Извлечение информации из текстов на конференциях серии «Диалог»: взгляд соседа по лестничной клетке

Information extraction at Dialogue conferences: a neighbour's view

Хорошевский В. Ф. (khor@ccas.ru)

Вычислительный центр им. А. А. Дородницына РАН, Москва

В работе представлен ретроспективный анализ публикаций конференций серии Диалог за 2000–2009 г. г. в области извлечения информации из текстов. Для построения общей «картины мира» конференций этой серии используются статистические методы, методы онтологического инжиниринга и семантической кластеризации и классификации, а для выявления значимых семантических пространств — технологии извлечения информации из текстов.

1. Немного истории вместо введения

Как известно, история конференций серии Диалог своими корнями уходит в 80-е годы прошлого столетия, когда в нашей стране под эгидой Научного совета по проблеме «Искусственный интеллект» при Президиуме АН СССР был сформирован междисциплинарный проект «Диалог», в рамках которого стали регулярно проводиться небольшие, но очень значимые по составу участников научные семинары. У истоков организации этих микроконференций стояли такие известные ученые, как Ю. Д. Апресян, А. П. Ершов, А. Е. Кибрик, Г. С. Поспелов и др., а безусловными лидерами, вокруг которых и «раскрутилась» вся эта активность, стали Д. А. Поспелов и А. Е. Кибрик. В организации первых встреч активное участие принимал Л. И. Микулич, но вскоре всю тяжесть проведения конференций по проекту «Диалог» взвалили на свои плечи А. С. Нариньяни, Д. Бухштаб и Н. Лауфер, благодаря энергии и труду которых каждый год в Эстонии (как правило, с помощью И. Сильдмяэ под Тарту) или в Подмосковье встречались «стенка на стенку» специалисты по компьютерной лингвистике и искусственному интеллекту. В те годы такие междисциплинарные встречи были уникальным явлением, но именно они, на мой взгляд, и позволили лингвистам и программистам, по сути дела, понять друг друга и проблемы в области создания интеллектуальных диалоговых систем.

Следующей значимой вехой в истории проекта «Диалог» стали, на мой взгляд, семинары «Диалог-2», к участию в которых, в основном, благодаря Д. А. Поспелову, были привлечены когнитивные психологи. К сожалению, такого всплеска междисциплинарных проектов, который обозначился в начале 90-х годов по результатам встреч на семинарах «Диалог-1», не произошло, хотя и «Диалог-2» сыграл свою роль в «наведении мостов» между специалистами по компьютерной лингвистике, когнитивной психологии и искусственному интеллекту.

В настоящее время мы являемся свидетелями (и участниками) нового этапа развития проекта «Диалог», ежегодные конференции которого стали значимой точкой в календаре междисциплинарных конференций по компьютерной лингвистике и интеллектуальным технологиям. В связи с этим, а также в связи с тем, что в качестве «горячей темы» очередной 16-й Международной конференции «Диалог» обозначена тема извлечения информации из ЕЯ-текстов, представляются интересными проведение ретроспективного анализа публикаций в трудах конференций этой серии на данную тему и попытка оценки текущего состояния исследований в данной области в нашей стране.

2. Модели и методы анализа научных сообществ

Следует сразу отметить, что обсуждение исследований и разработок в области анализа научных (и не только научных) сообществ выходит за рамки настоящей работы. Поэтому здесь мы отсылаем заинтересованного читателя к работам[1,2,3,4,5,6,7,8,9] и лишь отметим, что в нашей стране это направление активно развивалось в 70–80-х годах прошлого века. Сейчас эти работы становятся востребованными в связи с практическими задачами выявления «незримых коллективов» и формальной оценки их результативности, как в целом, так и на уровне отдельных членов этих коллективов, а также прогнозирования развития таких сообществ.

[1] *Налимов В. В.* Количественные методы исследования процесса развития науки // Вопросы философии. 1966. № 12. С. 38–47.

[2] *Налимов В. В., Кордон И. В., Корнеева А. Я.* Географическое распределение научной информации // Информационные материалы Научного совета по комплексной проблеме "Кибернетика" АН СССР. 1971. № 2 (49). С. 3–37.

[3] *Nalirnov V. V.* Faces of Science. // Philadelphia, ISI Press, 1981. 297 p.

[4] *Noma E.* Co-citation analysis and the invisible college // J. Amer. Soc. Info. Sci. 1984. Vol. 35. P. 29.

[5] Евстигнеев В. А. Методы теории графов в наукометрии: исследование структуры пространства журналов и незримых коллективов в программировании. // Новосибирск, 1987. (Препр. АН СССР. Новосиб. филиал. ИТМ и ВТ им. С. А. Лебедева; № 4).

[6] *Маршакова И. В.* Система цитирования научной литературы как средство слежения за развитием науки. // М.: Наука, 1988. 288 с,

[7] *Хайтун С. Д.* Проблемы количественного анализа науки. // М.: Наука, 1989. 280 с.

[8] *Паринов С. И.* Контуры единой сетевой инфраструктуры научного сообщества // Технологии информационного общества — Интернет и современное общество: труды V Всероссийской объединенной конференции. СПб., 25–29 ноября 2002 г. СПб.: Изд-во С.-Петерб. ун-та, 2002. С. 118

[9] Егоров В. С., Пожидаев А. В., Чернобровская Т. Н. Систематизация и использование сведений о научных мероприятиях в автоматизированной технологии ВИНИТИ. // НТИ. Сер. 1. — 2006. — №4.– С. 17–23.

3. Статистическая картина мира конференций «Диалог»

3.1. Общий портрет направления

Общая структура исследований и разработок в области извлечения информации из текстов на естественных языках, как она представляется автору на основании анализа публикаций в этой области и собственного опыта, может быть описана рубрикатором, представленным в Табл. 1. Понятно, что данный рубрикатор не претендует на полноту и законченность, но вместе с тем дает основу для оценки ситуации в нашей стране в этой стратегически важной области компьютерной лингвистики и искусственного интеллекта. В связи с этим представляется интересным ретроспективный анализ тех публикаций в трудах конференций серии Диалог, которые (по экспертной оценке автора) относятся к данному направлению. Для проведения такого анализа из трудов конференций за 2000–2009 г. г., представленных на сайте [1], были выбраны работы, которые вписываются в предложенную классификацию. Такое отображение конечно фасетное, так как одна работа (как правило) затрагивает несколько направлений. Полученные результаты представлены в Табл. 1 (наиболее важные, по мнению автора, направления выделены курсивом).

Как показывает анализ представленных выше материалов, из более чем 800 работ, опубликованных в трудах конференций серии Диалог за 2000–2009 г. г., 201 работа (около 25 %) связана тем или иным образом с извлечением информации из текстов. При этом направлению «Модели и методы» посвящены 133 работы (≈ 66.2 %), направлению «Инструментальные средства» — 18 работ (≈ 9 %) и направлению «Интеллектуальные приложения» — 50 работ (≈ 24.8 %). Таким образом, теория «опережает» приложения почти в 3 раза, а инструментарий — почти в 7 раз. На наш взгляд, эти цифры отражают общее состояние исследований в данной области в нашей стране.

Табл. 1.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Модели и методы	**6**	**9**	**18**	**11**	**17**	**14**	**13**	**15**	**13**	**17**
• Ресурсы										
• Словари	1		5	1			2	2	3	2
• Тезаурусы	2	2	2	1	2		1	2	2	2
• Онтологии		2	3		1		4	1	1	2
• Подходы к извлечению информации из текстов										
• Статистический	1	1	3	2	4	2	1	2	2	2
• «Легкий» (Shallow)	1	2	1		3	4	3	3	2	3

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
• «Тяжелый» (Deep)	1	2	4	5	3	4	2	3		2
• Гибридный					1	2		2	1	1
• Машинное обучение										
• Автоматическая генерация правил					1					1
• Онтологический инжиниринг				2	2	2			2	2
Инструментальные средства	**1**	**3**	**4**	**2**	**1**	**1**	**1**	**1**	**0**	**4**
• Средства и языки представления знаний		2	2		1					2
• Платформы	1		2	2		1	1			
• Автоматическое конфигурирование										1
• Отладка		1								1
• Тестирование и оценивание результатов								1		
• Автоматическая генерация процессоров и IE-систем										
Интеллектуальные приложения	**2**	**5**	**6**	**7**	**5**	**7**	**7**	**3**	**4**	**4**
• Семантическое аннотирование		1		1						
• Семантическая кластеризация/классификация		2	4				1		2	3
• Семантический поиск		2	1	1	1	1	1		1	
• Семантическая навигация				1	1					
• Семантическое дайджестирование	1									
• Семантическое реферирование	1				1	2	1			
• Аналитика на знаниях				3	2	3	3	3	1	
• Порталы знаний			1	1		1	1			1
• Семантический Веб										

3.2. Авторский индекс и организации-участники

Для более детальной оценки активности российских специалистов, представленных в конференциях серии Диалог в области извлечения информации из текстов, представляет интерес анализ индекса авторов и организаций этих конференций (Табл. 2).

Табл. 2.

	2000	2001	2002	2003	2004
Специалисты	21	21	43	28	29
Организации	8	11	22	15	17
	2005	**2006**	**2007**	**2000**	**2009**
Специалисты	35	32	36	26	24
Организации	15	15	15	14	18

Как показывает статистический анализ, всего в составе авторских коллективов в трудах конференций Диалог за 2000–2009 г. г. опубликованы работы 295 специалистов из 150 организаций. При этом уникальных авторов — 170, из них представивших более 2 публикаций — 58 чел., более 3 публикаций — 23 чел. и более 4 публикаций — 14 чел.

Наиболее активные авторы конференций серии Диалог по данной тематике представлены в диаграмме на Рис. 1.

Не менее интересен и состав организаций, где, как показывают публикации трудов конференций серии Диалог, ведутся исследования и разработки в области извлечения информации из текстов. Детальный геоландшафт организаций, работающих в данной области, будет представлен в докладе и, частично, дан ниже при анализе семантического портрета конференций серии Диалог. Пока же отметим, что, как и следовало ожидать, наиболее активны в этой области организации Москвы, Новосибирска, Санкт-Петербурга и Казани.

4. Семантическая картина мира конференций «Диалог»

4.1. OntosMiner/SGE — инструментарий анализа

Понятно, что статистические данные, приведенные выше, дают определенное представление о картине мира конференций серии Диалог, но нуждаются в дальнейшей детализации и анализе, который является достаточно трудоемким и потому не может быть проведен без соответствующих инструментальных средств. Детальное обсуждение результатов работы команды Ontos из российской IT-компании

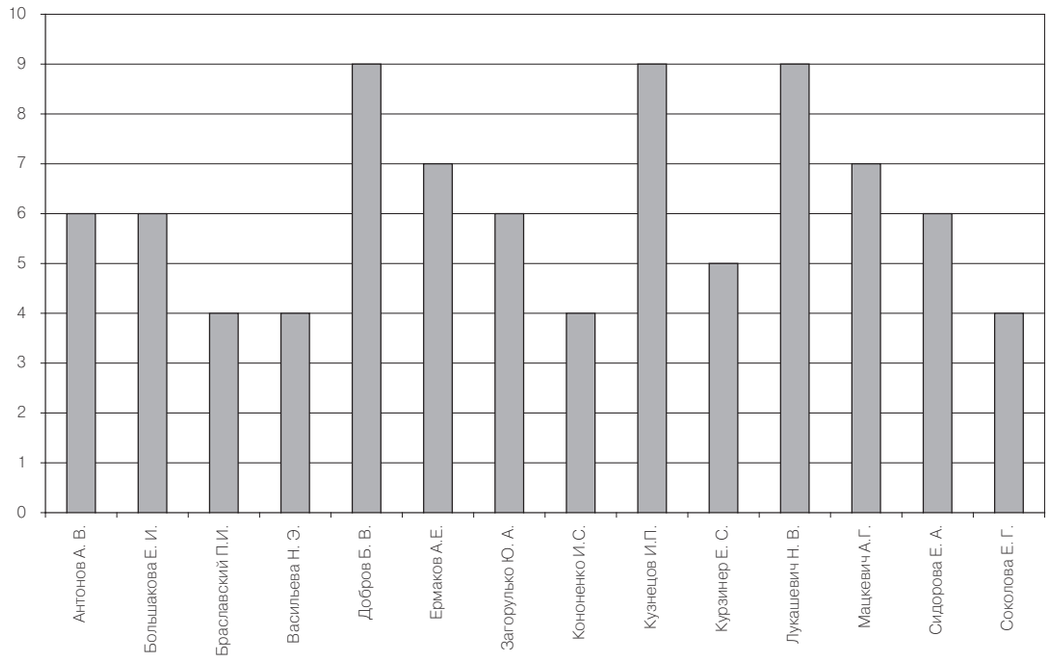

Рис. 1. Наиболее активные авторы конференций Диалог по тематике извлечения информации из текстов

«Авикомп Сервисез» в области извлечения информации из ЕЯ-текстов и соответствующих технологий выходит за рамки настоящей работы. Поэтому ниже кратко описывается только лингвистический процессор OntosMiner/SGE (Shadow Groups Extraction), который использовался в рамках настоящего исследования для построения семантической картины мира конференций Диалог за 2000–2009 г. г.

Концепция разработки системы OntosMiner/SGE лежит в общем русле работ по проекту OntosMiner [2–4] и состоит в следующем:

- Обработка текстов осуществляется под управлением модели предметной области, представленной в виде соответствующей онтологии.
- Для анализа текстов используется shallow approach, в основе которого лежит система шаблонов-образцов на специальном языке представления знаний.
- Результаты обработки отдельных текстов (документов) представляются в виде когнитивных карт (специальных семантических сетей), которые загружаются в базу знаний, где происходит формирование общей семантической сети коллекции документов.
- В качестве инструментальной среды для разработки и реализации системы OntosMiner/SGE использован стандартный инструментарий проекта Ontos, который является развитием среды GATE [5].

В соответствии с общей структурой IE-систем семейства OntosMiner в системе OntosMiner/SGE используется ресурсная цепочка в составе Tokenizer — Morph Tagger — Sentence Splitter — NE Transducer — Semantic Tagger — XML Generator. В качестве первых трех и последней компоненты в системе OntosMiner/SGE используются стандартные ресурсы семейства OntosMiner. Компонента NE Transducer реализуется на основе реинжиниринга соответствующих стандартных модулей, а стандартная компонента Semantic Tagger расширена за счет реализации новых семантических отношений и скомплексирована таким образом, чтобы исключить из нее те стандартные модули, которые обрабатывают семантические отношения, не представленные в данной предметной области.

Общий объем лингвистического процессора OntosMiner/SGE составил 73 специальных и около 600 общих правил, интегрированных из других процессоров семейства, которые компилируются, в соответствии с общей технологией Ontos, в Java-код, а затем и в соответствующую систему Java-классов. Скорость обработки одной статьи стандартного для этих конференций объема (5–9 стр.) составляет, в зависимости от насыщенности ее объектами и отношениями, 1–3 сек на персональном компьютере с процессором Intel® Core™2 Duo с тактовой частотой 2 Ггц и основной памятью 2 Гб.

4.2. Предметная онтология ShadowGroups и корпус текстов

Анализ структуры научных статей, опубликованных в трудах конференций Диалог показал, что с точки зрения целей и задач настоящей работы основными объектами, которые целесообразно извлекать из соответствующих текстов, являются

именованные сущности типа Person, Organization, Location, Paper, Domain, Term, OrgTeam и др., а также семантически значимые отношения между ними типа BeAuthor, BeCoauthor, MemberOf, ReferenceTo, ReferencedBy и т. п. С учетом этого была разработана предметная онтология ShadowGroups Ontology, представленная на Рис. 2.

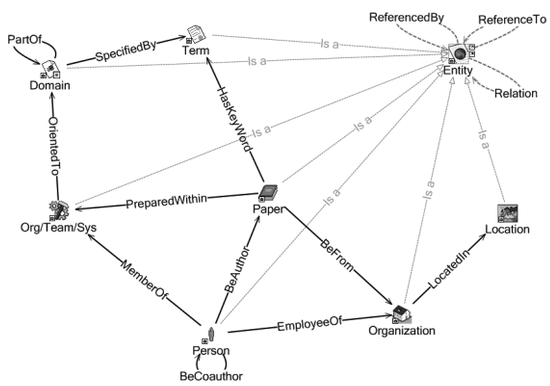

Рис. 2. Онтология предметной области ShadowGroups

Как указывалось выше, в качестве корпуса документов использовались электронные копии трудов конференций Диалог за 2000–2009 г. г., полученные с соответствующего Интернет-сайта. Как правило, в каждом выпуске было около 120 научных статей, покрывающих достаточно широкий спектр исследований и разработок в области теоретической и прикладной лингвистики, а также машинной обработки ЕЯ, но далеко не все они имеют непосредственное отношение к извлечению информации из текстов. Поэтому для дальнейшей обработки из трудов конференций была выбрана 141 статья. Выбранные для дальнейшей обработки статьи были приведены к единому текстовому формату, поскольку на сайте они представлены в форматах txt, doc, pdf и html.

Обработка статей осуществлялась с помощью процессора OntosMiner/SGE под управлением онтологии ShadowGroups, а для объединения результатов в общее семантическое пространство использовались специальные правила идентификации одинаковых объектов, представленных в текстах различным образом. Для визуализации результатов обработки использовалось desktop-приложение LightOntos for Workgroups.

Полученные результаты обсуждаются в оставшихся разделах настоящей работы.

4.3. Ландшафты конференций серии Диалог

Для построения геоландшафта конференций серии Диалог в общей когнитивной карте были оставлены для визуализации только экземпляры понятий Organization и Location. Полученные результаты представлены на Рис. 3.

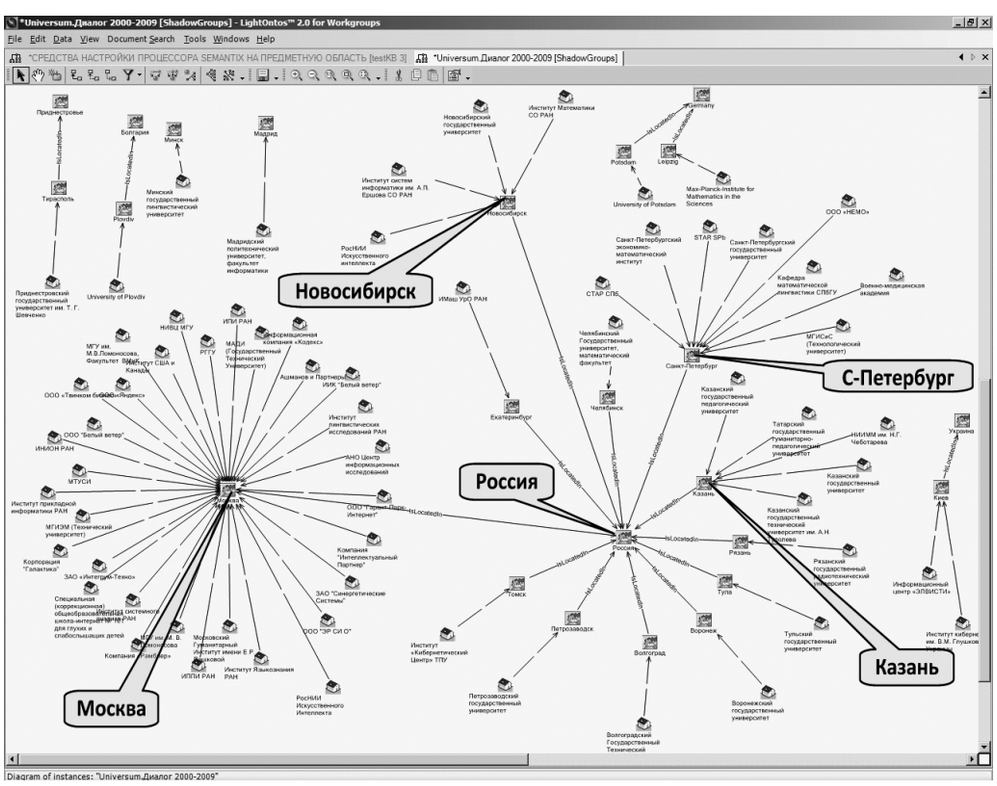

Рис. 3. Геоландшафт конференций серии Диалог

Как показывает анализ этих результатов, наиболее активны на конференциях серии Диалог организации из Москвы (32), Санкт-Петербурга (8), Казани (5) и Новосибирска (4). Из текстов научных статей выделены 26 геообъектов, организации были связаны с геоименами семантическими отношениями LocatedIn, а геоимена с геоименами более высокого уровня (города расположены в странах, если эта информация представлена в статьях) — семантическими отношениями IsLocatedIn. Точность выделения объектов типа Organization составила 83 %, а полнота — 89 %. При этом основные ошибки и пропуски связаны с неполным распознаванием объекта (не всегда выделяется кафедры и факультеты ВУЗов). Точность выделения объектов типа Location составила 93 %, а полнота –96 %. Основные ошибки в определении геоимен связаны, в основном, с ложными срабатываниями правил на фрагментах текстов типа «ПРЕДМЕТНАЯ ОБЛАСТЬ» в названиях статей. Точность и полнота выделения отношений, при условии правильного выделения объектов, — 100 %.

Полученные результаты дают общее представление о функционале процессора OntosMiner/SGE, но с точки зрения целей настоящей работы, на наш взгляд, интереснее обсуждение тематических ландшафтов конференций Диалог разных лет и более детальный анализ авторских коллективов и взаимоотношений между ними.

Всего из корпуса статей, представленных в 2000–2009 г. г. на конференциях Диалог, было выделено объектов типа Person — 1678, Paper — 842, Organization — 204 и Location — 151, между которыми установлены семантические отношеня BeAuthor, BeCoauthor, ReferenceTo и симметричные им отношения ReferencedBy. Характеристики точности и полноты обработки геообъектов даны выше. Точность выделения объектов типа Person составила 91 %, а полнота — 87 %, объектов типа Organization — 83 % и 89 %, а объектов типа Paper — 96 % и 79 %, соответственно. При этом основные ошибки на объектах типа Person были обусловлены необходимостью обработки в рамках одного процессора и русских, и иностранных авторов, использованием модулей из общей библиотеки OntosMiner, которые не адаптировались специально к задаче обработки статей, а также неаккуратностью авторов статей при их оформлении, особенно в коллекциях первых лет. Ошибки на объектах типа Organization, в основном, были связаны с неполнотой общих правил (не идентифицировались факультеты и кафедры как часть организации), а ошибки на объектах типа Paper — разнообразием способов оформления ссылок в коллекциях разных лет. Точность и полнота выделения отношений, при условии правильного выделения необходимых для их сборки объектов, — 100 %.

Как следует из анализа частных диаграмм, конференция Диалог-2000 является наименее насыщенной с точки зрения извлечения информации из текстов. При этом работы разных авторских коллективов, по существу не связаны ни между собой (фокусом когнитивной карты является геообъект «Россия»), ни ссылками на авторитеты, в том числе и западные. Эта же тенденция инкапсуляции ярко выражена в когнитивной карте 2003 года и практически сохраняется в когнитивных картах 2001, 2002 и 2004 г. г. И только в когнитивных картах 2005 г. и, особенно, 2009 г. появляются взаимные ссылки разных авторских коллективов. Авторские ландшафты 2006–2008 г. г. занимают промежуточное положение, поскольку на них видны процессы формирования новых авторских коллективов, с одной стороны, и постепенное появление ссылок между разными авторскими коллективами — с другой.

Вместе с тем, автономный анализ когнитивных карт конференций Диалог разных лет не позволяет понять, существуют ли в области извлечения информации из текстов в нашей стране признанные авторитеты и «скрытые» коллективы. Поэтому в следующем разделе настоящей работы обсуждаются результаты, полученные после объединения всех частных когнитивных карт в единое семантическое пространство.

4.4. Извлечение информации из текстов — скрытые коллективы

Для построения общего семантического пространства конференций серии Диалог отдельные когнитивные карты были обработаны алгоритмами автоматической идентификации одних и тех же объектов и отношений между ними, а полученные результаты верифицировались экспертным путем, а затем проводилось человеко-машинное «схлопывание» оставшихся дублей объектов типа Person, Paper и Organization. Неполнота автоматического объединения объясняется тем, что авторы статей достаточно часто фиксируют в ссылках один и тот же источник различным образом (где-то есть издательство, где-то нет, где-то не указаны страницы, а где-то они есть), названия одних и тех же организаций указываются то полностью, то их аббревиатурами, английские транскрибирования одних и тех же русских, и даже английских, авторов в разных ссылках указываются по разному.

После проведения автоматической идентификации одинаковых объектов в семантическом пространстве осталось объектов типа Person — 919, Paper — 816, Organization — 105 и Location — 25, а после экспертного «выравнивания» — объектов типа Person — 854, Paper — 743, Organization — 71 и Location — 13, а также соответствующие семантические отношения между ними (Рис. 4).

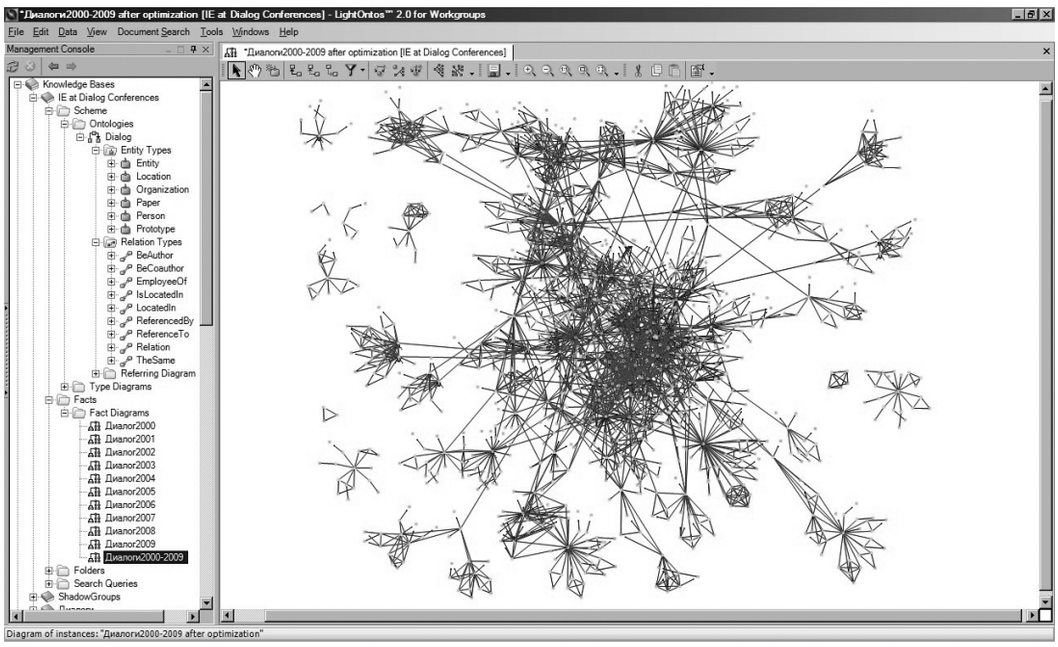

Рис. 4. IE-ландшафт конференций серии «Диалог»

Какие предварительные выводы можно сделать из анализа этой диаграммы? Во-первых, это значительная ее связность при сохранении некоторого числа полностью инкапсулированных авторских коллективов. Во-вторых, появление в диаграмме ярко выраженных и, по-видимому, как-то связанных «скоплений», что позволяет предположить наличие «скрытых коллективов».

Вместе с тем, мелкий масштаб, а также визуализация всех объектов и связей на одной диаграмме не позволяет провести детальный ее анализ. Поэтому для дальнейшего анализа «скроем» на диаграмме все объекты типа Paper, Organization и Location, отношения типа ReferencedBy, поскольку на диаграмме у них есть симметричные отношения типа ReferenceTo а также удалим изолированные подграфы. Диаграмма, полученная в результате таких преобразований, представлена на Рис. 5 и отражает возможные коллективы и связи между ними.

Для дальнейшего упрощения диаграммы, представленной на Рис. 5, уберем с карты все «желтые» авторские коллективы, тех, кто не попал в «подсветку», а также тех авторов, на которых ссылаются специалисты только одной команды. В результате получим диаграмму, представленную на Рис. 6.

Анализ персоналий потенциальных авторитетов в области IE по данным конференций серии Диалог показывает, что их можно разделить на три категории. В первой — общепризнанные авторитеты в области лингвистики, которые практически не работают в области извлечения информации из текстов и потому ссылки на них можно назвать «ритуальными» [6–10 и др.]. Во второй — ссылки на учебники и фундаментальные работы, которые можно отнести к образовательным [11–14 и др.]. И, наконец, в третьей категории — работы, которые повлияли на текущие исследования коллективов нашей страны в данной области в теоретическом и/или практическом плане [15–18 и др.].

Не менее, а может и более интересной является диаграмма, представленная на Рис. 7, где детализируется взаимодействие между разными авторскими коллективами и внутри каждого из них.

Как показывает анализ этой диаграммы, все «команды» в значительной мере «замкнуты» на себя и при этом подавляющая часть авторов «грешат» автоссылками. Самая многочисленная «команда» представляет несколько тесно связанных между собой организаций и коллективов из Новосибирска (кластер Загорулько [19,20 и др.]), который, при этом, никак не связан с другой новосибирской командой (кластер Загоруйко [21]), которая ссылается только на кластер Кузнецова (ИПИ РАН) [22,23]. Кластер Загорулько ссылается на кластер Ермакова, например, на [24], на кластер Добров-Лукашевич [25] (АНО Центр информационных исследований), например, на [25–27], который, в свою очередь, тесно связан общими исследованиями и статьями [28] с кластером Невзоровой (Казань), а взаимными ссылками — с кластером Ермакова (RCO) [24,28], где есть ссылка на кластер Антонова (корпорация «Галактика») [30,31] и кластер Кузнецова [22], а из кластера Кузнецова на кластер Ермакова [24]. Из диаграммы на Рис. 7 следует, что наиболее «открыт» миру IE-кластеров нашей страны кластер Большаковой [32,33 и др.] (ВМиК МГУ), где имеются несколько внешних ссылок [26, 27 и др.]. И, наконец, единственный кластер, на который ссылаются три других кластера — это кластер Браславского (ИМаш УрО РАН) и работы [34,35].

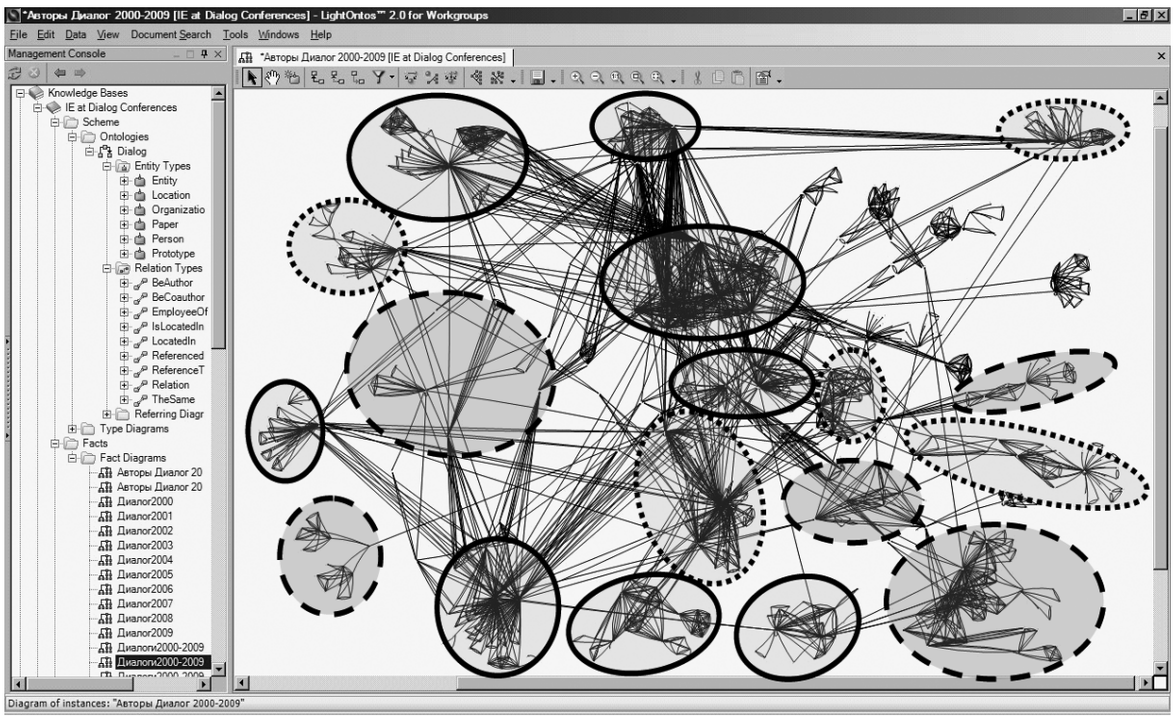

Рис. 5. Диаграмма связей авторских коллективов в области IE[10]

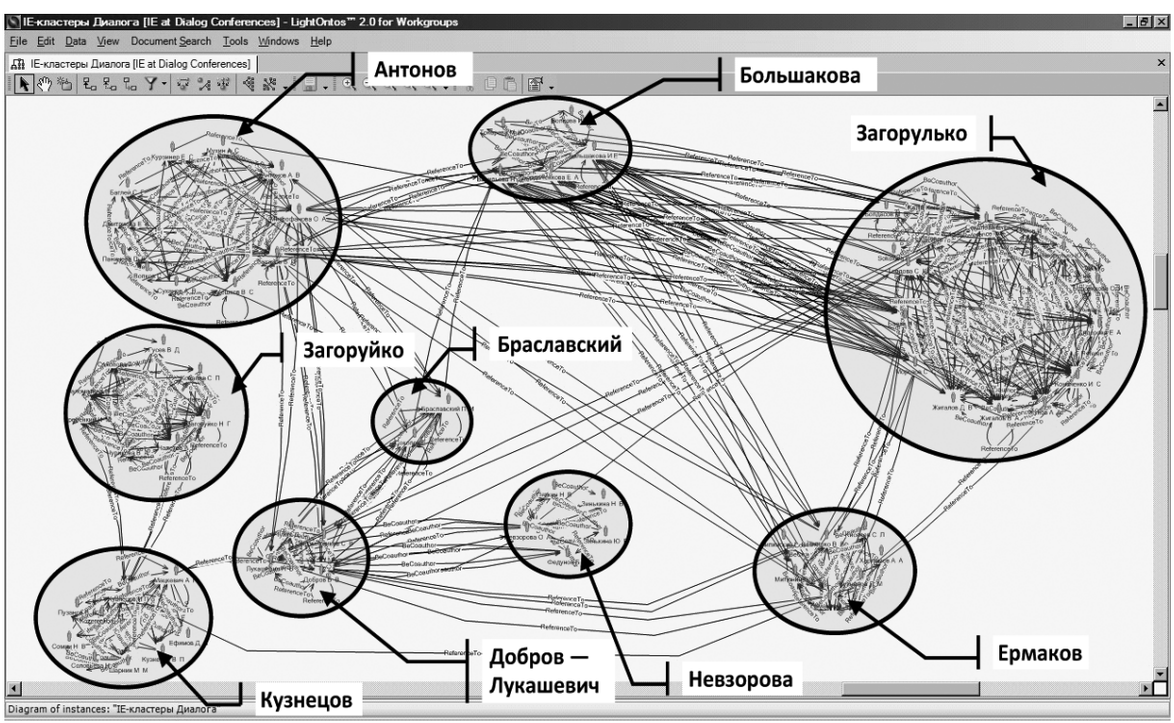

Рис. 6. Диаграмма потенциальных авторитетов и активных IE-кластеров[11]

---

[10] Точечной обводкой выделены известные коллективы, которые, как представляется, занимаются смежными и/или частными, по отношению к общей проблеме извлечения информации из текстов, исследованиями, пунктирной — потенциальные фокусы «скрытых» коллективов (авторитеты в области IE), а сплошной — коллективы, активно работающие в этой области.

[11] Пунктирной обводкой, как и раньше, выделены потенциальные авторитеты, а сплошной — наиболее активные кластеры, которые для удобства дальнейшего обсуждения связаны с фамилиями их лидеров.

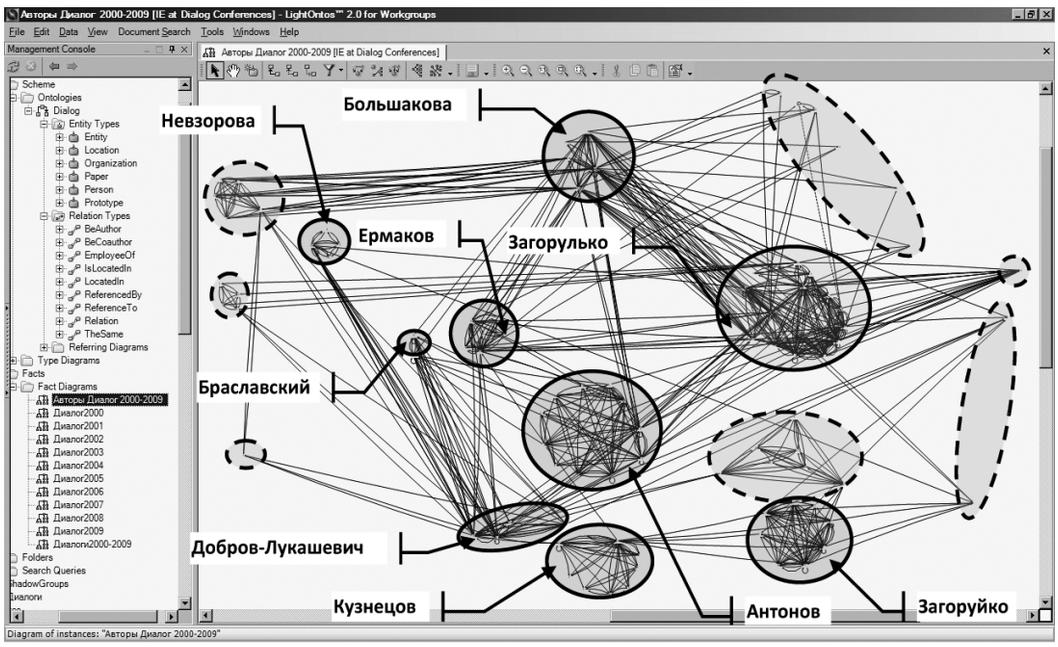

Рис. 7. Диаграмма взаимодействия активных IE-кластеров

Анализ перечисленных выше кластеров и работ российских специалистов, в них представленных, показывает, что в конференциях серии Диалог представлены только 3 коллектива (ИПИ РАН, RCO и ВМиК МГУ), основная деятельность которых связана с разработкой и реализацией систем типа IE, и 2 коллектива (АНО Центр информационных исследований и РосНИИ искусственного интеллекта), которые, скорее, можно отнести к области онтологического инжиниринга и, в частности, использования его результатов для извлечения информации из текстов. Подробное обсуждение исследований и разработок, которые ведутся в этих коллективах выходит за рамки настоящей статьи, но в целом можно констатировать, фронт этих работ явно недостаточен для решения такой важной и сложной проблемы как извлечение информации из текстов.

5. Несколько замечаний вместо заключения

В настоящей работе дан ретроспективный анализ публикаций конференций серии Диалог за 2000–2009 г. г. в области извлечения информации из текстов. Представлены статистическая и семантическая картины публикаций этих конференций в области извлечения информации из текстов.

Как показал проведенный анализ, в рамках конференций серии Диалог исследованиям и разработкам в этой области всегда уделялось определенное внимание. Вместе с тем, детальный анализ публикаций позволяет высказать несколько замечаний по поводу ситуации, которая сложилась в рамках исследований и разработок по извлечению знаний из ЕЯ-текстов в нашей стране:

1. Критическая масса специалистов и организаций, активно работающих в этой области, явно недостаточна по сравнению с общемировыми тенденциями, а география исследований и разработок ограничена традиционными центрами.
2. Теоретические исследования интересуют российских специалистов существенно больше, чем использование их результатов для создания практически значимых систем.
3. Практические разработки, как правило, доводятся лишь до уровня прототипов, а в публикациях отражаются лишь их отдельные аспекты, которые не дают возможности оценить уровень соответствующих систем.
4. Российские авторы публикаций в этой области лучше знают работы зарубежных коллективов, чем отечественных.
5. Практически все российские авторские коллективы «грешат» автоссылками, что, в конечном счете, показывает отсутствие признанных авторитетов и, как следствие, «скрытых коллективов» в этой области.

Учитывая все вышесказанное, представляется, что для российских специалистов в области извлечения информации из текстов время полномасштабного и скоординированного развертывания исследований и разработок в этой важнейшей области компьютерной лингвистики и интеллектуальных технологий еще впереди. И конференции проекта Диалог, как и конференции других смежных направлений, могут и должны здесь сыграть важную роль.

Литература

1. *Труды* международных конференций по компьютерной лингвистике и интеллектуальным технологиям: «Диалог 2000» — «Диалог 2009». http://www.dialog-21.ru
2. *Хорошевский В. Ф.* OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов. // Труды конференции КИИ-2004, Тверь, Россия, 2004.
3. *Минор С. А., Старостин А. С.* Ontos: технология извлечения знаний из неструктурированных текстов и семантическое индексирование. // Доклад на международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007». (Бекасово, 30 мая — 3 июня 2007 г.).
4. *Efimenko I., Hladky D., Khoroshevsky V., Klintsov V.* Semantic Technologies and Information Integration: Semantic Wine in Media Wine-skin. // Proc. of the 2nd European Semantic Technology Conference (ESTC2008), Vienna, 2008.
5. *Cunningham H., Maynard D., Bontcheva K., Tablan V.* GATE: an Architecture for Development of Robust HLT Applications // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.
6. *Апресян Ю. Д.* Избранные труды, том II . Интегральное описание языка и системная лексикография. // М., 1995.
7. *Иорданская Л.* (1992) Коммуникативная структура и ее использование в системе текстовой генерации // Международный форум по информации и документации. Т. 17, № 2.
8. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл — Текст» // М .: Наука ,1974 .
9. *Нариньяни А. С.* Автоматическое понимание текста — новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. — Москва, 1997.
10. *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы // Избранные труды. Новосибирск: ВО «Наука», 1994.
11. *Gruber T. R.* A translation approach to portable ontologies // Knowledge Acquisition, 1993, V. 5(2), P. 199–220
12. *Sowa, J. F.* Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA. 2000 .
13. *Вежбицка А.* Метатекст в тексте // Новое в зарубежной лингвистике. Вып. VIII . М.: Прогресс, 1978.
14. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем. Учебник. // СПб.:Питер, 2001.
15. *Lin D.* Using syntactic dependency as local context to resolve word sense ambiguity // Proceedings of the 35th annual meeting on Association for Computational Linguistics. Madrid , Spain , 1997.
16. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. // М., "Русский язык", 1977.
17. *Большаков И. А.* Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ сер. 2. 1994. — N 1.
18. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004 («Верхневолжский», 2–7 июня 2004 г.) / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. М.: Наука, 2004.
19. *Kononenko I., Kononenko S., Popov I., Zagorul'ko Yu.* Information Extraction from Non-Segmented Text (on the material of weather forecast telegrams). // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2, 2000.
20. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.
21. *Загоруйко Н. Г., Налетов А. М., Гребенкин И. М.* На пути к автоматическому построению онтологии. // Труды конференции Диалог-2003.
22. *Кузнецов И. П.* Семантические представления. // М. Наука. 1986. 290 с.
23. *Кузнецов В. П., Мацкевич А. Г.* Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста. // Труды международного семинара Диалог-98 по компьютерной лингвистике и ее приложениям. Том 2. Казань 1998.
24. *Ермаков А. Е.* Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Труды международной конференции Диалог'2007 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2007.
25. *Добров Б. В., Лукашевич Н. В.* Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'06, Бекасово, 31 мая — 4 июня 2006 г., 2006.
26. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции», 2003.

27. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 — М.: Наука, 2002.
28. *Добров Б. В., Лукашевич Н. В., Невзорова О. А., Федунов Б. Е.* Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. М.: 2004. № 6.
29. *Ермаков А. Е., Плешко В. В., Митюнин В. А.* RCO Pattern Extractor : компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. М.: 2003.
30. *Антонов А., Курзинер Е.* Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии, Труды Международного семинара Диалог-2002.
31. *Баглей С. Г., Антонов А. В., Мешков В. С., Суханов А. В.* Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006». М.: 2006.
32. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
33. *Большакова Е. И., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит, 2006. Браславский П. И., Соколов Е. А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2006». М.: 2006.
34. *Браславский П. И., Соколов Е. А.* Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
35. *Браславский П. И., Соколов Е. А.* Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.