Морфология и синтаксис в задаче семантической кластеризации.

Михайлов Д. В., Емельянов Г. М.

Новгородский Государственный Университет имени Ярослава Мудрого

Цель работы.

Разработка *модели* процесса *выделения и классификации* синтаксических отношений на множестве семантически эквивалентных фраз заданного Естественного Языка (ЕЯ) для решения задачи повышения точности синтансического анализа.

Задачи исследования.

1) Разработка концептуальной модели процесса семантической кластеризации ЕЯ-текстов по результатам их синтаксического разбора.
2) Определение границ проблемной области установления семантической эквивалентности ЕЯ-текстов.
3) Разработка математической модели процесса выделения закономерностей сосуществования словоформ в линейном ряду.
4) Выработка рекомендаций по качественному анализу моделей морфологии и синтаксиса естественного языка для практических задач обработки текстов.

Семантическая кластеризация текстов по результатам синтаксического разбора: постановка задачи.

Дано :

G — множество анализируемых текстов заданного Естественного Языка.

Требуется :

1) По результатам синтаксического разбора каждого $T_i \in G$ выявить:

   - множество $V(T_i)$ *ситуаций*, описываемых T_i;
   - множество $M(T_i)$ *объектов* и/или *понятий*, значимых в ситуациях из множества $V(T_i)$;
   - тернарное отношение $I \subseteq G \times M \times V$, ставящее в соответствие каждому $m \in M$, $M = \bigcup_i M(T_i)$, ту ситуацию $v \in V$, $V = \bigcup_i V(T_i)$, в которой он фигурирует относительно T_i.

2) На основе выявленного отношения I выделить в G группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях.

Синтаксический контекст существительного как основа формирования признаков текста.

Определение 1. *Под синтаксическим контекстом существительного понимается последовательность соподчиненных слов вида*

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

Здесь :

- v_1 — *предикатное слово (глагол или отглагольное существительное) и обозначает некоторую ситуацию;*
- m_{ki} — *существительное и обозначает некоторый объект, значимый в ситуации* v_1*;*
- k — *порядковый номер последовательности среди выявленных из текста* T_i*;*
- $n(k,i)$ — *количество соподчиненных существительных* $\{v_2, \ldots, v_{n(k,i)}, m_{ki}\}$.

Кроме того, для всех $\{v_l, v_{l+1}\} \subset S_{ki}$ существует *синтаксическое отношение* R_q:

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

где q — тип отношения, он характеризуется предлогом для связи главного слова с зависимым и падежом зависимого.

Транзитивность отношения R_q дает основание утверждать, что любое существительное из $\{v_2, \ldots, v_{n(k,i)}, m_{ki}\}$ обозначает некоторый объект, значимый в ситуации v_1. При этом q определяет его роль относительно v_1.

Анализ Формальных Понятий и концептуальная кластеризация.

Пусть множество G анализируемых текстов есть множество формальных объектов. Тогда множество M объектов (понятий), фигурирующих в текстах из G, есть множество формальных признаков. Множество ситуаций V, в которых эти объекты (понятия) фигурируют, есть множество значений формальных признаков.

Отношению $I \subseteq G \times M \times V$ ставится в соответствие формальный контекст:

$$K = \big(G, M, V, I\big).$$

Определение 2. *Под Формальным Понятием (ФП) понимается пара* (A, B)*:* $A \subseteq G$*,* $B \subseteq M \times V$*,* $A = B'$*,* $B = A'$*, причем*

$$A' = \big\{(m, v) \colon m \in M,\ v \in V \;\big|\; \forall\, T_i \in A \colon m(T_i) = v\big\},$$
$$B' = \big\{T_i \in G \;\big|\; \forall\, (m, v) \in B \colon m(T_i) = v\big\}.$$

Определение 3. *ФП* (A_1, B_1) *является подпонятием для ФП* (A_2, B_2)*, если* $A_1 \subseteq A_2$*, а* $B_2 \subseteq B_1$ *:* $(A_1, B_1) \leq (A_2, B_2)$*. При этом* (A_2, B_2) *называют суперпонятием для ФП* (A_1, B_1)*, а отношение* $\leq$ *— отношением порядка для ФП.*

Определение 4. *Множество* $\Re(G, M, V, I)$ *всех ФП контекста вместе с отношением* $\leq$ *называется решеткой Формальных Понятий.*

Замечание. *Каждое слово в множестве* M *представлено вместе с предлогом, посредством которого оно связывается с другим словом, синтаксически главным по отношению к нему.*

Синтаксические отношения в ситуации языкового употребления.

Определение 5. *Ситуация языкового употребления есть описание социального опыта человека средствами заданного Естественного Языка (ЕЯ).*

Фиксируемый при этом языковой контекст представляется тройкой:

$$S = (O, R, T),$$

где O есть множество объектов, ассоциируемых с ситуацией, R — множество отношений между $o \in O$, T — множество форм языкового описания S.

Пусть T состоит из тех ЕЯ-фраз исходного множества G, которые описывают одну и ту же ситуацию действительности (относительно языкового контекста S).

В силу произвольности R предположим, что его *элементами* являются *отношения подчинения* в рамках *синтаксического контекста существительного.*

В этом случае все ЕЯ-фразы из T строго синонимичны, а

$$O = \bigcup_{T_i \in T} \{M(T_i) \cup V(T_i)\}.$$

Здесь $V(T_i)$ и $M(T_i)$ содержат словесные обозначения, соответственно, для ситуации S и для объектов (понятий), ассоциируемых с S.

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

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

Синтагматические зависимости как основа выявления синтаксических отношений.

Пусть T — множество фраз заданного Естественного Языка (ЕЯ), описывающих некоторую ситуацию S. Рассмотрим $T_i \in T$ как множество символов.

Для любого $T_i \in T$ 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 — изменяемая часть.

Пусть W_{ij} — буквенный состав слова, j — его порядковый номер в ЕЯ-фразе.

Тогда

$$W_{ij} = W_{ij}^C \cup W_{ij}^F,$$

где $W_{ij}^C \subset T_i^C$ — неизменная, $W_{ij}^F \subset T_i^F$ — флективная часть. На множестве T_i^F выражаются синтагматические зависимости.

Определение 6. *Синтагматические зависимости определяют сосуществование словоформ в линейном ряду и задаются синтаксическими отношениями.*

Таким образом, на основе попарного сравнения W_{ij} различных T_i требуется найти:

1) W_{ij}^C и W_{ij}^F каждого W_{ij} при $\left|W_{ij}^C\right| \to \max$;
2) Синтаксическое отношение R_q, определящее допустимость сочетания слов с буквенным составом флексий W_{ij}^F и W_{ik}^F, $k \neq j$.

Модель линейной структуры фразы Естественного Языка.

Пусть J — индексное множество для неизменных частей всех слов, употребленных во всех синонимичных фразах множества T.

Определение 7. *Моделью L линейной структуры предложения $T_i \in T$ назовем последовательность индексов неизменных частей слов, присутствующих в T_i.*

Пусть $h(j, L(T_i))$ — позиция индекса j в модели $L(T_i)$, где $j \in J$.

Тогда множество связей относительно $L(T_i)$

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

Определение 8. *Связь*

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

является допустимой для модели $L(T_i)$, если $\exists \{T_l, T_m\} \subset T$, $l \neq m$, такие, что и $L(T_l)$, и $L(T_m)$ имеют подпоследовательностью либо $\{j, k\}$, либо $\{k, j\}$.

Положим, что для любого $T_i \in T$ все $d_{qi} \in D(T_i)$ удовлетворяют Определению 8.

Определение 9. *Будем считать, что модель $L(T_i)$ проективна относительно множества синтаксических связей для T, если*

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

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

Классификация синтаксических связей на основе графа синтагм.

Пусть $\bigcup_i D(T_i)$ есть множество связей, допустимых для всех моделей $L(T_i)$ линейных структур синонимичных фраз T_i, описывающих некоторую ситуацию.

Положим также, что J есть индексное множество, на котором задаются модели. При допустимости связи для $\{j,k\} \subset J$ пара (j,k) соответствует одной синтагме.

Определение 10. *Множество пар (j,k), сгруппированных по некоторому общему для них индексу k, есть элемент множества V^J вершин графа синтагм (V^J, I^J). При этом 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 $j \neq m$.*

Пусть

$$G^F = \left\{ f_{ij} \colon f_{ij} = \odot \left(W_{ij}^F \right) \right\}, \; I^F = \left\{ (f_{ij}, f_{ik}) \colon s(j,k) = \text{true}, \; \{j,k\} \subset J \right\}.$$

Здесь $\odot$ есть последовательная конкатенация символов флективной части слова.

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

1) $s(j_1, j_1) = \text{true}$;
2) $s(j_1, j_2) = \text{true}$, если:
   — либо $\exists\, E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists\, j_3 \in J$, для которого $s(j_2, j_3) = \text{true}$;
   — либо $\exists \left(E_1, E_2\right) \in I^J$: $\exists\, j_3 \in J$, при этом $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, а $s(j_3, j_2) = \text{true}$.

Отношению I^F ставится в соответствие формальный контекст:

$$K^F = \left(G^F, M^F, I^F\right), \text{ в котором } M^F = G^F.$$

Будем называть K^F *формальным контекстом сочетаемости флексий.*

Расщепленные Предикатные Значения.

Пусть S_1 и S_2 — пара множеств *последовательностей* соподчиненных слов, каждая из которых есть *синтаксический контекст* некоторого существительного.

Введем функции: $prep$, которая ставит в соответствие слову предлог для связи с зависимым словом; $case$, которая возвращает падеж для слова-существительного; $norm$, которая возвращает начальную форму слова.

Определение 11. *Парой* $\{S_1, S_2\}$ *описывается Расщепленное Предикатное Значение (РПЗ) либо конверсив, если для* $\forall\, S_{k1} \in S_1$ *найдется* $S_{j2} \in S_2$ *такое, что при этом возможны следующие случаи взаимного соответствия* S_{k1} *и* S_{j2}.

Случай (1).

$$S_{k1} = \left\{v'_{11}, v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\},\ S_{j2} = \left\{v'_{21}, v'_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\}.$$

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

Случай (2).

$$S_{k1} = \left\{v'_{11}, v'_{12}, v_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\},\ S_{j2} = \left\{v'_{21}, v'_{k2}, v_{k3}, \ldots, v_{kn(k,1)}, m_{k1}\right\}.$$

Здесь $norm(v_{k2}) = norm(v'_{k2})$, $case(v_{k2}) \neq case(v'_{k2})$ (в общем случае), но при этом для S_{j2} существует $S'_{k1} \in S_1$, $S'_{k1} \neq S_{k1} : \{S'_{k1}, S_{j2}\}$ соответствует *Случаю 1*, а для S_{k1} существует $S'_{j2} \in S_2$, $S'_{j2} \neq S_{j2} : \{S_{k1}, S'_{j2}\}$ также соответствует *Случаю 1*.

Само РПЗ есть пара $\{v_{11}, v_{12}\}$, где $v_{11} = norm(v'_{11})$ и $v_{12} = norm(v'_{12})$.

Поиск флексий слов Расщепленных Предикатных Значений.

Пусть $W_{ij} \subset T_i$ — буквенный состав слова при рассмотрении ЕЯ-фразы T_i как множества символов, а W_k^P — буквенный состав слова, неизменная часть которого не может быть найдена во всех ЕЯ-фразах заданного синонимического множества.

Рассмотрим

$$T_i^{\odot} = \left\{ w_{ij} \colon w_{ij} = \odot(W_{ij}) \right\}.$$

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

$$P_i^{\odot} = \left\{ u_k \colon u_k = \odot\left(W_k^P\right), \bigcup_k W_k^P = T_i^P \right\}.$$

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

Пусть T — множество синонимичных ЕЯ-фраз.

Лемма 2. *Слово $u_k \in P_i^{\odot}$ входит в состав Расщепленного Предикатного Значения (РПЗ), если $\exists\, T_j \in T\colon L(T_j) \neq L(T_i)$, а $u_k \in P_j^{\odot}$.*

При этом $\nexists\, T_k \in T$, для которого $P_k^{\odot} \subset P_i^{\odot}$, а $L(T_k) \neq L(T_j)$ и $L(T_k) \neq L(T_i)$.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих условию *Леммы 2*.

Теорема 1. *Для построения формального контекста сочетаемости флексий при наличии РПЗ необходимо и достаточно найти множество $T' \subset T$:*

$$T' = \left\{ T_i \colon \left|P_i^{\odot\prime}\right| \to \max \right\}.$$

Формальный контекст сочетаемости флексий при наличии РПЗ.

Пусть (V^J, I^J) — граф синтагм, J — индексное множество, на котором задаются модели $L(T_i)$ линейных структур синонимичных фраз множества T. Рассмотрим

$$I_1^J = \left\{(j,k)\colon \exists\, E \in V^J, (j,k) \in E\right\}.$$

Посредством I_1^J связываются объекты и атрибуты в *формальном контексте сочетаемости флексий*. Назовем $\left(V_1^J, I_1^J\right)$, $V_1^J = J$, деревом-прецедентом для T.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих условию *Леммы 2*, а T' — подмножество T, удовлетворяющее условию *Теоремы 1*.

Для $u_k \in \bigcup_i P_i^{\odot\prime}$: $T_i \in T'$ неизменная и флективная части формируются сравнением буквенного состава слова u_k со всеми $u_j \in \bigcup_l P_l^{\odot}$: $T_l \in (T \setminus T')$. Здесь $\forall\, P_l^{\odot}$ образуют слова, у которых изначально не найдена неизменная часть.

При этом необходимо, чтобы

$$2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|,$$

где индексы C относятся к составу неизменной, а F — флективной части слова.

Дерево $\left(V_1^J, I_1^J\right)$ преобразуется следующим образом:

- корень изменяется с $k = 0$ на значение k для слова $u_k \in P_i^{\odot\prime}$, имеющего максимальную встречаемость в различных ЕЯ-фразах из T;
- правое поддерево перевешивается на узел j для слова $u_j \in P_i^{\odot\prime}$ наименьшей встречаемости;
- для $\forall\left\{u_l, u_m\right\} \subset P_i^{\odot\prime}$ дочерним будет узел слова меньшей встречаемости.

Экспериментальная апробация : исходные данные.

Вопрос теста:

«Каковы негативные последствия переобучения при скользящем контроле ?»

Полученные правильные ответы:

«Нежелательное переобучение приводит к заниженности эмпирического риска.»
«Нежелательное переобучение, следствием которого является заниженность эмпирического риска.»
«Заниженность эмпирического риска является следствием нежелательного переобучения.»
«Заниженность эмпирического риска, являющаяся следствием нежелательного переобучения.»
«Эмпирический риск, заниженность которого является следствием нежелательного переобучения.»
«Эмпирический риск, заниженный вследствие нежелательного переобучения.»
«Эмпирический риск, к заниженности которого ведет нежелательное переобучение.»
«Риск, заниженный как следствие переобучения.»
«Эмпирический риск по причине, обусловленной нежелательным переобучением, может оказаться заниженным.»
«Эмпирический риск в силу обстоятельств, связанных с нежелательным переобучением, может оказаться заниженным.»
«Эмпирический риск по причине, вызванной нежелательным переобучением, может быть заниженным.»
«Эмпирический риск, к заниженности которого приводит нежелательное переобучение.»
«Нежелательное переобучение служит причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска, причиной которой является нежелательное переобучение.»
«Заниженность эмпирического риска является результатом нежелательного переобучения.»
«Нежелательное переобучение, с которым связана заниженность эмпирического риска.»
«Эмпирический риск, с переобучением связана его заниженность.»
«Заниженность эмпирического риска связана с переобучением.»
«Заниженность эмпирического риска, являющаяся результатом нежелательного переобучения.»
«Нежелательное переобучение, результатом которого является заниженность эмпирического риска.»
«Нежелательное переобучение, результат которого есть заниженность эмпирического риска.»
«Нежелательное переобучение, приводящее к заниженности эмпирического риска.»
«Нежелательное переобучение, служащее причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска относится к следствию нежелательного переобучения.»
«Заниженность эмпирического риска связана с нежелательным переобучением.»
«Нежелательное переобучение является причиной заниженности эмпирического риска.»
«Заниженность эмпирического риска, причиной которой служит нежелательное переобучение.»

Результат : максимально проективные ЕЯ-фразы с минимумом слов, не нашедших прообразы по буквенному составу.

Таблица 1. Правильные ответы $T_i \in T'$.

основа	флективная часть + предлог					
заниженн	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
переобучени	я	е	я	е	ем	е
явля	естя	—	ется	ется	—	—
следстви	ем	—	—	—	—	—
служ	—	ит	—	—	—	—
причин	—	ой	—	ой	—	—
результат	—	—	ом	—	—	—
связан	—	—	—	—	а:с	—
привод	—	—	—	—	—	ит:к

Здесь T' — множество ЕЯ-фраз, удовлетворяющих условию *Теоремы 1.*

Классификация синтаксических отношений для результирующего множества ЕЯ-фраз.

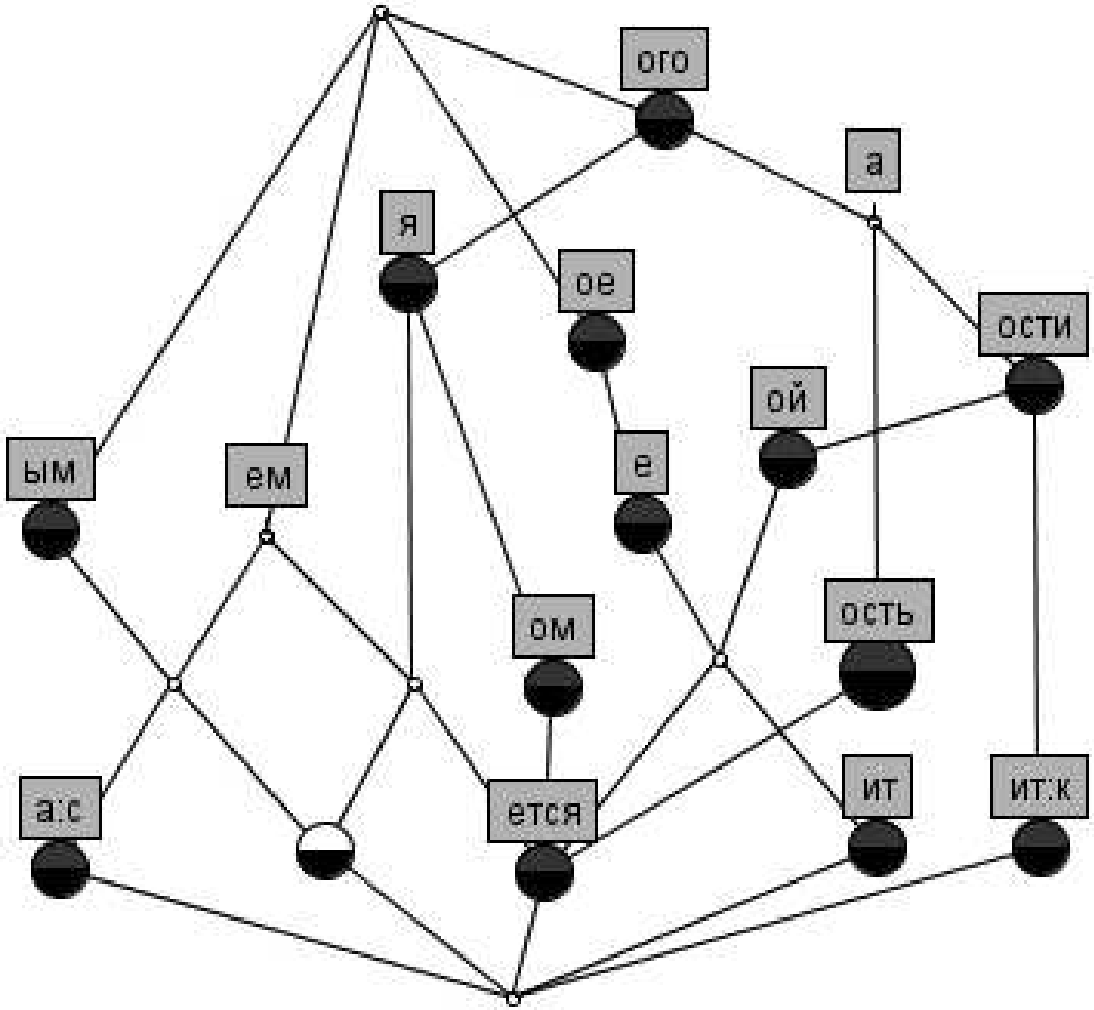

Рис. 1. Синтаксические отношения на основе сочетаний флексий.

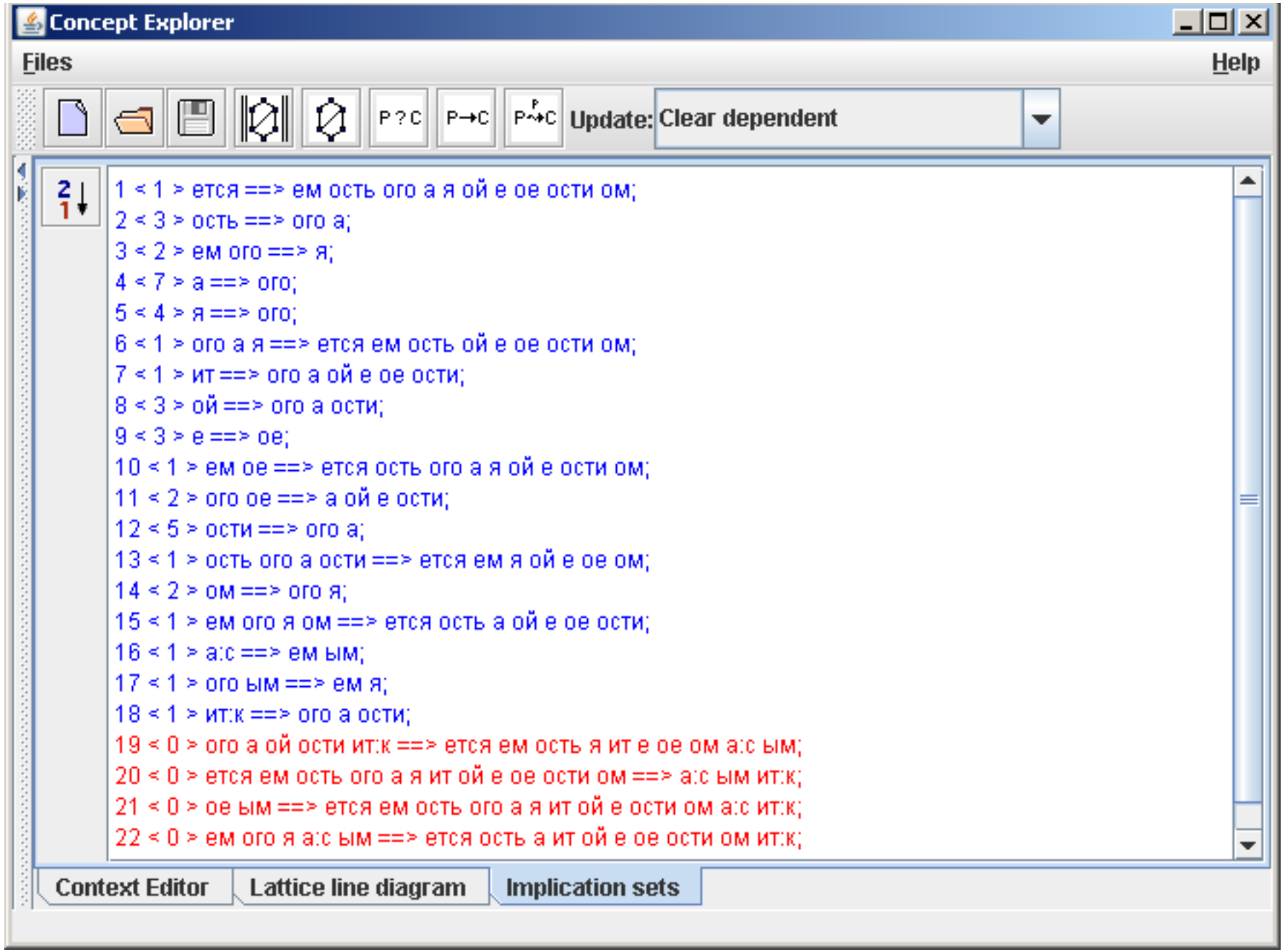

Рис. 2. Базис импликаций на основе результирующего множества ЕЯ-фраз.

Выделение морфологических классов слов.

Пусть $K^F = \left(G^F, M^F, I^F\right)$ есть *формальный контекст сочетаемости флексий* из множества G^F, $M^F = G^F$, $I^F \subseteq G^F \times M^F$, а $\mathcal{L}$ есть базис импликаций для K^F.

Положим также, что синтаксический контекст существительного m_{ki} задается задается последовательностью соподчиненных слов:

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

Правило 1. *Формальное Понятие* (A^F, B^F) *:* $A^F \subseteq G^F$, $B^F \subseteq M^F$, *соответствует предикатному слову* v_1 *в составе* S_{ki}, *если* $\exists\,(Pr \rightarrow Cs) \in \mathcal{L}$: $|Pr| = 1$, *а* $Pr \cup Cs = B^F$.

При этом наличие $(Pr_1 \rightarrow Cs_1) \in \mathcal{L}$: $Pr \subset Cs_1$ *допускается только тогда, когда* $Pr_1 \cup Cs_1 = B^F$.

Правило 2. *Формальное Понятие* (A^F, B^F) *соответствует прилагательному для существительного* m_{ki}, *относительно которого задается синтаксический контекст, если* B^F *есть множество признаков некоторого элемента множества* G^F *и* $\nexists\,(Pr \rightarrow Cs) \in \mathcal{L}$: $Pr \cup Cs = B^F$.

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

Выводы.

- Основу формирования решетки для *формального контекста сочетаемости флексий* составляют максимально проективные ЕЯ-фразы, которые наиболее точно описывают заданную ситуацию, а значит и более четко передают смысл. Морфологические зависимости, выделяемые по сходству характера флексии зависимого слова, соответствуют наиболее вероятным синтаксическим связям относительно множества форм языкового описания ситуации.

- Разработанная *модель процесса выявления закономерностей сосуществования словоформ в линейном ряду* дает возможность автоматически выделить лучший способ выражения нужной мысли в заданном Естественном Языке. Это позволит минимизировать количество ошибок синтаксического анализа при использовании его как инструмента формирования объектов и признаков.

- Предложенная в работе *методика выделения и классификации синтаксических отношений на основе множества семантически эквивалентных ЕЯ-фраз* позволяет автоматизировать разработку синтаксических стратегий и правил, что *особенно актуально* при исследовании случаев применения определенных грамматических конструкций в тематическом корпусе текстов. Качественные оценки формируемых знаний здесь могут быть даны на основе мер схожести решеток по аналогии с мерами схожести для Формальных Понятий.