ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

М. А. Белюга

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СЕМАНТИЧЕСКИ ПОДОБНЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ: ПОСТАНОВКА ЗАДАЧИ, ЕЕ ПРИНЦИПИАЛЬНАЯ СХЕМА РЕШЕНИЯ И ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

В статье рассмотрены некоторые аспекты задачи автоматического распознавания семантически релевантных текстовых документов (ТД). Предложена принципиальная схема решения задачи, которая предполагает построение семантического индекса ТД, полученного в результате глубокого семантического анализа текста с учетом синонимичных и иерархических отношений, с целью последующего распознавания семантической релевантности ТД. Дано описание лингвистического обеспечения задачи.

Развитие современных информационных технологий и особенно сети Интернет крайне обострило проблему обработки больших объемов информации и, в первую очередь, текстовой – в настоящее время поисковые серверы оперируют уже десятками миллиардов документов. В этой связи особую актуальность приобретает проблема распознавания и измерения степени смыслового подобия текстов, поскольку именно эти аспекты очень востребованы в задачах информационного поиска, кластеризации и категоризации ТД, распознавания плагиата, цифрового и дистанционного обучения и многих других.

В последнее время исследователями предпринимаются многочисленные попытки разработки универсального алгоритма выявления схожих или идентичных по смыслу текстов. Многие существующие на сегодняшний день решения формируют интегральную оценку смыслового подобия текстов на основании их лексического сходства, с учетом порядка следования слов, их канонических форм и синонимов. Однако уровень семантики не отдельного слова, а хотя бы целого предложения все еще остается недостаточно проработанным ввиду трудоемкости и наукоемкости этой области знаний о языке. Прогресс в этом ее направлении – это прямой путь к классу совершенно новых и эффективных, по-настоящему интеллектуальных информационных технологий.

Среди перечисленных выше прикладных задач, оперирующих понятием смыслового подобия текстов, одной из самых актуальных, безусловно, является задача информационного поиска, основным типом которого является поиск релевантных документов. В широком смысле релевантность в информационном поиске представляет собой семантическое соответствие поискового запроса (входного документа) документу из поискового пространства. На практике она обычно определяется путем сравнения поискового образа запроса (ПОЗ) с поисковым образом документа (ПОД) по определенному

алгоритму. ПОД и ПОЗ являются формальным представлением соответствующих документов и строятся с помощью так называемой процедуры индексирования [1, с. 71], в большей или меньшей степени использующей, в том числе, и автоматический лингвистический анализ текста.

Нельзя не отметить, что далеко не все документы в выдаче удовлетворяют информационной потребности пользователя. Как правило, они лишь формально соответствуют поисковому предписанию. Документы, действительно соответствующие потребности пользователя, называются пертинентными [2]. А сама информационная потребность представляет собой весьма сложное психическое явление, и проблема повышения степени пертинентности выдачи оказывается не только трудной для достижения, но и ее даже трудно четко поставить как практическую задачу. В нашем случае речь идет не о классической задаче информационного поиска, а, в конечном счете, о задаче оценки смысловой релевантности (иногда говорят сходства семантики) двух заданных ТД. Заметим, что в качестве таковых могут рассматриваться и их отдельные фрагменты. Понятно, что одинаковые документы будут абсолютно релевантны, в противном случае речь идет об их относительной релевантности. Таким образом, здесь нет аспекта пользователя, а, значит, и пертинентности, а релевантность рассматривается с точки зрения самой системы. Безусловно, существуют различные подходы к решению рассматриваемой задачи, и есть конкретные практические результаты.

В [3] авторы условно подразделяют все реализованные на тот момент системы распознавания смысловой релевантности текстов на три большие группы на основании природы мер подобия, лежащих в основе алгоритма:

1) системы, основанные на знаниях и признаках, соответственно получаемых и формируемых из строк ТД (string-based similarity systems); включают в себя системы посимвольного и пословного сравнения;

2) системы, основанные на информации о словах, получаемых из корпусов текстов (corpus-based similarity), это, например, данные о совместной встречаемости слов, контекстная информация или даже данные о числе пользователей, выбравших тот или иной результат поиска по тому или иному запросу и проч.;

3) системы, основанные на знаниях и признаках, соответственно получаемых и формируемых из семантических сетей, тезаурусов, онтологий, ЛБЗ и т.д. (knowledge-based similarity systems).

Останавливаясь подробнее на системах последней группы, отметим, что одним из самых популярных источников такого рода информации здесь является семантическая сеть английского языка WordNet [4], называемая также лексической базой данных, тезаурусом, в которой существительные, прилагательные, глаголы и наречия сгруппированы в ряды когнитивных синонимов – синсеты (synset), каждый из которых соответствует отдельному концепту. Синсеты связаны между собой посредством концептуально-семантических и лексических отношений. В зависимости от характера отношений между синсетами, на которые опирается та или иная мера

подобия данной группы подходов, все меры подобия этой группы можно условно отнести к собственно мерам семантического сходства или к мерам семантической соотнесенности. Так, семантически сходные концепты связаны между особой отношениями подобия, тогда как семантически соотнесенные концепты могут быть связаны и рядом других связей: род-вид, часть-целое, отношения антонимии и т.д. [5]. Выделяют 6 мер семантического сходства, три из которых основаны на содержащейся в синсетах информации ([6; 7; 8]), а три оставшиеся – на длине пути от синсета к синсету ([9; 10]). При этом авторы акцентируют внимание на том, что подобие слов может оцениваться на двух уровнях: лексическом и семантическом. Слова считаются «лексически подобными», если они состоят из одинаковых или подобных цепочек символов. Тогда как «семантически подобные» слова – это слова, которые:

- обладают синонимичным или антонимичным смыслом;
- или состоят в гипонимо-гиперонимических отношениях;
- или могут использоваться в тексте сходным образом (в одинаковых или подобных контекстах).

Таким образом, системы первой группы (string-based algorithms) можно отнести к группе систем установления лексического подобия текстов, и только системы второй (corpus-based algorithms) и особенно третьей (knowledge-based algorithms) групп правомерно относить к группе систем установления их семантического сходства.

Также отмечается, что в последнее время развиваются системы, опирающиеся на различные комбинации способов оценивания семантического подобия и формирующие интегральную оценку семантического подобия текстовых фрагментов на основании частных мер подобия разных уровней [11; 12]. Так, в [13] представлен метод комплексной оценки семантического подобия предложений или иных коротких фрагментов текста на основании информации о семантике слов и о их порядке в предложении. Однако авторам [14] удалось достичь еще лучших результатов за счет метода, который получил название Semantic Text Similarity (STS). Данный метод опирается на информацию о семантике слов предложения и о его синтаксической структуре: авторы принимали во внимание две обязательные функции (подобие строк и подобие слов по смыслу) и одну необязательную функцию (общность порядка слов двух текстовых фрагментов).

Представляется, что системы указанной выше третьей группы могут быть усилены за счет использования более глубокого семантического анализа текста, и здесь, в первую очередь, необходимо обратить внимание на следующее. Ориентация на промышленный характер приложений, а именно они крайне востребованы, неизбежно требует сохранения процедуры автоматического индексирования ТД с целью построения их формального представления, поскольку в этом случае достигается существенная минимизация общего времени решения задачи. Под таковым представлением понимается основное смысловое содержание ТД, выраженное с помощью

заданной системы знаний, а ее выбор является определяющим для эффективного решения целевой задачи. С практической точки зрения, как показывает опыт решения многих современных актуальных задач, связанных с автоматической обработкой текста, эффективным является подход, опирающийся на распознавание знаний основных трех типов: объектов, фактов (семантических отношений между объектами типа С-А-О, где С – субъект, А – акция, О – объект) и отношений типа Причина-Следствие на множестве фактов, полных и неполных, которые фактически отображают закономерности внешнего мира (предметной области) [15]. Учитывая, что собственно объект можно рассматривать как один из предельных случаев факта С-А-О, то поисковый образ документа или, опять-таки, его отдельного фрагмента, можно, таким образом, рассматривать как множество фактов. И в дальнейшем множество меток (тегов), фиксирующих распознаваемые в ТД факты, а также результаты его базового лингвистического анализа, будем называть семантическим индексом (SI) документа. Отметим, что при неодходимости он может быть дополнен и метками атрибутивных знаний [16], тем более, что их автоматическое распознавание осуществимо по той же технологии, что и знаний основных типов.

В нашем случае речь идет о небольших по объему (от одной до нескольких страниц) текстах d1 и d2: текстах социальных сетей, рекламных текстов, рефератах текстов и т.п. В принципе, можно исходить из того, что на входе задачи есть два текста D1 и D2 произвольной длины. Тогда, в общем случае будем полагать, что d1 = R(D1), а d2 = R(D2), где R(D1) и R(D2) – получаемые автоматически рефераты текстов, соответственно D1 и D2. Требуемый для этих целей алгоритм может быть основан, например, на методе машинного обучения. Необходимыми для решения целевой задачи, т.е. задачи распознавания смысловой релевантности текстов d1 и d2, их формальными представлениями являются получаемые автоматически семантические индексы, соответственно SI(d1) и SI(d2). Предполагается, что эти индексы погружаются в заданную онтологию внешнего мира (предметной области) с целью онтологического обобщения входящих в них фактов, в общем случае, по всем трем компонентам. Речь идет об отношении синонимии и отношении is A (общее – частное), для которого заранее задаются границы обобщения. На заключительном этапе производится сравнение множеств онтологически обобщенных фактов семантических индексов SI(d1) и SI(d2) и принятие на основе сформулированных заранее критериев решения о семантической схожести ТД d1 и d2, абсолютной или относительной, если таковые имеют место. Очевидно, что в последнем случае можно даже показать, относительно каких фактов имеет место смысловое подобие ТД. Причем, при использовании многоязычных онтологий, речь уже идет о возможности решения задачи в многоязычной информационной среде, т.е. о так называемой cross-language фунциональности соответствующей системы.

В соответствии с принципиальной схемой решения задачи, предложенной выше в п.2, в системе автоматического распознавания семантически подобных ТД на верхнем уровне организационной иерархии можно выделить три основных системных модуля:

1) модуль автоматического реферирования текста;
2) модуль построения ПОД, который в свою очередь подразделяется на модуль семантического индексирования документа и модуль онтологического обобщения фактов семантического индекса документа;
3) модуль сравнения поисковых образов документов.

Руководствуясь тем фактом, что многие решения задач автоматической обработки ТД, построенные с использованием методов машинного обучения, показывают высокую эффективность, целесообразно в основе модуля автоматического реферирования текста использовать вероятностно-статистическую модель, а именно нейронную сеть. Ключевым ресурсом для обучения нейронной сети в данном случае является предварительно построенный корпус текстов с соответствующими рефератами. Кроме того, использование методов машинного обучения предполагает построение признакового описания текстовых фрагментов, в данном случае предложений, получаемого в результате их количественного, а также автоматического лингвистического анализа.

Функциональность второго из упомянутых выше модулей обеспечивается базовым лингвистическим процессором (БЛП) и его лингвистической базой знаний (ЛБЗ), а также некоторой семантической сетью, тезаурусом и т.п.

Исходя из постановки задачи и принципиальной схемы ее решения, в качестве БЛП в нашем случае выбран ЛП IHS Goldfire[1], реализующий обработку ТД, начиная от предварительного их форматирования и заканчивая лексико-грамматическим, синтаксическим и семантическим анализом, что целиком согласуется с требованиями, предъявляемыми к разрабатываемому алгоритму. Данный лингвистический процессор ориентирован на промышленную обработку ТД для целого ряда естественных языков (включая английский) и автоматическое распознавание в ТД знаний упомянутых ранее основных трех типов, и имеет лучшие на настоящее время показатели эффективности среди аналогичных разработок. Его ЛБЗ включает требуемые знания о естественном языке в виде совокупности различных словарей, грамматик, корпусов текстов, классификаторов свойств языка и распознающих лингвистических моделей его анализа на различных уровнях глубины ЕЯ.

Что касается задачи обобщения фактов семантического индекса документа, то в основу его лингвистического обеспечения положена семантическая сеть английского языка WordNet, которая была существенно доработана с точки зрения уточнения и пополнения существующих подсетей, а также добавления новых, например, для такой части речи, как предлоги.

[1] https://www.ihs.com/products/design-standards-software-goldfire.html

Функциональность третьего из указанных ранее модулей системы фактически не требует разработки своего лингвистического обеспечения, поскольку здесь речь идет только об алгоритме сравнения ПОД.

Существующие алгоритмы решения рассматриваемой задачи фактически ориентированы на нахождение лексически эквивалентных фрагментов с учетом простейших морфологических преобразований и отношений синонимии. Предлагаемое решение задачи, с одной стороны, основано на развитом лингвистическом анализе ТД и позволяет распознавать семантически подобные ТД как в целом, так и относительно отдельных их фактов, с другой стороны, используя многоязычный вариант базовой семантической сети, обеспечивает соответствующей системе функциональность cross-language. Причем существующая возможность перехода от входного ТД к его реферату существенно минимизирует общее время решения задачи. Все это актуально для многих важных приложений.

ЛИТЕРАТУРА

1. *Солтон, Дж.* Динамические библиотечно-информационные системы = Dynamic information and library processing / Дж. Солтон; пер. В. Р. Хисамутдинов. – М. : Мир, 1979.
2. *Михайлов, А. И.* Основы информатики / А. И. Михайлов, А. И. Чёрный, Р. С. Гиляревский. – 2-е изд., перераб. и доп. – М. : Наука, 1968. – 756 с.
3. *Gomaa, W. H.* A Survey of Text Similarity Approaches / W. H. Gomaa, A. A. Fahmy // International Journal of Computer Applications. – Vol. 68, No. 13. – 2013. – P. 13–18.
4. *Miller, G. A.* WordNet: An online lexical database / G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller // International Journal of Lexicography, Vol. 3, No. 4. – Oxford University Press. – 1990. – P. 235–244.
5. *Patwardhan, S.* Using measures of semantic relatedness for word sense disambiguation / S. Patwardhan, S. Banerjee, T. Pedersen // Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City. – Springer-Verlag Berlin, Heidelberg. – 2003. – P. 241– 257.
6. *Resnik, P.* Using information content to evaluate semantic similarity in a taxonomy / P. Resnik // Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada. – Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. – 1995. – P. 448–453.
7. *Lin, D.* Extracting Collocations from Text Corpora / D. Lin // Workshop on Computational Terminology, Montreal, Canada. – 1998. – P. 57–63.
8. *Jiang, J.* Semantic similarity based on corpus statistics and lexical taxonomy / J. Jiang, D. Conrath // Proceedings of the International Conference on Research in Computational Linguistics, Taiwan. – 1997. – P. 19–33.
9. *Leacock, C.* Combining local context and WordNet sense similarity for word sense identification / C. Leacock, M. Chodorow // WordNet, An Electronic Lexical Database. – The MIT Press – 1998. – P. 265–283.

10. *Wu, Z.* Verb semantics and lexical selection / Z. Wu, M. Palmer // In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico. – 1994. – P. 133–138.
11. *Davide, B.* IRIT: Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method / B. Davide, T. Ronan, A. Nathalie, M. Josiane // First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada, June 7–8. – Association for Computational Linguistics, Stroudsburg, PA, USA. – 2012. – P. 552–556.
12. *Bar, D.* UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures / D. Bar, C. Biemann, I. Gurevych, T. Zesch // First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada, June 7–8. – Association for Computational Linguistics, Stroudsburg, PA, USA. – 2012. – P. 435–440.
13. *Li, Y.* Sentence similarity based on semantic nets and corpus statistics / Y. Li, D. McLean, Z. Bandar, J. O'Shea, K. Crockett // IEEE Transactions on Knowledge and Data Engineering. – Vol. 18, No. 8. – IEEE Educational Activities Department Piscataway, NJ, USA. – 2006. – P. 1138–1149.
14. *Islam, A.* Semantic text similarity using corpus-based word similarity and string similarity / A. Islam, D. Inkpen // ACM Transactions on Knowledge Discovery from Data. – Vol. 2, No. 2. – ACM, New York, NY, USA. – 2008. – P. 1–25.
15. *Совпель, И. В.* Система автоматического извлечения знаний из текста и ее приложения / И. В. Совпель // Искусственный интеллект. ІПШІ МОН і НАН Украіни «Наука і освіта». – 2004. – № 3. – С. 668–677.
16. *Постаногов, Д. Ю.* Автоматическая обработка естественного языка в задаче инженерии знаний и доступа к ним: дис. ... к-та техн. наук: 05.13.17 / Д. Ю. Постаногов. – Минск, 2012. – 136 л.

This article explores some aspects of automatic recognition of semantically relevant texts. We propose a method to detect the semantic relevance of textual documents using their sematic indices, which are built as a result of deep analysis of semantics taking into account synonymous and hierarchical relationships. The paper also describes linguistic resources needed for the problem solution.

Поступила в редакцию 28.04.18

М. В. Чернышевич

ПРИНЦИПИАЛЬНАЯ СХЕМА РЕШЕНИЯ ЗАДАЧИ АСАТ И ЕГО ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

В данной работе предложена принципиальная схема системы автоматического сентимент-анализа текстов на английском языке, определены необходимые ее компоненты и реализовано решение трех основных подзадач системы АСАТ. Подробно описано лингвистическое обеспечение, разработанное в рамках данной работы, включающее аннотированные корпуса, разнообразные лексические ресурсы и лингвистические паттерны.