УДК 004.9

Ю. С. Акинина, **А. А. Бонч-Осмоловская**, **И. О. Кузнецов**
В. П. Клинцов, **С. Ю. Толдова**

Научно-образовательный центр Высшей школы экономики
ул. Мясницкая, 20, Москва, 101000, Россия

E-mail: Julis5@yandex.ru

РОЛЬ ОБЩЕЙ И СПЕЦИФИЧЕСКОЙ ЛЕКСИКИ ПРИ ИЗВЛЕЧЕНИИ ИНФОРМАЦИИ ИЗ ТЕКСТА НА ПРИМЕРЕ АНАЛИЗА СОБЫТИЯ «ВВОД НОВЫХ ТЕХНОЛОГИЙ»

Рассматриваются подходы к выделению ключевых слов, необходимых для автоматического извлечения информации о фрейме и его участниках. В частности, на примере фрейма «Ввод новый технологий» исследуется ситуация, когда один и тот же фрейм связан с разными лексическими единицами в разных областях знаний. В работе оценивается вклад общей и специфической лексики в представление фрейма в конкретной предметной области.

Ключевые слова: автоматический анализ текста, извлечение информации, фреймовая модель события, компьютерная лингвистика.

Введение

Настоящее исследование проводилось в рамках подготовки решения [1] для программно-аналитического комплекса, включающего автоматическое извлечение информации о компаниях-инноваторах в разных отраслях экономики. Такой мониторинг в настоящее время является весьма востребованным в аналитических отделах, занимающихся развитием бизнеса и маркетинговым анализом.

Целью исследования является оценка эффективности методов извлечения информации об актантах фрейма «Ввод новых технологий» из текстов различающейся тематики. Под фреймом в данном случае понимается абстрактная концептуальная структура, позволяющая представить отношения участников в рамках события определенного типа (подробнее об этом см. ниже). Существенная особенность фрейма «Ввод новых технологий» состоит в том, что он применим практически ко всем областям экономической деятельности, однако на лексическом уровне может иметь очень специфическое выражение. Так, в примере 1 событие ввода новых технологий не выражается эксплицитным образом с помощью универсальных ключевых слов, но описывается на уровне специализированной лексики. Иначе говоря, вывод о том, что в данном случае речь идет о новых технологиях может сделать только эксперт:

[1] Работы над программно-аналитическим комплексом выполняются по субсидии, предоставленной по Постановлению № 218 Правительства Российской Федерации «О мерах государственной поддержки развития кооперации российских высших учебных заведений и организаций, реализующих комплексные проекты по созданию высокотехнологичного производства».

(1) Компании «Пилот» и «Транзакционные системы» выпускают комплексное решение для автоматической обработки международных платежных карт на POS-системах.

Задача исследования состояла в том, чтобы выяснить, насколько качество выделения данного события и одного из его участников (инициатора ввода новых технологий) зависит от учета отраслевой терминологической специфики текстового наполнения.

В рамках исследования были проведены несколько экспериментов. Первый эксперимент оценивал, насколько успешно происходит извлечение актора-инноватора при обучении универсальной лексике со значением инноваций. Во втором эксперименте использовался экспертный специфический список слов для одной из отраслей. В третьем – были использованы оба списка слов.

Результаты экспериментов показали, что использование экспертного списка не влияет существенным образом на качество извлечения актанта-инноватора.

Материалы и методы исследования

Онтологические и концептуальные проблемы. Понятие фрейма (сценария) используется в лингвистике для анализа структуры событийного ряда. В настоящем исследовании мы опираемся на понимание фрейма, используемого в проекте Framenet (подробнее об этом см.: [1; 2]. Под фреймом понимается концептуальная структура, определяющая некоторый тип события, его участников и их свойства. Элементы фрейма (участники события) противопоставлены по качеству своего участия в событии. Для описании функции элементов фрейма используется понятие семантической роли – каждый участник имеют одну и только одну семантическую роль в событии.

Фрейм «Ввод новых технологий» включает предикат, который вводит событие инновации, актора-инноватора (название компании или организации, осуществляющей ввод новых технологий) и объект (технологию) внедрения.

Детальный анализ текстов показывает, что роль агентивного участника может быть разной: ситуации, которые можно обозначить как «ввод новой технологии», на самом деле неоднородны по своей структуре, и агенты-инноваторы могут выполнять разные функциональные действия. Компании, которые можно считать инноваторами, могут принадлежать к следующим типам:

1) компания, которая вводит на своем предприятии некую новую технологию собственного изобретения;

2) компания, которая вводит на своем предприятии новую технологию, уже существующую на рынке;

3) компания, которая производит новую технологию;

4) компания-подрядчик, которая специализируется на внедрении и поддержке новой технологии на другом предприятии.

Другая проблема, с которой сталкивается разработчик онтологической структуры фрейма, – *неопределенность понятия* «*новая технология*». Во-первых, технология может быть современной – новой для рынка в целом – либо широко распространенной, давно существующей на рынке, но новой для конкретной компании. Во-вторых, расплывчато само понятие технологии: неясно, например, как квалифицировать новое оборудование, закупленное компанией, либо новые услугу или сервис, требующие технологических разработок. По итогам консультации с экспертами-экономистами было принята наиболее абстрактная трактовка фрейма, включающая в понятие новых технологий все вышеперечисленное. Класс компаний-инноваторов не делился на подклассы.

Для решения поставленной исследовательской задачи требуется также определить допустимую степень представленности фрейма в предложении. Под инновационным событием рассматривались факты, т. е. сообщения о событиях, позитивных с точки зрения истинностной шкалы и имеющих референтного агентивного актанта. Такие предложения, как (2), в которых агентивный актант не выражен, не рассматривались как релевантные.

(2) Если говорить о тенденциях, то идет планомерное внедрение новых технологий.

Материалы исследования. Целью представляемого исследования являлась оценка эффективности разных методов извлечения названий компаний-инноваторов в текстах разной те-

матики. Для экспериментов были выбраны четыре отрасли – банки, ритейл, электроэнергетика и строительство. В качестве материала использовались тексты новостей, извлеченные из специализированных интернет-ресурсов. Из них были составлены четыре корпуса объемом около 400 документов каждый: было отобрано примерно 200 текстов с новостными сообщениями о вводе новой технологии в каждой из выбранных отраслей, в 200 остальных содержались другие новости различных типов, выбранные случайным образом. Затем корпуса были размечены с помощью системы GATE [3]: тексты, описывающие факт введения новой технологии, отмечались особым тэгом (аннотацией), в атрибуте которого указывалось название организации-инноватора. При этом у документа создавалось столько аннотаций, сколько организаций-инноваторов было упомянуто в тексте. В неочевидных для неспециалиста случаях (например, считается ли определенное оборудование, установленное в процессе реконструкции, принципиально новым для некоторого промышленного предприятия) решение принимал эксперт-экономист.

Полученная разметка считалась эталонной и в дальнейшем использовалась для сравнения с результатом работы лингвистического процессора.

Лингвистический процессор. Данное исследование осуществлялось с помощью лингвистического процессора ONTOSMINER [4], использующего в основном эвристический подход к анализу языковых данных. ONTOSMINER позволяет создавать и применять шаблонные правила, а также подключать словари для автоматической обработки текста. В результате обработки текст получает аннотации стандартного вида, которые могут быть сопоставлены с эталонным корпусом при помощи автоматического модуля сравнения аннотаций AnnotationDiff [5] в системе GATE.

Описание и результаты экспериментов

Эксперимент 1: *обучение с помощью общей лексики.* Первый эксперимент оценивал эффективность использования общей неспецифицированной лексики для выделения компаний-инноваторов в текстах четырех тематик: электроэнергетика, строительство, банки, ритейл. Для выделения лексики были использованы обучающие корпуса, включающие одинаковое количество текстов четырех областей. В текстах были размечены компании-инноваторы – акторные актанты фрейма «Ввод новый технологий». Далее рассматривался контекст каждого из упоминаний компаний-инноваторов. Контекст ограничивался рамками одного предложения. Мы исходили из предположения, что слово, которое часто встречается в контексте ввода новых технологий, независимо от тематической принадлежности текста можно считать маркером *инноваторства*.

Таким образом, для каждого из корпусов текстов были получены списки частотности слов, употребленных в контексте компании-инноватора. Слова были предварительно нормализованы. Далее из списков исключались те слова, которые имеют высокую частоту употребления в русском языке в целом, а также те, которые встречаются только в одном списке (т. е. специфичные для определенной отрасли). В итоге из четырех списков полуавтоматическим образом был составлен общий список, содержащий такие слова, которые часто встречаются в текстах всех тематик, но при этом не являются частотными для русского языка в целом (табл. 1).

Как можно заметить, не все из этих слов обладают одинаковой различительной силой. Так, слово *инновационный* является хорошим маркером новых технологий, и на основании его употребления в отдельно взятом контексте мы можем предположить, что с большой вероятностью речь идет о вводе новых технологий. С другой стороны, слова *технический*, *проект*, *внедрение* по отдельности не позволяют вынести однозначного суждения о природе описываемой в тексте ситуации.

В рамках эксперимента мы разделили ключевые слова на *сильные* и *слабые*. Сильное ключевое слово является маркером инноваций само по себе. Слабое же может свидетельствовать о факте ввода новых технологий лишь в комбинации с другими слабыми или сильными ключевыми словами. К сильным ключевым словам относятся, например, *разрабатывать*, *изобретение*, *оптимизация*, *ноу-хау*. К слабым – *презентовать*, *запустить*, *впервые*, *решение*.

Таблица 1

Список слов маркеров инноваторства
с указанием их частотности

Слово	Частотность
Новый	363
Модернизация	359
Первый	242
Оборудование	172
Установка	169
Технология	158
Реконструкция	141
Система	117
Инновационный	110
Реализация	108
Технический	104
Проект	103
Роснано	96
Цифровой	79
Внедрение	79
Перевооружение	61
Современный	61

В зависимости от класса ключевым словам приписываются разные *веса*. Далее веса суммируются внутри каждого предложения, и если результирующий вес оказывается выше заданного порога, предложение считается «содержащим информацию о вводе новых технологий». В результате такой классификации мы получаем набор предложений, в которых, как мы ожидаем, говорится об инновациях. Всем организациям, попавшим в маркированный контекст, приписывается статус «Инноватора». В рамках нашего эксперимента были вычислены следующие значения весов: 0,9 для сильных ключевых слов; 0,3 для слабых ключевых слов. Порог был установлен в значении 0,8. Таким образом, чтобы предложение было помечено как содержащее указание на факт ввода новых технологий, необходимо, чтобы в нем встретилось как минимум 1 раз сильное ключевое слово или 3 и более раз – слабое.

С помощью лингвистического процессора был написан модуль, выделяющий предложения с «высоким инновационным контекстом» (т. е. таким, в котором сумма весов ключевых слов превышает заданный порог). Такие предложения рассматривались системой как предложения, содержащие фрейм «Ввод новых технологий», а компании, упомянутые в этом предложении, как компании-инноваторы.

Отдельно следует отметить, что мы стремились изолировать оценку качества работы модуля от влияния ошибок, возникающих при работе базового лингвистического процессора. В рамках выбранной модели эксперимента качество выделения фрейма «Ввод новых технологий» зависит от качества выделения объекта «Организация». Поэтому было принято решение подавать на работу модуля тексты с уже размеченными организациями и исправленными ошибками. Таким образом, результаты, представленные ниже в табл. 2, характеризуют работу правил независимо от работы процессора в целом.

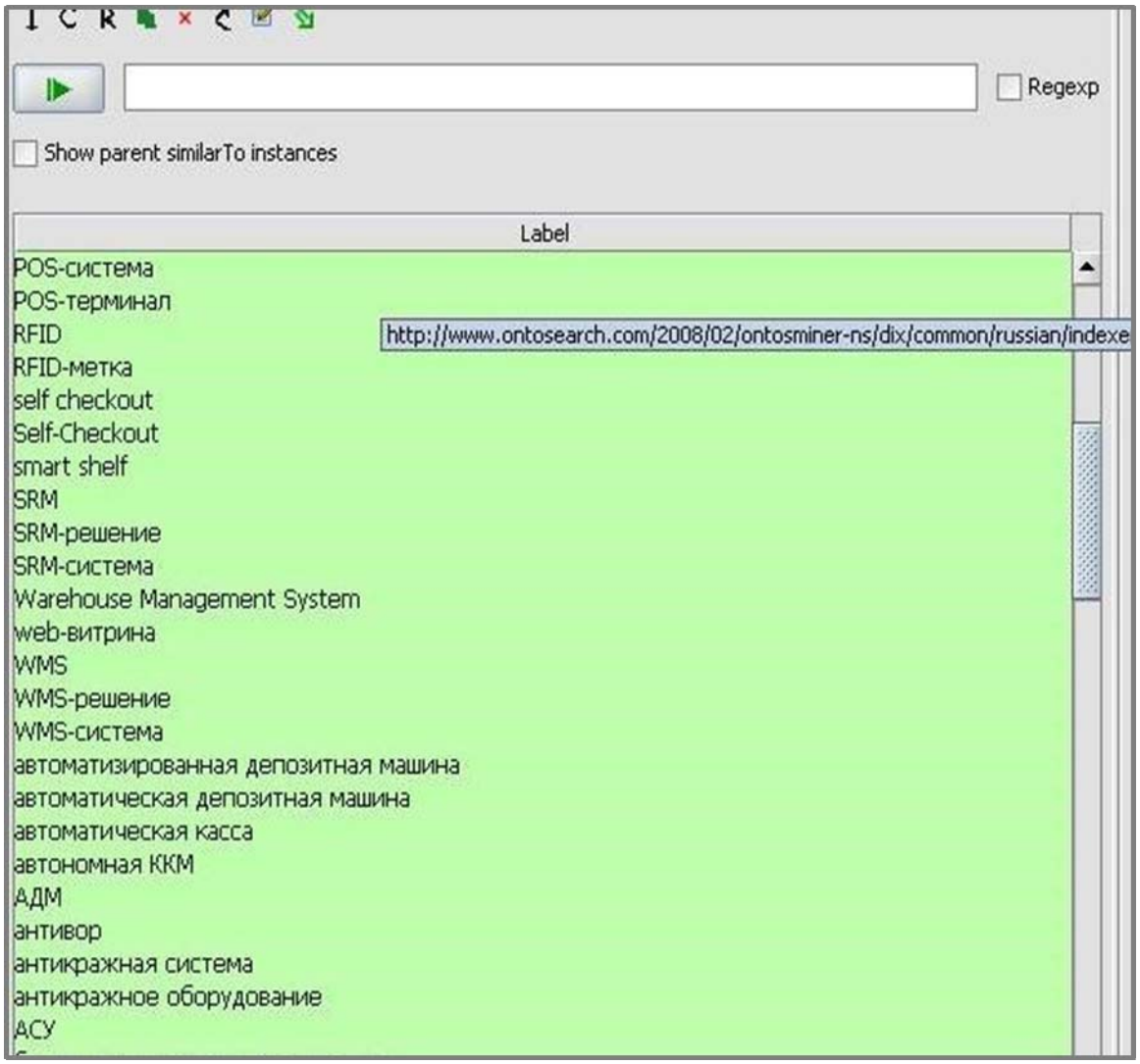
В табл. 2 представлены абсолютные результаты (correct – количество объектов, правильно выделенных системой, missing – количество пропущенных объектов, spurious – количество раз, когда объекты были выделены неверно). Эти результаты дают возможность для подсчета основных метрик оценки качества работы системы: полноты, точности и их гармонического среднего. Полнота R показывает долю выделенных компаний инноваторов из всех размеченных в тексте. Точность P показывает долю правильно выделенных компаний-инноваторов. F-мера является гармоническим средним R и P и высчитывается по следующей формуле:

$$F = \frac{2PR}{P+R}.$$

Таблица 2

Результаты оценки качества извлечения компаний-инноваторов
с помощью общих ключевых слов в тематически разных корпусах

Ключевое слово	Correct	Missing	Spurious	*R*	*P*	*F*
Ритейл	333	92	77	0,78	0,81	0,79
Банки	304	80	174	0,79	0,64	0,71
Строительство	169	94	62	0,64	0,73	0,68
Электроэнергетика	327	93	261	0,56	0,78	0,65



Пример страницы экспертного списка в области новых технологий в ритейле

Как видно из табл. 2, итоговые результаты могут весьма сильно различаться от отрасли к отрасли, тем не менее даже самый низкий результат в области электроэнергетики имеет неплохие показатели точности. Иначе говоря, общая лексика позволяет выделять основные, хотя и не все события, описывающие ввод новых технологий.

Эксперимент 2: *обучение с помощью экспертного списка.* Гипотеза, которая проверялась с помощью второго и третьего экспериментов, состояла в том, что результаты извлечения компаний-инноваторов могут быть существенно улучшены при использовании лексики, специфической для каждой отрасли. Для второго и третьего экспериментов были выбраны тексты, посвященные ритейлу. Соответственно результаты данных экспериментов оценивают только качество извлечения фрейма «Ввод новых технологий» для тематической области ритейла.

Во втором эксперименте был использован экспертный список, состоящий из выданных экспертом названий новых технологий, используемых в ритейле (см. рисунок). Все слова экспертного списка рассматривались как сильные маркеры инноваций и получали вес, равный 0,9. Соответственно компании, встречающиеся в контексте слов экспертного списка в пределах одного предложения, признавались системой компаниями-инноваторами.

По итогам эксперимента были получены следующие результаты, отраженные в табл. 3.

Эксперимент 3: *обучение с помощью смешанного списка.* В ходе третьего эксперимента оценивалось совместное использование специфических маркеров из второго эксперимента и общей лексики, выделенной в первом эксперименте. В табл. 4 сравниваются результаты всех трех экспериментов.

Обсуждение результатов

Гипотеза о том, что использование специфической отраслевой лексики и экспертных списков может существенно улучшить результаты извлечения компаний-инноваторов, не нашла подтверждения. Как видно из табл. 4, комбинированное использование двух подходов не дает существенного улучшения показателя *F* по сравнению с результатами работы модуля общей лексики (0,82 вместо 0,79).

Действительно, с помощью комбинированного метода удается несколько повысить полноту извлечения фрейма (0,83 вместо 0,78). В то же время уровень точности остается на том же уровне, что и в первом эксперименте: комбинированный метод воспроизводит ошибки и модуля общих слов, и экспертного списка.

Результаты исследования могут быть интерпретированы следующим образом: наиболее развернутое представление события в тексте с указанием всех значимых актантов происходит с участием общезначимой лексики. Экспертная лексика сама по себе редко маркирует события, см. пример 3, где непрерывным подчеркиванием обозначена общая лексика, а пунктиром выделена лексика из экспертного списка.

Таблица 3

Результаты эксперимента с экспертным списком

Эксперимент	Correct	Missing	Spurious	*R*	*P*	*F*
1	333	92	77	0,78	0,81	0,79
2	179	250	21	0,41	0,89	0,56

Таблица 4

Сравнение результатов трех экспериментов
по методам извлечения информации о событии «ввод новых технологий»

Эксперимент	Correct	Missing	Spurious	*R*	*P*	*F*
1. Общая лексика	333	92	77	0,78	0,81	0,79
2. Специальная лексика	179	250	21	0,41	0,89	0,56
3. Два списка	360	71	82	0,83	0,81	0,82

(3) В сети магазинов товаров для дома «Уютерра» стартовал проект внедрения аналитической системы SAP BusinessObjects.

Таким образом, выводы, полученные в результате эксперимента, могут иметь и более широкую интерпретацию. Существует достаточно много областей знаний (например, спорт, медицина, экономика и др.), в которых, с одной стороны, явным образом выделяются тематические категории со своей специфической лексикой, а с другой стороны, имеется универсальный обобщенный событийный ряд, определяющий отношения между ключевыми объектами в каждой такой категории. Поэтому встает вопрос о том, в какой степени алгоритмы извлечения должны опираться на специфическую лексику каждой тематики или же на универсальную лексику, описывающую общие события. Результаты проведенного исследования показывают, что в основном эта задача может быть решена с помощью использования общей лексики. Объединение общей и специфической лексики в третьем эксперименте улучшает показатели полноты, но требует дополнительных решений, повышающих точность и блокирующих суммирование ошибок при выделении объектов.

Список литературы

1. *Goddard C.* Semantic Analysis: A Practical Introduction. Oxford University Press, 1998.

2. *Ляшевская О. Н., Кузнецова Ю. Л.* Русский фреймнет: к задаче создания корпусного словаря конструкций // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог-2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, 2009. Вып. 8 (15). С. 306–312.

3. *Cunningham H., Maynard D., Bontcheva K., Tablan V.* GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

4. *Efimenko I. V., Khoroshevsky V. F., Klintsov V. P.* Ontosminer Family: Multilingual IE Systems // 9th International Conference «Speech and Computer» SPECOM'2004. Russia, St. Petersburg, 2004. P. 716–720.

5. *Cunningham H., Maynard D., Bontcheva K.* Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science, 2011.

Материал поступил в редколлегию 10.10.2012

Yu. S. Akinina, A. A. Bonch-Osmolovskaya, I. O. Kuznetsov, V. P. Klintsov, S. Yu. Toldova

THE ROLE OF GENERAL AND SPECIFIC VOCABULARY IN EXTRACTING FACTUAL INFORMATION FROM TEXTS, THE CASE OF INNOVATION-EVENT

This paper discusses approaches to the selection of keywords, used for information extraction of event frames. In particular, the innovation event is associated with different lexical items in different areas of knowledge. The paper evaluated the contribution of general and specific vocabulary in the representation of the frame in a particular subject area.

Keywords: automatic text analysis, information retrieval, frame model of the event, computer linguistics.