В.В. Борисов, С.Н. Андреев, Я.А. Федулов

Анализ сложных лингвистических объектов на основе нечетких оценочных моделей[1]

Аннотация. В работе обоснована актуальность создания нечетких оценочных моделей (нечетких моделей сходства), ориентированных на решение широкого класса задач лингвистического анализа в условиях неопределенности. Выполнен анализ и предложена классификация нечетких оценочных моделей в зависимости от характера агрегирования показателей. Предложен способ, разработаны нечеткие оценочные модели и рассмотрены подходы к решению задач анализа сложных лингвистических объектов с использованием этих моделей.

Ключевые слова: лингвистический анализ, нечеткая оценочная модель.

Введение

Современное состояние исследований в лингвистике характеризуется активным использованием и развитием различных математических методов и моделей, а также их комбинированием для решения таких комплексных задач как атрибуция текстов, автоматическое определение гендерной принадлежности (гендера), классификация индивидуальных стилей (идиостилей), представляющих собой комплекс объективно выделимых формальных характеристик авторских текстов, классификация самих текстов и др. [1-5]. К особенностям этих исследований относится необходимость учета эвристичности представления и субъективности оценки анализируемой информации, а также разнокачественности оцениваемых показателей и их измерения с использованием различных измерительных шкал. Это обусловливает необходимость использования методов интеллектуального анализа данных, и, прежде всего, методов нечеткого анализа и моделирования [6]. Существуют следующие подходы к созданию нечетких моделей в области лингвистического анализа:

1) Адаптация нечетких моделей, которые изначально создавались для решения задач из других предметных областей (технической диагностики, поддержки принятия решений, интеллектуального управления, анализа сложных систем). К достоинствам данного подхода относится использование накопленного научно-методического потенциала. В современных исследованиях в эмпирических науках превалирует именно эта тенденция.

2) Наделение известных моделей для эмпирических исследований нечеткостью для учета различного рода неопределенности. Несмотря на продуктивность данного подхода, возникают сложности введения нечеткости в эти модели.

3) Создание нечетких моделей, изначально ориентированных на решение задач данной эмпирической науки в условиях неопределенности. Несмотря на то, что этот подход является, по нашему мнению, наиболее плодотворным, он представляется и наиболее сложным вследствие необходимости одновременного учета как особенностей решаемых задач, так и свойств разрабатываемых нечетких моделей.

В статье рассматриваются вопросы построения нечетких моделей в рамках третьего под-

[1] Работа выполнена в рамках базовой части государственного задания Минобрнауки России № 2014/123 на выполнение государственных работ в сфере научной деятельности, проект № 2493, а также гранта РГНФ №14-04-00266 «Взаимодействие элементов стихотворного текста».

хода на примерах определения степени сходства оригинального текста и его переводов по отдельным группам показателей; определения степени сходства сопоставляемых фрагментов текста; упорядочения объектов и их признаков по отдельным критериям оценки.

1. Подходы к созданию оценочных моделей для лингвистического анализа

Оценочные модели при решении большинства задач лингвистического анализа предназначены: во-первых, для непосредственной оценки объектов; во-вторых, для ранжирования (упорядочения) объектов по некоторым показателям оценки. Эти задачи взаимосвязаны, но не всегда взаимозаменяемы. Модель оценки, предназначенная для решения задач первого класса, может использоваться даже в отсутствие сопоставляемых альтернативных объектов. При этом оцененные объекты могут быть проранжированы. С другой стороны, способы ранжирования объектов, зачастую неприменимы для оценивания отдельных объектов.

В общем виде задача оценивания ставится следующим образом. Имеется набор показателей оценки p_i, $i = 1, \ldots, n$. Задано множество лингвистических объектов $A = \{a_1, \ldots, a_j, \ldots, a_m\}$. Для $\forall\, a_j \in A$ требуется получить значения показателей $p_1(a_j), ..., p_n(a_j)$, которые задаются в диапазоне [0, 1] и характеризуют степень выполнения соответствующего критерия.

Как правило, модели оценки имеют сложную структуру и представляются в виде иерархически связанных показателей. Рассмотрим в качестве примера двухуровневую систему показателей. Пусть показатели p_i, $i = 1, \ldots, n$ являются частными показателями нижнего уровня иерархии, и пусть задан обобщенный показатель P. Значение показателя P для каждого из оцениваемых объектов можно получить путем свертки значений частных показателей:

$$\forall a_j \in A, \quad P(a_j) = h(p_1(a_j), ..., p_n(a_j)).$$

где h – некоторое отображение, удовлетворяющее следующим аксиомам:

A1) $h(0, ..., 0) = 0, h(1, ..., 1) = 1$ – граничные условия.

A2) Для любых пар $(p_i(a_j), p_i'(a_j)) \in [0,1]^2$, если $\forall i,\ p_i(a_j) \geq p_i(a_j)$, то $h(p_i(a_j), ..., p_i(a_j)) \geq$ $\geq h(p_i'(a_j), ..., p_i'(a_j))$ – аксиома неубывания.

A3) h – симметричная функция своих аргументов.

A4) h – непрерывная функция [7].

Задачу оценки лингвистических объектов можно разделить на два этапа: во-первых, получение оценок частных показателей лингвистических объектов; во-вторых, получение обобщенной оценки лингвистического объекта. Первый этап выполняется, как правило, традиционно. Второй же этап получения обобщенной оценки характеризуется рядом специфических особенностей.

Существуют следующие основные подходы для формирования обобщенного показателя, а фактически, к созданию оценочных моделей сложных лингвистических объектов:

- при условии равнозначности частных показателей;
- на основе рекурсивного агрегирования частных показателей;
- при условии неравнозначности частных показателей;
- на основе использования нечетких квантификаторов при свертке частных показателей.

Рассмотрим эти подходы более детально.

Обобщенный показатель формируется при условии равнозначности частных показателей

В рамках этого подхода возможны следующие варианты оценивания: 1-й вариант – значение обобщенного показателя определяется наихудшим значением частных показателей; 2-й вариант – значение обобщенного показателя определяется наилучшим значением частных показателей; 3-й вариант – компромиссные стратегии; 4-й вариант – гибридные стратегии [8].

Для 1-го варианта оценивания, помимо аксиом A1)–A4), должна выполняться аксиома:

A5)

$$\forall p_i(a_j), \quad h(p_1(a_j), ..., p_n(a_j)) \leq \min(p_1(a_j), ..., p_n(a_j)),$$

т.е. значение обобщенного показателя должно быть не больше минимального значения любого из частных показателей.

Для 2-го варианта оценивания, помимо аксиом A1)–A4), должна выполняться аксиома:

А6)
$\forall p_i(a_j),\ h(p_1(a_j),...,p_n(a_j)) \geq \max(p_1(a_j),...,p_n(a_j))$,
т.е. значение обобщенного показателя должно быть не меньше максимального значения любого из частных показателей.

Для компромиссных стратегий (3-й вариант), в дополнение к аксиомам А1)–А4), должна выполняться следующая аксиома:

А7) $\forall p_i(a_j)$,

$\min(p_1(a_j),...,p_n(a_j)) < h(p_1(a_j),...,p_n(a_j)) <$

$< \max(p_1(a_j),...,p_n(a_j))$, т.е. значение обобщенного показателя находится на промежуточном уровне между значениями частных показателей.

Для гибридной стратегии (4-й вариант) значение обобщенного показателя может быть получено с использованием свертки значений частных равнозначных показателей на основе симметрической суммы, такой, как медиана $\mathrm{med}(p_1(a_j),...,p_n(a_j); 0,5)$.

Другой разновидностью гибридной стратегии, двойственной к ранее рассмотренной, является использование ассоциативных симметрических сумм (исключая медиану), а именно:

если $\max(p_1(a_j),...,p_n(a_j)) < 0,5$,

то $h(p_1(a_j),...,p_n(a_j)) \leq \min(p_1(a_j),...,p_n(a_j))$;

если $\min(p_1(a_j),...,p_n(a_j)) > 0,5$,

то $h(p_1(a_j),...,p_n(a_j)) \geq \max(p_1(a_j),...,p_n(a_j))$.

Данный подход позволяет при условии равнозначности частных показателей идентифицировать адекватные операции для их свертки.

Нахождение (идентификация) адекватных операций свертки частных показателей, «покрывающих» все множество гибридных стратегий между их «противоположными» 1-м и 2-м вариантами, обеспечивает параметризированное семейство операций типа:

$$h(p_1(a_j),...,p_n(a_j)) = \mathrm{I}(p_1(a_j),...,p_n(a_j))^{\gamma} \cdot \mathrm{U}(p_1(a_j),...,p_n(a_j))^{(1-\gamma)},$$
$$\gamma \in [0,1],$$

где I, U – некоторые операции пересечения и объединения, соответственно; γ – показатель, характеризующий степень компромисса между 1-м и 2-м вариантами [8].

Обобщенный показатель формируется на основе рекурсивного агрегирования частных показателей

Часто выполнить оценивание сразу всех частных показателей сложно из-за размерности задачи оценивания. Тогда операцию $h(p_1,...,p_n)$ получения обобщенной оценки можно реализовать рекурсивно для числа равнозначных частных показателей $n > 2$ при сохранении аксиомы коммутативности А3 [8]:

$$h(p_1(a_j),...,p_n(a_j)) = h(h(...(h(p_1(a_j), p_2(a_j)),...), p_{n-1}(a_j)), p_n(a_j)).$$

Обобщенный показатель формируется при условии неравнозначности частных показателей

В рамках этого подхода возможны следующие варианты оценивания: 1-й вариант – установление пороговых значений для частных показателей; 2-й вариант – взвешивание частных показателей; 3-й вариант – асимметричное задание обобщенного показателя.

В случае установления пороговых значений для частных показателей (1-й вариант) для всех частных показателей задаются пороговые значения, достижение которых определяет приемлемость обобщенной оценки анализируемого объекта.

В случае взвешивания частных показателей (2-й вариант) каждому частному показателю назначается весовой коэффициент с последующим агрегированием взвешенных частных показателей с использованием операции свертки. Наиболее распространенной является выражение взвешенного суммирования частных показателей:

$$P(a_j) = \sum_{i=1}^{n} w_i p_i(a_j), \quad \sum_{i=1}^{n} w_i = 1.$$

Однако возможность определения весов w_i проблематична, если эти веса относятся к разнокачественным показателям. В этом случае целесообразным является замена в последнем выражении частных показателей на их однородные характеристики, например, на степени удовлетворенности (принадлежности) соответствующему свойству объекта. Процедуру взвешивания частных показателей можно распространить и на другие типы операций свертки.

Третий вариант асимметричного задания обобщенного показателя обусловлен его сложной структурой и необходимостью агрегирова-

ния частных показателей на нижнем уровне иерархии с использованием, например, «И–ИЛИ»-дерева.

Обобщенный показатель формируется на основе использования нечетких квантификаторов при свертке частных показателей

Нечеткие квантификаторы задают оценку достижения приемлемых результатов, например, таких как «большинство», «мало», «по меньшей мере, половина». Нечеткие квантификаторы могут задавать, в том числе, оценку того, насколько достигнуты пороговые значения частных показателей (или большинства этих показателей), определяющие приемлемость обобщенной оценки анализируемого объекта. В этом случае помимо нечеткого квантификатора задается и распределение весов этих частных показателей [7].

Важность выбора наилучшего подхода к формированию обобщенного показателя становится более очевидной, когда некоторую обобщенную оценку, в свою очередь, комбинируют с другими обобщенными оценками.

Несмотря на разнообразие подходов к созданию оценочных моделей, они не учитывают согласованность частных неравнозначных показателей при выборе операции свертки для вычисления значения обобщенного показателя. Это, в итоге, не позволяет создавать адекватные оценочные модели со сложной структурой для решения задач лингвистического анализа.

2. Способ и нечеткие оценочные модели для решения задач лингвистического анализа

Основными требованиями, предъявляемыми к оценочным моделям для решения задач лингвистического анализа, являются следующие:

- возможность формирования обобщенного показателя на основе изменяющихся наборов частных показателей;
- возможность объединения разнородных показателей (как количественных, так и качественных), различающихся по измерительным шкалам, диапазонам значений;
- учет различной значимости частных показателей в обобщенной оценке;
- учет согласованности частных показателей;
- гибкая настройка (адаптация) модели оценки при добавлении или исключении показателей и изменении параметров (согласованности и значимости показателей).

Результаты проведенного анализа существующих подходов к созданию оценочных моделей для лингвистического анализа позволяют сделать вывод о том, что наиболее целесообразным для решения задач лингвистического анализа является создание моделей оценки, основанных на нечетком подходе.

Предлагаемые способ и модель позволяют определить на основе операций над частными показателями $h(p_k, p_l), k, l \in \{1, ..., n\}, k \neq l$ (для 2-местных операций) и весовых коэффициентов этих частных показателей w_i $(i = 1, ..., n, \sum_{i=1}^{n} w_i = 1)$ операцию $h(p_1, ..., p_n)$ получения обобщенной оценки P. Для идентификации операций над этими частными показателями p_k и p_l должны быть определены попарные степени их согласованности $c_{k,l}$ $(k, l = 1, ..., n)$.

В зависимости от особенностей решаемой задачи оценки, согласованность может трактоваться как корреляция, взаимовлияние частных показателей, одновременная достижимость значений сопоставляемых частных показателей, смежность проявления частных показателей или их конкретных значений.

Так как для определения степени согласованности частных показателей могут использоваться различные методы, то целесообразно задать множество критериальных уровней согласованности этих показателей, упорядоченных в порядке возрастания согласованности: $\tilde{C} = \{NC$ – «Отсутствие согласованности», LC – «Низкая согласованность», MC – «Средняя согласованность», HC – «Высокая согласованность», FC – «Полная согласованность»$\}$. Тогда, полученные попарные степени согласованности частных показателей $c_{k,l}$ $(k, l = 1, ..., n)$, вне зависимости от метода их получения, сопоставляются с критериальными уровнями согласованности из множества $\tilde{C}$:

$$c_{k,l} \Leftrightarrow \tilde{c}_i \in \tilde{C} = \{NC, LC, MC, HC, FC\}.$$

В результате этого сопоставления множество частных показателей будет разбито на подмножества, соотносимые с отдельным критериальным уровнем согласованности. В свою очередь, все критериальные уровни согласованности частных показателей из множества $\tilde{C}$

Табл. 1. Сопоставление операций свертки с критериальными уровнями согласованности показателей

№	Операция свертки показателей p_k и p_l	Критериальный уровень согласованности показателей	Описание критериального уровня согласованности
1	$\min(p_k, p_k)$	*NC*	Отсутствие согласованности
2	$\text{med}(p_k, p_k; 0,25)$	*LC*	Низкая согласованность
3	$\text{med}(p_k, p_k; 0,5)$	*MC*	Средняя согласованность
4	$\text{med}(p_k, p_k; 0,75)$	*HC*	Высокая согласованность
5	$\max(p_k, p_k)$	*FC*	Полная согласованность

сопоставляются с операциями свертки этих показателей, удовлетворяющими предъявленным выше аксиомам: нормировки, неубывания, непрерывности, бисимметричности (ассоциативности). В работах [9, 10] обоснован набор операций свертки частных показателей, с одной стороны, удовлетворяющий представленным выше аксиомам, а с другой, сопоставленный с указанными критериальными уровнями согласованности показателей (Табл.1).

До сих пор рассуждения велись в предположении, что оценочная модель имеет двухуровневую иерархическую структуру, состоящую из обобщенного показателя и совокупности частных показателей. Однако система показателей может быть организована и в более сложную структуру. А именно, частные показатели могут быть разделены на группы, которые, в свою очередь, разделяются на подгруппы. И в результате группирования показателей структура оценочной модели может включать в себя три и более уровней.

Для решения различных задач лингвистического анализа могут быть предложены различные способы группирования показателей, например [4]:

- группировка показателей по их уровневому статусу:
  - синтаксические показатели;
  - показатели поэтического синтаксиса;
  - частеречные (морфологические) показатели: локализованные (в исходе сроки рифмованные, т.е. участвующие в создании рифмы; в исходе строки нерифмованные, т.е. не участвующие в создании рифмы)[1]; нелокализованные;
  - ритмические показатели;
  - рифменные показатели;
  - строфические показатели (а именно, количество строк в строфе; схема организации рифмы в строфе, когда рифмуются соседние строки, либо первая строка рифмуется с последней, либо четные строки рифмуются с четными и нечетные с нечетными и т.д.);
- группировка показателей по их локализации:
  - локализованные: ритмические показатели относительно начала строк[2]; морфологические показатели относительно конца строк[3];
  - нелокализованные: синтаксические показатели; показатели поэтического синтаксиса; строфические показатели.

В случае группирования оценка формируется для каждой группы показателей с учетом согласованности этих показателей внутри группы (подгруппы). Результаты оценки по каждой группе (подгруппе) показателей могут представлять самостоятельную ценность. Помимо этого результаты оценки по всем группам показателей могут быть агрегированы в обобщенную оценку с учетом степеней (уровней) согласованности между этими группами. Процедура

[1] Рифма здесь понимается традиционно, как совпадение звуковой формы слов, начиная с ударной гласной.

[2] Локализованные ритмические показатели относительно начала строк основаны на расхождении ритма и метра. Метр понимается как идеальная схема чередования ударных и безударных слогов; ритм – реальное распределение ударений в стихотворной строке. Эти показатели обычно включают в себя пропуски ударений, наличие дополнительных ударений, наличие дополнительных слогов по сравнению с метрической схемой.

[3] Эти морфологические показатели показывают, какие части речи находятся в конце строки, какие грамматические формы у этих слов.

получения обобщенной оценки на основе агрегирования результатов оценки по всем группам показателей аналогична процедуре оценки для отдельной группы показателей с учетом степени (уровня согласованности групп показателей между собой).

Рассмотрим в качестве примера создание оценочной модели для группы частеречных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”. Группу этих частеречных показателей можно разбить на три подгруппы:

- подгруппа частеречных локализованных рифмованных показателей (относительно строк): p_1 – число рифмующихся существительных; p_2 – число рифмующихся глаголов; p_3 – число рифмующихся прилагательных; p_4 – число рифмующихся наречий; p_5 – число рифмующихся местоимений;
- подгруппа частеречных локализованных нерифмованных показателей (относительно строк): p_6 – число существительных в последней позиции в строке, и не участвующих в образовании рифмы; p_7 – число глаголов, в последней позиции и не рифмующихся; p_8 – число прилагательных, в последней позиции и не рифмующихся; p_9 – число наречий, в последней позиции и не рифмующихся; p_{10} – количество местоимений, в последней позиции и не рифмующихся;
- подгруппа частеречных нелокализованных показателей (относительно всех слов части): p_{11} – число существительных; p_{12} – число глаголов; p_{13} – число прилагательных; p_{14} – число наречий.

В качестве метода для определения степени согласованности между показателями внутри указанных подгрупп воспользуемся результатами корреляционного анализа с использованием коэффициента корреляции Пирсона. В Табл. 2 приведены результаты сопоставления коэффициентов корреляции показателей из подгруппы частеречных локализованных рифмованных показателей с критериальными уровнями согласованности.

Аналогичным образом строятся матрицы согласованности для частеречных локализованных нерифмованных показателей (Табл. 3) и для частеречных нелокализованных показателей (Табл. 4).

Требование учета различной значимости частных показателей $\{p_1, p_2, ..., p_n\}$ в каждой группе (или подгруппе) обеспечивается за счет введения вектора весовых коэффициентов:

$$W = \{w_1, w_2, ..., w_n\}, \forall i : w_i \in [0,1], \sum_{i=1}^{n} w_i = 1.$$

Весовые коэффициенты $\{w_1, w_2, ..., w_n\}$ могут быть: либо определены в процессе эксперимента; либо непосредственно назначены экспертом; либо получены на основе обработки попарных сравнений значимости показателей, также выполненных экспертом. Последний из указанных подходов к определению весовых коэффициентов показателей наиболее удобен, так как ориентирован на сравнение каждый раз только пары показателей с последующей обработкой результатов попарных сравнений. Это существенно упрощает задачу эксперту в условиях большого количества сопоставляемых показателей.

Табл. 2. Матрица согласованности частеречных локализованных рифмованных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”

	p_1	p_2	p_3	p_4	p_5
p_1	–	*HC*	*LC*	*MC*	*LC*
p_2	*HC*	–	*LC*	*MC*	*NC*
p_3	*LC*	*LC*	–	*HC*	*MC*
p_4	*MC*	*MC*	*HC*	–	*MC*
p_5	*LC*	*NC*	*MC*	*MC*	–

Табл. 3. Матрица согласованности для частеречных локализованных нерифмованных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”

	p_6	p_7	p_8	p_9	p_{10}
p_6	–	*MC*	*LC*	*LC*	*HC*
p_7	*MC*	–	*LC*	*LC*	*MC*
p_8	*LC*	*LC*	–	*MC*	*MC*
p_9	*LC*	*LC*	*MC*	–	*LC*
p_{10}	*HC*	*MC*	*MC*	*LC*	–

Табл. 4. Матрица согласованности для частеречных нелокализованных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”

	p_{11}	p_{12}	p_{13}	p_{14}
p_{11}	–	*MC*	*MC*	*LC*
p_{12}	*MC*	–	*LC*	*MC*
p_{13}	*MC*	*LC*	–	*LC*
p_{14}	*LC*	*MC*	*LC*	–

Рассмотрим получение вектора весовых коэффициентов основе метода попарных сравнений значимости показателей [11]. Сначала на основе попарных сравнений значимости показателей из группы (подгруппы) формируется положительно определенная, обратно-симметричная матрица (Табл. 5), в которой элемент v_i/v_j обозначает степень значимости i-го показателя по сравнению с j-м показателем.

Для задания степеней значимости v_i и v_j используется следующие оценки: 1 – одинаково значимы; 3 – ненамного значимее; 5 – существенно значимее; 7 – значительно значимее; 9 – абсолютно значимее; 2, 4, 6, 8 – промежуточные значения.

На основе этих сравнений рассчитываются весовые коэффициенты показателей в подгруппе:

$$w_i = \frac{\sqrt[n]{\prod_{l=1}^{n}(v_k/v_l)}}{\sum_{k=1}^{n}\sqrt[n]{\prod_{l=1}^{n}(v_k/v_l)}}.$$

Для рассматриваемого примера получены следующие весовые коэффициенты показателей:

- для подгруппы частеречных локализованных рифмованных показателей: $\{w_1 = 0{,}12; w_2 = 0{,}29; w_3 = 0{,}25; w_4 = 0{,}15; w_5 = 0{,}19\}$;
- для подгруппы частеречных локализованных нерифмованных показателей: $\{w_6 = 0{,}11; w_7 = 0{,}30; w_8 = 0{,}23; w_9 = 0{,}14; w_{10} = 0{,}22\}$;
- для подгруппы частеречных нелокализованных показателей: $\{w_{11} = 0{,}25; w_{12} = 0{,}3; w_{13} = 0{,}3; w_{14} = 0{,}15\}$;

Результаты оценки согласованности и значимости показателей внутри каждой группы (подгруппы) показателей структурно можно представить в виде нечеткого неориентированного графа $\tilde{G}$ с нечеткими узлами и нечеткими ребрами:

$$\tilde{G} = (\tilde{P}, \tilde{R}),$$

где $\tilde{P} = \{(p_i / w_i)\}, i \in \{1, ..., n\}, p_i \in P$ – нечеткое множество вершин, причем, каждая вершина p_i взвешена соответствующим значением $w_i \in [0,1]$; $\tilde{R} = \{(p_k, p_l)/\tilde{c}_i)\}$, $k, l = 1, ..., n$, $\tilde{c}_i \in \tilde{C} = \{NC, LC, MC, HC, FC\}$ – нечеткое множество ребер, причем, каждая дуга (p_k, p_l) сопоставлена соответствующему критериальному уровню согласованности $\tilde{c}_i$.

Табл. 5. Матрица попарных сравнений значимости показателей из группы (подгруппы)

		Номер признака				
		1	…	*i*	…	*n*
Номер признака	1	1	…	v_1/v_i	…	v_1/v_n
	…	…	…	…	…	…
	i	v_i/v_1		1	…	v_i/v_n
	…	…	…	…	…	…
	n	v_n/v_1	…	v_n/v_i	…	1

На Рис. 1 представлен нечеткий неориентированный граф, соответствующий результатам оценки согласованности и значимости внутри подгруппы частеречных локализованных рифмованных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”.

Рассмотрим способ построения оценочной модели на примере модели оценки для указанной выше подгруппы частеречных локализованных рифмованных показателей $\{p_1, p_2, p_3, p_4, p_5\}$.

Предлагаемый способ основан на поочередном поиске в нечетком графе полных подграфов $\tilde{G}' = (\tilde{P}', \tilde{R}')$, дуги которого сопоставлены с одним из критериальных уровней согласованности $\tilde{c}_i \in \tilde{C}$. Порядок поиска таких подграфов определяется направлением просмотра уровней согласованности показателей.

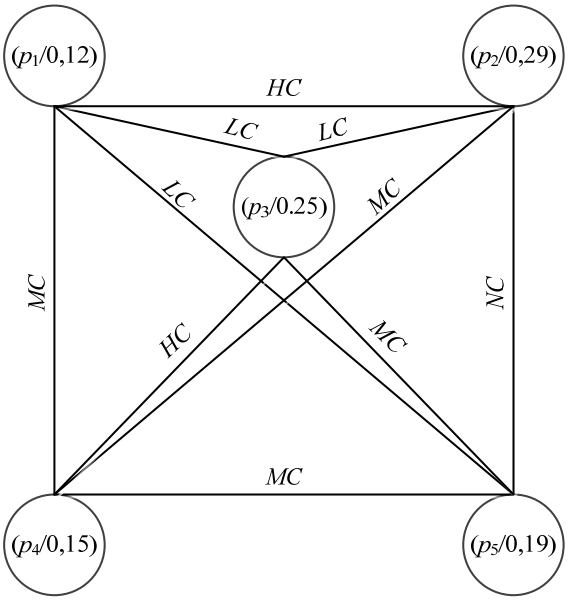


Рис. 1. Нечеткий неориентированный граф для подгруппы частеречных локализованных рифмованных показателей поэмы С. Кольриджа “The Rime of the Ancient Mariner”

Возможны два направления просмотра уровней согласованности показателей: во-первых, от наименее согласованных к наиболее согласованным показателям; во-вторых, от наиболее согласованных к наименее согласованным показателям. Порядок просмотра от наименее согласованных к наиболее согласованным показателям позволяет «не потерять» «хорошие» оценки плохо согласованных показателей, поскольку операции свертки показателей при этом обычно ориентированы на вариант оценивания, при котором значение обобщенного показателя определяется наихудшим значением частных показателей. Порядок просмотра от наиболее согласованных к наименее согласованным показателям позволяет полноценно учесть «плохие» оценки хорошо согласованных показателей, операции свертки по которым ориентированы на вариант оценивания, при котором значение обобщенного показателя определяется наилучшим значением частных показателей [9].

Рассмотрим поочередный поиск подграфов $\tilde{G}' = (\tilde{P}', \tilde{R}')$ при условии просмотра от наиболее согласованных к наименее согласованным показателям из подгруппы $\{p_1, p_2, p_3, p_4, p_5\}$. Процедура начинается с определения наибольшего уровня согласованности для этих показателей (Рис. 1). Пары показателей p_1 и p_2, а также p_3 и p_4 характеризуются уровнем согласованности *HC* – «Высокая согласованность», наибольшим для нечеткого графа на Рис. 1. Эти пары вершин образуют два полных подграфа максимального размера.

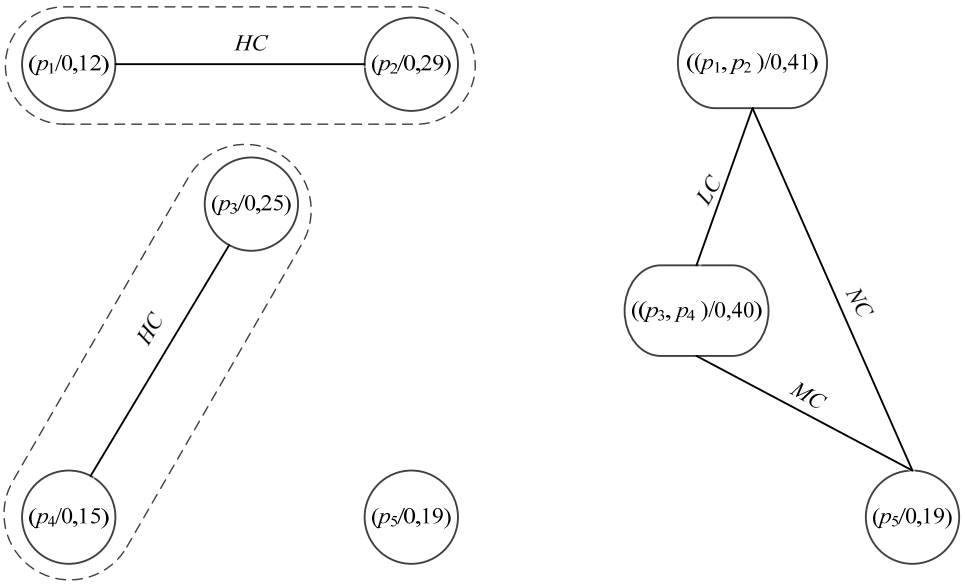
В случае нескольких таких подграфов проводится сортировка этих полных подграфов максимального размера в порядке убывания суммарного веса входящих в них вершин (можно установить и иной порядок просмотра подграфов). На Рис. 2, *a* оставлены только те ребра нечеткого графа, которые соответствуют рассматриваемому уровню согласованности *HC*.

Последовательно просматривая эти подграфы, объединяем все вершины этих подграфов в одну на основе операции свертки, соответствующей рассматриваемому уровню согласованности.

Если идентифицируемая операция свертки является не многоместной $h(p_1, ..., p_q)$, $q \in \{1, ..., n\}$, а двуместной неассоциативной операцией $h(p_k, p_l), k, l \in \{1, ..., n\}, k \neq l$, то для нее следует установить порядок перечисления показателей p_k и p_l, например, по убыванию весов.

Для случая идентифицированной бисимметричной операции также можно начинать с вершин, обладающих наибольшим весом. Тогда полагая, что $w_i \leq w_{i+1}, i \in \{1, ..., q-1\}$, получим:

$$h^*(p_1, ..., p_q) = h(h(...(h(p_1, p_2), ...), p_{q-1}), p_q).$$



a) *b*)

Рис. 2. Подграфы нечеткого графа, соответствующие уровню согласованности *HC* – (*a*); нечеткий граф с объединенными вершинами (p_1, p_2) и (p_3, p_4) – (*b*)

Для свертки показателей p_1 и p_2, а также p_3 и p_4 выбираются операции $\text{med}(p_1, p_2; 0,75)$ и $\text{med}(p_3, p_4; 0,75)$ из Табл.1, соответствующие уровню согласованности *НС*.

На Рис. 2, *б* показан нечеткий граф, образованный объединением вершин p_1 и p_2, а также p_3 и p_4 с использованием указанных операций. Веса новых вершин нечеткого графа образуются суммированием весов объединяемых вершин.

Затем в нечетком графе удаляются ребра, связанные с объединяемыми вершинами. После этого определяются уровни согласованности ребер, смежных с этими объединяемыми вершинами. Эти уровни согласованности определяются в зависимости от выбранной стратегии оценивания. Для рассматриваемого примера выберем вариант оценивания, для которого значение обобщенного показателя определяется наихудшим значением частных показателей. Далее процедура повторяется.

Затем определяется следующий по порядку наибольший уровень согласованности между показателями (вершинами нечеткого графа) из множества $\{(p_1, p_2), (p_3, p_4), p_5\}$. Очевидно, что пара показателей (p_3, p_4) и p_5 будет характеризоваться наибольшим уровнем согласованности *МС*. Для свертки этих показателей, в соответствии с Табл. 1, используется операция $\text{med}((\text{med}(p_3, p_4; 0,75), p_5; 0,5)$. На Рис. 3 показан нечеткий граф, образованный после очередного объединения вершин.

На заключительном этапе идентифицируется операция для свертки показателей (объединенных вершин нечеткого графа) (p_1, p_2) и $((p_3, p_4), p_5)$, в соответствии с их уровнем согласованности *NC*.

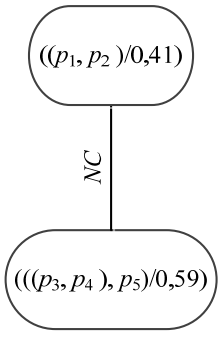


Рис. 3. Нечеткий граф после очередного объединения вершин

Таким образом, оценочная модель для подгруппы частеречных локализованных рифмованных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner” представляется в виде:

$$P_{члр} = \min((\text{med}(p_1, p_2; 0,75)),$$
$$(\text{med}((\text{med}(p_3, p_4; 0,75)), p_5; 0,5))).$$

Аналогичным образом строятся оценочные модели и для двух других подгрупп показателей.

На Рис. 4 представлен нечеткий неориентированный граф для подгруппы частеречных локализованных нерифмованных показателей $\{p_6, p_7, p_8, p_9, p_{10}\}$. Повторив этапы способа, получаем оценочную модель для подгруппы частеречных локализованных нерифмованных показателей в следующем виде:

$$P_{члн} = \text{med}((\text{med}((\text{med}(p_6, p_{10}; 0,75)), p_7; 0,5)),$$
$$(\text{med}(p_8, p_9; 0,5)); 0,25).$$

На Рис. 5 представлен нечеткий неориентированный граф для подгруппы частеречных нелокализованных показателей $\{p_{11}, p_{12}, p_{13}, p_{14}\}$. Оценочная модель для подгруппы частеречных нелокализованных показателей представляется в виде:

$$P_{чн} = \text{med}((\text{med}((\text{med}(p_{11}, p_{12}; 0,5)), p_{13}; 0,5)), p_{14}; 0,5).$$

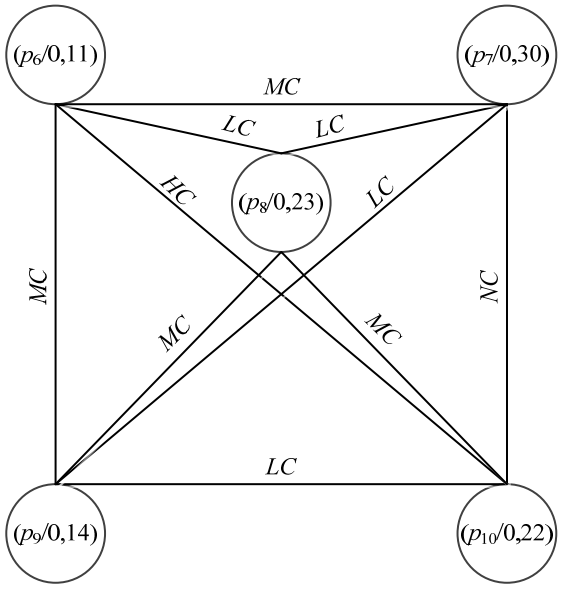


Рис. 4. Нечеткий граф для подгруппы частеречных локализованных нерифмованных показателей

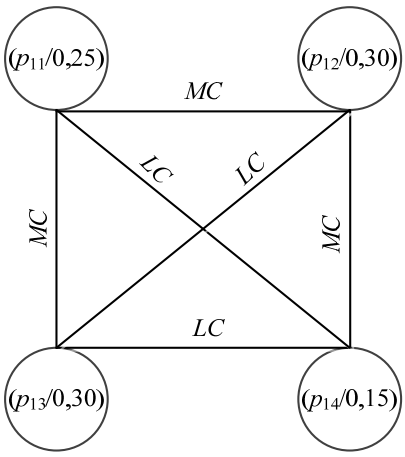


Рис. 5. Нечеткий граф для подгруппы частеречных нелокализованных показателей

Результаты оценки лингвистических объектов по каждой подгруппе показателей могут представлять самостоятельную ценность. В этом случае объекты будут оцениваться на основе вектора оценок, каждый элемент которого представляет собой оценку, получаемую с использованием оценочной модели по соответствующей подгруппе показателей. Например, вектор оценок частеречных (морфологических) показателей какого-либо лингвистического объекта a_j, j =1, …, m включает в себя совокупность оценок, полученных с использованием оценочных моделей для соответствующих подгрупп частеречных показателей $P_{члр}$, $P_{члн}$ и $P_{чн}$:

$$(P_{члр}(a_j), P_{члн}(a_j), P_{чн}(a_j));$$

$$P_{члр}(a_j) = \min((\text{med}(p_1(a_j), p_2(a_j); 0,75)),$$
$$(\text{med}((\text{med}(p_3(a_j), p_4(a_j); 0,75)), p_5(a_j); 0,5)));$$

$$P_{члн}(a_j) = \text{med}((\text{med}((\text{med}(p_6(a_j), p_{10}(a_j); 0,75)),$$
$$p_7(a_j); 0,5)), (\text{med}(p_8(a_j), p_9(a_j); 0,5)); 0,25);$$

$$P_{чн}(a_j) = \text{med}((\text{med}((\text{med}(p_{11}(a_j), p_{12}(a_j); 0,5)),$$
$$p_{13}(a_j); 0,5)), p_{14}(a_j); 0,5),$$

где $P_{члр}(a_j)$, $P_{члн}(a_j)$, $P_{чн}(a_j)$ – оценки лингвистического объекта a_j по соответствующим подгруппам частеречных показателей, полученные с использованием оценочных моделей $P_{члр}$, $P_{члн}$ и $P_{чн}$.

Значимость показателей учитывается непосредственно при расчете оценок конкретного лингвистического объекта с использованием разработанных оценочных моделей.

Результаты оценки по указанным подгруппам показателей могут быть агрегированы в обобщенную оценку по всей группе частеречных показателей. Для этого на основе предлагаемого способа требуется разработать оценочную модель, в которой $P_{члр}$, $P_{члн}$ и $P_{чн}$, в свою очередь, выступают в качестве «частных» показателей, а обобщенным показателем является показатель $P_{ЧП}$, формируемый с учетом их согласованности и значимости:

$$P_{ЧП} = H^*(P_{члр}, P_{члн}, P_{чн}),$$

где H^* – обозначение операции (совокупности операций) свертки, идентифицируемой в соответствии с уровнями согласованности $P_{члр}$, $P_{члн}$ и $P_{чн}$.

Очевидно, что определение уровней согласованности между $P_{члр}$, $P_{члн}$ и $P_{чн}$, как, впрочем, и задание коэффициентов их значимости, носит экспертный характер. Причем, для определения согласованности $P_{члр}$, $P_{члн}$ и $P_{чн}$ можно воспользоваться нечеткой когнитивной моделью [12], а для задания коэффициентов их значимости – методом попарных сравнений [11].

Допустим, установлено, что $P_{члр}$, $P_{члн}$ и $P_{чн}$ согласованы с уровнем согласованности *НС*. И для них заданы следующие весовые коэффициенты: $w_{члр}$ = 0,35; $w_{члн}$ = 0,25; $w_{чн}$ = 0,40. Тогда оценочную модель для всей группы частеречных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner” можно представить в следующем виде:

$$P_{ЧП} = \text{med}((\text{med}(P_{члр}, P_{чн}; 0,75)), P_{члн}; 0,75).$$

Выполним анализ лингвистических объектов с использованием разработанных моделей.

Подход к определению сходства (степени сходства) оригинального текста и его переводов

Сущность этого подхода заключается в определении значений сопоставляемых лингвистических объектов, например, a_1, a_2, a_3 – фрагментов текстов оригинала поэмы С. Кольриджа “The Rime of the Ancient Mariner” и двух ее переводов Н. Гумилева и В. Левика, соответственно.

При этом используются представленные выше оценочные модели для подгрупп показателей $P_{члр}$, $P_{члн}$ и $P_{чн}$ и для всей группы частеречных

показателей $P_{ЧП}$, разработанные для оригинального текста. Получаемые при этом результаты оценки позволяют не только проранжировать сопоставляемые лингвистические объекты, но и установить их степень сходства между собой.

В Табл. 6 представлены результаты оценки семи глав оригинала поэмы С. Кольриджа "The Rime of the Ancient Mariner" и двух ее переводов Н. Гумилева и В. Левика, полученные с использованием разработанной оценочной модели для группы частеречных показателей.

В соответствии с анализом динамики этих показателей можно сделать следующие выводы.

Во-первых, в оригинальном тексте С. Кольриджа на обобщенный показатель $P_{ЧП}$ наибольшее влияние оказывает показатель $P_{члн}$. В то же время и в случае перевода Н. Гумилева, и в случае перевода В. Левика, на обобщенный показатель $P_{ЧП}$ определяющее влияние оказывает показатель $P_{чн}$. Этим можно объяснить бо́льшие значения показателей $P_{ЧП}$ как для перевода Н. Гумилева, так и для перевода В. Левика, по сравнению со значениями этого показателя для оригинального текста С. Кольриджа.

Во-вторых, для оригинального текста С. Кольриджа значение обобщенного показателя $P_{ЧП}$ остается стабильным практически на протяжении всей поэмы, за исключением тенденции к его уменьшению, начиная с 6-й главы. И для 7-й главы значение показателя $P_{ЧП}$ принимает минимальное значение. Та же тенденция свойственна и обоим переводам. Однако для перевода Н. Гумилева эта тенденция является менее выраженной, чем для перевода В. Левика.

В-третьих, на протяжении всей поэмы можно отметить существенно большую степень корреляции изменений значения показателя $P_{ЧП}$ для оригинального текста С. Кольриджа и перевода Н. Гумилева по сравнению со степенью корреляции изменений значения показателя $P_{ЧП}$ для оригинала и перевода В. Левика.

Аналогичным образом можно выполнить сравнительный анализ динамики изменения показателей $P_{члр}$, $P_{члн}$, $P_{чн}$ для оригинального текста и его переводов.

В соответствии со сравнительным анализом динамики изменения показателей $P_{члр}$, $P_{члн}$, $P_{чн}$ можно отметить, что практически для них всех характерна существенно большая степень корреляции изменений значений для оригинального текста С. Кольриджа и перевода Н. Гумилева по сравнению со степенью корреляции изменений значений этих показателей для оригинала и перевода В. Левика.

Исключение составляют результаты анализа динамики изменения показателей $P_{члр}$ для глав 6 и 7. Для этих глав характерна бо́льшая степень корреляции изменений значений показателя $P_{члр}$ для оригинального текста С. Кольриджа и перевода В. Левика по сравнению со степенью корреляции изменений значений этого показателя для оригинала и перевода Н. Гумилева.

Оценка с использованием разработанных моделей проводилась с учетом следующих допущений:

- оценочные модели для оригинального текста не изменяются в рамках всего произведения. Можно построить различные оценочные

Табл. 6. Результаты оценки глав поэмы С. Кольриджа "The Rime of the Ancient Mariner" и переводов Н. Гумилева и В. Левика с использованием оценочной модели

Глава	Текст С. Кольриджа				Перевод Н. Гумилева				Перевод В. Левика			
	$P_{члр}(a_1)$	$P_{члн}(a_1)$	$P_{чн}(a_1)$	$P_{ЧП}(a_1)$	$P_{члр}(a_2)$	$P_{члн}(a_2)$	$P_{чн}(a_2)$	$P_{ЧП}(a_2)$	$P_{члр}(a_3)$	$P_{члн}(a_3)$	$P_{чн}(a_3)$	$P_{ЧП}(a_3)$
1	0,061	0,250	0,203	0,250	0,049	0,250	0,296	0,296	0,036	0,250	0,323	0,323
2	0,05	0,250	0,235	0,250	0,017	0,217	0,322	0,322	0,046	0,215	0,329	0,329
3	0,037	0,173	0,232	0,232	0,099	0,222	0,288	0,288	0,060	0,060	0,316	0,316
4	0,103	0,221	0,213	0,221	0,103	0,191	0,267	0,267	0,029	0,176	0,345	0,345
5	0,051	0,245	0,193	0,245	0,059	0,220	0,279	0,279	0,059	0,250	0,314	0,314
6	0,048	0,192	0,218	0,218	0,163	0,250	0,285	0,285	0,076	0,229	0,273	0,273
7	0,080	0,196	0,193	0,193	0,035	0,204	0,273	0,273	0,092	0,227	0,272	0,272

модели для различных фрагментов оригинального текста, а оценку проводить для каждого фрагмента текста с использованием «своей» оценочной модели;

- сопоставление проводилось после нормирования показателей относительно числа строк во фрагменте текста (за исключением подгруппы частеречных нелокализованных показателей). Очевидно, что результаты оценки будут более точными, если нормировать значения показателей относительно их максимально возможных значений с одновременным учетом весов показателей, задаваемых в явном виде;
- использовалась стратегия просмотра «от наиболее согласованных к наименее согласованным показателям». Использование иной стратегии может привести к другим результатам.

Полученные результаты оценки с учетом указанных допущений позволяют сделать выводы о том, что, в целом, перевод Н. Гумилева по группе частеречных показателей ближе к оригинальному тексту С. Кольриджа, чем перевод В. Левика, за исключением 6-й и 7-й глав.

Подход к определению степени сходства сопоставляемых фрагментов текста

Здесь в качестве сопоставляемых лингвистических объектов могут выступать различные фрагменты (главы) как одного текста, так и различных текстов одного автора. В этом случае разработанная модель может служить, например, для анализа динамики изменения авторского стиля в рамках отдельного произведения.

Подход к сопоставлению структур оценочных моделей для оригинального текста и переводов

Оценочные модели могут разрабатываться не только для оригинального текста, но и для его переводов. Результаты сопоставления структур оценочных моделей для оригинального текста и его переводов также представляют интерес, так как в них заложена информация о взаимосвязях и согласованности оцениваемых показателей.

Заключение

В статье обоснована актуальность создания нечетких оценочных моделей, ориентированных на решение задач лингвистического анализа. Выполнен анализ и предложена классификация оценочных моделей сложных лингвистических объектов в зависимости от характера агрегирования показателей.

Предложен способ и разработаны нечеткие оценочные модели для решения задач лингвистического анализа. Приведены примеры использования разработанных нечетких оценочных моделей для группы частеречных показателей для поэмы С. Кольриджа “The Rime of the Ancient Mariner”, а также для подгрупп показателей: частеречных локализованных рифмованных показателей; частеречных локализованных нерифмованных показателей; частеречных нелокализованных показателей.

Рассмотрены следующие подходы к решению задач лингвистического анализа с использованием разработанных оценочных моделей: к определению сходства (степени сходства) оригинального текста и его переводов; к определению степени сходства сопоставляемых фрагментов текста; к сопоставлению структур оценочных моделей для оригинального текста и его переводов.

Литература

1. Juola P. Authorship attribution. Foundations and Trends in Information Retrieval. Vol. 1. Is. 3. Hanover, MA, USA: Now publishers Inc., 2006. PP. 233–334.
2. Hoover D.L. Corpus stylistics, stylometry and the styles of Henry James// Style. Vol. 41(2), 2007. P. 160–189.
3. Mikros G.K. Content words in authorship attribution: An evaluation of stylometric features in a literary corpus. Studies in Quantitative Linguistics 5. Issues in Quantitative Linguistics / Ed. Reinhard Köhler. RAM-Verlag. 2009. PP. 61–75.
4. Andreev S.N. Literal vs. liberal translation – formal estimation// Glottometrics. Vol. 23, 2012. PP. 62–69.
5. Köhler R., Altmann G. Problems in Quantitative Linguistics 4. 2014. RAM-Verlag.
6. Borisov V.V. Hybridization of Intellectual Technologies for Analytical Tasks of Decision-Making Support// Journal of Computer Engineering and Informatics. Vol. 2, Iss. 1, 2014. PP. 148–156.
7. Dubois D., Prade H. Possibility theory. Applications to the representation of knowledge in Informatics. Moscow: Radio and communication, 1990. (in Russian – translation).
8. Dubois D., Prade A., Theorie des possibilites. Applications a la representation des connaissances en informatique. Masson, 1988.
9. Борисов В.В., Зернов М.М., Федулов Я.А. Способ нечеткого многокритериального оценивания с учетом согласованности параметров оценки // Информ. бюллетень Смоленского регионального отделения АВН. Вып. 28. Смоленск: ВА войсковой ПВО ВС РФ, 2013. – С. 122–135.
10. Андреев С.Н., Борисов В.В., Федулов Я.А. Стратегии нечеткого оценивания в задачах лингвистического

анализа// Сб. тр. 14 Междунар. конф. «Системы компьютерной математики и их приложения», СКМП–2012. – Смоленск: Изд-во СмолГУ, 2013. – Вып.14. – С. 60–63.
11. Saaty T.L. The Analytic Hierarchy Process. McGraw-Hill International. NY, U.S.A. 1980.
12. Borisov V.V., Fedulov A.S. Generalized Rule-Based Fuzzy Cognitive Maps: Structure and Dynamics Model// Lecture Notes in Computer Science. V. 3316, 2004. PP. 918–922.

Борисов Вадим Владимирович. Профессор кафедры вычислительной техники филиала ФБГОУ ВПО «Национальный исследовательский университет «МЭИ» в г. Смоленске. Окончил Московский энергетический институт в 1986 году. Доктор технических наук, профессор. Автор 250 печатных работ и 8 монографий. Область научных интересов: нечеткий и нейро-нечеткий анализ, моделирование сложных систем и процессов, ассоциативные системы хранения и обработки информации. E-mail: vborisov@etna-it.ru

Андреев Сергей Николаевич. Профессор кафедры иностранных языков ФБГОУ ВПО Смоленский государственный университет. Окончил Смоленский государственный педагогический институт в 1970 году. Доктор филологических наук, профессор. Автор 140 печатных работ и двух монографий. Область научных интересов: словообразование, математическая лингвистика, стилеметрия. E-mail: smol.an@mail.ru

Федулов Ярослав Александрович. Аспирант кафедры вычислительной техники филиала ФБГОУ ВПО «Национальный исследовательский университет «МЭИ» в г. Смоленске. Окончил филиал ФБГОУ ВПО «НИУ «МЭИ» в г. Смоленске в 2012 году. Автор 8 печатных работ. Область научных интересов: нечеткий анализ, моделирование сложных систем и процессов, интеллектуальная поддержка принятия решений. E-mail: fedulov_yar@mail.ru