УДК 159

Метод LDA в психологической оценке понимания текстов

LDA method in psychological assessment of text comprehension

Свердлов С.А., Уральский Федеральный Университет имени первого Президента России Б.Н. Ельцина (Екатеринбург), s.a.sverdlov@mail.ru

Sverdlov S.A., Ural Federal University named after the first President of Russia B.N. Yeltsin (Yekaterinburg), s.a.sverdlov@mail.ru

DOI: 10.51379/KPJ.2021.147.3.037

Ключевые слова: понимание текстов, репрезентации, психология мышления, психология чтения, латентное размещение Дирихле, когнитивные процессы, образование, когнитивно-репрезентативные структуры.

Keywords: understanding of texts, representations, psychology of thinking, psychology of reading, latent Dirichlet placement, cognitive processes, education, cognitive-representational structures.

Аннотация. Проблема понимания текстов и способов его оценки является одной из фундаментальных для современной психологии, педагогики, лингвистики. На настоящий момент имеется большое количество попыток выработки метода оценки понимания текстов, который бы отражал объективную картину интериоризации материала текста читателем, и вместе с тем подходил бы для стандартизации и применения на широком диапазоне текстов различного содержания и структуры, по возможности с минимальными затратами ресурсов оценщика и самого читателя. В настоящей статье раскрываются современные подходы к оценке понимания текстов, их преимущества и недостатки, и формулируется новый перспективный метод оценки понимания текстов, использующий в своём основании модель Латентного размещения Дирихле. Данный метод затем проходит исследование валидности через сравнение результатов его применения и наиболее часто используемых современных методов оценки понимания текстов, делается вывод о его применимости в реальных условиях и перспективах использования в спектре прикладных задач.

Abstract. The problem of understanding texts and methods of assessing it is one of the fundamental for modern psychology, pedagogy, linguistics. At the moment, there are a large number of attempts to develop a method for assessing the understanding of texts, which would reflect an objective picture of the internalization of the material of the text by the reader, and at the same time would be suitable for standardization and application on a wide range of texts of different content and structure, if possible with minimal expenditure of resources of the evaluator and the reader himself. This article reveals modern approaches to assessing text understanding, their advantages and disadvantages, and formulates a new promising method for assessing text understanding, based on the Dirichlet Latent Placement model. This method then goes through a validity study through a comparison of the results of its application and the most commonly used modern methods for assessing text understanding, a conclusion is made about its applicability in real conditions and the prospects for use in a range of applied problems.

Введение. Проблема оценки понимания текстов неотъемлема для психологии, педагогики, лингвистики, спектра когнитивных наук. Психологическое исследование понимания текстов раскрывает внутренние процессы, происходящие в рамках чтения текстов субъектами, и даёт возможность моделирования данных процессов для углубления представлений о вкладе личностных свойств, накопленного опыта и свойств когнитивных процессов на результат формирования репрезентативных

структур при чтении текстов. Непосредственное практическое приложение проблемы оценки понимания текстов является неотъемлемым для построения качественного образовательного процесса, особенно в контексте высшего образования. Разработка прикладных методов оценки понимания текстов также создаст предпосылки для формирования нового поколения программ индивидуального сопровождения при обучении, которые смогут достичь максимальных результатов благодаря

отслеживанию общих тенденций и индивидуальных показателей понимания учебного материала.

Целью настоящей статьи является консолидация представлений об имеющихся методах оценки понимания текстов и их практических свойствах, а также выработка и апробация метода, сочетающего наибольшее количество преимуществ, способного к расширению и стандартизации на широком диапазоне текстов.

Наиболее часто используемым методом исследования понимания текстов является тестирование. Учитывая, что при оценке знаний тестирование широко используется в педагогике – неудивительно, что данный способ быстро начал использоваться и в оценке понимания конкретных текстов. Для отделения фонового знания от результатов прочтения текста в научных исследованиях используется предварительное тестирование и тестирование после прочтения, а экспериментальный план с тестированием только после прочтения используется лишь при оценке понимания текстов, наличие фоновых знаний в которых маловероятно. Вместе с тем, тестовые методики отличает их сфокусированность конкретных тематиках и текстах, что значительно урезает их возможности с точки зрения стандартизации и расширения. Использование тестовых методик требует большой предварительной подготовки и значительного объема работ по подсчету результатов, что особенно заметно в случаях неавтоматизированного тестирования. Произвести прямое сравнение результатов оценок понимания двух разных текстов таким методом в большинстве случаев представляется невозможным.

Другим подходом к оценке степени понимания текстов является метод пропозиционального анализа, предложенный В. Кинчем [1]. Данный метод опирается на выделение пропозиций текста и исследование структуры их взаимосвязей. Преимущество сравнения на семантическом уровне состоит в том, что сравнение не зависит от поверхностных характеристик, таких как замена слов при пересказе, или стилистические изменения в воспроизведенном тексте. Однако, автоматический поиск структуры пропозиций в тексте на настоящий момент невозможен, а ручное нахождение пропозициональной структуры текстов может занимать много времени и требовать больших усилий, что может ограничить размер исследуемых текстов. В большинстве исследований Кинча

использовались тексты, содержащие менее 1000 слов [2;3]. Это накладывает серьезные ограничения на исследуемые тексты. Другим значимым обстоятельством является то, что обозначенные выше модели и их деривативы опираются непосредственно на структуру исследуемого текста, и поэтому оценивание может быть зависимым от специфики пропозициональной архитектуры текстов, что приводит к трудностям в соотнесении результатов оценивания понимания ряда текстов между собой и стандартизации данных методов для работы с широким спектром текстов.

В связи с данными факторами большей перспективой обладают модели, способные опираться на семантические структуры самого языка, а не конкретного текста, для построения модели репрезентаций текста и соотнесения артефактов испытуемых, формируемых после прочтения текста, с оригинальной семантической структурой. Поэтому далее мы рассмотрели вычислительные модели, которые содержат в себе преимущества рассмотренных выше моделей исследования понимания текстов, и опираются уже на семантику и структуру языка, используя большие данные для обработки текстов без необходимости их ручной подготовки.

Рассмотрим наиболее широко распространённый ввиду его результативности и гибкости [4] метод, основанный на формировании распределения слов текстов в пространстве их семантических связей. Латентный семантический анализ (LSA) — это статистическая модель использования слов, которая позволяет сравнивать семантическое сходство между частями текста. Это и является одним из наиболее значимых преимуществ методов латентного размещения – способность не зависеть как от конкретных словоформ, так и от структуры внешнего текста или одного из рассматриваемых фрагментов для их сравнения между собой. Таким образом, метод LSA автоматически захватывает более глубокую ассоциативную структуру, чем простые корреляции и кластеры. Практическое приложение LSA для оценки результата обучения на основе текстовых материалов уже применяется рядом исследователей [4-6]. Выполняя автоматический анализ текстов, которые были прочитаны испытуемыми, производное семантическое пространство можно использовать для сопоставления фрагментов текстов и артефактов активности испытуемых так же, как анализ пропозициональных структур, с меньшей зависимостью от и отсутствием необходимости в ручном поиске пропозициональных структур и

соотнесения фактологических единиц, которые формируют сходные пропозиции в общей структуре текста, однако сильно дифференцируются семантически.

Для рассмотрения в качестве наиболее перспективного метода оценки понимания текстов был избран метод латентного размещения Дирихле, являющийся продолжением эволюции моделей латентного размещения. Этот метод отличается от LSA использованием в своем основании размещения Дирихле как математического основания для создания семантических представлений текстов и языка. Данный метод на настоящий момент рассматривается как наиболее подходящий для целей определения степени понимания текстов, так как он сочетает в себе все обозначенные преимущества вычислительных моделей оценки понимания текстов, имеет опору на структуру и семантическое распределение слов всего языка, а также более соответствует семантическому распределению слов в языке, чем LSA-модели и показывает более точные результаты при использовании в задачах обработки естественных текстов. Также этот метод даёт возможность при оценке понимания текстов рассмотреть ряд дополнительных параметров, которые способны детализировать особенности понимания текстов испытуемыми и расширить поле для построения выводов на основании собранных данных, формируя как холистическую оценку понимания текстов так и аналитические шкалы для раскрытия полной картины свойств понимания текстов субъектами

LDA (Latent Dirichlet Allocation) — это алгоритм мягкой кластеризации, который может создавать вероятностные поля соотношения параметров на основании разметки входящих данных [7]. Латентное распределение Дирихле на любом текстовом корпусе представляет собой генеративную вероятностную модель, состоящего из слов и/или фраз (терминов). На основании множества документов формируется модель, состоящая из пространств терминов, отражающих их семантические взаимосвязи в рамках обработанных для построения модели документов. В результате данная модель

отражает тематико-семантическое пространство корпуса текстов, которые были ей обработаны, как в разрезе семантических связей самих слов и фраз, так и их тематического контекста. Причём, в отличие от других моделей семантического распределения, таких как LSI, модель LDA показывает вероятностные связи между элементами в семантическом поле, что значительно увеличивает точность расчётов в рамках задач обработки полисемантических пространств слов натурального языка. В результате использования глобального корпуса текстов русского языка, мы получили модель LDA общим объёмом более 600 миллионов слов, что по закону больших чисел формирует наиболее приближенную к семантике используемого языка структуру в виде векторного представления семантических связей между словами, основанного на построении латентного размещения Дирихле в пространстве тематической и дистантивной связи слов используемого корпуса.

Нами было проведено сравнение результатов данного метода оценки степени понимания текстов и методов классического тестирования, пропозиционального анализа и LSA для выявления валидности использования LDA в решении данной задачи. Производилось сравнение результатов выборки из 100 испытуемых при прочтении 3 текстов по показателям ответов на 10 стандартизованных вопросов, коэффициента понимания текстов методом пропозиционального анализа пересказа текста по формуле Воронина-Ицковича [7], применения на пересказ текста алгоритма LSA и косинусной семантической дистанции в модели LDA. Выборка была подобрана в соответствии с демографическим распределением современного высшего образования согласно статистике бюллетеня Образования в России [9].

При статистическом исследовании результатов оценок понимания текстов данными методами с помощью критерия Спирмена была получена следующая картина распределения (жирным выделены статистически значимые корреляции), см. рисунок 1:

Variable	Test 10 questions	Propositional VIC	LSA	LDA
Test 10 questions	1	0.17128	0.23485	0.35023
Propositional VIC	0.17128	1	0.27118	0.48640
LSA	0.23485	0.27118	1	0.68145
LDA	0.35023	0.48640	0.68145	1

Рисунок 1. – Корреляционная матрица группы начинающих получение высшего образования

Согласно наблюдаемой нами картине взаимосвязей результатов тестовых методик, оценка текста на основании косинусной семантической дистанции в модели Latent Dirichlet Allocation показала способность соответствовать оценкам методик, валидированных на задаче оценки понимания текстов. Также отмечается, что корреляционная картина оценки понимания текста с помощью LDA значимо коррелирует со всеми методиками, что дает основания полагать преимущество наилучших свойств каждой из методик оценки понимания текстов в исследованном нами алгоритме.

Выводы. Современные методы оценки понимания текстов представлены в основном тестовыми методиками, методом пропозиционального анализа и его деривативами,

а также семейством методов латентно-семантических методов моделирования структуры текстов, согласно результатам ведущих зарубежных [10-16] и отечественных исследователей методов оценки понимания текстов [4;8;17]. Рассмотренный в настоящей работе метод оценки понимания текстов на основании косинусной семантической дистанции в модели LDA показывает способность к оценке понимания текстов на основании сравнительного исследования, что подтверждается его основанностью на семантической структуре текста, отсутствию привязки к структуре конкретного текста и опоре на глобальный текстовый корпус, независимостью от факторов синонимии и обобщения/дробления высказываний при пересказе, что раскрывается теоретическом исследовании.

Литература:

1. Воронин В.М., Курицин С.В., Наседкина З.А. Автоматический анализ объяснений учащимися нарративного текста [Электронный ресурс] / В.М. Воронин, С.В. Курицин, З.А. Наседкина // Гуманизация образования. - 2016. - №2. - Режим доступа: <https://cyberleninka.ru/article/n/avtomaticheskij-analiz-obyasneniy-uchaschimisya-narrativnogo-teksta>
2. Образование в России – 2017. Статистический бюллетень. – М.: Московский технологический университет, 2017. – 444 с.
3. Знаков В.В. Субъектно-аналитический подход в психологии понимания / В.В. Знаков // Психологические исследования. - 2015. - Т. 8. - № 42. - С. 12.
4. Знаков В.В. Три традиции психологических исследований – три типа понимания / В.В. Знаков // Вопросы психологии. - 2009. - № 4. - С. 14-23.
5. Шаповал С.А. Понимание текстов как результат решения учебных филологических задач: автореф. дис. ... канд. психол. наук: 19.00.07 / Шаповал Светлана Анатольевна. - М., 2006.
6. List A., Alexander P.A. (2017). Analyzing and integrating models of multiple text comprehension. Educ. Psychol. 52, 143–147. doi: 10.1080/00461520.2017.1328309
7. List A., Alexander P.A. Cognitive affective engagement model of multiple source use. Educ. Psychol. 52, 182–199, 2017. doi: 10.1080/00461520.2017.1329014
8. Strømsø H.I. Multiple models of multiple-text comprehension: a commentary. Educ. Psychol. 52, p. 216–224, 2017. doi: 10.1080/00461520.2017.1320557
9. Hahnel C., Schoor C., Kröhne U., Goldhammer F., Mahlow N., Artelt C. The role of cognitive load for university students' comprehension of multiple documents. Zeitschrift für Pädagogische Psychologie. 33, 105–118, 2019. doi: 10.1024/1010-0652/a000238
10. Hahnel C., Kröhne U., Goldhammer F., Schoor C., Mahlow N., Artelt C. Validating process variables of sourcing in an assessment of multiple document comprehension. Br.J. Educ. Psychol. 89, 524–537, 2019. doi: 10.1111/bjep.12278
11. Florit E., Cain K., Mason L. Going beyond children's single-text comprehension: the role of fundamental and higher-level skills in 4th graders' multiple-document comprehension. Br.J. Educ. Psychol. 90, 449–472, 2019. doi: 10.1111/bjep.12288
12. Kintsch W., Praful M. "The Construction of Meaning" Topics in Cognitive Science 3 (2): 346–370, 2011.
13. Elleman A.M., Oslund E.L. Reading Comprehension Research: Implications for Practice and Policy. Policy Insights from the Behavioral and Brain Sciences. 2019;6(1):3-11. doi:10.1177/2372732218816339
14. Воронин В.М., Ицкович М.М. Понимание и ценности (норма и патология): монография. / В.М. Воронин, М.М. Ицкович. - Екатеринбург: Издательский Дом «Ажур», 2018. - 306 с.
15. Апресян Ю.Д., Богуславский И.М. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю.Д. Апресян, И.М. Богуславский, Б.Л. Иомдин и др. // Национальный корпус русского языка: 2003—2005. - М.: Индрик, 2005. - С. 193-214.
16. Образование в цифрах: 2019: краткий статистический сборник / Н.В. Бондаренко, Л.М. Гохберг, Н.В. Ковалева и др.; Нац. исслед. ун-т «Высшая школа экономики». – М.: НИУ ВШЭ, 2019. – 96 с.
17. Blei D.M., Ng A.Y., Jordan M.I. 2003. Latent Dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

References:

1. Voronin V.M., Kuritsin S.V., Nasedkina Z.A. Automatic analysis of explanations by students of a narrative text [Electronic resource] / V.M. Voronin, S.V. Kuritsin, Z.A. Nasedkina // Humanization of education. - 2016. - № 2. - Access mode: <https://cyberleninka.ru/article/n/avtomaticheskii-analiz-obyasneniy-uchaschimysya-narrativnogo-teksta>
2. Education in Russia - 2017. Statistical Bulletin. - M.: Moscow Technological University, 2017. - 444 p.
3. Znakov V.V. Subject-analytical approach in the psychology of understanding / V.V. Znakov // Psychological research. - 2015. - T. 8. - № 42. - P. 12.
4. Znakov V.V. Three traditions of psychological research - three types of understanding / V.V. Znakov // Questions of psychology. - 2009. - № 4. - S. 14-23.
5. Shapoval S.A. Understanding of texts as a result of solving educational philological problems: author. dis. ... Cand. psychol. Sciences: 19.00.07 / Shapoval Svetlana Anatolyevna. - M., 2006.
6. List A., Alexander P. A. (2017). Analyzing and integrating models of multiple text comprehension. Educ. Psychol. 52, 143-147. doi: 10.1080 / 00461520.2017.1328309
7. List A., Alexander P.A. Cognitive affective engagement model of multiple source use. Educ. Psychol. 52, 182-199, 2017. doi: 10.1080 / 00461520.2017.1329014
8. Strømsø H.I. Multiple models of multiple-text comprehension: a commentary. Educ. Psychol. 52, p. 216-224, 2017. doi: 10.1080 / 00461520.2017.1320557
9. Hahnel C., Schoor C., Kröhne U., Goldhammer F., Mahlow N., Artelt C. The role of cognitive load for university students' comprehension of multiple documents. Zeitschrift für Pädagogische Psychologie. 33, 105-118, 2019. doi: 10.1024 / 1010-0652 / a000238
10. Hahnel C., Kröhne U., Goldhammer F., Schoor C., Mahlow N., Artelt C. Validating process variables of sourcing in an assessment of multiple document comprehension. Br.J. Educ. Psychol. 89, 524-537, 2019. doi: 10.1111 / bjep.12278
11. Florit E., Cain K., Mason L. Going beyond children's single-text comprehension: the role of fundamental and higher-level skills in 4th graders's multiple-document comprehension. Br.J. Educ. Psychol. 90, 449-472, 2019. doi: 10.1111 / bjep.12288
12. Kintsch W., Praful M. "The Construction of Meaning" Topics in Cognitive Science 3 (2): 346-370, 2011.
13. Elleman A.M., Oslund E.L. Reading Comprehension Research: Implications for Practice and Policy. Policy Insights from the Behavioral and Brain Sciences. 2019; 6 (1): 3-11. doi: 10.1177 / 2372732218816339
14. Voronin V.M., Itskovich M.M. Understanding and values (norm and pathology): monograph. / V.M. Voronin, M.M. Itskovich. - Yekaterinburg: Publishing House "Azhar", 2018. - 306 p.
15. Apresyan Yu.D., Boguslavsky I.M. Syntactically and semantically annotated corpus of the Russian language: current state and prospects / Yu.D. Apresyan, I.M. Boguslavsky, B.L. Iomdin et al. // National corpus of the Russian language: 2003-2005. - M.: Indrik, 2005. - S. 193-214.
16. Education in figures: 2019: a short statistical collection / N.V. Bondarenko, L.M. Gokhberg, N.V. Kovalev and others; Nat. issled. University Higher School of Economics. - M.: NRU HSE, 2019. - 96 p.
17. Blei D.M., Ng A.Y., Jordan M.I. 2003. Latent Dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993-1022.

19.00.01 – Общая психология, психология личности, история психологии

Сведения об авторе:

Свердлов Сергей Александрович (г. Екатеринбург, Россия), аспирант, Уральский Федеральный Университет имени первого Президента России Б.Н. Ельцина, e-mail: s.a.sverdlov@mail.ru