УДК 004.89

МОДЕЛИ ИСПОЛЬЗОВАНИЯ И ИНТЕРПРЕТАЦИИ ОЦЕНОЧНОЙ ИНФОРМАЦИИ В ПРОГНОЗИРОВАНИИ: ВРЕМЯ, СОСТОЯНИЕ, ВЕРОЯТНОСТЬ

И. В. Ефименко (*iefimenko@hse.ru*)
НИУ «Высшая школа экономики», Москва

В докладе обсуждаются вопросы использования лингвистических шкал в прогнозировании, в частности, вопросы анализа и интерпретации (квантификации) оценочной информации, представленной в текстовых документах через явную или имплицитную апелляцию к лингвистичекой шкале. Представлены модели, разработанные для анализа временных горизонтов, вероятностных оценок и характеристик объектов научно-технического развития.

Введение

Данная работа реализована в рамках государственного контракта № 07.524.12.4018 на выполнение опытно-конструкторских работ по теме «Исследование и разработка моделей долгосрочного технологического прогнозирования и программного комплекса *Интерактивная дорожная карта с обратной связью*» (далее ПК ИДК) по заказу Министерства образования и науки Российской Федерации.

В рамках решения общей задачи автоматизации построения дорожных карт в качестве одной из ключевых подзадач можно назвать выявление и систематизацию основных типов оценочной информации в документах, анализ которых значим для проведения прогнозных исследований. Фрагменты текста, относящиеся к такого рода информации, содержат оценки – явные или имплицитные, формальная интерпретация которых является важной с точки зрения:

- Непосредственно проведения прогнозных исследований (например, в качестве инструментов поддержки экспертов);
- Интерпретации результатов исследований и экспертных мнений.

В качестве базы для проведения настоящего исследования и разработки моделей интерпретации и квантификации оценочной информации были использованы коллекции документов различных

жанров на английском и русском языках: материалы российских и зарубежных форсайт-проектов, проектов в научно-технической сфере, российских и зарубежных конференций, отчеты международных организаций (UN, EU, OECD, CERN, BRICS и др.) и отдельных стран, аналитические отчеты, новости, блоги, RSS-ленты, научные статьи, патенты и ряд других типов документов.

1. Систематизация оценочной информации в прогнозных исследованиях: основные типы оценок

На верхнем уровне целесообразно выделить следующие классы информационных объектов, сведения о которых представлены в значимых для прогнозирования документах в неформализованном виде (при этом такого рода сведения могут быть преобразованы к виду, предполагающему отображение данных на метрические и неметрические шкалы различного типа): временные горизонты; вероятностные оценки (вероятность «в житейском смысле», правдоподобие, plausibility); оценки других типов (хорошо/плохо, быстро/медленно, сильно/слабо, много/мало и др.) для описания характеристик объектов, состояние которых может изменяться в процессе научно-технического развития (НТР).

Смежной областью исследования является анализ явных или имплицитных положительных/отрицательных оценок, в т.ч., комбинируемых с оценками одного из указанных выше типов. Частично методы и задачи анализа имплицитного оценочного компонента обсуждаются в разделе, посвященном квантификации оценок.

2. Время и вероятность

В результате проведенных исследований было принято решение о нецелесообразности на настоящем этапе выполнения полномасштабного временного анализа [Ефименко, 2007].

Временной анализ выполняется с учетом ряда исходно заданных ограничений. В частности, анализируются только лингвистические маркеры, относящиеся к будущему времени, некоторые из которых могут одновременно являться маркерами вероятностных оценок. Отображение временных оценок, присутствующих в тексте, на календарную шкалу, выполняется следующим образом:

1. Вводится супертип объекта «Временной горизонт», для которого задаются подтипы «Краткосрочная перспектива», «Среднесрочная перспектива», «Долгосрочная перспектива»;
2. На основе анализа текстовых коллекций формируется лингвистическая онтология оценок временных горизонтов (слова, словосочетания и другие лингвистические средства – «уже завтра», «in less than a year», «in the medium term» и т.п.);

3. На основе интервьюирования носителей русского/английского языка (для интервьюирования по материалам английского языка были приглашены также русскоязычные респонденты с высоким уровнем владения английским как иностранным) объекты лингвистической онтологии распределяются по трем подтипам объекта «Временной горизонт»;
4. Разрабатывается онтология «События НТР» (*Внедрение новой технологии*, *Выход на рынок нового продукта* и т.п.);
5. На основе интерьвюирования экспертов в предметной области для каждого типа событий определяется период времени на календарной шкале («в течение ближайших 3 лет», «от 5 до 10 лет», «после 2030 года» и т.п.), соответствующий краткосрочной, среднесрочной или долгосрочной перспективе;
6. Выполняется отнесение анализируемого события, для которого в тексте представлен лингвистический маркер временной оценки, к одному из заданных в онтологии «События НТР» типов.
7. Выполняется отображение маркера временной оценки на календарную шкалу с учетом типа события и временного горизонта, фиксируемого с использованием данного маркера.

Шаги 1-5 выполняются на подготовительном этапе (на этапе разработки); 6-7 – на этапе решения конкретной пользовательской задачи. При этом шаги 1-3 и 7 являются независимыми от предметной области и выполняются без учета ее специфики. Для шагов 4 и 6 влияние предметной области незначительно. Основное внимание особенностям предметной области уделяется на этапе интервьюирования экспертов.

Для вероятностных оценок алгоритм отображения на шкалу (от 0 до 1) в целом аналогичен. Одно из основных отличий состоит в том, что шкала вероятности, в отличие от временной шкалы, является закрытым интервалом. При интервьюировании носителей языка используется несколько различных вариантов деления шкалы вероятности на отрезки.

3. Анализ характеристик объектов научно-технического развития

3.1. Предварительные замечания

В качестве одной из важнейших задач при разработке ПК ИДК и других программных решений, применимых в прогнозировании, был выделен анализ неформализованной оценочной информации, описывающей характеристики объектов, состояние которых может изменяться в результате НТР. Основными подзадачами являются:

– Выявление в результате анализа корпуса текстов основных характеристик такого рода объектов, их спецификация и разработка их онтологических моделей, учитывающих, в т.ч.,

зависимость/независимость характеристики от предметной области дорожного картирования;
- Разработка методов автоматизированного преобразования оценочной информации к квантитативному виду.

3.2. Основные супертипы объектов НТР

В части первой из указанных выше подзадач были выделены два основных супертипа объектов, в отношении которых предполагается проводить анализ оценочной информации в текстах, основанный на использовании лингвистических шкал:
- Потребности (человека, общества);
- Собственно объекты НТР (направления исследований и разработок, технологии, продукты, рынки и т.п.), т.е. объекты, которые создаются или трансформируются в результате НТР.

Каждый из супертипов является основой для самостоятельной онтологической модели. Соответствующие онтологии взаимосвязаны между собой, т.к., во-первых, существует объективное взаимное влияние потребностей человека и общества и объектов НТР, во-вторых, одни и те же концепты (например, природные ресурсы) могут выступать как в роли объекта НТР, так и в качестве потребности.

В результате исследования было принято решение о целесообразности анализа любой присутствующей в текстах оценочной информации для объекта «Потребности» (например, фиксации недостатка некоторого ресурса), которая в терминах лингвистических шкал может быть представлена как фиксация точки или отрезка на лингвистической шкале. Примером соответствующего фрагмента текста является следующий: «The US Export-Import Bank is to sign a $2 billion deal with South Africa to fund a green energy scheme in the *electricity-short country*». Для объектов НТР в рамках разрабатываемой модели исследуются случаи изменения характеристик объектов, которые соответствуют фактам движения точки, фиксирующей значение характеристики, по лингвистической шкале.

Как показывает исследование документов указанных выше жанров, сам факт наличия в тексте сведений об изменении некоторой характеристики, которое может быть представлено как движение точки по лингвистической шкале, является одним из индикаторов значимости фрагмента текста для задач прогнозирования.

В силу имеющихся ограничений на объем статьи далее рассматривается алгоритм разработки модели в части объектов НТР. Для информационных объектов типа «Потребности» в следующих подразделах приводятся примеры в рамках описания подходов к квантификации оценочной информации.

3.3. Онтологическая модель объектов НТР

Для построения модели характеристик объектов НТР был разработан и использован следующий алгоритм:

1. Проанализирован корпус текстов различных жанров на английском и русском языках, сформированный для выбранной в качестве примера предметной области Green Energy («Экологически чистая энергия»);
2. Сформирована система лингвистических шаблонов, описывающих способы выражения фактов изменения характеристик объектов, соответствующих движению по лингвистической шкале (случаи типа «It will help … to reduce the use of diesel-powered gensets currently used to power these towers», «Using a hybrid of techniques means that at any point in time power will be drawn from the most suitable source with less wastage», «UK is "leading from the front" in a global revolution towards cleaner sources of energy» и др.);
3. Проведена повторная автоматизированная обработка корпуса с использованием разработанных шаблонов, в результате чего были обеспечены расширение набора шаблонов и их дополнительная структуризация (детализация, обобщение и др.);
4. Обработаны корпуса текстов для дополнительных предметных областей с использованием системы шаблонов, сформированной в результате предыдущих шагов;
5. В результате обработки сформирована коллекция документов, состоящих только из фрагментов текстов, где в той или иной форме представлены сведения об изменении некоторой характеристики, которое можно представить как факт движения точки по лингвистической шкале;
6. По результатам анализа полученной коллекции сформирована онтологическая модель основных характеристик, значение которых изменяется в процессе НТР. Указанная модель была сформирована с применением методов автоматизированного построения онтологий [Хорошевский, 2012] и верифицирована ведущими экспертами НИУ ВШЭ в области прогнозирования в сфере НТР. В качестве типов объектов в онтологию входят исследуемые характеристики и объекты НТР, в качестве их атрибутов – направления изменений, интерпретируемые как положительный результат (т.е. изменение значений характеристики в указанном направлении является изначальной целью субъектов НТР), в качестве семантических отношений – взаимосвязи между характеристиками и взаимосвязи характеристик с объектами НТР.

Как показали полученные результаты, зависимость от предметной области характеристик, подлежащих анализу, существует, но незначительна. Во-первых, большинство характеристик – *Стоимость*

(Затраты), Эффективность, Срок создания, Срок службы, Размер (для *Устройства/Продукта*), *Объем* (для *Ресурса*) и т.п. – являются общими для различных предметных областей. Во-вторых, характеристики, изменение которых является непосредственной целью отдельных предметных областей (например, *Уровень заболеваемости* и *Уровень смертности* для Медицины), значимы также для других предметных областей как опосредованная цель или дополнительный показатель эффективности (например, *Уровень заболеваемости* и *Уровень смертности* для предметной области «Экологически чистая энергия»).

4. Квантификация оценочной информации

4.1. Основные методы квантификации оценочной информации

Основными методами, используемыми в рамках разрабатываемой модели для квантификации оценочной информации, являются:

- Анализ коллекции документов, содержащей фрагменты текста с имплицитными оценочными компонентами для квантитативной информации (случаи типа «*only* 2.0 meters below sea level»). Использование качественной интерпретации квантитативной информации;
- Анализ коллекции документов, содержащей как фрагменты текста, где представлены оценки, выраженные количественно, так и фрагменты текста с оценками, основанными на использовании лингвистических шкал, в отношении одного и того же объекта оценивания и/или взаимосвязанных с ним объектов.

4.2. Квантификация с использованием качественной интерпретации количественной информации

В процессе квантификации информации с использованием имплицитных качественных оценок для количественной информации обеспечивается:

- Формирование гипотезы о длине шкалы для определенной характеристики определенного типа объектов;
- Разбиение шкалы на отрезки, соответствующие различной степени выраженности характеристики. Количество отрезков может устанавливаться как параметр;
- Возможность последующей частичной интерпретации оценок для соответствующих объектов и их характеристик без квантитативной составляющей («significantly deep»).

При этом предполагается, что в исходной коллекции документов, используемой для формирования гипотезы о длине и структуре шкалы, оценки без квантитативной составляющей, подлежащие интерпретации, могут отсутствовать. Таким образом, квантификация оценочной

информации с применением рассматриваемого метода происходит в соответствии с представленным ниже алгоритмом.

Предварительные шаги (на этапе разработки):

- Формирование лингвистической онтологии имплицитных оценочных компонентов (слов, словосочетаний и других лингвистических маркеров). В качестве основы для создания указанной онтологии были использованы, в частности, результаты работ в области лингвистической экспертизы и речевого воздействия (работы А. Н. Баранова и др.).
- Интервьюирование носителей языка (для интервьюирования по материалам английского языка были приглашены также русскоязычные респонденты с высоким уровнем владения английским как иностранным). Общее количество респондентов – 100 человек (люди с высшим образованием в возрасте от 21 до 80 лет, 50% мужчин, 50% женщин). Для интервьюирования был подготовлен вопросник, где были представлены 5 порядковых шкал (длина шкалы: 10, 20, 50, 100, 1000 единиц), поделенные каждая на 3 и 5 равных отрезков. Задача интервьюируемого состояла в отнесении каждого из лингвистических маркеров к одному или нескольким из отрезков для каждого из вариантов (в силу большого объема для различных групп интервьюируемых были предложены различные группы маркеров);
- Обработка результатов интервью и отображение лингвистических маркеров на отрезки порядковых шкал для каждого из вариантов (5 шкал, 2 варианта деления).

Шаги, выполняемые на этапе решения конкретной пользовательской задачи:

- Поиск в обрабатываемых коллекциях и накопление любых фрагментов текста с имплицитными оценочными компонентами для квантитативной информации (без спецификации объекта оценивания);
- Упорядочивание оценок (на основе количественных данных) для каждого из объектов оценивания, в отношении которого были найдены сведения, содержащие одновременно количественные данные и имплицитный оценочный компонент;
- Формирование гипотезы о длине шкалы с использованием модели, полученной на предварительном этапе;
- При поступлении качественной информации об имеющихся объектах оценивания (их характеристиках) – интерпретация и квантификация поступивших оценок.

Точность оценок, полученных с применением указанного метода, зависит от объема сведений о различных количественных значениях, представленных в коллекции документов. По мере накопления данных точность результатов увеличивается. В общем случае полученные

результаты могут быть использованы для предварительной оценки, а также применяться при автоматизированном формировании опросников для экспертов.

Анализ комбинации качественных и количественных оценок (второй метод), в силу ограничений на объем статьи, будет рассмотрен в докладе. Как показали полученные результаты, в качестве единицы измерения для одной и той же характеристики могут выступать самые различные единицы (не только те, которые объективно используются для измерения значений данной характеристики). Так, например, недостаток электроэнергии может быть выражен по результатам анализа текстовых коллекций в Гвт, количестве человек, испытывающих ее нехватку, проценте населения и др.

В зависимости от решаемой задачи может применяться один из методов или их комбинация.

Заключение

Областями применения полученных в настоящей работе результатов в рамках проекта «Исследование и разработка моделей долгосрочного технологического прогнозирования и программного комплекса *Интерактивная дорожная карта с обратной связью*» являются:

- Разработка методов расчета элементов ДК;
- Автоматизированное построение анкет экспертов;
- Интерпретация и консолидация экспертных мнений.

Представляется, что результаты применимы также в ряде других областей, в частности, в области экспертной журналистики.

Благодарности. Работа выполняется НИУ ВШЭ по заказу Министерства образования и науки РФ в рамках государственного контракта № 07.524.12.4018 от 16.05.2012 «Исследование и разработка моделей долгосрочного технологического прогнозирования и программного комплекса *Интерактивная дорожная карта с обратной связью*».

Список литературы

[Ефименко, 2007] Ефименко И. В. Модель времени в системах извлечения информации из письменного дискурса. Автореферат кандидатской диссертации, МГУ, Москва, 2007.

[Хорошевский, 2012] Хорошевский В.Ф., Выявление новых технологических трендов: проблемы и перспективы // Труды КИИ-2012.