УДК 004.6

И.А. Черенков, С.В. Орехов

Национальный технический университет «ХПИ», Харьков

ДОБЫЧА ДАННЫХ ИЗ ТЕКСТОВЫХ НОВОСТЕЙ НА ПРИМЕРЕ РЫНКА ПОЛИМЕРОВ

В статье рассматриваются вопросы добычи данных из текстовых новостей на основе синтаксического анализа. Сформулирована модель анализа текстовой новости. Детально рассмотрен процесс формирования множества синтаксических моделей на примере. Выделен комплексный подход добычи данных из текстовой новости на основе морфологического и синтаксического анализов.

Ключевые слова: *добыча данных, синтаксический анализ, синтаксические модели, новостной поток.*

Введение

Для реализации автоматического ценового прогнозирования на основе новостного потока необходимо разработать подход, позволяющий добывать данные из текста новости в автоматическом режиме и с высокой точностью [1]. Сложность добычи данных из текстовых новостей обусловливается характером представления информации посредством разговорного языка, что создаёт значительную неопределённость при автоматическом анализе и затрудняет как непосредственное извлечение информации, так и последующую её обработку.

Существующие подходы добычи данных из текстовой новости базируются, в первую очередь, на морфологическом анализе, в частности, извлечении данных на основе частоты вхождения слов. Реализация такого подхода предполагает формирование для конкретной предметной области множеств лексем и семантических полей [3]. Лексема – абстрактное множество словоформ одного слова. Семантическое поле – множество лексем имеющих общий семантический признак. Качество добычи данных на основе морфологического анализа определяется качеством формирования множеств лексем и семантических полей для конкретной предметной области. Рассмотренный под-

ход универсален, однако, обладает недостаточной точностью, что делает нецелесообразным его применение в автоматическом прогнозировании. Причина низкой точности такого подхода объясняется тем, что данные не хранятся обособленно в лексемах, но и определяются синтаксической структурой текста.

Постановка задачи. Выработаем подход извлечения данных из текстовой новости с применением синтаксического анализа при условии уже проведенного морфологического анализа на примере рынка полимеров. Под добычей данных из текстовой новости следует понимать определение типа события и характеризующей его информации, отражённой в текстовой новости.

Подход

Большинство существующих подходов при добыче данных из текстовой информации не предполагают проведение синтаксического анализа, довольствуясь морфологическим анализом, в частности, удельным весом слова в тексте. Для таких подходов характерна большая неточность в добываемых данных, поскольку одна и та же лексема может входить в разные множества синтаксических полей, а одно и то же синтаксическое поле, в свою очередь, может входить в разные множества полей разных типов события. Также теряется информация, хранимая в синтаксической структуре текста. В общем случае применение синтаксического анализа для текстовой информации осложняется тем фактом, что потенциальное множество синтаксических моделей слишком велико и разнообразно. Однако, множество синтаксических моделей, необходимых для анализа текстовых новостных, данных может быть существенно снижено за счёт сужения предметной области до конкретной области, а также формирования множества моделей при анализе новостных объектов исключительно для анализа названий и лидов новости, содержащих 90% важной информации о событии, необходимой и достаточной для дальнейшего автоматического прогнозирования [1, 2].

Разработка множества синтаксических моделей для анализа новостей, в качестве подготовительного этапа, требует идентификацию типов событий и характеризующих их параметров, что достигается посредством разработки онтологии конкретной предметной области (рис. 1), т.е. комплексного описания множества объектов и связей между ними.

Разработка онтологии конкретной предметной области является обязательным этапом анализа и позволяет сформировать множества семантических полей, соответствующих им лексем и семантических моделей. Упрощённая онтология событий рынка полимеров в спецификации OWL приведена на рис. 1.

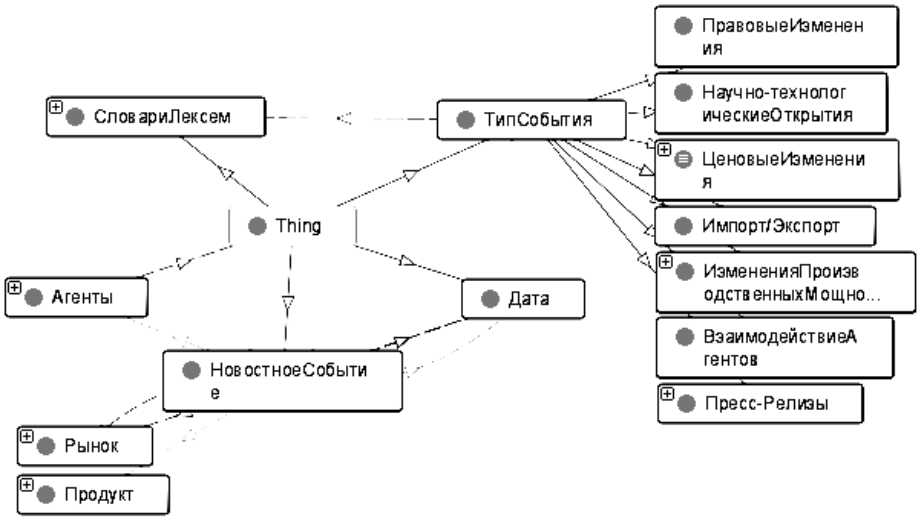


Рис. 1. Онтология событий рынка полимеров

Событие включает данные об агентах, участвующих в событии, дату события, рынок, продукт и тип события. Тип события может соответствовать одной из категорий, в частности, категории научных открытий, изменений цены, пресс релизов и т.д. На основе онтологии возможно формирование множества моделей новостей, соответствующих конкретным типам событий. Пользуясь методологией грамматик непосредственных составляющих [4], можно описать подобную модель на примере рынка полимеров рис. 2. Модель будет включать множества синтаксических моделей фраз, делящихся на глагольные фразы, фразы существительных, а также множества лексем на основе морфологического анализа и дополнительные уникальные признаки типа события.

Анализ на основе синтаксических моделей подразумевает формирование множества синтаксических моделей фраз для каждого типа события. Добыча данных из новости на основе синтаксических моделей, в рамках методологии грамматик непосредственных составляющих, подразумевает определение правильной конструкции фраз с помощью соответствующих множеств лексем в рамках семантического поля типа события, позволяющих в априори извлечь данные.

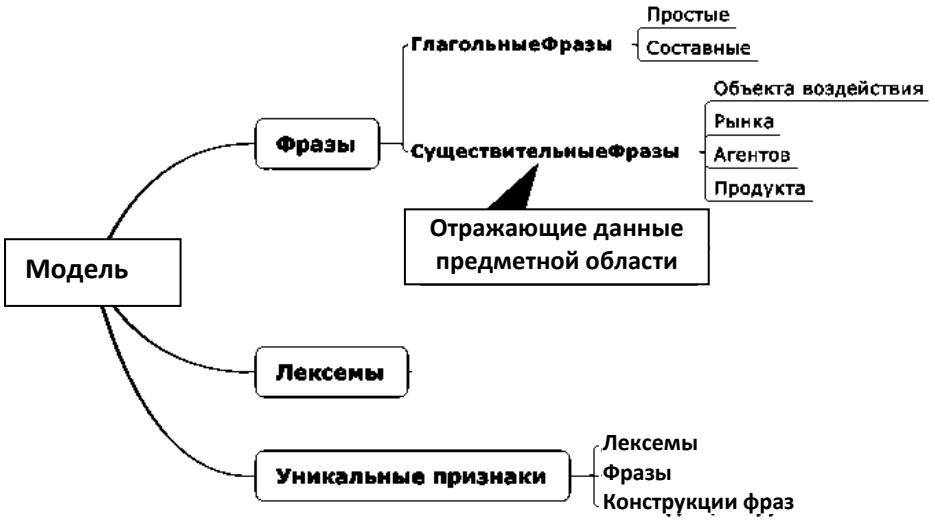


Рис. 2. Модель анализа события рынка полимеров

Рассмотрим пример формирования множества синтаксических моделей для типа событий научно-технологических изменений рынка полимеров. Обозначим гипотезу H_0 как верно подобранную синтаксическую модель новости, с помощью которой возможно извлечение данных. H_1 – модель подобрана неверно, и данные извлечь нельзя. Общая модель новости научно-технологических нововведений приведена на рис. 3, где: SP(Sentence Phrase) – предложение фраза, VP(Verb Phrase) – глагольная фраза, NP(Noun phrase) – существительное фраза. Зачастую в качестве подлежащего выступает фраза рынка (места) события, представленная в скрытом виде в качестве названия фирм, значительно реже – объекта воздействия, в таком случае сказуемое выступает в пассивном залоге.

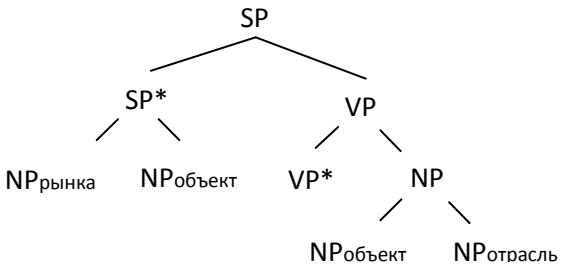


Рис. 3. Общая модель новости научно-технологических нововведений

Важной особенностью глагольной фразы является тот факт, что глагол может быть представлен чаще в прошедшем, реже в настоящем времени, т.к. научные открытия публикуются только после получения конкретного результата.

Глагольная фраза (сказуемого) представлена рис. 4 и включает случаи простой глагольной фразы из изолированного глагола и составной глагольной фразы.

Рис. 4. Глагольная фраза:
а – простая фраза, б – сложная

Сложная глагольная фраза включать только случай составного глагола.

Простая фраза рынка события состоит исключительно из существительного, обозначающего название фирмы, рис. 5. Сложная фраза рынка события включает следующие случаи: предлога в совокупности с существительным, обозначающим географический признак («в Украине») рис. 5 (б); прилагательного, обозначающего географический признак, с существительным, обозначающим субъектов научно-исследовательской деятельности («европейский учёные, разработчики») рис. 5 (с).

Фраза объекта, на который направлено воздействие в событии, зачастую неидентифицируема, поскольку отражает новое изобретение, и включает следующие случаи: простая фраза объект из одного существительного, обозначающего технологию или изобретение, словарь таких лексем заранее сформировать не представляется возможным рис. 6 (а).

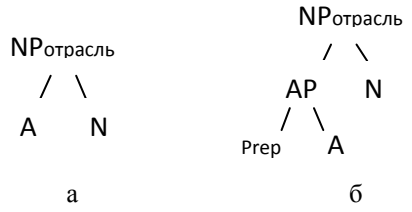
Рис. 5. Фраза существительное рынка новости:
а – простая фраза, б – сложная с предлогом, с – сложная с прилагательным

а б

Рис. 6. Фраза существительное объекта воздействия: а – простая фраза, б – сложная

Сложная фраза состоит из прилагательного рис. 6 (б), обозначающего отрасль, и существительного, обозначающего технологию.

Фраза отрасли также является трудно-идентифицируемой и присутствует только в сложной форме. Первый случай, в котором в качестве существительного идёт отраслевое название, а прилагательное его уточняет, т.е. является производным слов тех же отраслевых названий.



а б

Рис. 7. Фраза существительное отрасли, а – сложная, б – сложная с уникальным признаком

Второй случай с вложенной фразой прилагательным. В данном случае используется всегда один предлог: «для». При этом слово, отображающее отраслевой указатель, может быть как исключительно прилагательное, так и существительное. При этом структура всегда одна и та же prep+A+N рис. 7 (б).

Полученные синтаксические модели позволяют обрабатывать новости, описывающие события типа научных нововведений. Очевидно, что подобные модели должны быть составлены для каждого типа события для обеспечения возможности обработки всего новостного потока. Однако, окончательное подтверждение нулевой гипотезы при добыче данных из текстовой новости возможно при помощи уникальных признаков. В качестве уникальных признаков могут выступать (рис. 2): лексемы, фразы, уникальные символы. В рассмотренном примере к уникальным признакам можно отнести случаи изображенные на рис. 6 (б) и рис. 7 (б) и словарь лексем существительных рис. 5(c), а также обязательное отсутствие будущего времени глагола в новости.

Таким образом, общий подход по автоматической добыче данных из новостей состоит из двух ключевых этапов: построения онотологии предметной области и формирования на её основании синтаксических моделей по каждому типу события, выделения уникальных признаков модели. Алгоритм автоматической добычи данных из новости примет вид. 1. Выдвижение нулевой гипотезы о типе события. 2. Поиск лексем, входящих в семантическое поле типа события. 3. Попытка идентифицировать синтаксические модели 4. Проверить текст новости на соответствие уникальным признакам. 5. Подтвердить или опровергнуть нулевую гипотезу. 6. Извлечь данные при выполнении нулевой гипотезы.

Выводы

Предложенный подход позволяет добывать данные и текстов новостей, однако, не является гибким, т.к. формируемые модели всецело зависят от экспертного мнения, в то же время качество извлечения данных значительно выше, чем при использовании гибких алгоритмов. Т.о., применение данного подхода целесообразно в тех случаях, когда необходима наибольшая точность в добыче данных, а сами синтаксические модели, используемые в тексте, постоянны.

Список литературы

1. Черенков И.А. Прогнозирование на основе новостного потока посредством ассоциативных правил / И.А. Черенков // Энергосбережение. Энергетика. Энергоаудит. . – 2012. – №11 (105). – С. 38-42.

2. Черенков И.А. Обоснование прогнозирования цен полимеров посредством новостного потока / И.А Черенков, С.В. Орехов // Восточно-европейский журнал передовых технологий. – 2010. – № 5/7 (47). – С. 18-21.

3. Черенков И.А. Автоматический поиск данных из новостей на примере рынка полимеров. / И.А Черенков, С.В. Орехов // Системы обработки информации: –Х.: ХУПС, 2011. – Вип. 8(98). – С. 156-159.

4. Мельчук И.А. Опыт теории лингвистических моделей смысл-тектс. Семантика. Синтаксис / И.А. Мельчук. – М.: Высш. шк., 1999. – 345 с.

Поступила в редколлегию 29.11.2012

Рецензент: д-р техн. наук, проф. Е.Л. Пиротти, Национальный технический университет «ХПИ», Харьков.

ДОБУВАННЯ ДАНИХ З ТЕКСТОВИХ НОВИН НА ПРИКЛАДІ РИНКУ ПОЛІМЕРІВ

І.О. Черенков, С.В. Орєхов.

У статті розглядаються питання добування даних з текстових новин на основі синтаксичного аналізу. Сформульована модель аналізу текстової новини. Детально розглянуто процес формування множин синтаксичних моделей на прикладі. Виділений комплексний підхід добування даних з текстової новини на основі морфологічного та синтаксичного аналізів.

Ключові слова: *добування даних, синтаксичний аналіз, синтаксичні моделі, новинний потік.*

NEWS DATA MINING BASED ON EXAMPLE OF POLYMER MARKET

I.A. Cherenkov, S.V. Orekhov

Subject of this paper is data mining of text news based on syntactic models. The model was created for analysis of textual news. The process of forming a set of syntactic models was showed in detail on example. Complex approach was proposed of textual data mining news on the basis of morphological and syntactic analyzes.

Keywords: *data mining, parsing, syntactic model, the news flow.*