СЕМАНТИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ТЕКСТОВ ПРЕДМЕТНЫХ ЯЗЫКОВ (МОРФОЛОГИЯ И СИНТАКСИС)

Дмитрий Владимирович Михайлов[1] *(доцент, e-mail: Dmitry.Mikhaylov@novsu.ru),*
Геннадий Мартинович Емельянов[1] *(профессор, e-mail: Gennady.Emelyanov@novsu.ru)*
[1] *Государственное образовательное учреждение высшего профессионального образования «Новгородский государственный университет имени Ярослава Мудрого»*

Аннотация

Рассматривается задача семантической кластеризации текстов предметного Естественного Языка. Предложен подход к выработке критериев качества синтаксического анализа как инструментального средства выделения объектов и признаков. Особое внимание уделяется Расщепленным Значениям и конверсивам в составе синтаксических контекстов существительных.

Ключевые слова: понимание текстов, естественный язык, предметная область, семантическая эквивалентность, кластеризация знаний, теория решеток.

Введение

Одним из наиболее перспективных путей повышения качества распознавания текстов Естественного Языка (ЕЯ) является привлечение семантической информации. Знания о семантике ЕЯ, ее связи с синтаксисом и морфологией в наибольшей степени востребованы при установлении эквивалентности смысла распознаваемого текста заданному смысловому эталону для случая, когда указанный эталон описывается конечным пользователем на некотором предметно-ориентированном подмножестве ЕЯ.

Следует отметить, что в общих чертах установить факт Семантической Эквивалентности (СЭ) означает доказать идентичность ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами.

Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система Exactus [1].

Тем не менее, существуют задачи сравнения смысла, отличные от традиционного для поисковых систем взаимодействия "запрос-ответ". Примером является интерпретация текста ответа на тестовое задание открытой формы в системе автоматизированного контроля знаний [2]. Необходимо не столько отобразить ответ на предметную область, сколько оценить его близость ответу, "правильному" с точки зрения преподавателя, конструировавшего тест. Анализ близости высказываний здесь требует учета лексико-функциональной синонимии, в частности - расщепленных значений и конверсивов [3]. В более общем случае многих обучаемых мы имеем задачу текстовой кластеризации [4].

По оценке Г.С. Осипова [5], требуется более детальное исследование свойств семантических связей и в самой коммуникативной грамматике.

Как было показано нами ранее [4], формализация понятий Предметной Области, представляющих участников тех или иных ситуаций, предполагает исследование сочетаемости соответствующих существительных со словами, синтаксически главными по отношению к ним. Актуальным здесь является задействование методов машинного обучения как в процессе формирования указанных связей, так и в целом для изучения взаимодействия семантики, синтаксиса и морфологии при установлении СЭ.

Основной причиной ограниченности использования методов распознавания и обучения в лингвистических процессорах является сложность моделирования неограниченного усложнения предложения естественного языка. Вместе с тем, разумные ограничения на предметную область ЕЯ-высказываний в совокупности с ограничениями ситуационного плана позволяют эффективно исследовать законы изменения буквенного состава слов анализом близости символьных последовательностей. Тем более, что одним из показателей морфологической зависимости в языках с развитой морфологией является флексия - изменяющаяся при склонении или спряжении часть слова, находящаяся в конце словоформы. Так, в русском языке из флексий вычисляется большая часть грамматических категорий, а сами флексии приписываются грамматическим значениям. Это позволяет в ряде случаев обнаруживать зависимость между словоформами, отсутствующими в словаре. Взаимовлияние морфологического и синтаксического анализа состоит в том, что грамматическое значение как основа поиска морфологической зависимости может быть однозначно проинтерпретировано только вследствие фиксации того синтаксического отношения, которое служит средством выражения этой зависимости [6].

Разработка математической модели процесса выделения и обобщения синтаксического отношения в языке с развитой морфологией является *целью* настоящей работы.

Грамматика русской морфологии и флективные классы

Предлагаемое решение проблемы основано на закономерностях выражения смысла в ЕЯ его носителем.

Как уже обсуждалось нами ранее [7], языковой опыт человека можно разделить в соответствии с разделением концептуальной картины мира. При этом основополагающим является понятие ситуации употребления ЕЯ как основы его генезиса.

Под ситуацией употребления ЕЯ понимают описание нового социального опыта (содержания совместных действий) средствами этого ЕЯ. Формально фиксируемый ситуацией S языковой контекст представляется тройкой:

$$S = (O, R, T), \quad (1)$$

где O есть множество объектов-участников S, R - множество отношений между $o \in O$, T - множество форм языкового описания S.

Предположим, что в качестве T выступает множество синонимичных (с точки зрения носителя ЕЯ) фраз, причем каждая из них описывает одну ситуацию действительности (относительно языкового контекста S). Положим выбор $T_i \in T$ для описания ситуации S равновероятным.

При использовании последовательности соподчиненных слов

$$S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\} \quad (2)$$

в качестве основы выделения $o \in O$ в множество R войдут синтаксические отношения R_q:

$$v_l R_q v_{l+1}, \ldots, v_{n(k,i)} R_q m_{ki} \quad (3)$$

для всех S_{ki}, $i = 1, \ldots, |T|$.

Здесь:

- v_1 – предикатное слово (глагол, либо отглагольное существительное), которое обозначает ситуацию;
- m_{ki} – существительное и обозначает некоторое понятие, значимое в v_1;
- $\forall v_l \in \{v_2, \ldots, v_{n(k,i)}\}$ – существительное;
- k – порядковый номер последовательности среди выявленных из T_i;
- $n(k,i)$ – количество соподчиненных существительных последовательности;
- q – тип отношения R_q, он характеризуется падежом зависимого слова и предлогом для связи главного и зависимого слова. При этом q соответствует имени синтагмы, которая определяет бинарное отношение вида (3).

Поскольку S есть (по определению) полное и независимое описание контекста, то имеем задачу:

Задача 1. На основе ЕЯ-фраз из T найти R, рассматривая отношения между $o \in O$ в качестве признаков последних относительно (1).

Рассмотрим $T_i \in T$ с точки зрения символов, которые его составляют. Для $\forall T_i \in T$ справедливо:

$$T_i = T_i^C \cup T_i^F,$$

где T_i^C – общая неизменная часть для всех $T_i \in T$, T_i^F – флективная часть.

На множестве T_i^F выражаются синтагматические зависимости, которые задаются с помощью R. Если $T_i = \bigcup_j W_{ij}$, то, соответственно,

$$W_{ij} = W_{ij}^C \cup W_{ij}^F. \quad (4)$$

Здесь W_{ij} – буквенный состав слова, $W_{ij}^C \subset T_i^C$ – неизменная, $W_{ij}^F \subset T_i^F$ – флективная часть.

Таким образом, попарным сравнением W_{ij} различных T_i требуется найти:

1). W_{ij}^C и W_{ij}^F каждого W_{ij} при $|W_{ij}^C| \to \max$;

2). Отношение R_q, определяющее допустимость сочетания (W_{ij}^F, W_{ik}^F), $k \neq j$.

Введем в рассмотрение индексное множество J для неизменных частей всех слов, употребленных во всех фразах из T.

Определение 1. Моделью L линейной структуры предложения $T_i \in T$ будем называть упорядоченную совокупность индексов $j \in J$ неизменных частей слов, присутствующих в T_i.

При этом порядок индексов в L идентичен порядку следования соответствующих слов в T_i. Поэтому $L(T_i)$ позволяет однозначно восстановить ЕЯ-фразу T_i на множестве всех слов для всех фраз из T. И наоборот, для $\forall T_i \in T$ на индексном множестве J можно однозначно построить $L(T_i)$.

Для построения множества R в (1) необходимо найти совокупность указанных моделей, удовлетворяющих требованиям проективности. С учетом линейной природы синтагм дополним ограничения на проективность [8], используемые в системах анализа текстов, следующим образом.

Пусть $h(j, L(T_i))$ – позиция индекса j в модели $L(T_i)$. Тогда множество связей относительно $L(T_i)$

$$D: T_i \to \{ (h(j, L(T_i)), h(k, L(T_i))) : j \neq k \}.$$

Определение 2. Связь

$$d_{qi} = (h(j, L(T_i)), h(k, L(T_i)))$$

является допустимой для модели $L(T_i)$, если

$$\exists \{T_l, T_m\} \subset T,\ l \neq m,$$

причем и $L(T_l)$, и $L(T_m)$ содержат в качестве подпоследовательности либо $\{j,k\}$, либо $\{k,j\}$. При этом пара индексов (j,k) соответствует одной синтагме, а индекс q – типу синтаксического отношения, которое ей соответствует.

Положим, что для $\forall T_i \in T$, $i = 1, \ldots, |T|$, все $d_{qi} \in D(T_i)$ удовлетворяют *Определению 2*.

Определение 3. Будем считать, что модель $L(T_i)$ проективна относительно R в (1), если

$\sum_{q=1}^{|D(T_i)|} \Delta_{qi} \le |L(T_i)|$, где

$\Delta_{qi} = |h(j, L(T_i)) - h(k, L(T_i))|$.

На основе $\bigcup_i D(T_i)$ формируется граф синтагм (V^J, I^J). Элементами множества вершин V^J этого графа являются множества пар (j,k), $\{j,k\} \subset J$, сгруппированных по некоторому общему для них индексу k. Множества E_1 и E_2, входящие в V^J, будут соединены ребром из I^J, если $\exists\{j,k,m\} \subset J$: $(j,k) \in E_1$, $(k,m) \in E_2$ и $j \ne m$.

Анализом (V^J, I^J) строится дерево-прецедент (V_1^J, I_1^J) для $\bigcup_i T_i$, $i = 1,\ldots,|T|$. Формально

$$V_1^J = J,\ I_1^J = \{(j,k) : \exists E \in V^J, (j,k) \in E\}. \quad (5)$$

При этом индекс $k \in V_1^J$ соответствует корню дерева (V_1^J, I_1^J), если $\exists E_1 \in V^J$, в котором пары индексов сгруппированы по k, $|E_1| > 1$, а k не содержится ни в одной паре индексов для $\forall E_2 \in V^J$: $E_1 \ne E_2$.

Содержательно корень соответствует предикатному слову в (1), которое (по определению) обозначает ситуацию. Поскольку исследуемая проблема точности синтаксического анализа, в частности, при использовании технологии прикладного морфологического анализа без словаря [6], характерна для ситуаций (1) с двумя и более участниками, то число дочерних узлов у корня полагается больше одного.

Будем использовать маршруты в дереве (5) для выделения классов отношений из R в (1) согласно сформулированной нами *Задаче 1*. Данная задача наиболее естественно решается методами Анализа Формальных Понятий (АФП, [9]).

Рассмотрим множество флексий как множество формальных объектов:

$$G^F = \{ f_{ij} : f_{ij} = \bullet(W_{ij}^F) \},$$

где $i = 1,\ldots,|T|$, а $\bullet$ есть операция конкатенации, которая последовательно выполняется над символами из W_{ij}^F в (4).

Введем в рассмотрение формальный контекст:

$$K^F = (G^F, M^F, I^F), \quad (6)$$

в котором $M^F = G^F$, а $I^F \subseteq G^F \times M^F$. При этом:

$$I^F = \{ (f_{ij}, f_{ik}) : s(j,k) = true, \{j,k\} \subset J \}.$$

Отношение s определяется рекурсивно на основе (V^J, I^J):

1). $s(j_1, j_1) = true$;

2). $s(j_1, j_2) = true$ в одном из следующих двух случаев:

– $\exists E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists j_3 \in J$, для которого $s(j_2, j_3) = true$;

– $\exists (E_1, E_2) \in I^J$: $\exists j_3 \in J$, при этом $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, а $s(j_3, j_2) = true$.

Модель (6) выделяет классы в R по характеру изменения флективной части зависимого слова в $\forall R_q \in R$ с учетом бинарности R_q.

Рассмотрим задачу поиска флексий для слов в составе расщепленных значений и конверсивов.

Введем следующие функции: $prep : w_{ij} \to p_y$, которая ставит в соответствие каждому $w_{ij} = \bullet(W_{ij})$ предлог p_y для связи с зависимым словом; $case : w_{ij} \to c_y$, которая ставит в соответствие каждому именному w_{ij} обозначение его падежа $c_y \in \{"nom","gen","dat","acc","ins","loc"\}$. Соответствие между словом и его начальной формой зададим функцией $norm$.

Опираясь на описанные в [3] правила конверсивных замещений и обобщая введенное нами в [4] понятие Расщепленного Предикатного Значения (РПЗ), сформулируем определение конверсива следующим образом.

Определение 4. Пусть S_1 и S_2 – пара множеств последовательностей вида (2). Применительно к $\{S_1, S_2\}$ имеет место конверсив, если для $\forall S_{k1} \in S_1$ найдется $S_{j2} \in S_2$ такая, что возможны следующие случаи взаимного соответствия S_{k1} и S_{j2}.

Случай 1.

$$S_{k1} = \{ v_{11}', v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1} \},$$

$$S_{j2} = \{ v_{21}', v_{k2}', v_{k3}, \ldots, v_{kn(k,1)}, m_{k1} \}.$$

При этом $norm(v_{11}') \ne norm(v_{21}')$, $norm(v_{k2}) = norm(v_{k2}')$, причем в общем случае $prep(v_{11}') \ne prep(v_{21}')$, а $case(v_{k2}) \ne case(v_{k2}')$.

Случай 2.

$$S_{k1} = \{ v_{11}', v_{12}', v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1} \},$$

$$S_{j2} = \{ v_{21}', v_{k2}', v_{k3}, \ldots, v_{kn(k,1)}, m_{k1} \}.$$

Здесь $norm(v_{k2}) = norm(v_{k2}')$, $case(v_{k2}) \neq case(v_{k2}')$ (в общем случае), но при этом для S_{j2} $\exists S_{k1}' \in S_1$, $S_{k1}' \neq S_{k1}$, такая, что $\{S_{k1}', S_{j2}\}$ соответствует *Случаю 1*, а для S_{k1} $\exists S_{j2}' \in S_2$, $S_{j2}' \neq S_{j2}$: $\{S_{k1}, S_{j2}'\}$ также соответствует *Случаю 1* взаимного соответствия последовательностей вида (2).

Замечание 1. Положим $v_{21} = norm(v_{21}')$ в S_{j2} для *Случая 1* и *Случая 2*, $v_{11} = norm(v_{11}')$ и $v_{12} = norm(v_{12}')$ в S_{k1} для *Случая 2*, соответственно. По аналогии с РПЗ будем называть $\{v_{11}, v_{12}\}$ Расщепленным Конверсивом для v_{21}.

Замечание 2. Рассматриваемые конверсивные замены включают в себя как простые перестановки актантов исходного слова на другие места без расщепления последнего, так и замены РПЗ на их нерасщепленные семантические эквиваленты с последующей перестановкой актантов. В качестве замен без расщепления могут быть рассмотрены и синонимические замещения, описываемые Лексической Функцией *Syn* [3]. Здесь для *Случая 1* мы имеем: $k = j$, $prep(v_{11}') = prep(v_{21}')$, а $case(v_{k2}) = case(v_{k2}')$.

Как следует из *Определения 4*, для слов в составе РПЗ и конверсивов не может быть найдено представление (4) попарным сравнением буквенного состава слов во всех $T_i \in T$.

Рассмотрим

$$T_i^{Cnc} = \{ w_{ij} : w_{ij} = \bullet(W_{ij}) \}.$$

Положим также, что $\exists T_i^P \subset T_i$, определяющее последовательность:

$$P_i^{Cnc} = \{ u_k : u_k = \bullet(W_k^P), \bigcup_k W_k^P = T_i^P \},$$

где $W_k^P \in T_i$ – последовательность символов слова, для которого не найдено представления (4).

Лемма 1. Последовательность P_i^{Cnc} содержит предикатное слово, если $\exists \{j, 0, k\} \subset L(T_i)$: $\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{Cnc}$, где $\{u_1, \ldots, u_p\} = P_i^{Cnc}$, $p = |P_i^{Cnc}|$.

Доказательство следует из определения корня дерева (V_1^J, I_1^J) и сделанного допущения о числе участников ситуации (1) с учетом проективности $L(T_i)$.

Пусть для последовательности P_i^{Cnc} выполняется условие *Леммы 1*.

Лемма 2. Слово $u_k \in P_i^{Cnc}$ принадлежит РПЗ, если $\exists T_j \in T$: $L(T_j) \neq L(T_i)$, а $u_k \in P_j^{Cnc}$, где P_j^{Cnc} также отвечает условию *Леммы 1*. При этом $\neg \exists T_k \in T$: $P_k^{Cnc} \subset P_i^{Cnc}$, а $L(T_k) \neq L(T_j)$ и $L(T_k) \neq L(T_i)$.

Доказательство следует из доказанной *Леммы 1* и определения множества ребер в графе (V^J, I^J).

Замечание 3. При выполнении условия *Леммы 2* u_k может быть в том числе и зависимым словом в составе РПЗ.

Пусть $P_i^{Cnc'}$ – последовательность слов, удовлетворяющих условию *Леммы 2*.

Теорема 1. Для формирования контекста (6) при наличии РПЗ либо конверсива необходимо и достаточно найти множество $T' \subset T$:

$$T' = \{ T_i : |P_i^{Cnc'}| \to \max \}. \quad (7)$$

Доказательство теоремы следует из доказанной *Леммы 2*.

Помимо выполнения условия *Теоремы 1*, ключевым требованием при отборе $T_i \in T$ является минимум слов, не представляемых соотношением (4). Для $\forall u_k \in \bigcup_i P_i^{Cnc'}$, $T_i \in T'$ представление (4) формируется сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{Cnc}$: $T_l \in (T \setminus T')$. При этом необходимо, чтобы $2|W_k^C| > |W_k^F| + |W_j^F|$, где $W_k^P = W_k^C \cup W_k^F$, а $W_j^P = W_j^C \cup W_j^F$.

Замечание 4. Если $P_i^{Cnc'} \cap P_i^{Cnc} \neq \varnothing$, то $\forall u_m \in (P_i^{Cnc} \setminus P_i^{Cnc'})$ есть предлог и представляется вместе со словом, стоящим слева от него в P_i^{Cnc}.

С учетом $P_i^{Cnc'}$ дерево (5) преобразуется следующим образом:

1) Корень изменяется с $k = 0$ на значение k для $u_k \in P_i^{Cnc'}$, имеющего максимальную встречаемость в различных T_i^{Cnc}.
2) Левое поддерево остается без изменений.
3) Правое поддерево перевешивается на узел j для $u_j \in P_i^{Cnc'}$ наименьшей встречаемости.
4) В паре $\{u_l, u_m\} \subset P_i^{Cnc'}$ дочерним будет узел для слова с меньшей встречаемостью.

В итоге основу формирования контекста (6) составляют те T_i, которые наиболее полно описывают ситуацию (1).

Рассмотрим свойства контекста K^F, актуальные для выделения морфологических классов слов из T'.

Пусть L – базис импликаций [9], а $\Re^F$ – решетка Формальных Понятий (ФП) для контекста K^F.

Утверждение 1. ФП $\left(A^F, B^F\right): A^F \subseteq G^F, B^F \subseteq M^F$ соответствует $v_1 \in S_{ki}$ в (2), если $\exists\left(\Pr \to Cs\right) \in L: \left|\Pr\right| = 1$ и $\Pr \cup Cs = B^F$. При этом наличие импликации $\left(\Pr_1 \to Cs_1\right) \in L: \Pr \subset Cs_1$ допускается только тогда, когда $\Pr_1 \cup Cs_1 = B^F$.

Утверждение 2. Применительно к $m_{ki} \in S_{ki}$ в (2) ФП $\left(A^F, B^F\right)$ соответствует прилагательному, если B^F есть множество признаков некоторого элемента множества G^F и $\neg\exists\left(\Pr \to Cs\right) \in L: \Pr \cup Cs = B^F$.

В противном случае ФП $\left(A^F, B^F\right)$ соответствует существительному из $\{v_2, \ldots, m_{ki}\} \subset S_{ki}$.

Синтаксические отношения выделяются анализом наименьшей верхней грани каждой пары ФП в $\Re^F$ и образуют классы по сходству характера флексии зависимого слова. Отдельному классу соответствует область в решетке, а Наименьшее Общее Суперпонятие [9] этой области – прецеденту класса.

Оценка выделенных классов отношений дается в сопоставлении с контекстом вида (6) по результатам работы программы синтаксического анализа. *Актуальной* здесь *является* автоматическая лингвистически интерпретируемая классификация выявляемых конверсивов и РПЗ.

Введем в рассмотрение формальный контекст:

$$K^{Conv} = \left(G^{Conv}, M^{Conv}, I^{Conv}\right), \quad (8)$$

в котором согласно *Определению 4*

$$G^{Conv} = \left\{v_{21} : v_{21} = norm\left(v_{21}'\right)\right\},$$

$$M^{Conv} = \left\{v^{Conv} : v^{Conv} = \begin{Bmatrix} v_{11} \\ v_{12} \bullet " : " \bullet v_{11} \end{Bmatrix}\right\}.$$

Здесь:

- $v_{11} = norm\left(v_{11}'\right)$, $v_{12} = norm\left(v_{12}'\right)$;
- операция конкатенации имеет место для *Случая 2* из рассматриваемых *Определением 4*;
- Отношение $I^{Conv} \subseteq G^{Conv} \times M^{Conv}$ ставит в соответствие каждому варианту конверсивной замены $v_{21} \in G^{Conv}$ заменяемый конверсив $v^{Conv} \in M^{Conv}$.

Пусть $\Re^{Conv}$ есть решетка ФП для контекста (8). Введем индексы: 1 – для контекстов, формируемых с применением предложенной нами модели, 2 – для контекстов, формируемых с применением программы синтаксического анализа. Положим, что $\Re_2^{Conv}$ и $\Re_2^F$ формируются на основе неструктурированного текста заданной тематики, включающего подмножество T относительно языкового контекста ситуации (1). Мощность этого подмножества зависит от репрезентативности текста [4].

Под показателем репрезентативности здесь следует понимать количество форм языкового описания заданной ситуации, использованных при формировании $\Re_1^F$ и $\Re_1^{Conv}$, которые присутствуют в анализируемом тексте.

Каждая область решетки $\Re^{Conv}$ (вне зависимости от исходных данных для построения) при единственности Наибольшего Общего Подпонятия и Наименьшего Общего Суперпонятия получает содержательную интерпретацию группы смысловых отношений со сходным составом аргументов и сходным характером перестановок аргументов (типом конверсии).

Введем в рассмотрение базисы импликаций: L_1^{Conv} – базис импликаций для K_1^{Conv}, L_2^{Conv} – для K_2^{Conv}.

Утверждение 3. Будем считать классификацию отношений из R в (1) на основе контекста (6) допустимой применительно к случаю наличия в T фраз, отвечающих *Определению 4*, если $\Re_1^F \subset \Re_2^F$ и $\exists\left(\Pr_1^{Conv} \to Cs_1^{Conv}\right) \in L_1^{Conv}: \exists\left(\Pr_2^{Conv} \to Cs_2^{Conv}\right) \in L_2^{Conv}$, где $\Pr_1^{Conv} \cap \Pr_2^{Conv} \neq \varnothing$ и $Cs_1^{Conv} \cap Cs_2^{Conv} \neq \varnothing$.

При этом случай $\Re_1^F = \Re_2^F$ не обязательно соответствует тексту с максимальной репрезентативностью по сформулированному нами критерию. Встречаемость тех или иных сочетаний флексий находится в зависимости и от количества описываемых текстом ситуаций. В частности, текстом может описываться несколько ситуаций, близких рассматриваемой по составу участников и их ролевой ориентации. Анализ взаимной близости самих ситуаций в этом случае – тема отдельного обсуждения.

Экспериментальная апробация

Исходными данными для формирования контекстов K_1^F и K_1^{Conv} были правильные ответы на тестовое задание открытой формы.

Вопрос теста: «Каковы негативные последствия переобучения при скользящем контроле?»

В итоге было получено двадцать семь вариантов правильного ответа на данный вопрос (рис.1).

На рис.2 представлена решетка $\Re_1^F$ для T' (табл.1). Формирование контекстов K_2^F и K_2^{Conv} также производилось по вариантам правильных ответов на тесты открытой формы, но более широкой тематики проблем качества обучения алгоритмов. Морфологический и синтаксический анализ текста осуществляется программой Cognitive Dwarf [10].

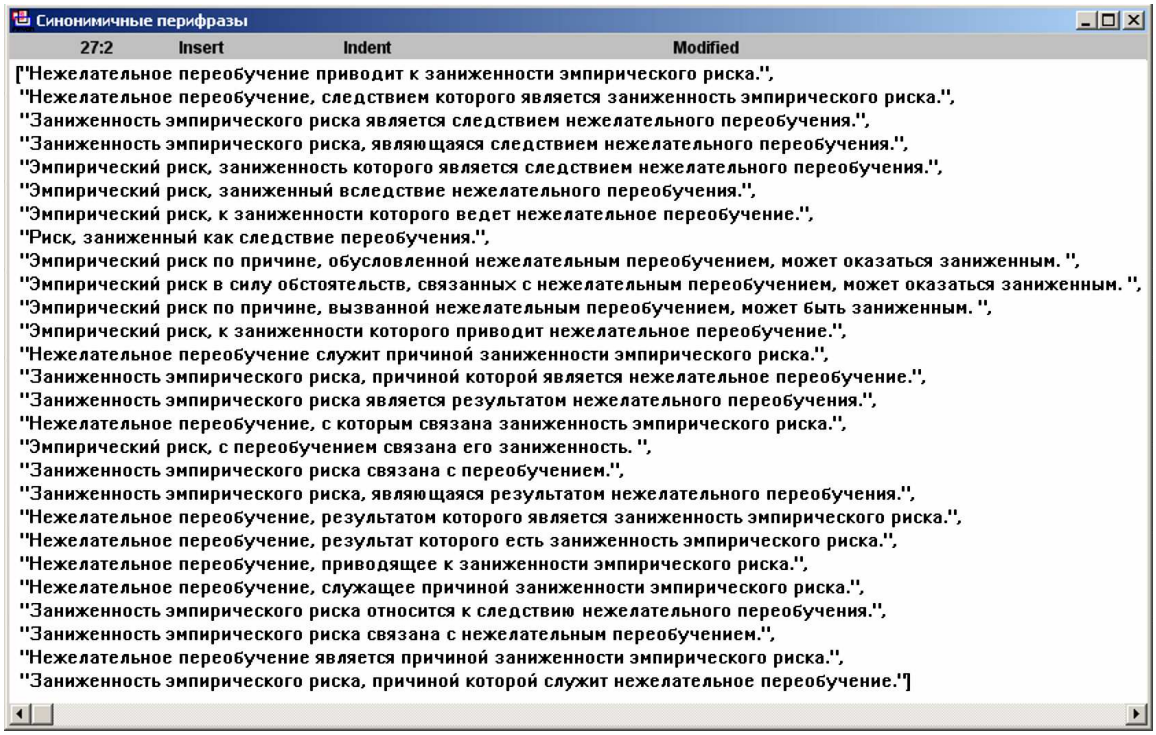

Рис.1. Исходные данные для формирования K_1^F *и* K_1^{Conv}

Таблица 1. Правильные ответы $T_i \in T'$ *в (7) из представленных на рис. 1*

основа	флективная часть + предлог					
занижен	ость	ости	ость	ости	ость	ости
эмпирическ	ого	ого	ого	ого	ого	ого
риск	а	а	а	а	а	а
нежелательн	ого	ое	ого	ое	ым	ое
переобучении	я	е	я	е	ем	е
являя	ется	—	ется	ется	—	—
следствии	ем	—	—	—	—	—
служ	—	ит	—	—	—	—
причин	—	ой	—	ой	—	—
результат	—	—	ом	—	—	—
связан	—	—	—	—	a:c	—
привод	—	—	—	—	—	ит:r

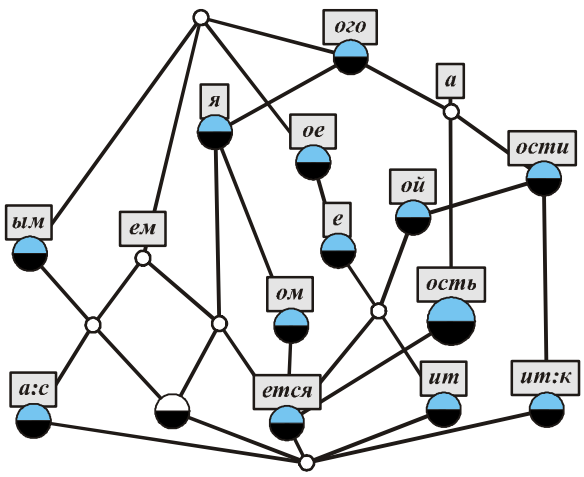

Рис.2. Синтаксические отношения на основе сочетаний флексий

В целях краткости изложения здесь не приводится $\mathfrak{R}_2^F$. Решетки $\mathfrak{R}_1^{Conv}$ и $\mathfrak{R}_2^{Conv}$ представлены на рис.3 и 4. Визуализацию решетки диаграммой линий [9] в настоящей работе выполняет ПО Concept Explorer [11], реализующее методы АФП. Область в $\mathfrak{R}_1^{Conv}$ (рис.3), отвечающая условию *Утверждения 3*, обозначена прямоугольником.

Заключение

Сферой рассмотрения настоящей работы были классы отношений для слов с изменяемой частью в конце словоформы. Тем не менее, чрезвычайно интересным является дальнейшее развитие предложенного в работе метода применительно к изменениям в составе основы слова. Здесь следует отметить беглые гласные, чередования гласных и согласных в составе основы, а также вариантные формы основ.

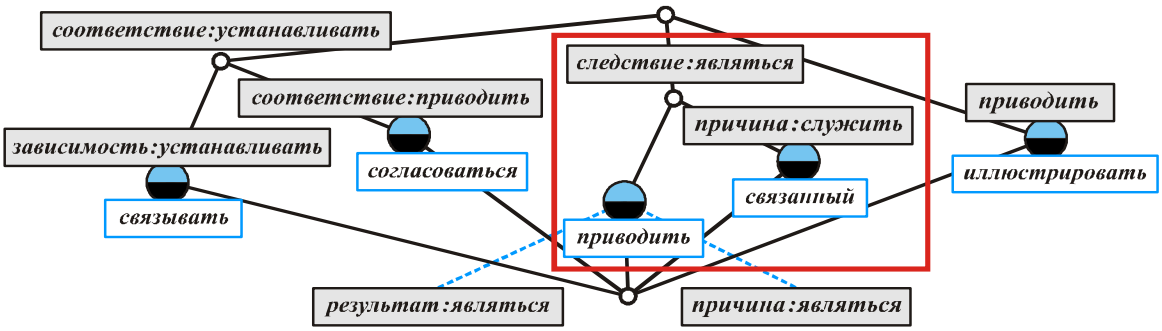

Рис.3. Группировка РПЗ и конверсивных замен по результатам Cognitive Dwarf

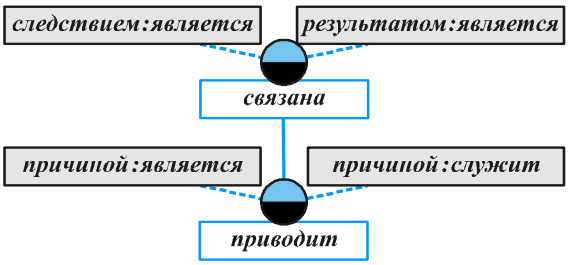

Рис.4. РПЗ и конверсивы в составе фраз из T' (табл.1)

В частности, отдельного рассмотрения заслуживает включение в синтаксические контексты вида (2) имен числительных, для которых особенно актуально явление чередования в основах. Пример: «триста», «трехсот», «тремстам», «триста», «тремястами», «трехстах».

В связи с этим другое немаловажное направление дальнейших исследований – распознавание слов-паронимов в составе синонимичных фраз. Наиболее плодотворные результаты данное исследование даст совместно с количественным изучением вариативности на уровне морфем и лексем русского языка [12].

Благодарности

Работа выполнена при поддержке РФФИ (проект №06-01-00028).

Литература

1. **Тихомиров, И.А.** Интеграция лингвистических и статистических методов поиска в поисковой машине "Exactus" [Электронный ресурс] / И.А. Тихомиров, И.В. Смирнов // Межд. Конф. по компьютерной лингвистике "Диалог-2008". http://www.dialog-21.ru/dialog2008/materials/html/80.htm (дата обращения: 18.11.2009).
2. **Васильев, В.И.** Методологические правила конструирования компьютерных тестов [Текст] / В.И. Васильев, А.Н. Демидов, Н.Г. Малышев, Т.Н. Тягунова - М.: МГУП, 2000. – 64 с.
3. **Мельчук, И.А.** Опыт теории лингвистических моделей "Смысл⇔текст": Семантика, синтаксис [Текст] / И.А. Мельчук. – М.: Шк. "Языки русской культуры", 1999. – 345 с.
4. **Mikhailov, D.V.** Formation and clustering of Russian's nouns's contexts within the frameworks of Splintered Values [Текст] / D.V. Mikhailov, G.M. Emelyanov, N.A. Stepanova // 9th Int. Conf. "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-9-2008). – Nizhni Novgorod. – NNSU. – 2008. – Vol.2. – P. 39-42.
5. **Осипов, Г.С.** Приобретение знаний интеллектуальными системами: Основы теории и технологии [Текст] / Г.С. Осипов. – М.: Наука, 1997. – 112 с.
6. **Ножов, И.М.** Синтаксический анализ [Электронный ресурс] / И.М. Ножов // Компьютерра. – 2002. – №21 (446). http://www.computerra.ru/offline/2002/446/18250/ (дата обращения: 18.11.2009).
7. **Емельянов, Г.М.** Концептуально-ситуационное моделирование процесса перифразирования высказываний Естественного Языка как обучение на основе прецедентов [Текст] / Г.М. Емельянов, А.Н. Корнышов, Д.В. Михайлов // Искусственный интеллект. – 2006. - №2. – С. 72-75.
8. **Кибрик, А.Е.** Очерки по общим и прикладным вопросам языкознания / А.Е. Кибрик. – М.: КомКнига, 2005. – 332 с.
9. **Ganter, B.** Formal Concept Analysis – Mathematical Foundations [Текст] / Ganter B. and Wille R. - Berlin : Springer-Verlag, 1999. - 284 с.
10. Программный пакет синтаксического разбора и машинного перевода [Электронный ресурс] // http://cs.isa.ru:10000/dwarf/ (дата обращения: 18.11.2009).
11. The Concept Explorer [Электронный ресурс] // http://conexp.sourceforge.net (дата обращения: 18.11.2009).
12. **Гусев, В.Д.** Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) [Электронный ресурс] / В.Д. Гусев, Н.В. Саломатина // Межд. Конф. по компьютерной лингвистике "Диалог-2004". http://www.dialog-21.ru/Archive/2004/Salomatina.htm (дата обращения: 18.11.2009).

References

1. **Tikhomirov, I.A.** Integration of linguistic and statistic methods in search engine "Exactus" [Electronic resource] / I.A. Tikhomirov, I.V. Smirnov // Computional linguistics and intellectual technologies: International Conference "Dialogue-2008". http://www.dialog-21.ru/dialog2008/materials/html/80.htm. - (in Russian, access date: 18.11.2009).
2. **Vasilev, V.I.** Methodological rules of designing of computer tests [Text] / V.I. Vasilev, A.N. Demidov, N.G. Malyshev, T.N. Tjagunova - Moscow: MSUPA, 2000. – 64 p. – (in Russian).
3. **Mel'chuk, I.A.** An Attempt at a Theory of "Meaning⇔Text" Linguistic Models: Semantics, Syntax [Text] / I.A. Mel'chuk. – Moscow: Languages of Slavonic Culture, 1999. – 345 p. – (in Russian).

4. **Mikhailov, D.V.** Formation and clustering of Russian's nouns's contexts within the frameworks of Splintered Values [Text] / D.V. Mikhailov, G.M. Emelyanov, N.A. Stepanova // 9th Int. Conf. "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-9-2008). – Nizhni Novgorod. – NNSU. – 2008. – Vol.2. – P. 39-42.
5. **Osipov, G.S.** Knowledge acquisition by intellectual systems: fundamentals of theory and technology [Text] / G.S. Osipov. – Moscow: Nauka, 1997. – 112 p. – (in Russian).
6. **Nozhov, I.M.** Syntactic analysis [Electronic resource] / I.M. Nozhov // Computerra. – 2002. – No21 (446). http://www.computerra.ru/offline/2002/446/18250/. - (in Russian, access date: 18.11.2009).
7. **Emelyanov, G.M.** Conceptually-situational modeling of process of synonymic transformation of the natural-language statements as machine learning on the basis of precedents [Text] / G.M. Emelyanov, A.N. Kornyshov, D.V. Mikhailov // Scientific-theoretical magazine «Artificial intelligence». – 2006. - No2. – P. 72-75. – (in Russian).
8. **Kibrik, A.E.** Sketches on the general and applied questions of linguistics [Text] / A.E. Кибрик. – Moscow: KomKniga, 2005. – 332 p. – (in Russian).
9. **Ganter, B.** Formal Concept Analysis – Mathematical Foundations [Текст] / Ganter B. and Wille R. - Berlin : Springer-Verlag, 1999. - 284 p.
10. Software package of syntactic analysis and machine translation [Electronic resource] // http://cs.isa.ru:10000/dwarf/. - (in Russian, access date: 18.11.2009).
11. The Concept Explorer [Electronic resource] // http://conexp.sourceforge.net. - (access date: 18.11.2009).
12. **Gusev, V.D.** Algorithm of revealing of set expressions with taking into account their variability (morphological and combinatorial) [Electronic resource] / V.D. Gusev, N.V. Salomatina // Computional linguistics and intellectual technologies: International Conference "Dialogue-2004". http://www.dialog-21.ru/Archive/2004/Salomatina.htm. - (in Russian, access date: 18.11.2009).

SEMANTIC CLUSTERING OF SUBJECT-ORIENTED LANGUAGES'S TEXTS (MORPHOLOGY AND SYNTAX)

Dmitrii Vladimirovich Mikhailov[1] (docent, e-mail: Dmitry.Mikhaylov@novsu.ru),
Gennadii Martinovich Emelyanov[1] (professor, e-mail: Gennady.Emelyanov@novsu.ru)
[1] State Educational Institution of Higher Vocational Education "Yaroslav-the-Wise Novgorod State University"

Abstract

The problem considered is the semantic clustering of texts in Subject-Oriented Natural Language. The approach offered is to elaborate perfomance criteria for syntactic analysis as a toolbox to reveal objects and attributes. Especial attention is given to the Splintered Values and conversives within nouns's syntactic contexts.

Key words: text mining, natural language, subject area, semantic equivalence, knowledge clustering, lattice theory.

Поступила в редакцию 20.11.2009 г.