Морфология и синтаксис в задаче семантической кластеризации*

Михайлов Д. В., Емельянов Г. М.
Dmitry.Mikhaylov@novsu.ru
Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

Рассматривается задача семантической кластеризации текстов Естественного Языка (ЕЯ). На основе теории Анализа Формальных Понятий предложен подход к выработке качественных оценок моделей морфологии и синтаксиса как инструментальных средств выделения объектов и признаков.

Одна из центральных задач понимания текста — выделение класса Семантической Эквивалентности (СЭ). В общих чертах установить факт СЭ означает доказать идентичность ролей сходных понятий относительно сходных ситуаций.

Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система Exactus [5].

Тем не менее, существуют задачи сравнения смысла, отличные от традиционного для поисковых систем взаимодействия «запрос–ответ».

Примером является тестовое задание открытой формы в системе контроля знаний [3]. Необходимо не столько отобразить ответ на предметную область, сколько оценить близость ответу, «правильному» с точки зрения разработчика теста. Анализ взаимной близости ответов здесь требует учета лексико-функциональной синонимии, в частности — расщепленных значений и конверсивов [3].

Актуальная *глобальная задача*, которой посвящена настоящая работа — автоматизация накопления знаний о взаимодействии семантики, синтаксиса и морфологии, необходимых для установления СЭ, непосредственно по ЕЯ-текстам.

Постановка проблемы

Сформулируем задачу СЭ следующим образом.

Пусть G есть множество ЕЯ-текстов. По результатам синтаксического разбора каждого $T_i \in G$ требуется 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.

Множества M и V выделяются на основе *синтаксических контекстов существительных* — последовательностей соподчиненных слов вида

$$S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}. \qquad (1)$$

Здесь $v_1 \in V(T_i)$ является предикатом (глаголом или словом, производным от него). Существительное m_{ki} обозначает некоторое понятие, значимое в ситуации v_1. Индекс k есть порядковый номер последовательности среди выявленных из T_i. Целочисленное значение $n(k,i)$ равно количеству 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:

$$v_l R_q v_{l+1}, \quad v_{n(k,i)} R_q m_{ki}, \qquad (2)$$

тип q которого определяется предлогом для связи главного слова с зависимым и падежом зависимого.

Транзитивность отношения R_q дает основание утверждать, что $\{v_2, \ldots, m_{ki}\} \subset M(T_i)$. В конечном итоге, тип указанного отношения между v_1 и словом справа от него в (1) определяет роль относительно v_1 для каждого $m \in \{v_2, \ldots, m_{ki}\}$.

На основе I выделяются группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях. Данная задача наиболее естественно решается методами Анализа Формальных Понятий (АФП, [2]). При этом для $A \subseteq G$ и $B \subseteq M \times V$ вводится пара отображений:

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

Пара множеств (A, B) таких, что $A' = B$ и $B' = A$, называется *формальным понятием* (ФП).

Тернарному отношению I здесь ставится в соответствие *формальный контекст* $K = (G, M, V, I)$, для которого строится *решетка ФП* $\mathrm{Re}(G, M, V, I)$.

Визуализация Re диаграмой линий [2] позволяет графически отображать группировку текстов.

Тем не менее, актуальной является проблема точности синтаксического анализа как инструмента выделения понятий и их признаков. Известные синтаксические анализаторы реализуют стратегию разбора на основе наиболее вероятных связей [1].

Вместе с тем, часто требуется исследовать природу выявляемых синтаксических связей. При неправильном разборе нужно установить причину использования той или иной стратегии (правила) с учетом особенности отражения ситуации, описываемой анализируемой фразой, в заданном ЕЯ.

Целью настоящей работы *является* разработка модели автоматического выделения и классификации наиболее вероятных синтаксических связей для множества СЭ-фраз.

*Работа выполнена при финансовой поддержке РФФИ, проект № 06-01-00028.

Методы решения

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

Как уже обсуждалось нами ранее [3], 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 [3].

Указанное описание выполняется в некоторой знаковой системе с целью обобщения и передачи знаний от человека к человеку.

Формально фиксируемый ситуацией S языковой контекст представляется тройкой:

$$S = (O, R, T), \tag{3}$$

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

Предположим, что T состоит из синонимичных фраз, каждая из которых описывает одну и ту же ситуацию действительности (относительно языкового контекста S). Выбор $T_i \in T$ для описания S является равновероятным. В силу произвольности R предположим, что его элементами являются синтаксические отношения вида (2).

При этом все ЕЯ-фразы из T являются строго синонимичными, а

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

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 (3).

Рассмотрим текст $T_i \in T$ как множество символов. Тогда 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. На множестве T_i^F выражаются синтагматические зависимости, которые определяют возможность сосуществования словоформ в линейном ряду и задаются с помощью R.

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

$$W_{ij} = W_{ij}^C \cup W_{ij}^F, \text{ где} \tag{4}$$

$W_{ij}^C \subset T_i^C$ — неизменная, $W_{ij}^F \subset T_i^F$ — флективная часть, изменяемая при склонении (спряжении).

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

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

Определение 1. *Моделью L линейной структуры предложения $T_i \in T$ назовем последовательность индексов $j \in J$ неизменных частей слов, присутствующих в T_i.*

При этом порядок индексов в L идентичен порядку следования соотсетствующих 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 формирования множества R в (3) необходимо найти совокупность указанных моделей, удовлетворяющих требованиям проективности [4].

Модель L следует считать проективной в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана L. Кроме того, если из позиции некоторого индекса выходят несколько стрелок, то эту позицию не должны накрывать стрелки, выходящие из позиций других индексов.

С учетом линейной природы синтагм дополним вышеуказанные требования следующим образом.

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

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

Определение 2. *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

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

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

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

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

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

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 $j \neq m$.

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

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

При этом индекс $k \in V_1^J$ соответствует корню дерева $\left(V_1^J, I_1^J\right)$, 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 $\forall\, E_2 \in V^J \colon E_1 \neq E_2$.

Содержательно корень соответствует предикатному слову в (1), которое (по определению) обозначает ситуацию. Поскольку исследуемая проблема точности синтаксического анализа характерна для ситуаций 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 (3) согласно сформулированной нами Задаче 1.

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 конкатенация, последовательно выполняемая над символами из W_{ij}^F в (4). 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 $s(j_3, j_2) = \text{true}$.

Введем в рассмотрение формальный контекст:

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

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

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

Рассмотрим задачу поиска флексий для слов в составе расщепленных значений и конверсивов. Будем рассматривать Расщепленное Предикатное Значение (РПЗ) — совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию. Для РПЗ, как и для конверсивов (слов, обозначающих ситуацию с точки зрения разных ее участников) представления вида (4) не могут быть найдены попарным сравнением буквенного состава слов во всех $T_i \in T$.

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

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 $\exists\, \{j, 0, k\} \subset L(T_i)$:*

$$\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot},$$

где $\{u_1, \ldots, u_p\} = P_i^{\odot}$, а $p = \left|P_i^{\odot}\right|$.

Лемма 2. *Слово $u_k \in P_i^{\odot}$ входит в состав 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, каждое из которых удовлетворяет условию Леммы 2.

Теорема 1. *Для формирования контекста (6) необходимо и достаточно найти множество $T' \subset T$:*

$$T' = \left\{T_i \colon |P_i^{\odot\prime}| \to \max\right\}. \quad (7)$$

Другой критерий отбора $T_i \in T$ основан на минимизации числа слов, не представимых как (4).

Для $u_k \in \bigcup_i P_i^{\odot\prime} \colon\ T_i \in T'$ представление (4) формируется сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{\odot} \colon T_l \in (T \setminus T')$. При этом необходимо, чтобы $2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|$, где $W_k^P = W_k^C \cup W_k^F$, а $W_j^P = W_j^C \cup W_j^F$.

Дерево (5) преобразуется следующим образом с учетом вышесказанного для всех $T_i \in T'$:

1) корень изменяется с $k = 0$ на значение k для $u_k \in P_i^{\odot\prime}$, имеющего максимальную встречаемость в различных $T_i^{\odot}$;
2) левое поддерево остается без изменений;
3) правое поддерево перевешивается на узел j для $u_j \in P_i^{\odot\prime}$ наименьшей встречаемости;
4) для всех $\{u_l, u_m\} \subset P_i^{\odot\prime}$ дочерним будет узел для слова с меньшей встречаемостью.

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

Экспериментальная апробация

На материале результатов теста открытой формы был проведен машинный эксперимент по выделению и классификации синтаксических отношений предложенным в работе методом.

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

Таблица 1. Правильные ответы $T_i \in T'$ в (7).

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

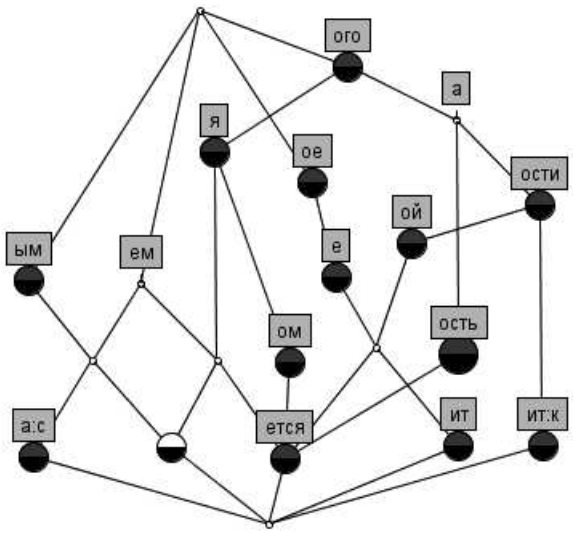

Рис. 1. Синтаксические отношения в решетке ФП.

Было получено двадцать семь вариантов правильного ответа, которые служили исходными данными при формировании контекста (6).

На рис. 1 представлена решетка Re^F для T'. При $P_i^{\odot\prime} \cap P_i^{\odot} \neq \varnothing \;\; \forall\, u_m \in \left(P_i^{\odot} \backslash P_i^{\odot\prime}\right)$ есть предлог и представляется вместе со словом $u_l \in P_i^{\odot\prime}$, стоящим слева от него в $P_i^{\odot}$, см. таблицу 1.

Содержательная интерпретация Re^F может быть получена выделением морфологических классов слов с учетом структуры последовательностей (1) согласно приведенным ниже правилам.

Пусть $\mathcal{L}$ — базис импликаций [2] для K^F из (6).

Правило 1. ФП (A^F, B^F) : $A^F \subseteq G^F$, $B^F \subseteq M^F$, соответствует предикатному слову в (1), если $\exists\,(Pr \to Cs) \in \mathcal{L}$: $|Pr| = 1$, а $Pr \cup Cs = B^F$.

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

Правило 2. ФП (A^F, B^F) соответствует прилагательному для m_{ki} в (1), если B^F есть множество признаков некоторого элемента множества G^F и $\nexists\,(Pr \to Cs) \in \mathcal{L}$: $Pr \cup Cs = B^F$.

В противном случае ФП (A^F, B^F) 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 наименьшая верхняя грань множества ФП этой области — прецеденту класса.

В примере на рис. 1 классы отношений соответствуют словоизменению прилагательных (нежелательн-ого, эмпирическ-ого) и существительных в составе генитивных конструкций (результат-ом переобучени-я, следстви-ем переобучени-я). Последний в силу транзитивности отношений (2) может включать сочетания существительного (вне генитивных конструкций) с глаголом.

Поскольку основу формирования решетки составляют те ЕЯ-фразы, которые максимально точно описывают ситуацию, а значит и более четко передают смысл, то выявленные отношения будут соответствовать искомым наиболее вероятным синтаксическим связям относительно (3).

Заключение

Предложенная в работе модель позволяет решить две важные задачи, актуальные для семантической кластеризации ЕЯ-текстов.

Во-первых, автоматически выделить лучший способ выражения нужной мысли в заданном ЕЯ, что позволит избежать ошибок синтаксического анализа при использовании его как инструмента формирования объектов и признаков.

Во-вторых, автоматизировать разработку синтаксических стратегий и правил 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 ФП [3].

Литература

[1] `http://www.aot.ru` — 2009.

[2] *Ganter B., Wille B.* Formal Concept Analysis — Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[3] *Mikhailov D. V., Emelyanov G. M., Stepanova N. A.* Formation and clustering of Russian's nouns's contexts within the frameworks of Splintered Values // 9th Int. Conf. PRIA-9-2008. — Nizhni Novgorod: NNSU, 2008. — Vol. 2. — P. 39–42.

[4] *Кибрик А. Е.* Очерки по общим и прикладным вопросам языкознания. — М.: КомКнига, 2005. — 336 с.

[5] *Осипов Г. С., Тихомиров И. А., Смирнов И. В.* Exactus — система интеллектуального метапоиска в сети Интернет // 10-я конф. КИИ-2006. — М.: Физматлит, 2006. — Т. 3. — С. 859–866.