. Отсюда, в частности, следует, что знания нуждаются в своей объективации, т.е. отчуждении от носителя знаний в некоторой внешней объективной форме. Так как во внутренней (идеаль-

ной) форме передачу знаний осуществить нельзя, то используются внешние формы в виде специальным образом обустроенных данных. Историческим примером такого обустройства является естественный язык (язык-речь и язык-письмо).

Для передачи знаний используется синтаксически и семантически разомкнутые формы представления, предполагающие существование некоторых подразумеваемых знаний как о предметной области, относительно которой эти знания выражаются, так и знаний о форме выражения передаваемых знаний. Только согласовав эти фоновые (подразумеваемые) знания и знание структуры и правил интерпретации форм представления знаний, появляется возможность адекватной передачи и усвоения новых знаний.

Для представления и передачи знаний используются данные, которые рассматриваются как последовательность состояний (временная или пространственная) некоторого материального объекта (процесса, явления). Данные воспринимаются субъектом как содержащие некоторую информацию. В зависимости от эмоционального состояния субъекта, особенностей его восприятия и имеющихся у него фоновых знаний одни и те же данные воспринимаются как содержащие разную информацию. Следовательно, информацию можно определить как результат интерпретации данных, осуществляемой при определенном, возможно неполном знании их структуры и правил интерпретации.

С прагматической точки зрения знания представляются как «данные, полученные в нужном месте и в нужное время для решения практической задачи» [12, с. 50]. С этой точки зрения знания и данные не отличаются по своей структуре и способу хранения. Данные становятся знаниями тогда и только тогда, когда они соответствующим образом проинтерпретированы машиной или человеком.

Иная точка зрения на знания используется в «искусственном интеллекте» [19, с. 224], где знания отличаются от данных своей структурой. Данные представляют собой знания, если они:

– упорядочены в соответствии с концептуальной моделью предметной области (предметной онтологией);

– представлены в одной из форм представления знаний (семантическими сетями, фреймами, сценариями, продукциями и др.);

– имеют процедуру получения новых (скрытых) знаний из имеющихся (задана эффективная процедура вывода на знаниях);

– хранятся таким образом, который обеспечивает высокую эффективность типовых операций над знаниями (поиск на графах, анализ иерархий, логический вывод и др.).

В итоге получаем, что данные должны храниться, давать возможность поиска, проверяться, поддерживаться и обновляться. Знания, в свою очередь, хранятся как данные, но в отличие от данных должны преобразовываться из одной формы представления в другую и иметь процедуру вывода знаний. Однако основная трудность при работе со знаниями заключается не в представлении и оперировании знаниями, а в их первичном извлечении.

4. Извлечение знаний

Рассмотрим известную классификацию методов извлечения знаний [9, с. 100]. Коммуникативные методы извлечения знаний (рис. 1) предполагают непосредственное взаимодействие людей, а именно взаимодействие инженера по знаниям – приемника знаний и эксперта в предметной области – источника знаний.

Коммуникативные методы разделяются на пассивные и активные: в пассивных методах ведущую роль играет эксперт, а в активных – инженер по знаниям. Полная автоматизация коммуникативных методов извлечения знаний в настоящее время проблематична по причине ключевой роли в этом процессе субъектов-носителей и субъектов – получателей знаний.

Текстологические методы подразумевают под источником знаний различного рода документы (методики, пособия, руководства, монографии, учебники и др.). Выделяются три класса текстологических методов, направленных на анализ специальной литературы, на анализ учебников и на анализ документов. Эти методы отличаются по объему фоновых знаний, которые требуются для извлечения знаний из анализируемых текстов.

Рис. 1. Классификация методов извлечения знаний

Наиболее простым методом является метода анализа учебников. Анализ документов, в отличие от анализа учебников, усложнен тем, что знания в них сильно сжаты: документы содержат мало рассуждений, пояснений и комментариев. В свою очередь специальная литература требует использования большого объема фоновых знаний и поэтому является самой трудной для текстологического извлечения знаний.

В отличие от коммуникативных текстологические методы позволяют автоматизировать процесс извлечения знаний. Причина тому – наличие у текстов некоторой единой формальной структуры. Однако группа текстологических методов на сегодняшний момент оказалась наименее изученной. Трудности, с которыми столкнулись исследователи, состоят в том, что до сих пор не решена задача семантического анализа текста, заключающаяся в извлечении смысла, содержащегося в тексте, и его преобразование в одну из известных форм представления знаний [5, 27].

Восприятие любого текста осуществляется на четырех уровнях понимания [17]:

– понимание контекстуальных значений слов и предложений (физическое восприятие текста и понимание прямого, «поверхностного» значения слов и предложений);

– понимание переносных и символических смыслов (соотнесение с контекстом, понимание «глубинных» значений слов и предложений, соотнесение с фоновым знанием, осознание смысла текста, его концепта);

– понимание характеров и настроений персонажей (интеллектуально-эмоциональное восприятие текста);

– понимание авторского отношения к излагаемому материалу.

Первый уровень – языковый, требует знание языка. Путем «соединения» смысла слов на этом уровне происходит понимание смысла предложения как семантической единицы текста.

Второй уровень – смысловой, или уровень смыслового понимания всего текста. На этом уровне происходит анализ связей между смысловыми единицами текста – предложениями.

Третий уровень – личностный. На этом уровне читатель соотносит смысл прочитанного текста со своим жизненным опытом и знаниями. Путем размышлений читатель пытается понять авторский замысел.

Четвертый уровень – рефлексивный. На этом уровне читатель строит образ автора и соотносит свое отношение к прочитанному тексту с отношением к этому тексту самого автора.

Очевидно, текстологические методы работают на языковом уровне и не претендуют на более высокие уровни «понимания» текста. Однако для многих прикладных областей, в том числе и для оценки сложности и трудности учебных задач, достаточным оказывается не полное или частичное извлечение знаний из текстов, а их качественная идентификация.

5. Идентификация знаний

Известно, что любой связный текст порождает некоторое семантическое пространство. «Семантическое пространство слов – это область существования и функционирования их лексических значений. Значения же слов существуют не изолированно, а находятся в определенных отношениях и связях друг

с другом, благодаря чему оказываются связанными и слова как языковые единицы» [31].

Семантические связи в тексте можно разделить на явные и неявные. Явные связи – это связи, непосредственно заданные в тексте и устанавливаемые на этапе его синтаксического анализа. К неявным связям относятся связи между словами, определяемые фоновым знанием и свойственные любому естественному языку [27].

Однако большая часть семантических связей выявляется на этапе синтаксического анализа, так как при написании текстов стараются минимизировать необходимый объем фоновых знаний. В противном случае тексты с большим числом неявных связей трудны для чтения и понимания.

Лексическим значением слова является выражаемое им понятие. «Понятие – мысль, которая выделяет из некоторой предметной области и собирает в класс (обобщает) объекты посредством указания на их общий и отличительный признак» [24].

Следует заметить, что отношения между понятиями, в отличие от отношений между словами, имеют некоторые специфические особенности: понятие может выражаться не одним, а множеством слов, два понятия могут быть связанны отношениями, различающимися в различных областях знаний, понятие может иметь отношения к самому себе. Также немаловажно то, что понятия, в отличие от понятий-слов, использованных в некотором тексте, не является завершенным результатом познания: понятия находятся в постоянном содержательном уточнении и изменении [16].

В качестве основного положения описываемого подхода примем предположение о том, что идентификация знаний, содержащихся в тексте, может быть осуществлена путем анализа последовательности его слов. В этом случае прикладные понятия, имеющие различное выражение в тексте, рассматриваются как разные понятия.

Тогда произвольный текст может быть представлен как последовательность понятий-слов, связанных между собой явными и неявными отношениями, а знания, содержащиеся в тексте, могут быть идентифицированы явными отношениями между его

словами с точностью до неявных отношений, порождаемых фоновым знанием. Например, такие абстрактные понятия как «Прямая» и «Точка» имеют явные отношения, выражаемые следующими предложениями: «точка лежит на прямой», «прямая проходит через точку», «точка принадлежит прямой».

В отличие от известных подходов, при которых для идентификации знаний отношения между понятиями предметной области пытаются выявить на основе семантического анализа текста с участием экспертов, идентификацию знаний, содержащихся в тексте, осуществим установлением отношений между его словами по результатам синтаксического анализа текста.

Таким образом, под идентификацией знаний будем понимать построение семантической сети текста на основе его синтаксического анализа. Основное положение такого подхода состоит в том, что синтаксическая сеть текста, построенная по результатам синтаксического анализа, является подсетью его семантической сети. В этом случае предполагается, что взаимосвязи понятий, полученные в результате синтаксического анализа текста в обязательном порядке должны допускаться и его семантической сетью.

6. Синтаксис предложений

Простое предложение русского языка имеет предикативную структуру и может быть представлено грамматическим предикатом, аргументами которого являются грамматический субъект и грамматический объект.

Обычно подлежащее выражает грамматический объект, сказуемое – грамматический предикат, а дополнение – грамматический субъект. В свою очередь сложное предложение состоит из простых и имеет в своем составе две или несколько предикативных единиц, образующих в смысловом, конструктивном и интонационном отношении единое целое [8].

Обычно простое предложение русского языка представляют в виде дерева, где каждая дуга идет от главного слова к зависимому слову и имеет имя синтаксического отношения [22, 38]. Для упрощения представления предложений сильно связанные слова объ-

единяют в синтаксические группы. В этом случае один из членов группы (слово) всегда выступает в роли представителя группы и подчиняет остальные ее члены. Процесс построения такого дерева называют синтаксическим анализом.

Процесс выделения в предложении синтаксических групп слов будем называть синтаксическим разбором предложения, а синтаксический анализ предложения определим как сопоставление содержащихся в нем групп слов предикативной структуре.

На рис. 2 представлен пример синтаксического разбора и анализа предложения «Перпендикулярные прямые образуют прямой угол».

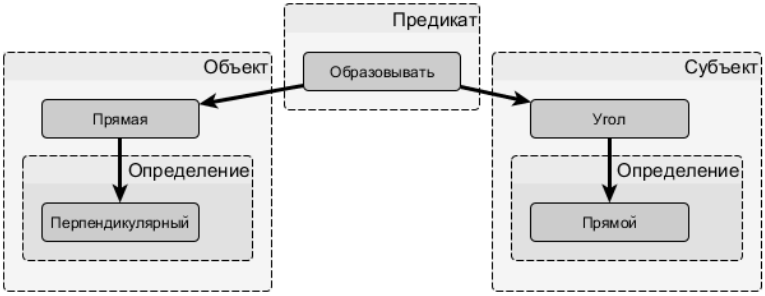

Рис. 2. Пример синтаксического разбора и анализа

В вершине дерева, полученного после синтаксического анализа простого предложения, всегда стоит грамматический предикат. Группа грамматического предиката является главной в предложении и подчиняет группы грамматического объекта и грамматического субъекта. В отличие от группы грамматического предиката, наличие групп грамматического объекта и грамматического субъекта является не обязательным.

Каждая из главных синтаксических групп слов может содержать в себе другие группы слов. Например, существуют группы определения, количественные, обстоятельственные, длительные, сочинительные, служебные и т.д. Как правило, деление на группы зависит от конкретной реализации синтаксического анализатора.

Будем предполагать, что грамматический объект и грамматический субъект предложения описывают понятия предметной области, а грамматический предикат – связь между ними. Под понятием будем понимать главное слово грамматической группы объекта или субъекта дополненное имеющимися определениями. Например, понятия «Треугольник», «Равнобедренный треугольник», «Равнобедренный прямоугольный треугольник» являются разными понятиями. Под предикатом будем понимать главное слово, входящее в группу предиката.

Иногда в тексте встречаются предложения, не имеющие объекта или субъекта. В таком случае для установления связи между объектом и субъектом вводится «пустое» понятие. Например, в предложении «Существует несколько типов фигур» присутствует объект, выраженный понятием «Типы фигур», но отсутствует субъект. Поэтому, понятие «Типы фигур» связывается с «пустым» понятием отношением «Существовать»

Таким образом, с помощью синтаксического анализа простого предложения можно выделить одно суждение, которое сообщает о взаимосвязанных понятиях и о характере их взаимосвязи.

Однако для идентификации знаний, содержащихся в тексте, требуется учет всех суждений. Объединение суждений осуществим на основе представления текста в виде семантической сети.

7. Семантическая сеть

В результате синтаксического анализа предложения может быть получена его семантическая сеть (мультиграф), состоящая из узлов (вершин), помеченных именами грамматических субъектов и объектов, и дуг, задающих отношения между ними и помеченных именами грамматических предикатов. Так как текст состоит из множества предложений, то объединим семантические сети этих предложений в единую семантическую сеть.

Пусть заданы две семантические сети: сеть текста S и сеть простого предложения S'. Сеть текста S зададим в виде упорядоченного множества из трех элементов:

(1) $S = (N, E, P)$,

где $N = \{n_i \mid i = 1, \ldots, Q\}$ – множество узлов сети с числом элементов Q; $E = \{(n_i, n_j, p_k) \mid n_i, n_j \in N; p_k \in P\}$ – множество ее дуг; P – множество двуместных предикатов. Дуги заданы упорядоченными множествами из трех элементов $(n_i, n_j, p_k) \in N \times N \times P$, где $n_i \in N$ – начальный узел, $n_j \in N$ – конечный узел, $p_k \in P$ – имя дуги, а × – знак операции декартового произведения множеств.

В свою очередь сеть предложения S' простая и состоит из двух узлов n_1, n_2 и одной дуги, помеченной именем некоторого предиката p:

$$S' = (N', E', P'),\ N' = \{n_1, n_2\},\ E' = \{(n_1, n_2, p)\}.$$

Тогда объединением сетей $S = (N, E, P)$ и $S' = (N', E', P')$ будет сеть $S'' = S \cup S'$ такая, что $S'' = (N'', E'', P'')$ и $N'' = N \cup N'$, $E'' = E \cup E'$, $P'' = P \cup P'$.

Аналогично можно определить пересечение сетей. Пересечением сетей $S' = (N', E', P')$ и $S'' = (N'', E'', P'')$ будет сеть $S = S' \cap S''$ такая, что $S = (N, E, P)$ и $N = N' \cap N''$, $E = E' \cap E''$, $P = P' \cap P''$.

В свою очередь разностью сетей $S' = (N', E', P')$ и $S'' = (N'', E'', P'')$ называется сеть $S = S' \setminus S''$ такая, что $S = (N, E, P)$ и $N = N' \setminus N''$, $E = E' \setminus E''$, $P = P' \setminus P''$.

Для упрощения семантические сети будем изображать в виде взвешенного ориентированного мультиграфа. В этом случае вместо имен грамматических предикатов будем указывать кратности дуг, равные числу различных предикатов, связывающих соответствующие узлы. Так как исследуются не статистические свойства текста, а знания, этим текстом выражаемые, то при задании кратности дуг не будем учитывать частоту повторения одного и того же предиката. В итоге имеем, что кратность дуги между двумя узлами сети равна числу различных грамматических предикатов, связывающих соответствующие слова-понятия.

На рис. 3 показан фрагмент семантической сети, полученный в результате синтаксического анализа параграфа «Основ-

ные свойства геометрических фигур» учебника по геометрии [26]. Семантическая сеть содержит 90 понятий анализируемой предметной области и 215 отношений, заданных на них.

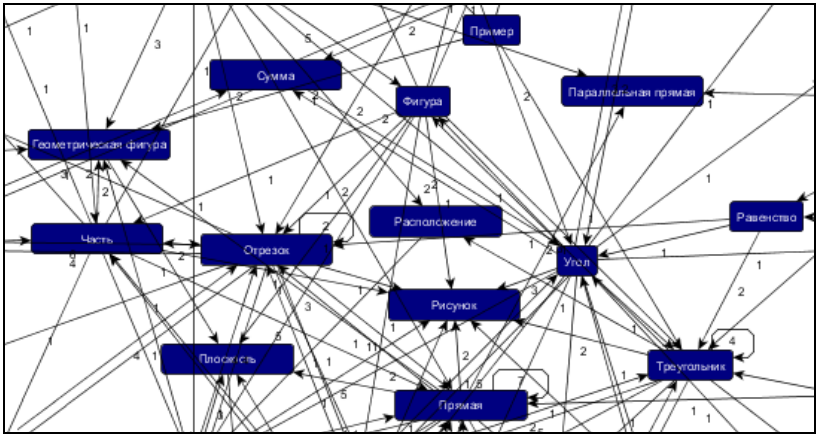

Рис. 3. Фрагмент семантической сети «Планиметрия»

Сеть была получена в результате выполнения следующих действий:

– поиска понятий, выраженных однословными и многословными терминами [6];

– установления синтаксических отношений между выявленными понятиями;

– разыменования найденных референтных связей.

Построенная семантическая сеть оказалась связным графом, у которого любой узел имеет как минимум один путь, связывающий его с любым другим узлом.

8. Семантическое расстояние

Два текста выражают близкие знания, если пересечение их семантических сетей соизмеримо с их объединением. В то же время два текста выражают разные знания, если пересечение семантических сетей мало по сравнению с их объединением.

Для количественной оценки близости текстологических знаний необходимо уметь определять расстояние между семантическими сетями и узлами одной семантической сети.

Известны два основных подхода для измерения семантических расстояний между словами: парадигматический и синтагматический [21]. Парадигматический подход базируется на измерении семантических расстояний в лексиконе языка, где лексикон определяется как набор классов слов, связанных парадигматическими связями, а под парадигматической связью понимается отношение между словами, имеющих общностью значений. Парадигматические отношения определяют онтологические свойства языка в целом и отражают те связи, которые существуют между выражаемыми ими явлениями действительности безотносительно какой-либо предметной области.

Синтагматический подход основан на измерении семантических расстояний между словами и текстами на основе статистических методов. При этом выделяются две группы методов: статистические и логико-статистические. К статистическим методам относятся методы, основанные на модели «ранг–частота» и описываемые законами Ципфа, Мандельброта и Бредфорда. Эти законы выражают динамическую зависимость частоты слова в тексте от его ранга, где ранг слова – порядковый номер слова в частотном словаре языка.

К логико-статистическим методам относятся дистрибутивно-статистический метод, метод гипертекстовой ссылки, частотно-семантический метод и метод компонентного анализа. Дистрибутивно-статистический метод позволяет получать количественную характеристику связанности слов в тексте на основе вычисления некоторой заранее заданной формулы, зависящей от статистических характеристик слов в этом тексте. Метод гипертекстовой ссылки устанавливает связь между понятиями-словами на основе общих слов в определении этих понятий, а частотно-семантический метод дополнительно учитывает и их частоту. В свою очередь метод компонентного анализа связывает понятия-слова путем разложения их значений на минимальные семантические составляющие – семы. Компонентный анализ основан на гипотезе о том, что значение всякой языковой

единицы состоит из сем и словарный состав языка может быть описан с помощью ограниченного их числа.

Первая мера семантической близости понятий-концептов была создана для оценки близости узлов в таксономиях [2], являющихся частным случаем семантической сети. Мера близости *sim* двух узлов n_1 и n_2 в таксономии представляет собой обратное значение длины кратчайшего пути между этими узлами $d(n_1, n_2)$:

$$sim(n_1, n_2) = \frac{1}{d(n_1, n_2)}.$$

В [35] предложена нормированная мера близости, которая определяется через логарифм отношения кратчайшего пути в таксономии между узлами n_1 и n_2 к ее удвоенному диаметру:

$$sim(n_1, n_2) = -\log\frac{d(n_1, n_2)}{2D},$$

где D – диаметр таксономии, или максимальное расстояние между ее узлами.

В другом типе мер используются семантические сети, построенные на базе определений понятий. В этом случае семантическая связь двух понятий-слов n_1 и n_2 прямо пропорциональна числу понятий-слов, входящих в определение первого и второго понятия:

$$sim(n_1, n_2) = |\, gloss(n_1) \cap gloss(n_2) \,|,$$

где $gloss(n)$ – множество понятий-слов в определении понятия-слова n.

В ряде случаев оказывается, что в пересечении определений сравниваемых понятий-слов может не оказаться ни одного общего понятия. Поэтому в [36] учитываются не только понятия-слова, которые учувствуют в определении каждого понятия, но и понятия-слова, которые связывают эти понятия в семантической сети текста.

Рассмотренные выше методы определения семантического расстояния между понятиями опираются на определенные типы отношений, характерные для таксономий. Выявление таких отношений в тексте является нетривиальной задачей, требую-

щей, в конечном итоге, привлечения экспертов предметной области [23].

Для рассматриваемого класса задач семантическое расстояние между понятиями определим иным образом. Пусть задана семантическая сеть S. Зафиксируем два произвольных ее узла n_i и n_j. Найдем $R(n_i, n_j)$ – множество путей без циклов (цепей) длины не более чем M, ведущих от узла n_i к узлу n_j. Тогда семантическое расстояние L между узлами n_i и n_j может быть вычислено по формуле:

$$(2) \quad L(n_i, n_j) = \sum_{r \in R(n_i, n_j)} \frac{\min(w_1^r, w_2^r, \ldots, w_{d(r)}^r)}{d(r)},$$

где $d(r)$– длина пути r, $d(r) \le M$; M – глубина связи; w_i^r – вес дуги i пути r, $i = 1, 2, \ldots, d(r)$; min – функция, возвращающая минимальное значение ее аргументов.

Из формулы (2) следует, что два узла отдалены друг от друга, если между ними имеется много путей (понятия слабо связаны). Отдаленность двух узлов тем больше, чем больше веса соединяющих их дуг (более вариативными являются связи между понятиями). Однако если в пути встречается дуга с небольшим весом, то этот путь вносит меньший вклад в удаленность узлов друг от друга. Но не все пути учитываются при подсчете расстояния между узлами: исключаются те пути, длина которых больше заданной глубины связи (трудно установить связь между понятиями, так как это требует использования большого числа предложений).

Следует обратить внимание на то, что семантическое расстояние (2) не является метрикой, так как для него не выполняются аксиомы тождества, симметрии и неравенства треугольника. Это связано с тем, что понятие может иметь отношение к самому себе, связи между двумя понятиями по определению не симметричные и близость двух понятий зависит не только от непосредственно связывающих их предикатов, но и от предикатов, задающих косвенную связь через другие понятия.

Два понятия, соединенные длинным путем, признаются слабо связанными. Поэтому при расчете семантических расстояний задается глубина учитываемых связей M. Для многих

практических применений M можно выбирать из диапазона от двух до семи. Стандартная интерпретация M – число суждений, которыми одновременно может оперировать обучающийся, или максимальное число суждений, встречающихся в его умозаключениях.

Таким образом, единицей измерения семантических расстояний является грамматический предикат, который может использоваться для выражения одного суждения. Следует заметить, что в дидактических исследованиях суждение также признается основной единицей измерения объема знаний [20].

Если два узла сети не связаны ни одним путем, то вычисление семантического расстояния между ними дает величину, равную нулю. Нулевое семантическое расстояние обозначает отсутствие связи между соответствующими понятиями и утверждает их содержательную неразличимость. С другой стороны, чем больше величина семантического расстояния между узлами сети, тем более отдаленными являются соответствующие понятия по содержанию.

Показательно значение семантического расстояния $L(n, n)$ между одним и тем же узлом n. Если $L(n, n)$ равно нулю, то соответствующее понятие может быть признано простым. Если величина $L(n, n)$ большая, то соответствующее понятие является сложным и может быть признано как не раскрытое по содержанию.

9. Измерение знаний

Помимо семантического расстояния между узлами сети для оценки трудности и сложности учебных задач требуется вычисление семантического расстояния между семантическими сетями и определение объема знаний, в них содержащихся.

Измерение объемов знаний до сих пор осуществляется методами, основанными на экспертных оценках результатов учебной работы обучающихся (субъективные методы) и на тестировании обучающихся (объективные методы). Теоретический фундамент этих методов заложен в современной теории педагогических измерений [1], где процесс обучения рассматривается как постоянное преодоление обучающимся грани между до-

ступной областью знаний (уровнем актуального развития) и потенциально доступной (зоной ближайшего развития). Задача педагогов состоит в том, чтобы подобрать трудные, но посильные задания, способствующие выявлению уровня актуального развития [13].

Известен также подход, согласно которому измерение знаний осуществляется на основе измерения емкости понятий, где под емкостью понятия понимается число связей этого понятия с другими понятиями, а сама единичная связь выступает в качестве единицы измерения [15]. В этом случае измерение объема знаний в тексте, теме, учебной дисциплине сводится к выявлению понятий предметной области и подсчету числа связей между ними экспертными методами.

Для определения близости двух семантических сетей используется поиск гомоморфизмов, преобразующего одну сеть в другую. Однако нахождение гомоморфизма позволяет определить только качественную «похожесть» сетей и не позволяет измерить объемы знаний, содержащихся в этих сетях.

Очевидно, что перечисленные методы непригодны для определения объема знаний в семантической сети текста, полученной на основе синтаксического анализа, так как в одном случае требуется привлечение экспертов (экспертная оценка, тестирование, подсчет емкости понятий), а в другом – отсутствует эффективно вычисляемое расстояние между семантическими сетями (поиск гомоморфизмов).

Под объемом знаний, содержащихся в семантической сети $S = (N, E, P)$, будем понимать величину, вычисляемую по следующей формуле:

$$(3) \quad K(S) = \sum_{n_i, n_j \in N} L(n_i, n_j),$$

где $K(S)$ – объем знаний в семантической сети S, а $L(n_i, n_j)$ – семантическое расстояние между узлами n_i и n_j, вычисляемое по формуле (2).

Формула (3) утверждает, что объем знаний в сети S есть сумма семантических расстояний между всеми парами ее узлов.

Как и у семантического расстояния, единицей измерения объема знаний является грамматический предикат.

Теорема 1. *Объем знаний* (3) *является аддитивной мерой на множестве семантических сетей.*

Для доказательства теоремы 1 сначала покажем, что мера пустой сети равна нулю. Действительно, если семантическая сеть S пуста, $S = (\varnothing, \varnothing, \varnothing)$, то из формул (3) и (2) непосредственно следует $K(S) = 0$. Также из формул (3) и (2) следует утверждение о том, что мера объединения двух сетей S_1 и S_2 таких, что $S_1 \cap S_2 = (\varnothing, \varnothing, \varnothing)$, равна сумме их мер: $K(S_1 \cup S_2) = K(S_1) + K(S_2)$. ♦

Таким образом, в отличие от других известных подходов, формула (3) позволяет объективно измерить объем знаний, содержащийся в произвольном тексте.

Известны также несколько подходов к определению расстояний между графами. Это использование высоты ориентированного графа, которая равна наибольшей длине пути от корня к листу в ярусно-параллельной форме его представления [30]. Также используется расстояние, получаемое на основе вычисления диаметра графа – максимального числа ребер, связывающих две его вершины [11]. Известно также расстояние между графами, получаемое путем вычисления реберной плотности – числовой величины, характеризующей близость графа к полносвязному [14].

Очевидно, что перечисленные подходы непригодны для определения расстояний между семантическими сетями. Семантическое расстояние между сетями S_1 и S_2 определим как объем знаний, содержащийся в симметрической разности этих сетей:

(4) $D(S_1, S_2) = K(S_1 \setminus S_2 \cup S_2 \setminus S_1)$.

Теорема 2. *Семантическое расстояние* (4) *является метрикой на множестве семантических сетей.*

Для доказательства теоремы 2 достаточно показать, что на множестве семантических сетей удовлетворяются аксиомы тождества, симметрии и треугольника. Пусть S_1 и S_2 – семантические сети. Если $S_1 = S_2$, то из (4) следует $D(S_1, S_2) = 0$. Теперь пусть $D(S_1, S_2) = 0$. Тогда из (4) следует, что $S_1 = S_2$. В итоге $D(S_1, S_2) = 0$ тогда и только тогда, когда $S_1 = S_2$. Аксиома тождества доказана. Аксиома симметрии также непосредственно

следует из (4): $D(S_1, S_2) = D(S_2, S_1)$. В свою очередь аксиома треугольника следует из формулы (4) и теоремы 1:

$$K(S_1 \setminus S_2 \cup S_2 \setminus S_1) + K(S_2 \setminus S_3 \cup S_3 \setminus S_2) \geq K(S_1 \setminus S_3 \cup S_3 \setminus S_1). \blacklozenge$$

Таким образом, множество семантических сетей текстов является метрическим пространством, а семантическое расстояние между двумя сетями равно суммарному объему знаний, в них содержащихся.

10. Оценка знаний

Оценка знаний предполагает сопоставление имеющихся знаний у обучающегося с эталонными. В нашем случае эталонными знаниями являются знания о предметной области, а знания обучающегося определяются по решенным им задачам. При успешном решении учебных задач обучающийся показывает приобретенные им знания.

Рассмотрим предлагаемый в настоящей статье подход для оценки объемов знаний, а также трудности и сложности учебных задач. Пусть имеются следующие семантические сети:

– S – семантическая сеть предметной области, полученная путем синтаксического анализа текстов, описывающих эту предметную область;

– T – семантическая сеть учебной задачи, полученная путем синтаксического анализа ее текста;

– O – семантическая сеть обучающегося, полученная путем последовательного объединения семантических сетей задач, успешно решенных обучающимся.

Результат сопоставления сети предметной области S и сети текущей задачи T определяет сложность этой задачи (объективная характеристика задачи), а сопоставление сети текущей задачи T и сети обучающегося O – трудность задачи (субъективная характеристика задачи). В свою очередь сопоставление сети предметной области S и сети обучающегося O позволяет определить объем еще не усвоенных им знаний (рис. 4).

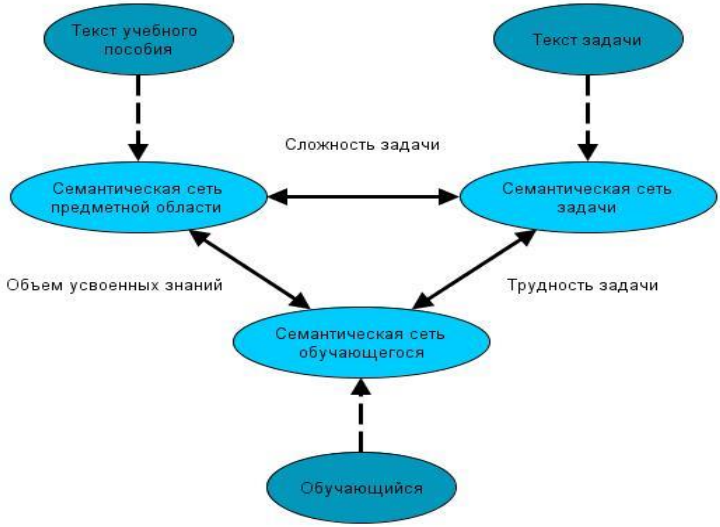

Рис. 4. Семантические характеристики

Учитывая специфику решаемой задачи, можно показать, что сети *S*, *T* и *O* согласованы, т.е. $T \subseteq S$, $O \subseteq S$, где отношение включения сетей $\subseteq$ определяется так: если $S' = (N', E', P')$, $S'' = (N'', E'', P'')$ и $S' \subseteq S''$, то $N' \subseteq N''$, $E' \subseteq E''$ и $P' \subseteq P''$.

В итоге имеем следующие формулы, позволяющие вычислить сложность и трудность учебной задачи *T*, а также объем еще не усвоенных знаний:

(5) $U(O) = K(S \setminus O)$, $u(O) = K(S \setminus O) / K(S)$;

(6) $C(T) = K(T \cap S)$, $c(T) = K(T \cap S) / K(S)$;

(7) $H(T) = K(T \setminus O)$, $h(T) = K(T \setminus O) / K(O)$;

(8) $\overline{H}(T) = K(T)$, $\overline{h}(T) = K(T) / K(O)$.

Объем не усвоенных знаний $U(O)$ и относительный объем не усвоенных знаний $u(O)$ вычисляются по формуле (5) на основе объемов знаний, содержащиеся в семантической сети предметной области *S* и семантической сети обучающегося *O*.

Сложность *C* учебной задачи *T* определяется по формуле (6) и равна объему знаний из семантической сети предметной

области *S*, содержащейся в семантической сети задачи *T*. Относительная сложность задачи *c* – это доля знаний предметной области, содержащихся в задаче *T*.

Трудность *H* учебной задачи *T* определяется по формуле (7) как объем новых знаний, содержащийся в семантической сети задачи *T* относительно семантической сети обучающегося *O*. В свою очередь относительная трудность задачи *h* – это отношение объема новых знаний в задаче *T* к знаниям, имеющимся у обучающегося.

Помимо трудности учебной задачи *T*, связанной с наличием в ней новых знаний, можно ввести в использование интегральные трудности учебной задачи $\overline{H}$ и $\overline{h}$ (8), учитывающие общий объем знаний в задаче. Интегральные трудности позволяет учесть число шагов решения (число суждений), которые выполнит обучающийся, чтобы проследить связи между всеми понятиями задачи.

Рассмотрим теперь построение описанных выше семантических сетей в процессе обучения. Как правило, курс обучения разбивается на части (разделы). После каждой части происходит проверка знаний, которая реализуется в виде решения соответствующих задач и их проверки (рис. 5).

Первый этап – построение семантической сети предметной области *S* по текущей части учебного материала. Второй этап – решение обучающимся контрольных задач и их проверка. При этом могут использоваться различные стратегии выбора контрольных задач, например, в порядке увеличения их трудностей или сложностей. После того как учебный материал усвоен обучающимся (объем не усвоенных знаний равен нулю или относительно мал), происходит построение семантической сети предметной области для новой части учебного материала. При этом сеть не строится заново, а достраивается.

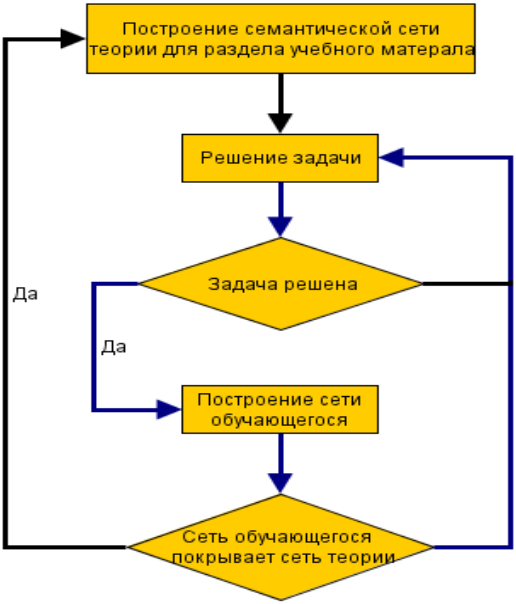

Рис. 5. Процесс обучения

11. Демонстрационный пример

Рассмотрим описанный выше метод оценки знаний на примере. Пусть имеется текст задачи: «На стороне AB треугольника ABC взята точка D. Чему равна сторона AB треугольника, если сторона AD равна 5 см, а сторона BD – 6 см?».

В данной задаче имеются три понятия и два отношения, одно из которых встречается дважды. На рис. 6 и 7 представлена сеть задачи, а также фрагменты сети обучающегося и сети предметной области. Фрагмент сети предметной области включает в себя все понятий сети задачи. Заметим, что из сети обучающегося следует, что обучающийся не знает, как соотносятся понятия «Сторона» и «Точка».

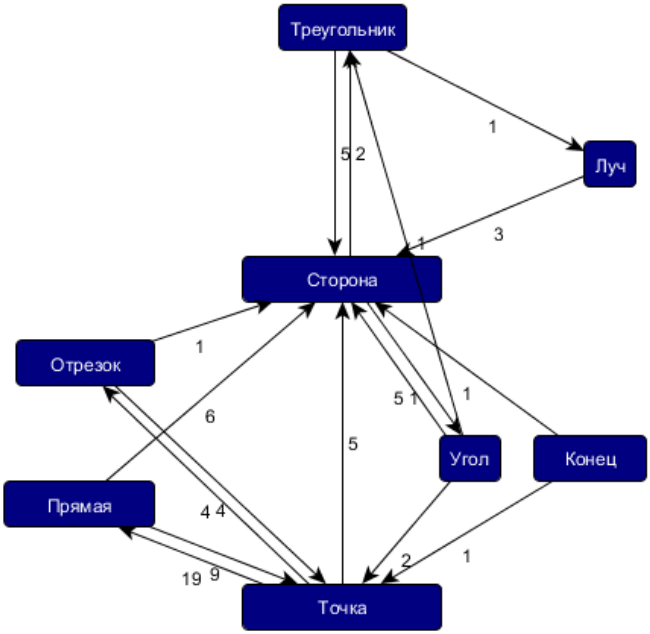

Рис. 6. Семантическая сеть предметной области S

Предварительно покажем, как производится вычисление семантических расстояний между понятиями. Для этого рассмотрим понятия «Точка» и «Сторона» в семантической сети предметной области *S*. Для упрощения вычислений зададим глубину прослеживаемых связей *M*, равную двум. Запишем множество путей *R*(Точка, Сторона), соединяющих выбранные узлы: {(Точка, Сторона), (Точка, Прямая, Сторона), (Точка, Отрезок, Сторона)}. Тогда по формуле (2) находим

$$L\,(\text{Точка, Сторона}) = \frac{\min(5)}{1} + \frac{\min(19,6)}{2} + \frac{\min(4,1)}{2} = 8{,}5.$$

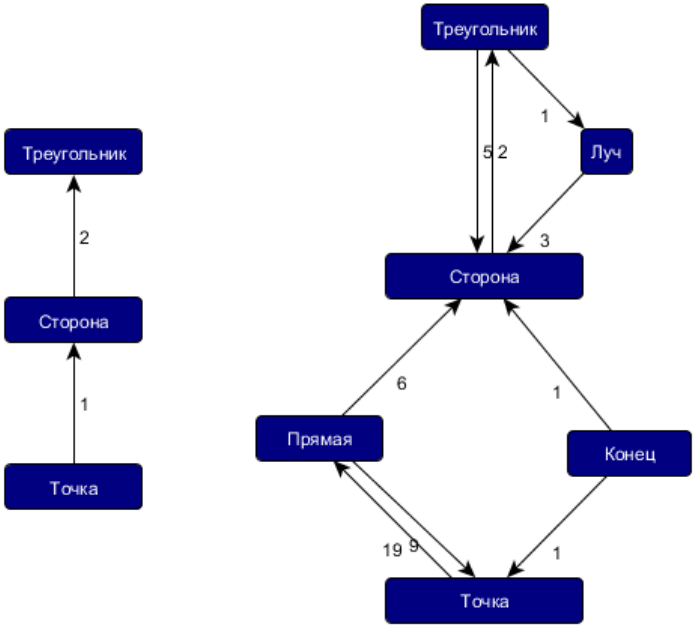

Рис. 7. Семантическая сеть задачи T и обучающегося O

Полученный результат может быть проинтерпретирован так: понятие «Точка» связано с понятием «Сторона» 8,5 различными предикатами. В свою очередь семантическое расстояние между понятием «Сторона» и «Точка» равно другому значению: *L* (Сторона, Точка) = 0,5.

Вычисление объемов знаний покажем на примере определения трудности *H*(*T*) учебной задачи *T*. Объем новых знаний, содержащийся в семантической сети задачи *T* относительно семантической сети обучающегося *O*, определяется семантической сетью *T**O*, которая в рассматриваемом примере состоит из узлов «Сторона» и «Точка» и дуги, их соединяющей. Объем знаний, содержащийся в этой сети, равен единице. Следовательно, трудность задачи *H*(*T*) также равна единице. Полученная величина *H*(*T*) определяет объем знаний, который получит обучающийся, правильно решив задачу *T*.

12. Вычислительный эксперимент

Для проведения вычислительного эксперимента в качестве синтаксического анализатора был выбран лингвистический процессор ЭТАП-3, разработанный в Институте проблем передачи информации им. А.А. Харкевича РАН [22]. По сравнению с другими анализаторами синтаксический анализатор ЭТАП-3 обладает рядом преимуществ: имеется общедоступный сервис для автоматического синтаксического анализа текстов, результаты анализа пересылаются в xml-формате (UNL), выполняется синтаксический анализ предложений без ограничения сложности.

С помощью ЭТАП-3 был произведен синтаксический анализ двух глав предметной области «Планиметрия»: «Основные свойства простейших геометрических фигур» и «Геометрические построения». Каждая глава разделена на три части: предметная область, вопросы и задачи. Например, в результате синтаксического анализа главы «Геометрические построения» найдено 242 понятий-слов и 817 отношений между ними.

Для проведения вычислительного эксперимента разработана программа (рис. 8), которая позволяет загружать файлы с результатами синтаксического анализа в формате UNL и формировать из них семантические сети. Для работы с семантическими сетями в программе реализованы следующие инструменты:

– добавление, удаление, объединение семантических сетей;

– перемещение семантической сети в разные группы (предметная область, решенные задачи, нерешенные задачи);

– просмотр статистических данных семантических сетей (частота слов и отношений, части речи);

– просмотр семантических сетей в графическом режиме.

На примере главы «Геометрические построения» смоделирован процесс решения и оценки задач. За базовую сеть обучающегося была взята сеть, составленная на основе семантических сетей вопросов, т.е. было предположено, что обучающийся ответил на все вопросы и ему предстоит решить 30 задач.

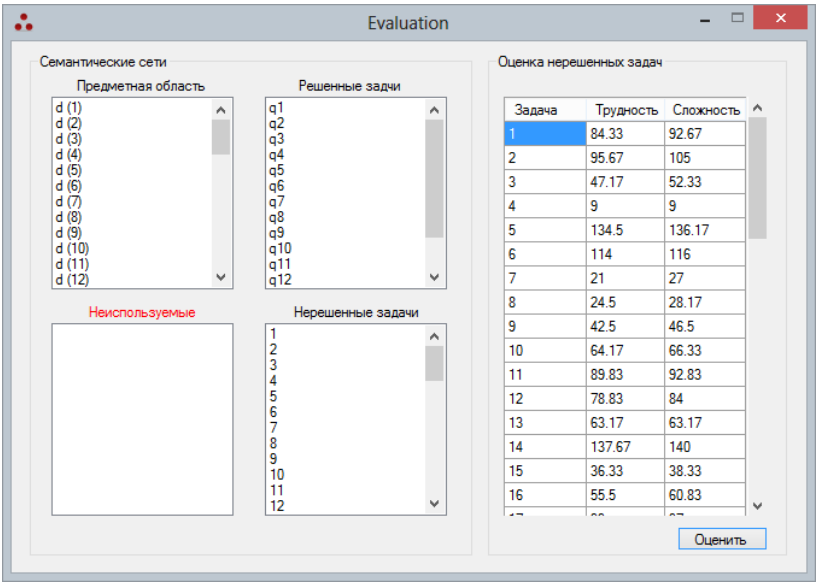

Рис. 8. Интерфейс программы

На рис. 9 показаны графики пошагового изменения трудности задач в процессе их решения. На каждом шаге для решения выбиралась задача с наименьшей трудностью. После выбора задачи ее семантическая сеть добавляется в семантическую сеть обучающегося, что вызывало уменьшение трудности остальных задач на следующих шагах обучения.

Из рис. 9 также видно, что решение задач с небольшой трудностью несущественно влияет на изменение трудностей остальных задач на следующих шагах обучения. Однако при выборе для решения более сложных задач наблюдается существенное уменьшение трудности еще не решенных задач.

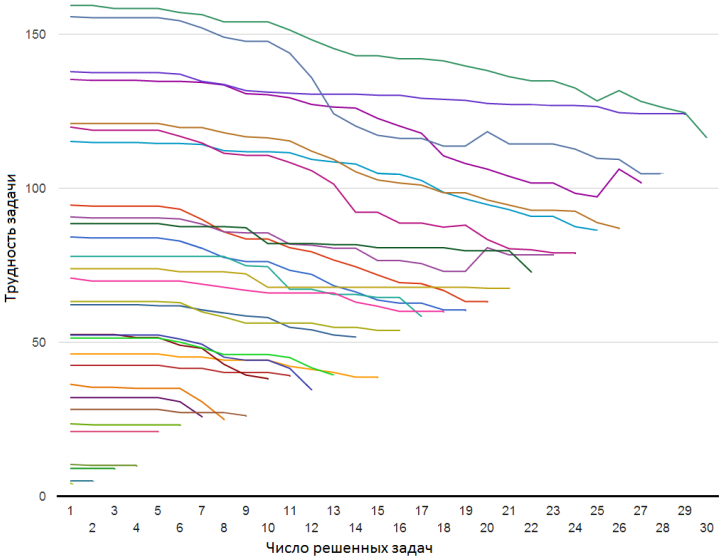

Рис. 9. Изменения трудности задач в процессе решения

13. Заключение

Оценка сложности и трудности учебных задач является одной из ключевых проблем в области автоматизации обучения. В данной работе было предложено решение этой задачи на основе синтаксического анализа текстов, направленного на выявление предикативной структуры составляющих его предложения, и построения по результатам этого анализа семантических сетей текстов.

Для оценки объемов знаний, содержащихся в семантических сетях, разработан математический аппарат, базирующийся на определения семантических расстояний между понятиями-словами, которые выявлены в процессе автоматического синтаксического анализа.

Показано, что объем знаний, содержащихся в семантической сети, является мерой на множестве семантических сетей, а

введенное расстояние между семантическими сетями превращает множество семантических сетей в метрическое пространство.

В отличие от других существующих методов определения объемов знаний, основанных на использовании онтологий и тезаурусов, разработанный метод отличается универсальностью применения, так как не привязан к конкретной предметной области и не требует привлечения экспертов для первичного ее описания. Основное условие применимости метода – предоставление описаний предметной области в виде множества текстов на естественном языке и наличие алгоритма, выявляющего предикативную структуру предложений.

Как и в теории информации Шеннона–Хартли [34, 37], в представленном методе измерения объемов знаний выполнено абстрагирование от психической природы изучаемых явлений и найден такой материальный объект, по характеристикам которого можно судить об интенсивности моделируемых психических процессов.

Так, в теории информации – это вероятность сообщения, или частота его предъявления испытуемому: чем более часто появляется некоторое сообщение, тем оно менее «неожиданно» для испытуемого и, как следствие этого, воспринимается им как содержащее меньший объем информации. Для учета особенностей восприятия сообщений объем информации в сообщении определен как отрицательный логарифм от его вероятности.

В разработанном методе измерения объемов знаний – это текст, предъявляемый испытуемому как множество предложений, имеющих предикативную структуру, и воспринимаемый им как содержащий некоторые знания: чем больше понятий и связывающих их предикатов содержится в тексте, тем больший объем знаний может быть воспринят из этого текста. Для учета активной природы знаний глубина прослеживаемых связей при подсчете объема знаний ограничена способностью испытуемого устанавливать мысленные связи между понятиями путем выполнения умозаключений с определенным числом исходных суждений в них.

Литература

1. АВАНЕСОВ В.С. *Знания как предмет педагогического измерения* // Педагогические измерения. – 2005. – №3. – С. 43–52.
2. АНИСИМОВ А.В., ЛИМАН К.С., МАРЧЕНКО А.А. *Методы вычисления мер семантической близости слов естественного языка.* – [Электронный ресурс]. – URL: http://lingvoworks.org.ua (дата обращения: 09.04.2013).
3. БАБАНСКИЙ Ю.К. *Оптимизация процесса обучения: Общедидактический аспект.* – М.: Педагогика, 1997. – 256 с. – С. 16–23.
4. БАЛЛ Г.А. *Теория учебных задач: психолого-педагогический аспект*. – М.: Педагогика. 1990. – 184 с.
5. БЕЛОНОГОВ Г.Г. *Компьютерная лингвистика и перспективные информационные технологии* // Русский мир. – 2004. – 248 с.
6. БРАСЛАВСКИЙ П.И., СОКОЛОВ Е.А. *Сравнение пяти методов извлечения терминов произвольной длины* // Труды международной конференции «Диалог 2008». – 2008. – С. 67–74.
7. ВАГИН В.Н., ГОЛОВИНА Е.Ю., ЗАГОРЯНСКАЯ А.А. ФОМИНА М.В. *Достоверный и правдоподобный вывод в интеллектуальных системах*. – М.: Физматлит, 2008. – 712 с.
8. ВАЛГИНА Н.С., РОЗЕНТАЛЬ Д.Э., ФОМИНА М.И. *Современный русский язык: Учебник* / Под ред. Н.С. Валгиной. – М.: Логос, 2002. – 528 с.
9. ГАВРИЛОВА Т.А., ХОРОШЕВСКИЙ В.Ф. *Базы знаний интеллектуальных систем.* – СПб.: Питер, 2000. – 384 с.
10. ДАНИЛЬЯН О.Г., ПАНОВА Н.И. *Современный словарь по общественным наукам.* – М.: Эксмо-Пресс, 2007. – 528 с.
11. ЕВСТИГНЕЕВ В.А. *Применение теории графов в программировании.* – М.: Наука, 1985. – 332 с.
12. ЕРМАКОВ А.Е. *Извлечение знаний из текста и их обработка: состояние и перспективы* // Информационные технологии. – 2009. – № 7. – С. 50–55.

13. ЕФРЕМОВА Н.Ф. *Тестовый контроль в образовании: Учебное пособие.* – М.: Университетская книги, 2007. – 540 с.
14. КАРПЕНКО А.П., СОКОЛОВ Н.К. *Меры сложности семантической сети в обучающей системе* // Вестник МГТУ им. Н.Э. Баумана, серия «Приборостроение». – 2009. – №1(74). – С. 50–66.
15. КАРПЕНКО М.П. *Проблема измерения знаний и образовательные технологии* // Журнал практического психолога. – 1997. – №4. – С. 74–79.
16. КРАВЦОВ Л.Г. Методологические проблемы психологического анализа мышления в понятиях // Мат. I Рос. конф. по когнитивной науке. Казань, Казанский гос. ун-т, 2004. [Электронный ресурс]. – URL: http://www.ksu.ru/ss/cogsci04/science/cogsci04/sod.php3 (дата обращения: 22.11.2013).
17. КРАСНЫХ В.В. *Основы психолингвистики и теории коммуникации.* – М.: Гносис, 2012. – 333 с.
18. ЛЕРНЕР И.Я. *Факторы сложности познавательных задач* // Новые исследования в педагогических науках. – 1970. – №1. – С. 86–91.
19. ЛЮГЕР Д.Ф. *Искусственный интеллект: стратегии и методы решения сложных проблем.* – М.: Вильямс, 2005. – 864 с.
20. МИКК Я.А. *Оптимизация сложности учебного текста*: *В помощь авторам и редакторам.* – М: Просвещение, 1981. – 33 с.
21. МИТРОФАНОВА О.А. *Семантические расстояния: проблемы и перспективы* // XXXIV Международная филологическая конференция: Вып. 21. Прикладная и математическая лингвистика. – СПб., 2005. – С. 59–63.
22. *Многоцелевой лингвистический процессор ЭТАП-3.* – [Электронный ресурс]. – URL: http://www.iitp.ru/ru/science/ (дата обращения: 09.04.2013).
23. НАЙХАНОВА Л.В. *Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования*. – Улан-Удэ: Изд-во БНЦ СО РАН, 2008. – 237 с.

24. *Новая философская энциклопедия*: В 4-х т. / Ин-т философии РАН; Нац. обществ.-науч. фонд; Предс. научно-ред. совета В.С. Степин. – М.: Мысль, 2010. – [Электронный ресурс]. – URL: http://iph.ras.ru/enc.htm (дата обращения: 09.04.2013).
25. *Новейший философский словарь*: 3-е изд., исправл. – Мн.: Книжный Дом, 2003. – 1280 с.
26. ПОГОРЕЛОВ А.В. *Геометрия: Учебник для 7–11 классов общеобразовательных учреждений.* – М.: Просвещение, 1995. – 383 с.
27. ПОПОВ Э.В. *Общение с ЭВМ на естественном языке*. – М.: Наука, 1982. – 360 с.
28. РОМАДИНА О.Г., РАКИТИНА Н.И. *Методика оценки сложности учебных задач по информатике* // Вопросы современной науки и практики. Университет им. В.И. Вернадского. – 2010. – №10–12(31). – С. 146–151.
29. РЫЖЕНКО Н.Г. *Графовое моделирование как средство определения сложности решений текстовых задач школьного курса математики* // Математика и информатика: Наука и образование. Вып. 1. – Омск: Изд-во ОмГПУ, 2001. – С. 99–103.
30. ФЕДОТОВ И.Е. *Некоторые приемы параллельного программирования: Учебное пособие.* – М.: Изд-во МГИРЭА, 2008. – 188 с.
31. ФИЛИППОВИЧ Ю.Н., ПРОХОРОВ А.В. *Семантика информационных технологий: опыты словарно-тезаурусного описания.* – М.: МГУП, 2002. – 368 с.
32. COLLINS-TOMPSON K., BENNETT P., WHITE R., CHICA S., SONTAG D. *Personalizing web search results by reading level* // Proc. 20th ACM international conference on Information and knowledge management. – New York, 2011. – P. 403–412.
33. DUBAY W. *The Principles of Readability* // Impact Information. – Costa Mesa, California, 2004. – 73 p.
34. HARTLEY R.V.L. *Transmission of Information* // Bell System Technical Journal. – July, 1928. – P. 535–563.

35. LEACOCK C., CHODOROW M. *Combining local context and WordNet similarity for word sense identification* // In: WordNet: An electronic lexical database / Ed. C. Fellbaum. – MIT Press, 1998. – P. 265–283.
36. PATWARDHAN S., BANERJEE S., PEDERSEN T. *Using measures of semantic relatedness for word sense disambiguation* // Proc. 4th International Conference on Intelligent Text Processing and Computational Linguistics. – 2003. – P. 241–257.
37. SHANNON C.E. *A Mathematical Theory of Communication* // Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423, 623–656.
38. *Syntactic and semantic parser based on ABBYY Compreno linguistic technologies* / Anisimovich K.V., Druzhkin K.Ju., Minlos F.R., Petrova M.A., Selegey V.P., Zuev K.A. // Компьютерная лингвистика и интеллектуальные технологии. – 2012. – С. 810–822. – [Электронный ресурс]. – URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf (дата обращения: 02.12.2013).

USING SYNTACTIC TEXT ANALYSIS TO ESTIMATE EDUCATIONAL TASKS' DIFFICULTY AND COMPLEXITY

Igor Naumov, Institute of Control Sciences of RAS, Moscow, Engineer (naigsa@gmail.com).
Valeriy Vykhovanets, Institute of Control Sciences of RAS, Moscow, Doctor of Science (Moscow, Profsoyuznaya st., 65, (495) 926-77-84).

Abstract: We suggest a routine for automatic assessment of complexity and difficulty of educational tasks. This routine is based on text parsing, phrases' predicative structures identification and semantic network construction. Then we develop a mathematical model which employs a notion on semantic distance between words to calculate the volume of knowledge in a semantic network. We show that the volume of knowledge in a semantic network is a measure in the space of all semantic networks, and the semantic distance makes this space the metric one.

Keywords: education process, measurements in teaching, the difficulty and complexity of tasks, parsing, semantic nets, semantic distance.

Статья представлена к публикации членом редакционной коллегии Д.А. Новиковым

Поступила в редакцию 16.06.2013.
Опубликована 31.01.2014.