DOI: 10.25586/RNU.V9187.18.04.P.91 УДК 681.3.07

О.В. Золотарев, М.М. Шарнин, С.В. Клименко, А.Г. Мацкевич

ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ АССОЦИАТИВНО-ИЕРАРХИЧЕСКОГО ПОРТРЕТА ПРЕДМЕТНОЙ ОБЛАСТИ[1]

В работе рассматриваются проблемы семантического моделирования, методики автоматизированного выявления иерархических, синонимических и ассоциативных связей из интернет-текстов и построение лингвостатистических портретов различных предметных областей. Исследование основано на гипотезе о том, что более общие термины имеют больше ассоциативных связей, а также о привлечении ассоциативных связей для определения значения, полный смысл которого выявляется с помощью контекстных окружений, что дает возможность автоматизации процесса разграничения значений и извлечения знаний из текстов. Решение проблемы строится на основе комплексного подхода, сочетающего методы статистики, корпусной лингвистики и дистрибутивной семантики, и реализуется в технологии, которая предполагает разработку лингвостатистических механизмов формирования ассоциативно-иерархического портрета предметной области (АИППО), представляющего собой словарь значимых терминов предметной области, элементы которого связаны ассоциативными и иерархическими связями.

Работы проводятся на основе анализа различных предметных областей, в частности – по автономным необитаемым подводным аппаратам (АНПА).

Ключевые слова: *ассоциативный портрет предметной области, онтология, иерархические связи, синонимические связи, ассоциативные связи, контекстное окружение, векторные пространства.*

O.V. Zolotarev, M.M. Sharnin, S.V. Klimenko, A.G. Matskevich

RESEARCH OF METHODS OF AUTOMATIC FORMATION OF ASSOCIATIVE AND HIERARCHICAL PORTRAIT OF THE SUBJECT AREA

The paper discusses the problems of semantic modeling techniques for automated detection of hierarchical, synonymous and associative relationships from online texts and the construction of linguistic and statistical portraits of various subject areas. The study is based on the hypothesis that the more general terms have more associative relations. The involvement of associative relationships for the definition of the full meaning is revealed by the context of the environments that gives you the ability to automate the process of differentiating between values and knowledge extraction from texts. The solution is based on an integrated approach that combines statistical methods, corpus linguistics and distributional semantics, and is implemented in a technology which involves the development of linguo-statistical mechanisms for the formation of associative-hierarchic portrait of the subject area (AHPSA), which is a dictionary of important terms of the subject area, elements of which are connected by the associative and hierarchical relationships.

Work is carried out on the basis of the analysis of different subject areas, in particular, Autonomous Unmanned Underwater Vehicle (AUUV).

Keywords*: associative-hierarchic portrait of subject area, ontology, hierarchical relationships, synonymous relationships, associative relationships, contextual environment, vector space.*

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований, гранты 15-07-06586, 16-07-00756, 16-29-09527, 18-07-00909 + 18-07-01111.

Методика построения АИППО

Описываемый в работе подход по построению ассоциативно-иерархического портрета предметной области (АИППО) основывается на проведении автоматического статистического анализа больших объемов текстов из Интернета [3–5]. Иерархические связи, входящие в АИППО, образуют полииерархию и классификатор, облегчающие поиск и навигацию в предметной области АНПА (ПО АНПА). Подобная методика позволяет решать широкий класс задач как в области когнитивной семантики, так и в сфере информационно-поисковых систем, так как АИППО может в большинстве случаев, связанных с контекстным поиском, заменить или дополнить тезаурус/онтологию предметной области, составление которого вручную представляет собой весьма трудоемкую задачу. Дополнительно проект затрагивает следующие задачи: мониторинг новых объектов, фактов и идей в ПО АНПА, автоматическая классификация новых объектов по классификатору АИППО, в частности вид/тип аппарата АНПА, его характеристики, компания-производитель, ее руководство, сотрудники, конкуренты, партнеры и т.д., как часто упоминается объект в различные периоды времени, тональность сообщений, источник информации, установление границ предметной области; развитие интеллектуальных интернет-технологий; автоматизированное формирование интерактивных предметно ориентированных энциклопедий; визуализация результатов интерактивного сетевого поиска (визуальные карты предметной области) [6; 8].

Методика построения ассоциативно-иерархического портрета предметной области (АИППО) основана на структуризации текстов предметной области и построении иерархии категорий, в которой для расчета иерархических связей между ЗС/ЗТ (значимые слова/значимые термины) используются методы тематического моделирования, такие, как LDA и hLDA. Выделенные по указанным методикам ассоциативные и иерархические связи между значимыми словосочетаниями и терминами позволяют разрабатывать более совершенные методы и метрики/меры подобия научных текстов.

Итак, методы тематического моделирования служат для построения тематической модели коллекции документов. Тематическая модель определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдается числовой вектор, составленный из оценок степени принадлежности данного документа к каждой из тем. Размерность этого вектора, равная числу тем, может либо задаваться на входе, либо определяться моделью автоматически.

Далее коллекция документов, относящихся к одной предметной области обрабатывается специальным программным обеспечением.

Для анализа и обработки больших текстовых массивов первоначально надо провести разметку текста. Для этого существует множество способов и инструментов, в том числе общедоступных. Например, язык разметки RDF (Resource Description Framework) – среда описания ресурса, разработка консорциума W3C для описания метаданных; OWL (Web Ontology Language) – язык описания онтологий для семантической разметки информации, представленной в сети Интернет и т.д. [9]. В нашей работе используется оригинальная, разработанная в ИПИ РАН, система разметки текстов – PullEnty, в результате работы которой в тексте выделяются именованные типизированные сущности (персоны, организации, местоположения, даты, связи, …), события, их характеристики, например время, когда это событие произошло, действия и т.д. [10]. Кроме этого в работе используется инструмент анализа семантики текста, основанный на дистрибутивной семантике и векторном представлении слов – Word-2Vec [1; 2], который в результате обработки текста сопоставляет каждому слову вектор. Векторное представление строится на основе контекстной близости: считается, что слова, находящиеся ближе друг к другу, будут иметь похожие значения координат

векторов слов. Степень сходства документов оценивается на основе косинусного расстояния:

$$\cos(\alpha) = \frac{\sum_{i=1}^{n} di \cdot bi}{\sqrt{\sum_{i=1}^{n} (di)^2} \sqrt{\sum_{i=1}^{n} (bi)^2}}.$$

Здесь d и b – текстовые документы из коллекции документов, n – количество слов в словаре, составленном из слов коллекции документов, исключая стоп-слова (предлоги, союзы, местоимения и т.д.). Каждый документ представляется как разреженный вектор (потому что только отдельные слова из словаря входят в каждый документ):

$$D = (d1, d2, \ldots di, \ldots, dn).$$

Здесь каждое значение di представляет частоту встречаемости i-го слова из словаря в тексте D, n – общее количество слов в словаре.

Модели векторных представлений, семантическое контекстное пространство

Модели векторных пространств находят всё более широкое применение в исследованиях, связанных с семантическими моделями естественного языка, и имеют разнообразный спектр потенциальных и действующих приложений. Данная область в настоящее время является одной из наиболее актуальных. Следует отметить модели Word Space Model и Semantic Space Model [Sahlgren, 2006], а также Word-Space Model [Ph.D. thesis, University of Stockholm, Stockholm] и работу [Lenci, A. Distributional semantics in linguistic and cognitive research // Rivista di Linguistica, 1, 2008, pp. 1–30]. В основе этих работ лежит пространство лексем. Семантическое пространство, которое базируется на распределении слов в корпусе текстов с целью представления их семантической связанности путем оценки пространственной близости. Используется понятие семантического контекстного пространства (СКП), где точки пространства соответствуют контекстным векторам значимых словосочетаний. Концепция семантических векторных пространств (СВП) впервые была реализована в информационно-поисковой системе SMART [Salton, 1971]. SMART был пионером многих концепций, которые успешно используются современными поисковиками [Manning, Raghavan, & Schutze, 2008]. Идея СВП состоит в представлении каждого документа из коллекции в виде точки в пространстве, т.е. вектора в векторном пространстве. Точки, расположенные ближе друг к другу в этом пространстве, считаются более близкими по смыслу. Пользовательский запрос рассматривается как псевдодокумент и тоже представляется как точка в этом же пространстве. Документы сортируются в порядке возрастания расстояния, т.е. в порядке уменьшения семантической близости от запроса, и в таком виде предоставляются пользователю. В настоящее время большинство поисковиков используют СВП для измерения степени близости запроса и найденных документов [Manning et al., 2008]. M. Baroni, A. Lenci (2010) предложили обобщенную модель, названную «дистрибутивная память», которая является обобщением ранее известных моделей векторных пространств (vector spaces), семантических пространств (semantic spaces), пространств слов (word spaces), семантических моделей корпусной статистики (corpus-based semantic models) и дистрибутивных семантических моделей (distributional semantic models). Такая модель описана в работе [M. Baroni, A. Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics // Computational Linguistics. V. 36, Issue 4, 2010, pp. 673–721]. Успех СВП для информационно-поисковых задач направил исследователей на применение СВП для других семантических задач.

Разработчики СВП отмечают, что основная проблема известных семантических пространств – это недостаточный учет порядка слов в контексте. Для решения этой проблемы следует перейти от контекста слов к контексту значимых словосочетаний. Более того, технология СВП развивалась для английского языка. Проект СКП пред-

полагает работу как с русским, так и с английским языками. Возможно в дальнейшем включение других языков. Построение семантического контекстного пространства СКП направлено на развитие методов СВП для решения следующих задач:

1. Выявление синонимии и семантической близости слов и словосочетаний путем оценки их встречаемости в различных контекстах.

2. Поиск категорий терминов и отношений с помощью лексико-синтаксических форм.

3. Выявление близких по смыслу отношений и их классификация методами статистического анализа контекстных зависимостей.

4. Автоматическая кластеризация слов по степени их близости в СКП и классификация слов путем использования лексико-семантических форм.

5. Автоматическая генерация тезаурусов методами статистической обработки терминов и разрешение неоднозначности слов путем использования контекста.

6. Расширение запросов за счет ассоциативных связей и извлечение знаний из текстов с использованием статистических методов и лингвистических моделей.

7. Оценка степени сходства лексических конструкций на основе их лексико-семантического анализа.

Описание разработанных программных средств

Коллективом разработана технология автоматического поиска научно-технических документов в Интернете и построения их коллекций. Также разработана технология автоматического выделения библиографических ссылок в найденных документах. Размер коллекции, построенной авторским коллективом из открытой информации в Интернете по тематике «Автономные необитаемые подводные аппараты» (АНПА), составляет более 200 документов.

Для пополнения коллекций текстов из Интернета разработан усовершенствованный метод (KeyCrawler-2) семантического поиска ЕЯ-текстов из Интернета с целью направленного извлечения из текстов информации для построения онтологий, в котором собственный поисковый робот в качестве начальных адресов в Интернете (URL) использует не только крупнейшие поисковики, но также каталоги электронных библиотек/магазинов и собственную разработку научного коллектива – энциклопедию ключевых понятий KEYWEN. KeyCrawler-2 по заданным ключевым терминам строит не только интернет-корпуса естественно-языковых текстов, но также строит тематические коллекции научных документов (PDF) с названиями, авторами и библиографическими ссылками. Метод KeyCrawler-2 апробирован на ряде предметных областей, включая АНПА. Разработан модуль статистического анализа для алгоритма формирования ассоциативно-иерархического портрета предметной области (АИППО).

Данный модуль статистического анализа включает в себя:

2.1) поиск иерархических связей терминов с помощью методов кластеризации в пространстве контекстных векторов из значимых словосочетаний (ЗС);

2.2) поиск ассоциативных и иерархических связей при помощи методов тематического моделирования, включая построение полииерархии тем и аннотации произвольного размера из ключевых фраз для корпуса/коллекции;

2.3) поиск иерархических связей с помощью методов кластеризации в пространстве тематических векторов по результатам тематического анализа (LDA);

2.4) поиск иерархических связей с помощью лингвистического процессора PullEnty, который выделяет различные типы объектов, например люди и организации, относящихся к заданной предметной области;

2.5) поиск иерархических и ассоциативных связей по различным лексическим шаблонам, например АНПА XXX (АНПА Ремус, АНПА Гавиа);

2.6) поиск ассоциативных связей по косинусной мере между контекстными векторами ЗС;

2.7) поиск связей перевода и ассоциативных связей при помощи программ перевода;

2.8) объединение найденных иерархических связей по методу патентных заявок USPTO 20100161671 и 61/096255, в которых представлено изобретение, использованное Шарниным М.М. для построения одной из крупнейших иерархий категорий для электронной энциклопедии Keywen.

С помощью метода KeyCrawler-2 составлена коллекция научно-технических публикаций (статьи, диссертации, монографии) по выбранным предметным областям для проведения экспериментов, размером более 10 000 документов, в том числе коллекция по теме АНПА содержит более 200 документов, по теме «компьютерная графика» – более 900 документов, в которых выявлены ссылки еще на 6 000 документов.

Заключение

В результате проделанной работы разработан макетный вариант онтологии предметной области АНПА, выделены концептуальные термины предметной области (классы, целевое назначение, конструктивный облик, навесное оборудование, режим использования, виды работ, профильные компании и учреждения), представлены системы понятий и коллекции экземпляров [7].

Благодарности

Мы благодарны РФФИ за поддержку и финансирование наших проектов.

Литература

1. *Matt J. Kusner, Yu Sun, Nisholas I. Kolkin, Kilian Q.* Weinberger. From Word Embeddings To Document Distances // Proceedings of the 32 nd International Conference on Machine. Learning. – Lille, France, 2015. JMLR: W&CP. – Vol. 37.

2. *Mikolov, Tomas, Wen-tau Yih, Geoffrey Zweig*. Linguistic Regularities in Continuous SpaceWord Representations // Proceedings of NAACL-HLT 2013, Atlanta, Georgia, 9–14 June 2013. Association for Computational Linguistics. – 2013. – P. 746–751.

3. *Золотарев О.В.* Новые подходы в формировании функциональной структуры предметной области / О.В. Золотарев // Двадцать лет постсоветской России: кризисные явления и механизмы модернизации : материалы XIV Всероссийской научно-практической конференции Гуманитарного университета : в 2 т. – Екатеринбург, 2011. – С. 639–643.

4. *Золотарев О.В., Шарнин М.М.* Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик // Труды Международной научной конференции СРТ2014 Международная научная конференция Московского физико-технического института (государственного университета) Института физико-технической информатики. Институт физико-технической информатики. – М., 2015. – С. 92–98.

5. *Шарнин М.М., Золотарев О.В., Сомин Н.В.* Извлечение и обработка знаний из неструктурированных текстов деловой сферы и социальных сетей // Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты : материалы Четвертой Международной научно-практической конференции. – М., 2015. – С. 364–371.

6. *Шарнин М.М., Шагаев И., Протасов В.И., Родина И.В., Золотарев О.В., Попова О.А*. Использование веб-семантики для совершенствования образовательных программ вузов // Rhema. Рема. – 2015. – № 2. – С. 97–112.

7. *Клименко И.С.* Теория систем и системный анализ : учебное пособие. – М., 2014.

8. *Золотарев О.В.* Методы выделения процессов, объектов, отношений из текстов естественного языка // Проблемы безопасности российского общества. – 2014. – № 3–4. – С. 276–283.

9. *Клименко С.В., Золотарев О.В., Шарнин М.М.* Использование онтологического подхода для анализа текстов естественного языка // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2017. – № 1. – С. 67–71.

10. *Золотарев О.В., Шарнин М.М., Клименко С.В., Кузнецов К.И.* Система PullEnty – извлечение информации из текстов естественного языка и автоматизированное построение информационных систем // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SCVRT2015-16) : труды Международной научной конференции : в 2 т. – М., 2016. – С. 28–35.

References

1. *Matt J. Kusner, Yu Sun, Nisholas I. Kolkin, Kilian Q. Weinberger*. From Word Embeddings To Document Distances // Proceedings of the 32 nd International Conference on Machine. Learning. – Lille, France, 2015. JMLR: W&CP. – Vol. 37.

2. *Mikolov, Tomas, Wen-tau Yih, Geoffrey Zweig.* Linguistic Regularities in Continuous SpaceWord Representations // Proceedings of NAACL-HLT 2013, Atlanta, Georgia, 9–14 June 2013. Association for Computational Linguistics. – 2013. – P. 746–751.

3. *Zolotarev, O.V.* Novye podkhody v formirovanii funktsional'noy struktury predmetnoy oblasti / O.V. Zolotarev // Dvadtsat' let postsovetskoy Rossii: krizisnye yavleniya i mekhanizmy modernizatsii : materialy XIV Vserossiyskoy nauchno-prakticheskoy konferentsii Gumanitarnogo universiteta : v 2 t. – Ekaterinburg, 2011. – S. 639–643.

4. *Zolotarev, O.V., Sharnin, M.M.* Metody izvlecheniya znaniy iz tekstov estestvennogo yazyka i postroenie modeley biznes-protsessov na osnove vydeleniya protsessov, ob'ektov, ikh svyazey i kharakteristik // Trudy Mezhdunarodnoy nauchnoy konferentsii CPT2014 Mezhdunarodnaya nauchnaya konferentsiya Moskovskogo fiziko-tekhnicheskogo instituta (gosudarstvennogo universiteta) Instituta fiziko-tekhnicheskoy informatiki. Institut fiziko-tekhnicheskoy informatiki. – M., 2015. – S. 92–98.

5. *Sharnin, M.M., Zolotarev, O.V., Somin, N.V.* Izvlechenie i obrabotka znaniy iz nestrukturirovannykh tekstov delovoy sfery i sotsial'nykh setey // Sotsial'nyy komp'yuting: osnovy, tekhnologii razvitiya, sotsial'no-gumanitarnye effekty : materialy CHetvertoy Mezhdunarodnoy nauchno-prakticheskoy konferentsii. – M., 2015. – S. 364–371.

6. *Sharnin, M.M., Shagaev, I., Protasov, V.I., Rodina, I.V., Zolotarev, O.V., Popova, O.A.* Ispol'zovanie veb-semantiki dlya sovershenstvovaniya obrazovatel'nykh programm vuzov // Rhema. Rema. – 2015. – № 2. – S. 97–112.

7. *Klimenko, I.S.* Teoriya sistem i sistemnyy analiz : uchebnoe posobie. – M., 2014.

8. *Zolotarev, O.V.* Metody vydeleniya protsessov, ob'ektov, otnosheniy iz tekstov estestvennogo yazyka // Problemy bezopasnosti rossiyskogo obshchestva. – 2014. – № 3–4. – S. 276–283.

9. *Klimenko, S.V., Zolotarev, O.V., Sharnin, M.M.* Ispol'zovanie ontologicheskogo podkhoda dlya analiza tekstov estestvennogo yazyka // Vestnik Rossiyskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". – 2017. – № 1. – S. 67–71.

10. *Zolotarev, O.V., Sharnin, M.M., Klimenko, S.V., Kuznetsov, K.I.* Sistema PullEnty – izvlechenie informatsii iz tekstov estestvennogo yazyka i avtomatizirovannoe postroenie informatsionnykh sistem // Situatsionnye tsentry i informatsionno-analiticheskie sistemy klassa 4i dlya zadach monitoringa i bezopasnosti (SCVRT2015-16) : trudy Mezhdunarodnoy nauchnoy konferentsii : v 2 t. – M., 2016. – S. 28–35.