ББК В187
УДК 519.76

Д. С. Ботов, Ю. Д. Кленин

ПОДХОД К ОПРЕДЕЛЕНИЮ МЕРЫ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

Предлагается подход к решению проблемы семантического анализа результатов обучения. Результаты обучения являются короткими текстами, состоящими из глаголов действия и терминов предметной области из различных областей человеческого знания. С учетом специфики представления результатов обучения рассмотрены существующие подходы к определению семантической близости, проанализированы различные способы использования внешних баз знаний и выбран ряд наиболее подходящих методов.

Введение

В условиях развития в России студентоориентированного подхода к разработке образовательных программ, опирающегося на формулирование ожидаемых результатов обучения (learning outcomes) в их гармоничной увязке с компетенциями новых образовательных стандартов (ФГОС ВО 3+) и профессиональных стандартов, все более важным становится создание средств интеллектуальной поддержки, позволяющих повысить эффективность разработки новых или актуализации существующих образовательных программ и индивидуальных образовательных траекторий [1].

При решении комплекса задач поиска и анализа разнородного образовательного контента необходимо определить подход к сопоставлению результатов обучения разного уровня (от отдельных тем, образовательных курсов до образовательных программ университетов) через определение меры семантической близости. Применение данного подхода на практике позволит эффективно в автоматизированном режиме подбирать, анализировать и синтезировать актуализированный образовательный контент исходя из заданного пользователем набора результатов обучения.

Понятие результата обучения и особенности описания

Результаты обучения — это формулировки того, что, как ожидается, будет знать, понимать и/или будет в состоянии продемонстрировать обучающийся после завершения процесса обучения [16]. Они состоят из:

– **action verb** — глагол, означающий действие, которое должен продемонстрировать обучающийся (например: знать, применять на практике, сопоставлять, анализировать, создавать и т. д.);

– **learning statement** — формулировка, состоящая из терминов предметной области, которая определяет, что из изученного будет продемонстрировано обучающимся в результате;

– **criterion** — критерии — дополнительные формулировки, определяющие условия в которых будет оцениваться выполняемое обучающими действие [17].

Текстовое представление каждого конкретного результата обучения имеет следующие особенности:

1) текст небольшой длины, состоящий всего из нескольких фраз, включающих в себя, как правило, не более двух десятков слов;

2) использование глаголов действия (Action Verb) из ограниченного списка слов, определяемых различными таксономиями образовательных целей [6; 7] вне зависимости от предметной области образовательной программы или курса;

3) использование терминов в Learning statement и Criterion из областей знания, определяемых предметной областью образовательной программы или курса. При этом используемые термины могут быть узкоспециализированы и редки в употреблении.

Данные особенности накладывают дополнительные требования на используемые алгоритмы и методы интеллектуального анализа текста (Text Mining). Поэтому следует отдельно рассматривать анализ семантики глагольной и терминологической составляющих результатов обучения с целью определения меры их семантической близости.

Анализ семантики терминологической составляющей результата обучения

Для моделирования семантики терминологической составляющей результатов обучения (Learning statement, Criterion) предлагается использовать векторную модель, в которой текст представляется в виде разреженного числового вектора в многомерном пространстве, измерениями которого являются какие-либо особенности текста. В большинстве случаев в качестве данных особенностей используются входящие в текст слова, поэтому данная модель имеет альтернативное название «Bag of Words» (дословно «мешок слов») (далее — модель BoW) [15]. Данная модель позволяет свести операции по обработке и анализу текста к операциям над векторами, значительно ускоряя обработку данных. Одной из наиболее часто используемых статистических мер для построения вектора является TF-IDF [2].

Существует несколько проблем применения модели BoW. Во-первых, использование BoW приводит к полному исключению синтаксической структуры текста из рассмотрения. Во-вторых, при решении конкретной задачи моделирования результатов обучения возникает другая проблема, которая заключа-

ется в небольших размерах моделируемых текстов при значительных количествах различных терминов определенной предметной области. Это означает, что смоделированные вектора будут крайне сильно разрежены и могут не иметь ни единого общего слова- термина, даже если они являются полными синонимами — так называемая проблема семантического разрыва (semantic gap).

В силу упомянутой разреженности текстовых векторов корпуса, анализ смоделированных данных в их чистом виде будет бесполезен [4]. Для решения данной проблемы используются методы внесения в короткие тексты семантических данных извне — «обогащения» (enrichment). Короткий текст результата обучения при семантическом моделировании можно обогатить как дополнительной информацией о терминах предметной области, так и другой информацией, содержащейся в анализируемых образовательных программах и курсах — например, тематика, предмет, учебная организация, преподаватель и т. д.

Обогащение коротких текстов можно осуществить путем использования внешних корпусов и онтологий [5], содержащих семантическую информацию о терминах предметной области.

Обзор возможности применения корпусов текстов и онтологий для обогащения семантики терминологической составляющей

Рассмотрим подробнее возможности применения различных источников для обогащения коротких формулировок.

WordNet — это лексическая база данных, содержащая слова различных тематик с их краткими определениями и отображающая отношения и связи с другими словами в базе [11]. WordNet позволяет использовать данные о взаимосвязях понятий между собой при определении близости объектов, однако непосредственное дополнение модели BoW новыми словами будет неэффективно вследствие сжатости и краткости даваемых определений. Кроме того, WordNet может быть сильно ограничен при использовании его для обогащения специализированных текстов, использующих узкоспециализированную, профессиональную терминологию.

Многие текстовые корпуса, такие как Oxford English Corpus или American National Corpus, содержат большие массивы текстов на различную тематику, включая данные, полученные из web, используемые для исследовательских целей, преимущественно в области лингвистики [12; 13]. Основным недостатком данных систем является принцип получаемой информации. Поскольку данные системы ищут употребления слов и словосочетаний в текстах, не являющихся непосредственными определениями или словарными статьями для конкретного термина, то полученные результаты могут оказаться вырванными из контекста или вносить в обогащаемый результат обучения шумовую информацию, слабо отражающую сущность самого термина.

Словарные корпуса представляют собой собрания определений слов по типу обычных словарей. В частности, Oxford Dictionaries базируется на упомянутом выше Oxford English Corpus и предоставляет пользователям широкий функционал по поиску синонимов, переводу слов, поиску примеров употребления, проверки грамотности и т. п. [12]. Словарные корпуса в задачах обогащения коротких текстов страдают от той же проблемы, что и WordNet, — предоставляемые ими определения слишком коротки, однако, в отличие от WordNet, онтологическая иерархия модели полностью или частично утеряна.

Google Ngram Viewer — это система поиска слов и словосочетаний в текстах, собранных сервисом Google Books. По своей сути является очень большим корпусом, оснащенным поисковыми мощностями систем Google. Данные системы имеют общий недостаток с другими текстовыми корпусами — тексты, в которых производится поиск, могут быть недостаточно информативны по отношению к конкретному термину или привносить шумовую информацию.

В работе Е. А. Черниковой [8] в рамках предложенного ею подхода к сравнению образовательных курсов и программ используется WordNet в качестве внешнего источника семантической информации, в основе подхода лежит собственная онтология образовательных курсов, включающая в себя в том числе онтологические концепты результатов обучения. Также собственные онтологии образовательных программ и курсов используются, например, в работах О. Н. Сметаниной [19], С. В. Тархова [18], А. Ю. Ужвы [20] и др. Однако использование собственной онтологии требует решения крайне трудоемкой задачи — разработки средств пополнения и постоянной актуализации базы знаний с учетом непрекращающегося роста количества и качества образовательного контента в среде Интернет.

В отличие от описанных выше работ, учитывая недостатки создания и наполнения собственной онтологии и использования WordNet или иных словарных корпусов в качестве внешних источников, в данном исследовании авторами предлагается использовать открытую базу знаний интернет-энциклопедии Wikipedia в качестве внешнего корпуса текстов для обогащения семантики терминологической составляющей результата обучения.

Wikipedia является бесплатным многопользовательским ресурсом, с большим числом статей, фокусирующихся на конкретных темах. Актуальность статей Wikipedia поддерживается сообществом пользователей. Основным недостатком является свободное редактирование данных Wikipedia пользователями энциклопедии, однако другие пользователи принимают активное участие в контроле за вносимыми изменениями, оперативно нивелируя наносимый статьям вред. В остальном Wikipedia объединяет

в себе лучшие стороны других внешних источников знаний, рассмотренных выше.

Наиболее очевидный способ обогащения коротких текстов — прямое обогащение их векторов модели BoW словами, полученными из связанных с ними более крупных текстов [3]. Поскольку тексты статей Wikipedia достаточно обширны, они позволят в значительной мере обогатить короткий текст дополнительной информацией с помощью данного метода. Иерархичность энциклопедии сродни онтологическим схемам систем типа WordNet с большим количеством дополнительной ссылочной информации, например, ссылки на внешние ресурсы и перекрестные ссылки между статьями.

Институтом системного программирования РАН разрабатывается система Texterra [21] с инструментарием Text Mining, в котором семантика статей Wikipedia используется для улучшения и повышения эффективности поиска и навигации в текстовых базах данных. Существуют также и проекты, совмещающие данные WordNet и Wikipedia для получения расширенных онтологий.

Методы расчета семантической близости с использованием Wikipedia

Рассмотрим две разновидности методов расчета семантической близости концептов Wikipedia — контентные (текстовые) и ссылочные (сетевые) [22; 23].

Контентные (текстовые) методы определяют меры семантической близости, основываясь на текстовом содержимом статей Wikipedia. В данной категории методов термины (концепты) и тексты соответствующих статей, как правило, представляются в виде векторов в пространстве терминов.

В простейшем случае может быть определена косинусная мера близости векторов статей, соответствующих концептам, на основе широко используемой статистической метрики TF-IDF (term frequency — inverse document frequency), предложенной Джонсом [2]. Применение TF-IDF позволяет определить значимость слова в векторе для определения общего смысла текста, основываясь на частоте употребления данного слова в конкретном тексте и во всем наборе текстов.

Для повышения качества анализа текстов используются более сложные методы семантического анализа: латентный (LSA) и явный (ESA).

Латентный семантический анализ (LSA), представленный в работе [25], применяется в задачах поиска, классификации и фильтрации информации. Основная идея заключается в том, что совокупность всех контекстов, в которых встречается искомое слово, определяет множество ограничений, которые позволяют определить семантическую близость слов и множеств слов между собой. Основные недостатки заключаются в смешении понятий семантического подобия и семантической связности, и требуется большое количество документов достаточного объема, большой размер характеристических векторов.

Явный семантический анализ (Explicit Semantic Analysis, ESA) [14] представляет собой алгоритм, обучающийся на предложенной ему базе знаний статей Wikipedia, а затем применяющий полученную информацию для определения близости подаваемых ему на вход слов и текстов. В процессе обучения ESA составляет инвертированный индекс, в котором каждому слову соответствует набор концептов, к которым оно относится, и их весов, что позволяет сразу избавиться от всех малозначащих связей между словами и концептами, отбрасывая те, вес которых ниже заданного порога.

Обучившись на базе корпуса Wikipedia, ESA может производить поиск концептов в текстах. Принимая входной текст как множество входящих в него слов $T = \{w_i\}$ и сопоставляя каждому слову его TF-IDF вес v_i, алгоритм ESA составляет вектор V семантической интерпретации текста T. Для этого ESA отображает слова текста в пространство концептов Wikipedia C, используя упомянутые выше инвертированные индексы:

$$V = \left\langle \sum_{w_i \in T} v_i \cdot k_j \right\rangle,$$

где k_j является весом концепта c_j в инвертированном индексе слова w_i.

Таким образом, на выходе получаются вектора, представляющие текст как взвешенный набор концептов, а не BoW. Затем семантическая близость рассчитывается как косинусное расстояние для полученных векторов V.

Ссылочные (сетевые) методы основаны на представлении концептов вершинами в графе и учитывают не только частотные характеристики слов, но и характеристики путей между понятиями, их положение в иерархии, в частности ближайшие общие понятия. Ссылки могут быть разных видов: обычные внутритекстовые ссылки, ссылки на основные статьи, категорийные ссылки, ссылки из списка тематически связанных статей и т. д. [22].

Существует ряд локальных ссылочных метрик, использующих структурную информацию внешних баз знаний (применимы не только к Wikipedia). Перечислим некоторые из них: кратчайший путь, вычисляемый как величина, обратная длине кратчайшего пути в графе; мера близости Резника (Resnik), основывающаяся на общей для двух понятий информации, содержащейся в их общих иерархических предках; меры близости Лин (Lin) и Джианг — Конрата (Jiang — Conrath), учитывающие влияние и сходств, и различий понятий на семантическую дистанцию между ними; мера Ликок — Чодороу (Leacock — Chodorow), нормализующая кратчайший путь относительно размеров графа; мера Ву — Палмер (Wu — Palmer), использующая глубины понятий и их

ближайшего общего предка; мера Seco et al., использующая число гипонимов для определения объёма информационного контента в понятии.

Меры на базе Wikipedia

Существует ряд подходов к определению мер близости понятий, использующих дополнительную информацию, содержащуюся в Wikipedia:

1) Wikipedia Link-based Measure [9] использует при определении близости двух понятий число гиперссылок на страницах энциклопедии, ведущих на страницы данных понятий:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))},$$

где множества *A, B* и *W* соответствуют множеству страниц, содержащих ссылку на страницу понятия *a*, множеству страниц, содержащих ссылку на понятие *b*, и множеству всех страниц Wikipedia;

2) упомянутая выше система Texterra при разрешении лексической многозначности применяет меру близости Дайса (Dice) [21]:

$$sim_{Dice}(x,y) = \\ = \frac{\sum_{z \in N(x) \cap N(y)} (w(x,z) + w(z,y))}{\sum_{z \in N(x)} w(x,z) + \sum_{z \in N(y)} w(z,y)},$$

где $N(c)$ определяет множество страниц, связанных ссылкой с концептом *c* в графе ссылок Wikipedia, а $w(a, b)$ — это вес ссылки от *a* к *b*. Критикой данного метода является ограниченность поиска только первыми окрестностями вершин;

3) «WikiRelate!» [10] использует как текстовую, так и ссылочную информацию энциклопедии для вычисления близости понятий. Данный алгоритм совместно использует сетевые меры близости Лико́к — Чодороу, Ву — Палмер, модифицированную (Seco et al.) меру информационной близости Резника и меру Extended Lesk для определения близости текстовой информации;

4) WikiWalk [24] комбинирует контентный и глобальный ссылочный подход, используя метод случайного блуждания в графе. Паре концептов сопоставляются ESA-вектора соответствующих им статей. Затем на каждом из них как на начальном распределении запускается алгоритм Personalized PageRank, рассчитывающий для каждой вершины графа вектор вероятностей переходов на другие страницы. Результирующие распределения сравниваются с помощью косинусной меры для получения оценки близости. Метод имеет существенно бóльшую временну́ю сложность, чем локальные ссылочные меры, однако дает лучшее качество определения семантической близости концептов.

Сопоставление с другим подходом к анализу терминологической составляющей результатов обучения

В работе Е. А. Черниковой, в рамках предложенного ею подхода к сравнению образовательных курсов, используется сочетание следующих мер при определении семантической близости результатов обучения: расстояние Левенштейна, мера Ву — Палмер, применённая поверх структуры онтологии WordNet, дистрибутивная мера близости DISCO2 [13], основывающаяся на идее взаимосвязи семантической близости понятий и близости распределения слов в текстах.

В отличие от работы Черниковой для анализа терминологической составляющей результатов обучения с целью повышения точности оценки предлагается использовать сочетание контентной меры (алгоритм ESA как наиболее эффективный) и ссылочной меры семантической близости.

Анализ семантики глаголов действия с использованием таксономии образовательных целей

В широко применяемой в сфере образования таксономии образовательных целей Блума (Bloom) [6] выделяется 6 категорий действий в пространстве познания: знание (Knowledge), понимание (Comprehension), приложение (Application), анализ (Analysis), синтез (Synthesis) и оценка (Evaluation). Кратволь (Krathwohl) в своей работе [7] переработал и изменил порядок следования категорий таксономии Блума и уточнил их наименование: вспоминание (Remembering), понимание (Understanding), приложение (Applying), анализ (Analyzing), оценка (Evaluating) и синтез (Creating или Synthesizing). Модифицированная таксономия Кратволя позволяет с большей точностью классифицировать результаты обучения исходя из используемых глаголов действий.

Модифицированная таксономия Кратволя применена для определения семантической близости глаголов действия в работе Е. А. Черниковой. Черникова вносит дополнительные изменения в таксономию и представляет её в виде двумерной матрицы, имеющей категории пространства познания в качестве одного измерения и сложность описываемого глаголом действия в качестве другого.

Черникова затем вводит меру близости глаголов действия:

$$sim_{CAVe} = \frac{\max(d_{EK}) - d_{EK}(AV_i, AV_j)}{\max(d_{EK})},$$

где d_{EK} — дистанция между глаголами в матрице таксономии, а $\max(d_{EK})$ — максимальная дистанция в матрице таксономии.

Используемая при расчетах дистанция d_{EK} — это нормализованное евклидово расстояние, содержащее коэффициент веса *w*, принимающий целочисленные значения, большие 6:

$$d_{EK}(AV_i, AV_j) = \\ = \sqrt{w^2(CPD_i - CPD_j)^2 + (C_i - C_j)^2},$$

где *CPD* и *C* — это измерения в матрице — категория пространства познания и сложность действия соответственно.

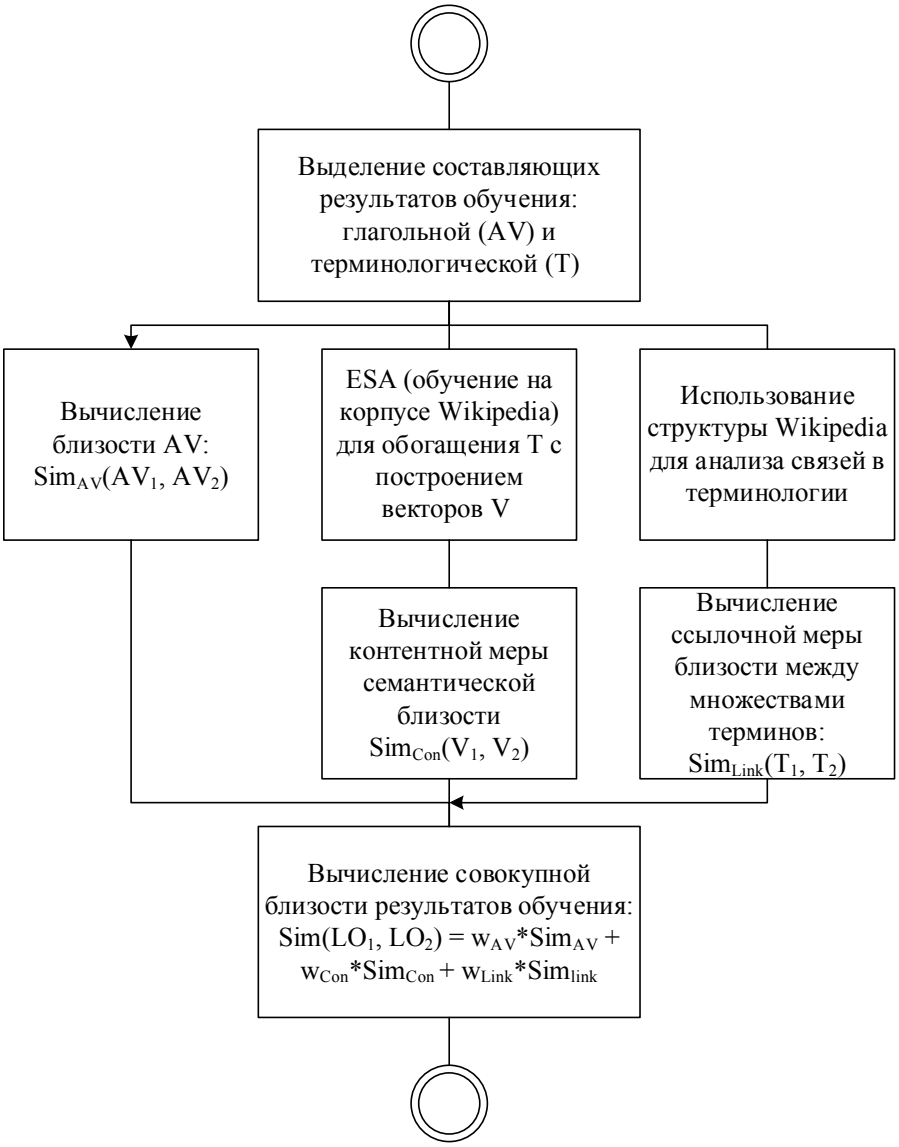
Ограничения на значения коэффициента выведены из нормализующего неравенства, требующего, чтобы максимальное расстояние между глаголами внутри одной категории пространства познания всегда было меньше минимального расстояния между глаголами в разных категориях.

Разработка алгоритма определения меры семантической близости результатов обучения

Схема разрабатываемого алгоритма для определения меры семантической близости для заданной пары результатов обучения представлена на рисунке.

На первом этапе предлагается разделить глагольную и терминологическую составляющую результата обучения для их последующей независимой обработки. Семантическая близость глагольных составляющих может быть вычислена посредством меры, предложенной Черниковой, в то время как для анализа терминологической составляющей предлагается использовать Wikipedia в качестве внешнего корпуса для обогащения короткой формулировки результата обучения путем отображения используемых терминов на содержание соответствующих статей Wikipedia. Для повышения точности расчета семантической близости планируется использовать сочетание контентных и ссылочных мер семантической близости.

Для расчета контентной меры предлагается использовать наиболее эффективный алгоритм семантического анализа — ESA, с построением характеристических векторов концептов обогащенного текста с последующим определением косинусного расстояния между векторами. Для расчета ссылочных мер близости предлагается экспериментально определить наиболее оптимальный алгоритм из рассмотренных.



Алгоритм вычисления близости результатов обучения

В итоге мера семантической близости для пары результатов обучения будет рассчитывается как средневзвешенная сумма значений меры семантической близости глагольной составляющей и контентной и ссылочной мер близости терминологической составляющей. Весовые коэффициенты определяются экспериментально.

Заключение

Дальнейшие шаги исследования включают в себя экспериментальную оценку качества предложенного подхода, выбор оптимальных алгоритмов оценки ссылочных мер близости терминов в аспекте точности и полноты определения мер семантической близости результатов обучения по сравнению с экспертными оценками на подготовленном корпусе образовательных программ российских и зарубежных университетов.

Отдельно стоит оценить временну́ю сложность рассматриваемых алгоритмов для различных входных данных.

Также необходимо выбрать подход к решению проблемы лексической многозначности при отражении термина из формулировки результата обучения на корпус статей Wikipedia.

После проведенных экспериментов на основании данного подхода к определению меры семантической близости результатов обучения будут определяться меры семантической близости образовательного контента на разных уровнях детализации (от отдельных тем до образовательных курсов или программы в целом), а также меры семантической близости результатов обучения, компетенций (из требований образовательных стандартов) и профессиональных требований к знаниям, навыкам и умениям (из требований профессиональных стандартов и требований вакансий на рынке труда).

Библиографический список

1. Botov, D. Educational Content Semantic Modelling for Mining of Training Courses According to the Requirements of the Labor Market / D. Botov, J. Klenin // Proceedings of the 1st International Workshop on Technologies of Digital Signal Processing and Storing. 2015. P. 214–218.

2. Jones, K. S. A statistical interpretation of term specificity and its application in retrieval / K. S. Jones // J. of Documentation. 1972. № 28. P. 11–21.

3. Banerjee, S. Clustering short texts using Wikipedia / S. Banerjee, K. Ramanathan, A. Gupta // Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007. P. 787–788.

4. Phan, X.-H. Learning to classify short and sparse text & web with hidden topics from large-scale data collections / X.-H. Phan, L.-M. Nguyen, S. Horiguchi // Proceeding of the 17th international conference on World Wide Web. ACM, 2008. P. 91–100.

5. Urena-Lopez, L. Integrating linguistic resources in TC through WSD / L. Urena-Lopez, M. Buenaga, J. Gomez // Computers and the Humanities. 2001. № 35 (2). P. 215–230.

6. Bloom, B. Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain / B. Bloom, M. Engelhart, E. Furst [et al.]. N. Y. : David McKay Company, 1956.

7. Krathvohl, D. Revising Bloom's Taxonomy / D. Krathvohl // Theory into Practice. 2002. № 41. P. 212–264.

8. Chernikova, E. A Novel Process Model-driven Approach to Comparing Educational Courses using Ontology Alignment / E. Chernikova. 2014.

9. Milne, D. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links / D. Milne, I. Witten // Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy. 2008. P. 25–30.

10. Strube, M. WikiRelate! Computing Semantic Relatedness Using Wikipedia / M. Strube, S. Ponzetto // Proceeding of the 21st national conference on Artificial intelligence. 2006. Vol. 2. P. 1419–1424.

11. Princeton University: About WordNet [Электронный ресурс]. URL: https://wordnet.princeton.edu

12. The Oxford English Corpus — Oxford Dictionaries [Электронный ресурс]. URL: http://www.oxforddictionaries.com/words/the-oxford-english-corpus

13. Kolb, P. DISCO: A Multilingual Database of Distributionally Similar Words / P. Kolb // Proceedings of KONVENS-2008. Berlin, 2008.

14. Gabrilovich, E. Computing semantic relatedness using Wikipedia-based explicit semantic analysis / E. Gabrilovich, S. Markovitch // Proceedings of the 20th international joint conference on Artifical intelligence. 2007. P. 1606–1611.

15. Aggarwal, C. Mining Text Data / C. Aggarwal, C. Zhai. Springer Publishing Company, 2012.

16. ECTS Users' Guide. Brussels: Directorate-General for Education and Culture, 2005.

17. LEARNING OUTCOMES / S. Lesch, G. Brown College [Электронный ресурс]. URL: http://liad.gbrownc.on.ca/programs/InsAdult/currlo.htm

18. Тархов, С. В. Методологические и теоретические основы адаптивного управления электронным обучением на базе агрегативных учебных модулей / С. В. Тархов. Уфа, 2009. 336 с.

19. Smetanina, O. N. Methodological bases of management of the educational route using intellectual information support : doctoral thesis / O. N. Smetanina. UGATU, Ufa, 2012.

20. Ужва, А. Ю. Автоматизированная разработка онтологической модели предметной области для поиска образовательных ресурсов с использованием анализа текстов рабочих программ [Электронный ресурс] / А. Ю. Ужва. URL: http://www.science-education.ru/107-8324/

21. Texterra: a toolkit for text mining [Электронный ресурс]. URL: www.modis.ispras.ru/texterra

22. Varlamov, M. Computing semantic similarity of concepts using shortest paths in Wikipedia link graph / M. Varlamov, A. Korshunov // Institute for System Programming of the Russian Academy of Sciences.

23. Велихов, П. Меры семантической близости статей Википедии и их применение к обработке текстов / П. Велихов // Информ. технологии и вычисл. системы. 2009. № 1.

24. Yeh, E. WikiWalk: random walks on Wikipedia for semantic relatedness / E. Yeh, D. Ramage, C. Manning, A. Soroa // Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. 2009. P. 41–49.

25. Deerwester, S. Indexing by Latent Semantic Analysis / S. Deerwester, S. Dumais, G. Furnas [et al.] // J. of the American Society for Information Science. 1990. № 41. P. 391–407.

Сведения об авторах

Ботов Дмитрий Сергеевич — преподаватель кафедры информационных технологий и экономической информатики, заведующий лабораторией проектного обучения Института информационных технологий Челябинского государственного университета. dmbotov@gmail.com

Кленин Юлий Дмитриевич — магистрант 1-го курса Института информационных технологий Челябинского государственного университета. jklen@ya.ru

---

D. S. Botov, J. D. Klenin

APPROACH TO SEMANTIC SIMILARITY MEASURE OF LEARNING OUTCOMES

This paper aims to find a solution to the problem of semantic analysis of learning outcomes. Learning outcomes are short texts, which consist of verb and term components, second of which includes terminology from any area of human knowledge. With these specifics in consideration, we reviewed existing approaches to semantic similarity evaluation, as well as analyzed a variety of algorithms, utilizing external knowledge bases and selected an array of most fitting methods.