OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

МОДЕЛИРОВАНИЕ ДИНАМИКИ ПРОЦЕССОВ ПОСЛЕДОВАТЕЛЬНОСТЬЮ ОДНОРОДНЫХ СЕМАНТИЧЕСКИХ СЕТЕЙ НА ОСНОВЕ АНАЛИЗА ПОСЛЕДОВАТЕЛЬНОСТИ ОПИСЫВАЮЩИХ ПРОЦЕСС ТЕКСТОВЫХ ВЫБОРОК

Харламов А.А.[*], Ермоленко Т.В. [**], Жонин А.А.[***]

[*] *Институт высшей нервной деятельности и нейрофизиологии РАН, г.Москва*

kharlamov@analyst.ru

[**] *Институт проблем искусственного интеллекта, г. Донецк, Украина*

etv@iai.dn.ua

[***] *Государственный научно-исследовательский институт информационных технологий и телекоммуникаций «Информика», г.Москва*

neurofish@yandex.ru

Описанный в статье подход к моделированию динамики процессов основан на технологии автоматического смыслового анализа текстовой информации. В процессе обработки текста формируется ассоциативная сеть, ключевые понятия которой, в том числе, лексические маркеры анализируемого процесса, ранжируются их смысловым весом. Взвешенный статусом маркера на шкале «хорошо-плохо», этот вес дает значение вклада маркера в характеристику состояния процесса. Изменение нормированной суммарной для всех маркеров характеристики процесса от временного среза к временному срезу и характеризует направленность процесса.
Ключевые слова: автоматическая обработка текста, ассоциативная (однородная семантическая) сеть, моделирование динамики процессов, лексические маркеры.

Введение

Огромный объем личной информации, в том числе, текстовой, которой делятся пользователи, отслеживается не только спецслужбами, друзьями или случайными знакомыми; за потенциальными или действующими клиентами следят банки и микрофинансовые организации, которые стремятся оценить кредитоспособность заемщика. Целью такого отслеживания является поиск и анализ сообщений по заданной тематике [Конторович, 2011].

Помимо соцсетей источниками информации для анализа могут быть блогеры, количество читателей постов которых сравнимо с аудиторией СМИ, и наконец, собственно СМИ.

Располагая инструментарием для автоматического определения эмоциональной и оценочной окраски текста, можно обследовать выборки текстов блогосферы значительного объема. Зная тематическую принадлежность или другие характеристики исследуемых текстов, можно определять, какие сегменты блогосферы связаны с выражением положительных или отрицательных оценок и эмоций. Таким образом, анализируя эмоциональную окраску последовательности (вчера, сегодня, завтра) текстовых выборок из социальных сетей, в результате можно получить динамику развития социального процесса (улучшение, ухудшение, стабильно).

Выявление в документе эмоционально окрашенной лексики и эмоциональной оценки объектов автором является основной задачей анализа тональности текста или Sentiment analysis – развивающегося направления компьютерной лингвистики. Эмоциональная оценка, выраженная в тексте, также называется тональностью, или сентиментом текста [Розин и др., 2011].

Большинство существующих методов эмоциональной оценки основано на использовании словаря эмоционально окрашенных слов [Kerstin Denecke, 2008], [Thelwall et al., 2010] и словаря символов, обозначающих эмоции. В словарных методах каждое слово обладает весом, характеризующим его эмоциональную окрашенность.

К статистическим методам выявления таких терминов относятся способы анализа текста,

реализованные на основе нейросетевых алгоритмов. Одним из хорошо зарекомендовавших себя методов статистического анализа, реализующего глобальный анализ текстов является метод на основе формализма искусственных нейронных сетей из нейроподобных элементов с временной суммацией сигналов, используемого для формирования статистического портрета текста; и формализма искусственных нейронных сетей Хопфилда, используемого для смысловой перенормировки весов ключевых понятий в тексте [Харламов и др., 2008], [Харламов и др., 2011]. Этот подход обладает достаточным быстродействием и не зависит от языка и предметной области.

В результате анализа текста из него автоматически извлекается индекс в виде сети основных понятий и их связей с весовыми характеристиками. В качестве смыслового портрета текста рассматривается не просто список ключевых слов, а сеть понятий – множество ключевых слов или устойчивых словосочетаний связанных между собой. Каждое понятие получает некоторый вес, отражающий значимость этого понятия в тексте. Связь между понятиями тоже имеет вес. Использование связей позволяет более точно взвешивать понятия текста. Для извлечения лексических маркеров используются лингвистические фильтры.

Текст как законченная мысль автора описывает некоторую ситуацию. Поэтому корректный смысловой анализ текста позволяет выявить ключевые понятия текста и их ранги (смысловые веса) в этом тексте. Если имеется множество текстов, описывающих динамику некоторой ситуации: один текст – вчера, второй – сегодня, третий – завтра. Тогда анализ каждого из этих текстов позволяет выявить ранги ключевых понятий, и состояний ситуации вчера-сегодня-завтра. Таким образом, ассоциативная (однородная семантическая) сеть, полученная на основе анализа текста, который описывает состояние ситуации в текущий момент времени, включающая в свой состав ключевые понятия текста в их взаимосвязях и с их (ключевых понятий и их связей) весовыми характеристиками (рангами), дополненная такими же сетями, полученными в предыдущие и последующие моменты времени, является моделью динамики ситуации. Объединение одноименных понятий в такой последовательности сетей (вдоль оси времени) горизонтальными связями, дополненное рангами этих понятий в соответствующие моменты времени, показывает динамику их участия в ситуации. При этом ближайшие ассоцианты ключевых понятий (понятия, отстоящие в сети от анализируемого понятия на одну связь) позволяют проинтерпретировать динамику ключевых понятий (влияние на него сопутствующих обстоятельств).

Последовательность таких семантических сетей, описывающих последовательные во времени выборки текстов, с их (сетей, следовательно, и текстов) суммарными численными характеристиками, является моделью процесса в его динамике. При этом из каждой сети фильтруется только та ее часть, которая содержит лексические и психолингвистические маркеры, характеризующие моделируемый процесс (и их ближайшие ассоцианты). Эти лексические и психолингвистические маркеры имеют свои смысловые веса в сети. Суммарные численные значения весов маркеров, взвешенных их статусом на шкале «хорошо-плохо» характеризуют состояние процесса в текущий момент времени. А в последовательности сетей возникает динамика смысловых рангов соответствующих понятий, что и моделирует динамику социального процесса: улучшение, ухудшение, стабильно.

1. Теоретические основы моделирования динамики процессов на основе анализа последовательности текстовых выборок

Предлагаемый в работе количественный анализ направленности процессов основывается на двух основных подходах.

Для анализа процесса используется автоматическое формирование семантической сети текста (корпуса текстов), содержащего информацию, касающуюся процесса. При этом автоматически извлекаемые из текста, среди прочих ключевых понятий, лексические и психолингвистические метки ранжируются в зависимости от их значимости в тексте. Суммарный ранг этих меток определяет состояние процесса (хорошо-нейтрально-плохо).

Рассматриваются последовательные срезы текстовых корпусов, с построением их семантических сетей (и ранжированием меток), что позволяет выявлять динамику исследуемого процесса как изменение суммарного ранга меток, характеризующих процесс, от среза к срезу.

Рассмотрим эти два подхода более подробно.

1.1. Формирование ассоциативной (однородной семантической) сети текста

Анализ статистики слов и их связей в тексте позволяет реконструировать внутреннюю структуру текста, и таким образом, сформировать описание семантики предметной области текста.

Статистический анализ выявляет ключевые понятия текста - слова или устойчивые словосочетания с их частотой встречаемости в тексте. Важной особенностью используемого подхода, является возможность автоматически устанавливать взаимосвязи между выявленными ключевыми понятиями текста. При выявлении связей учитывается статистика попарного появления слов в смысловых фрагментах исследуемого материала. Далее статистические показатели ключевых слов пересчитываются в семантические веса, при этом учитываются

подобные характеристики ключевых понятий с ними связанных, а также учитываются численные показатели связей.

После пересчета статистических характеристик текста в семантические, ключевые понятия, которые не релевантны структуре текста, получают малый вес, а наиболее представительные наделяются высоким рангом. Полученная семантическая сеть позволяет производить различные виды анализа текстовой информации. Сеть отражает внутреннюю структуру текста, значимость выделенных ключевых понятий, а также, показывает степень связанности понятий в тексте. Такое представление текста получается полностью автоматически.

Технология реализует следующую обработку информации:

- сегментацию слов и предложений текста на основе графематических правил;
- нормализацию грамматических форм слов и вариаций словосочетаний. Выявление корневых основ ключевых понятий;
- фильтрацию в тексте семантически несущественных, вспомогательных слов: удаляются предлоги, числительные и самые общеупотребимые слова с широким значением;
- выявление ключевых понятий текста (слов и словосочетаний), и их взаимосвязей в тексте;
- вычисление их относительной (в тексте) значимости – рангов ключевых понятий;
- формирование представления семантики текста в форме семантической сети ключевых понятий.

До собственно статистического (с элементами лингвистики) анализа текста осуществляется его первичная обработка. Задачей первичной обработки текста является подготовка его к статистическому анализу. Подготовка текста заключается в очистке его от нетекстовых символов, а также в корректной обработке таких единиц текста как аббревиатуры, инициалы, заголовки, адреса, номера, даты, указатели времени.

Сегментация предложений позволяет разбить текст на участки, которые могут содержать терминологические словосочетания предметной области и избежать выделения неадекватных словосочетаний на стыках таких участков. В результате предобработки (с использованием морфологического анализа) близкие по форме слова и словосочетания приводятся к одинаковой форме (нормализуются).

Ключевые понятия предметной области (слова и словосочетания) выделяются с использованием частотного анализа текста. В процессе формирования частотного портрета текста подсчитывается частота встречаемости слов в тексте.

Сформированное таким образом представление лексики текста подвергается затем пороговому преобразованию по частоте встречаемости. Порог отражает степень детальности описания текста. В процессе статистического анализа выделяются устойчивые термины и терминологические словосочетания, которые служат далее в качестве элементов для построения семантической сети. При этом в составе частотного портрета текста общеупотребительные слова, а также словосочетания, содержащие только общеупотребительные слова, опускаются.

Первичная (частотная) сеть формируется из выявленных на предыдущем этапе ключевых понятий за счет использования ассоциативных связей (попарной встречаемости) этих слов в смысловых фрагментах текста. В качестве критерия для определения наличия ассоциативной связи между парой понятий используется частота их совместной встречаемости в одном смысловом фрагменте текста (например, в предложении). Превышение частотой попарной встречаемости ключевых понятий некоторого порога позволяет говорить о наличии между понятиями ассоциативной (семантической) связи, а совместные вхождения понятий в предложения с частотой меньше порога считаются просто случайными.

Элементы полученного таким образом частотного портрета текста (однородной семантической) – ассоциативной – сети и их связи имеют числовые характеристики, отражающие их относительный вес в данном тексте, соответствующий частоте их встречаемости в тексте.

Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого семантического сгущения. В результате итеративной процедуры перенормировки наибольшие веса получают ключевые понятия, связанные с наибольшим числом других понятий с большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста. Полученные таким образом смысловые веса ключевых понятий показывают значимость этих понятий в тексте.

1.2. Выявление динамики процесса

Полученная сеть представляет собой семантический (структурный) портрет текста (корпуса текстов). Если текст, или корпус текстов описывает некоторую структуру (научную разработку, предметную область, социологическую ситуацию), то сформированная таким образом семантическая сеть представляет собой семантический срез этой структуры в момент написания текста.

Семантическая сеть, построенная на тексте, написанном позже, и описывающем ту же структуру, может отличаться от первой, поскольку представляет текст, релевантный состоянию описываемого процесса на момент времени более поздний, чем предыдущее. Сеть может содержать те же ключевые понятия, но может не содержать

некоторых из них, которые выбыли из описываемой структуры, а может включать в себя другие понятия, которые появились в описываемой текстом структуре за это время. И, главное, весовые характеристики содержащихся в сети понятий могут отличаться от их весовых характеристик, какие были в первой сети.

Соединим одинаковые ключевые понятия обеих сетей связями, толщина которых будет пропорциональна весу ключевого понятия. Если понятия в обеих сетях имеют одинаковый вес, связь имеет одинаковую толщину от сети к сети. Если понятия имеют разные веса, связь, их соединяющая, либо утолщается, либо утоньшается, демонстрируя динамику состояний ключевых понятий, и, таким образом, динамику состояний сети в целом.

Если мы возьмем тексты следующего временного среза, и построим еще одну сеть, и присоединим ее к двум предыдущим, то будем иметь картину разворачивания структуры (научной разработки, предметной области, социологической ситуации) во времени. И так сколько угодно шагов. Такая модель динамики процесса наглядна, удобна для исследования (сеть как статический смысловой срез исследуемой структуры представляет собой удобный для навигации по нему объект в силу ассоциативности связей между ключевыми понятиями), и обладает числовыми характеристиками, что делает ее удобной для аналитического исследования процессов, и, как следствие, удобной для автоматического анализа.

Наконец, для того, чтобы исследовать конкретный процесс в его динамике, выберем ключевые понятия семантической сети, которые являются лексическими и психолингвистическими метками этого процесса, и будем исследовать динамику развития количественных характеристик этих понятий. Как и другие ключевые понятия, эти понятия могут появляться вновь, появляться последовательно на разных временных срезах, и наконец, исчезать. Могут также меняться от временного среза к срезу их численные характеристики. То есть они ведут себя как обычные ключевые понятия в динамике.

Удалим все ключевые понятия всех сетей, кроме упомянутых меток. В этом случае оставшаяся часть модели динамики текстов становится моделью динамики исследуемого процесса. Причем, суммарная числовая характеристика оставшихся ключевых понятий сети характеризует состояние процесса в текущий момент времени, а их изменение, от временного среза к срезу, характеризует динамику процесса во времени.

1.3. Формализм подхода

Для формирования однородной семантической (ассоциатиативной) сети создается частотный портрет текста, содержащий информацию о частоте встречаемости ключевых понятий текста, представленных как корневые основы соответствующих слов, или их устойчивых сочетаний, встречающихся в тексте, а также об их совместной (попарной) встречаемости в смысловых фрагментах текста (например, в предложениях). Частотный портрет, таким образом, содержит информацию о частоте встречаемости ключевых понятий и их попарной (в терминах их ассоциативной связи) встречаемости в тексте. Использование хопфилдоподобного алгоритма позволяет перейти от частоты встречаемости к смысловому весу (вес связей при этом остается неизменным).

Эта обработка включает несколько этапов. На этапе первичной обработки из текста удаляется нетекстовая информация, текст сегментируется на слова и предложения, из текста удаляются стоп-слова, рабочие и общеупотребимые слова, а оставшиеся слова подвергаются морфологической обработке. Морфологическая обработка производится с использованием заранее подготовленного морфологического словаря (словаря флективных морфем) – словаря первого уровня - $\{B_i\}_1$. В результате формируется словарь второго уровня – $\{B_i\}_2$ – словарь корневых основ (и устойчивых словосочетаний).

На следующем этапе строится частотный портрет текста, то есть выявляются частоты p_i встречаемости корневых основ B_{i2} ключевых понятий (полученных в результате морфологического анализа) и их устойчивых сочетаний, и частоты p_{ij} их попарной встречаемости в предложениях текста. Одновременно формируется словарь третьего уровня $\{B_i\}_3$ - словарь пар слов.

На третьем этапе, частоты встречаемости перенормируются в смысловые веса с использованием хопфилдоподобной итеративной процедуры. В результате итеративной процедуры перенормировки наибольшие веса получают ключевые понятия, связанные с наибольшим числом других понятий с большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста.

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\overline{E}) \qquad (1)$$

здесь $w_i(0) = p_i$; $w_{ij} = p_{ij} / p_j$ и $\sigma(\overline{E}) = 1/(1+e^{-k\overline{E}})$ – функция, нормирующая на среднее значение энергии всех вершин сети $\overline{E}$, где p_i – частота встречаемости i-го слова в тексте, p_{ij} – частота совместной встречаемости i-го и j-го слов в фрагментах текста. В дальнейшем эти весовые характеристики корневых основ используется для выявления предложений текста,

содержащих наиболее важную информацию в тексте.

В результате получается так называемая ассоциативная (однородная семантическая) сеть N как совокупность несимметричных пар понятий $< c_i c_j >$, где c_i и c_j – ключевые понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста). Иначе семантическую сеть можно представить в виде множества звездочек $< c_i < c_j >>$, где $< c_j >$ - множество ближайших ассоциантов ключевого понятия c_i.

Под семантической сетью N. понимается множество несимметричных $< c_i c_j > \neq < c_j c_i >$ пар ключевых понятий $<< c_i c_j >>$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некоторой ситуации):

$$N \cong \{< c_i c_j >\}. \quad (2)$$

Семантическая сеть, описанная таким образом, может быть переописана как множество так называемых звездочек $< c_i < c_j >>$:

$$N \cong \{z_i\} = \{< c_i < c_j >>\} \quad (3)$$

Под звездочкой $< c_i < c_j >>$ понимается конструкция, включающая главное событие c_i, связанное с множеством событий-ассоциантов c_j, которые являются семантическими признаками главного события, отстоящими от главного события на одну связь. Связи направлены от главного события к событиям-ассоциантам.

Последовательность одноименных звездочек, принадлежащих разным временным смысловым срезам – семантическим сетям, называется элементарным процессом π:

$$\pi = z_i(t_1) \Rightarrow z_i(t_2) \Rightarrow z_i(t_3) \Rightarrow \ldots \quad (4)$$

где $z_i(t_k)$ – конкретная звездочка в момент времени t_k . Вес ключевого понятия в текущий момент времени, определяющий его ранг в семантической сети – $w_i(t_k)$.

События-ассоцианты c_j главного понятия звездочки c_i являются его семантическими признаками, и позволяют интерпретировать его содержательно на каждом шаге процесса.

2. Описание информационной модели оценки направленности процесса на основе однородной семантической (ассоциативной) сети

Информационная модель включает в свой состав:

- модуль поиска релевантной исходной информации (текстов из открытых источников);
- модуль извлечения из подготовленных текстов общей для них семантической сети;
- модуль оценки направленности анализируемого процесса.

Модуль поиска релевантных для обработки текстов осуществляет поиск по заданным источникам текстов, удовлетворяющих условиям поставленной задачи. В том числе, задается регион, для которого предполагается проведение исследований, временной промежуток $\Delta T_l = (t_{lend} t_{lbeg}), l = 1..L$, который принимается за l-й временной срез (один из L), а также лексические и психолингвистические маркеры M, характеризующие предметную область исследуемого процесса. Последние задаются экспертом, характерны строго для своего процесса (или нескольких процессов, если они исследуются совместно), и, в конченом итоге, определяют качество анализа. Зато информационная модель абсолютно не зависит от предметной области. Ей все равно, что представлять.

Процесс поиска текстов для каждого маркера $M_k, k = 1..K$ осуществляется отдельно, при заданных остальных (место и время) общих для всех маркеров параметрах поиска. Полученные на этапе поиска тексты обрабатываются (по отдельности) с формированием для каждого текста семантической сети N (см. следующий этап), исключительно для ранжирования текстов в корпусе текстов – результатов поиска относительно релевантности именно этому маркеру. Для этого в каждом тексте вычисляется смысловой вес $r_i = w_i$ заданного маркера. Он определяет релевантность текста этому маркеру.

Далее, для каждого маркера отбираются тексты, ранг которых по этому маркеру оказывается выше заданного порога $r_j \geq h_{отбора}$. Эти тексты, в совокупности по всем маркерам, составляют корпус текстов, подлежащих обработке на следующем этапе.

Модуль формирования семантической сети корпуса текстов N предусматривает несколько процедур в своем составе. Исходные тексты претерпевают предобработку, в процессе которой из них удаляется нетекстовая информация, удаляются также стоп-слова, рабочие и общеупотребимые слова, то есть слова, которые не несут смысла в этом корпусе текстов. Помимо этого, в процессе морфологического анализа, все словоформы

приводятся к своим корневым основам, чтобы увеличить достоверность результатов последующей статистической обработки

Процедура частотного анализа формирует частотный портрет текста в виде первичной ассоциаивной сети, в которой участвуют в качестве вершин оставшиеся после предобработки леммы слов, связанные между собой связями, полученными из анализа попарной встречаемости слов в смысловых фрагментах текста (например, предложениях). Как вершины сети, так и их связи имеют числовые характеристики – частоты их (вершин - p_i , и связей - p_{ii}) встречаемости в анализируемом тексте. Эти числовые характеристики нам понадобятся после перенормировки для оценки рангов наших маркеров описываемого текстом процесса.

Процедура перенормировки итеративно пересчитывает частотные характеристики вершин сети (ключевых понятий текста) в смысловые веса таким образом, что понятия (вершины сети), связанные с большим числом других понятий, увеличивают свой вес, в ущерб весам других понятий. Понятия, несущие в тексте максимальную смысловую нагрузку, становятся максимально весомыми. Они как бы стягивают на себя структуру текста. Становятся главными темами текста.

Далее, работает последняя процедура обработки информации, которая выявляет численные характеристики маркеров их суммированием для конкретного временного среза текстов, и выявляет динамику изменений полученных таким образом обобщенных характеристик процесса от временного среза к временному срезу.

На этом этапе выбранные экспертом лексические и психолингвистические маркеры исследуемого процесса M_k , которые были заданы экспертом на первом этапе работы модели, фильтруют полученную на предыдущем этапе семантическую сеть $N = \{< c_i < c_j >>\}$. Из нее удаляются все вершины, кроме понятий-маркеров c_i , а также ближайших ассоциантов маркеров – вершин семантической сети (понятий), отстоящих от понятий-маркеров на один шаг $< c_j >$, являющихся их семантическими признаками.

Каждому маркеру на каждом временном срезе ставится в соответствие его смысловой вес $w_i = 0..100$, полученный на предыдущем этапе, который становится рангом этого маркера M_i для этого временного среза l .

Для всех маркеров вычисляется произведение Π_i статуса маркера (на шкале «хорошо-плохо») $S_i = (-1,0,+1)$ на его ранг:

$$\Pi_i = S_i * r_i \qquad (5)$$

И полученные для каждого маркера M_i произведения Π_i суммируются по всем маркерам:

$$\Pi = \sum \Pi_i \qquad (6)$$

Таким образом, получается суммарная характеристика $\Pi(l)$ временного среза l оцениваемого процесса.

Далее строится график суммарных характеристик $\Pi(l)$ срезов. Или, в другом представлении, визуализируется ряд семантических сетей, включающих помимо маркеров также их семантические признаки (ближайшие ассоцианты) $c_i < c_j >$. При этом центральное понятие-маркер c_i в сети визуализируется размером, пропорциональным его рангу, цвет – соответствующий его статусу на шкале «хорошо-плохо», и все центральные понятия связываются между собой связями своего цвета вдоль оси времени.

Таким образом, на графике представляются основные тенденции каждого маркера во времени, а их ближайшие ассоцианты позволяют проинтерпретировать получившуюся картину.

3. Информационное моделирование оценки направленности процесса

В качестве примера представлено построение модели оценки направленности процесса на примере текстов по тематике внутренней политики РФ на основе новостных материалов портала newsru.com в свете отношений правительства и общества. Модель включает в свой состав:

- тексты из открытых источников, относящиеся к одной теме и разным временным срезам;
- семантические сети, построенные для корпусов текстов каждого среза;
- оценку направленности социального стресса, основанную на характеристиках лексических и психолингвистических маркеров.

Поиск релевантных текстов для обработки начинается с поиска по заданным источникам текстов, удовлетворяющих условиям поставленной задачи. Процесс поиска текстов для каждого маркера $M_k, k = 1..K$ осуществляется отдельно, при заданных остальных (место и время), одинаковых для всех маркеров, параметрах поиска. Полученные на этапе поиска тексты обрабатываются (по отдельности) с формированием семантической сети N (см. следующий этап), исключительно для ранжирования их в корпусе текстов – результатов поиска относительно релевантности именно этому маркеру. Для этого в каждом тексте вычисляется смысловой вес $r_j = w_i$ заданного маркера. Он определяет релевантность текста к этому маркеру.

Задачей моделирования оценки направленности процесса будет построение модели оценки направленности процесса социальной напряженности на примере упомянутых текстов по тематике внутренней политики РФ.

Перечень выбранных экспертным путѐм маркеров: "конфликт" (статус -2), "консенсус" (статус 1), "согласие" (статус 2), "бесконфликтность" (статус 2). Ранг отражает, с содержательной стороны, степень выраженности отношений согласия между правительством и обществом, чем выше ранг – тем более выражено согласие. С формальной вычислительной стороны ранг отражает вклад веса каждого маркера в итоговую интегральную оценку процесса социального стресса.

Тематические термины процесса: "правительство", "общество".

Временные срезы: сентябрь и октябрь 2013 года.

Выбранные тексты: выбрано 22 новостных текста за сентябрь 2013 и 17 новостных текстов за октябрь 2013. Не во всех текстах встречаются все выбранные маркеры. Надѐжность оценки выбранного среза прямо зависит от объема корпуса текстов с выбранными маркерами в данном срезе.

Далее все тексты для каждого среза объединяются в единый текст и обрабатываются модулем автоматического анализа. В результате обработки строится итоговая семантическая сеть, содержащая среди понятий маркеры, для которых были вычислены их семантические веса (ранги).

Наконец, выявляются численные характеристики маркеров, их суммированием для конкретного временного среза текстов, и выявлением динамики изменений полученных таким образом обобщенных характеристик процесса от среза к срезу.

Таблица 1 – Оценка направленности процесса

Маркер	Статус	Ранг 09.2013г.	Вклад 09.2013	Ранг 10.2013	Вклад 10.2013
конфликт	-2	65	-130	79	-158
консенсус	1	72	72	98	98
согласие	2	54	108	93	186
бесконфликтность	2	83	166	52	104
Итоговая сумма:			216		230

Итоговые суммы являются интегральной оценкой процесса социальной напряженности по данной тематике для данного временного среза. Однако бессмысленно рассматривать интегральные оценки в отрыве от динамики их изменения, поскольку сумма статусов участвующих в рассмотрении маркеров не сбалансирована и почти наверняка имеет отклонение в ту или иную сторону, что приведѐт к отклонению интегральной оценки в случайном направлении. Состав маркеров одинаков для разных временных срезов и, поэтому, разница интегральных оценок между срезами не подвержена этому эффекту - случайные компоненты вклада взаимно уничтожаются.

Среднеквадратичное отклонение интегральной оценки, вычисленной методом бутстрепа (формирование случайной подвыборки) составляет 8.7. Оценка динамики, произведенная на данной выборке с данным распределением, и с данным среднеквадратичным отклонением, является достоверной.

Разница интегральных оценок есть оценка направленности процесса социальной напряженности. Показателем качества модели является устойчивость оценки направленности относительно добавления или удаления маркера сходного с имеющимися типа. Это означает, что предпочтительны модели с большим числом оцениваемых маркеров, и большее число тематических текстов должно подвергаться анализу. Возможна "калибровка" модели с целью вычисления среднего отклонения интегральной оценки при наличии устойчивости модели (признаваемой экспертно). Для этого следует принять некоторый временный срез за "0", точку отсчета и значение интегральной оценки в этом срезе вычитать в дальнейшем из интегральных оценок прочих временных срезов. Во всяком случае, необходимо иметь ввиду, что данный пример имеет чисто технический характер демонстрации работы алгоритма.

Оценка модели: данная модель имеет несбалансированный вклад множества рассматриваемых маркеров (общий вклад статусов маркеров равен +3, средний +0.75) и на рассматриваемой выборке новостных текстов по тематике взаимоотношений общества и правительства (задаваемых ключевыми словами "правительство" и "общество") обнаруживает положительную направленность процесса социальной напряженности от временного среза – «сентябрь 2013г.» до временного среза – «октябрь 2013г.».

Заключение

Описанный в статье подход к моделированию динамики процессов основан на хорошо зарекомендовавшей себя технологии автоматического смыслового анализа текстовой информации. В процессе обработки текста формируется ассоциативная сеть, ключевые понятия которой, в том числе, лексические и психолингвистические маркеры анализируемого процесса, ранжируются их смысловым весом.

Умноженный на статус маркера на шкале «хорошо-плохо», этот вес дает значение вклада маркера в характеристику состояния процесса. Изменение суммарной для всех маркеров характеристики процесса от временного среза к временному срезу характеризует направленность процесса. Приведенный пример не демонстрирует качество обработки, а лишь иллюстрирует работу механизма оценки. Предлагаемый подход является для эксперта инструментом моделирования процесса, настраивая который, подстраивая под свои представления, под реальный процесс, он может добиться адекватности моделирования.

Работа была выполнена в рамках НИР «Исследование методов интеллектуального анализа полуструктурированной информации и информационного моделирования направленности процессов социального стресса на основе данных из открытых источников» (при финансовой поддержке Министерства образования и науки Российской Федерации по Госконтракту от 10 октября 2013г. № 14.514.11.4114).

Библиографический список

[Конторович, 2011] Конторович С.Д., Литвинов С.В., Носко В.И. Методика мониторинга и моделирования структуры политически активного сегмента социальных сетей [Электронный ресурс] / С.Д. Конторович, С.В. Литвинов, В. И. Носко // «Инженерный вестник Дона», 2011, №4. – Режим доступа: http://ivdon.ru/magazine/archive/n4y2011/642/2/1428 (доступ свободный) – Загл. с экрана. – Яз. рус.

[Розин и др., 2011] Розин М.Д., Свечкарев В.П., Конторович С.Д., Литвинов С.В., Носко В.И. Исследование социальных сетей как площадки социальной коммуникации рунета, используемой в целях предвыборной агитации [Электронный ресурс] / М.Д. Розин, В.П. Свечкарев, С.Д. Конторович, С.В. Литвинов, В.И. Носко // «Инженерный вестник Дона», 2011, №1. – Режим доступа: http://http://ivdon.ru/magazine/archive/n1y2011/397 (доступ свободный) – Загл. с экрана. – Яз. рус.

[Харламов и др., 2008] Харламов, А.А. Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды. / А.А. Харламов, В.В. Раевский // Речевые технологии, N 3, 2008. С. 27-35.

[Харламов и др., 2011] Харламов, А.А. Семантические сети как формальная основа решения проблемы интеграции интеллектуальных систем. Формализм автоматического формирования семантической сети с помощью преобразования в многомерное пространство / А.А. Харламов, Т.В. Ермоленко //Материалы международной научно-технической конференции OSTIS-2011, Минск БГУИР. С. 87-96.

[Kerstin, 2008] Kerstin, Denecke Using SentiWordNet for multilingual sentiment analysis / Denecke Kerstin // IEEE 24th International Conference on Data Engineering Workshop. 2008. Pp. 507-512.

[Thelwall et al., 2010] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. Sentiment strength detection in short informal text / M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, *A.* Kappas // Journal of the American Society for Information Science and Technology, 2010. Pp. 2544–2558.

MODELING OF PROCESS DYNAMICS BY SEQUENCE OF HOMOGENOUS SEMANTIC NETWORKS ON THE BASE OF TEXT CORPUS SEQUENCE ANALYSIS

Kharlamov A.A. *, Yermolenko T.V. **, Zhonoin A.A. ***

**Institute of Higher Nervous Activity and Neurophysiology of Russian Academy of Sciences, Moscow*

kharlamov@analyst.ru

*** Institute of Artificial Intelligence Problems, Donetsk, Ukrain*

etv@iai.dn.ua

**** State Institute of Information Technologies and Telecommunications «Informika» , Moscow*

neurofish@yandex.ru

The represented approach of dynamic process modeling is based on the technology of automatical semantic text analysis. An associative network is forming during text processing. Its key notions, including lexical and psycholinguistic markers of the analyzed process, are ranked by theirs semantic weights. The weight being multiplied by marker status value at the scale of "good-bad" gives its contribution to the process stage characteristic. Transformation (dynamics) of the accumulated for all of the markers process characteristics from one period of time to another one characterizes a direction of the process.

Introduction

Process dynamics modeling (social process for example) needs to form the process model. This model is usually formed by hand. Automatization of the process could cut the cost of the model. Just now everybody can extract the information about sentiments from texts about analyzed process. There is two more steps to model of the process dynamics. We need the instrument which automatically shows the process statics. And then we must to show the results on the timeline.

Main Part

Associative (homogeneous semantic) network is such an instrument. Its nodes are the main concepts of analyzed text – lexical and psycholinguistic markers of the process. Also the network includes weight characteristics of the concepts and their relationships. This weights are characterized rank of the concepts in the text. The network can be constructed automatically by TextAnalyst program instrument.

The sum of such concept ranks weighted by their status (at the scale "good-bad") characterizes quantitatively the order of the process at the given time period. The process markers are selected by expert. The expert also fixes their status at the scale "good-bad".

Such value sequence which extracted automatically at the base of text corpus characterizes analyzed process dynamics at the timeline (yesterday-today-tomorrow).

Conclusion

In the paper the approach to automatical modeling of process dynamics at the base of automatical text corpus analysis for the sequential points at the timeline are shown. Formalism of the approach on the base of automatical semantic text analysis and algorithm of process dynamic analysis are shown also. The algorithm action is explained by example.