И. В. Смирнов, Метод автоматического установления значений минимальных синтаксических единиц текста, *ИТиВС*, 2008, выпуск 3, 30–44

Метод автоматического установления значений минимальных синтаксических единиц текста

И.В. Смирнов

Аннотация. Рассматривается задача порождения правил установления значений минимальных синтаксических единиц текстов, возникающая в ходе их семантического анализа. Предлагается модификация ДСМ-метода порождения гипотез для обработки объектов с признаками сложной природы и метод порождения правил установления значений синтаксических единиц. Описаны методы установления значений и снятия семантической многозначности синтаксических единиц. Представлены результаты компьютерных экспериментов.

1. Описание проблемы

В связи с ростом объемов информации в современных сетях передачи данных появился ряд новых задач в области поиска и анализа информации. Среди них следует отметить задачи повышения точности поиска текстовых документов. Известно, что традиционные поисковые системы выдают много нерелевантных запросу пользователя документов. Это связано с тем, что традиционный подход к поиску основывается только на статистических характеристиках слов. Однако очевидно, что текст на естественном языке, выполняя задачу коммуникации – передачи информации, является *осмысленной* целостностью высказываний и слов. В связи с этим появились методы семантического поиска документов, при котором каждое предложение текста документа рассматривается как осмысленное высказывание, состоящее из конструктивных смысловых единиц. Семантический поиск позволяет находить документы, близкие запросу именно по смыслу, тем самым, повышая точность соответствия результатов поиска запросу пользователя.

Семантический анализ текста имеет своей целью извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов. Рассматриваемый нами подход к семантическому анализу текста опирается на лингвистическую теорию коммуникативной грамматики [1], разработанную в Институте русского языка им. В.В.Виноградова РАН, на основании которой делается предположение, что семантика, смысл текста на естественном языке передается с помощью ***синтаксем*** - минимальных синтактико-семантических единиц. Каждая синтаксема является минимальной единицей смысла. В предложениях для передачи осмысленной информации синтаксемы объединяются в более сложные конструкции по правилам коммуникативной грамматики.

При компьютерном семантическом анализе текста множество синтаксем каждого предложения отображается в неоднородную семантическую сеть [2] с синтаксемами в вершинах и семантическими связями на множестве синтаксем в качестве ребер. Представление семантики предложений с помощью формализма неоднородных семантических сетей позволяет легко выполнять различ-

ные действия над ними, например, сравнивать сети, находить меру близости между ними, т.е. вычислять смысловую близость текстов, по которым построены сети.

Семантический анализ текста оперирует в основном именными синтаксемами. Именная синтаксема представляется в тексте именной или предложной группой – словосочетанием с существительным или предлогом в качестве управляющего слова. Именная синтаксема характеризуется морфологической формой – предлогом, падежом, и категориально-семантическим классом существительного, от которого она образована. Морфологическая форма синтаксемы и категориально-семантический класс определяются с помощью лингвистического анализатора текста. Синтаксема характеризуется также синтаксической функцией, которую она может выполнять в предложении, и синтаксическим значением.

В ходе семантического анализа текста необходимо установить ***значения*** именных синтаксем, которые являются обозначениями смыслов, передаваемых текстом.

Морфологическая форма и категориально-семантический класс именной синтаксемы не однозначно задают её значение, а синтаксическую функцию, в которой выступает конкретная синтаксема, встречаемая в тексте в ходе анализа, автоматически определить невозможно. Поэтому обычно в анализ вовлекается контекст - глагол или отглагольное существительное, т.е. предикатное слово, при котором именная синтаксема встречается в предложении. Учет такого рода контекста требует создания специального словаря, описывающего наиболее частые сочетания определенного глагола с возможными синтаксемами при нем, и такой словарь был создан для глаголов и отглагольных существительных, наиболее часто встречаемых в текстах определенной тематики [3].

Словарь предикатных слов не может охватить все глаголы и отглагольные существительные, т.к. перечисление возможных синтаксем при глаголе является весьма трудоёмкой задачей, требующих больших затрат сил лингвистов. Поэтому часто при семантическом анализе невозможно опираться на предикатное слово, так как его нет в словаре предикатов, и в таких случаях для установления значения синтаксемы необходимо учитывать другой ее контекст.

В безглагольных предложениях синтаксемы присутствуют рядом с другими элементами предложения и несут своё значение только в данном контексте. Зависимость значения синтаксемы от собственных морфологических характеристик и характеристик соседних элементов предложения (не глаголов) является языковой закономерностью, которую необходимо обнаружить и зафиксировать для выполнения семантического анализа безглагольных предложений в дальнейшем. Такую закономерность для значений синтаксемы можно записать в виде правила, где в предпосылке правила находятся характеристики самой синтаксемы и окружающих её синтаксем и других элементов предложения, а в заключении правила находится значение, которое необходимо приписать целевой, рассматриваемой синтаксеме.

Построение правил установления значений синтаксем экспертом-лингвистом требует больших трудозатрат на просмотр текстов, где встречаются анализируемые синтаксемы, анализ контекста синтаксем, обобщение признаков, влияющих на значение синтаксемы в разных текстах. Поэтому встала задача автоматического построения контекстных правил, позволяющих устанавливать значения синтаксем на основании доступных характеристик самих синтаксем и других элементов предложения, соседствующих с рассматриваемыми синтаксемами.

В синтаксическом словаре Г.А. Золотовой [4] собраны синтаксемы с примерами их встречаемости в текстах классических произведений русской литературы и периодики. Приводимые в словаре предложения, в которых присутствуют синтаксемы, могут послужить источником создания обучающих примеров, на основании которых можно построить правила установления значений синтаксем. Выборка примеров встречаемости синтаксем имеет небольшой объем, что делает неэффективным использование статистических методов обнаружения закономерностей. Поэтому было решено использовать индуктивные методы, основанные на обобщении признаков, не требующие больших объемов обучающих выборок.

2. Метод порождения правил установления значений синтаксем

Предлагаемый метод порождения правил установления значений синтаксем основывается на ДСМ-методе порождения гипотез [5, 6], который применяется для выявления причинно-следственных закономерностей в некоторой предметной области. Задачей ДСМ-метода является обнаружение причин возникновения некоторого явления, или наличия свойств у объектов из некоторого множества. Решение этой задачи основывается на фактах или обучающем множестве объектов. Найденные причины используются для прогнозирования наблюдения явлений в дальнейшем.

Индуктивный вывод в ДСМ-методе основывается на принципе единственного сходства, сформулированном английским философом Д.С. Миллем:

Если какое-то обстоятельство постоянно предшествует наступлению исследуемого явления, в то время как иные обстоятельства изменяются, то это обстоятельство есть, вероятно, причина данного явления.

Формирование гипотез о причинах возникновения некоторого явления заключается в нахождении общих черт в характере проявления двух или более случаев этого явления, причем эти общие характеристики должны быть наибольшими, т.е. не содержаться в других общих характеристиках. На языке объектов, свойств и признаков формулировка задачи порождения гипотез выглядит так: построение гипотез о причинах обладания некоторым множеством объектов определенным свойством заключается в нахождении характеристики их сходства - максимального множества признаков, которое принадлежит двум или более объектам с данным свойством. Эта характеристика сходства будет тем самым обстоятельством, которое не меняется от случая к случаю при наблюдении явления, т.е. гипотетической причиной наличия свойства у объектов.

Полученные гипотезы используются для предсказания наличия свойств у новых объектов, про которые неизвестно, обладают они этим свойством или нет. Гипотезы ищутся среди признаков новых объектов. Если множество признаков объекта содержит все признаки, составляющие гипотезу о наличии свойства, то принимается, что объект обладает соответствующим свойством.

Далее с помощью теории множеств излагается суть ДСМ-метода и предлагаются его модификации для решения задачи порождения правил установления значений синтаксем.

2.1. ДСМ-метод порождения гипотез и его модификации

2.1.1. Формальное изложение метода

Пусть A – универсальное множество признаков. Оно содержит элемент $a_\varnothing$ такой, что $\forall a \in A, a \cup a_\varnothing = a$. Этот элемент мы назовем *пустым признаком*.

Любое подмножество A множества признаков будем называть *объектом*. Множество объектов O, таким образом, является множеством подмножеств множества признаков A. Объект $o \in O$, не содержащий признаков, будем называть *пустым объектом*, и обозначать $o_\varnothing$. $\forall o \in O, o \cup o_\varnothing = o$.

Под признаком можно понимать как и атрибут объекта рассматриваемой предметной области, так и любую характеристику, вычисляемую по атрибутам объекта. Таким образом, признак – некоторая характеристика, идентифицирующая объект, позволяющая отличить его от других объектов. Определение объекта через подмножество всевозможных признаков не нарушает общепринятого понимания сущности "объект".

Определение 1. *Фрагментом f объекта $o \in O$ называется подмножество его признаков.*

Характеристикой сходства s двух объектов o' и o", $o' \in O$, $o'' \in O$, – их общей частью - разумно назвать результат пересечения множеств признаков двух объектов $s = o' \cap o''$. Этот результат будет являться фрагментом первого и второго объектов, т.е. принадлежать множеству O, поэтому сходство более двух объектов можно найти поочередным пересечением множеств признаков объектов.

Определение 2. *Множество признаков называется характеристикой сходства множества объектов $O', O' \subset O$, если оно является фрагментом каждого объекта из O'.*

Операция вычисления сходства объектов $\cap$ находит характеристику сходства двух объектов $o' \in O$ и $o'' \in O$, и обладает следующими свойствами:

1. Коммутативность: $o' \cap o'' = o'' \cap o'$;
2. Ассоциативность: $o' \cap (o'' \cap o''') = (o' \cap o'') \cap o'''$.

Эти свойства позволяют находить характеристики сходства для более двух объектов поочередно в произвольном порядке, что используется при нахождении характеристик сходства объектов на практике.

Для любого непустого подмножества множества O может существовать непустая характеристика сходства, поэтому можно говорить о множестве S_O характеристик сходств для множества объектов O.

На множестве объектов определим отношение вложенности $\leq$:

Объект o' вложен в объект o'', если o' является фрагментом o'', т.е. $o' \leq o'' \Leftrightarrow o' \cap o'' = o'$.

Отношение вложенности является отношением частичного порядка на множестве объектов.

Определение 3. *Характеристика сходства **c** множества объектов $O' \subset O$ называется максимальной, если она не вложена в другую характеристику сходства, образованную теми же объектами, из которых образована **c**, и если в неё не вложена характеристика сходства, образованная другими объектами этого множества O'.*

Таким образом, максимальная характеристика сходства интерпретируется как то, в чем наиболее схожи объекты. С одной стороны, она не содержится в других характеристиках сходства, образованных теми же объектами, т.е. является наиболее полной, с другой стороны, в неё не вложены характеристики сходства других объектов, т. е. она является действительно той неизменной частью, которая одинакова у всех образующих её объектов.

Если предположить, что максимальная характеристика сходства может существовать для каждого подмножества множества O, тогда можно говорить о множестве $S_O max$ максимальных характеристик сходств.

Определение 4. *Пусть O^* - множество непустых подмножеств множества $O' \subset O$. Операцией вычисления сходства множества объектов $O' \subset O$ называется отображение $\Pi : O^* \to O$, такое, что $\Pi O' = S_O max$.*

Операция вычисления сходства отображает множество объектов в множество максимальных характеристик сходства всех его подмножеств.

Пусть P – множество свойств объектов. Введем двухместный предикат POSSESS(o, p), интерпретирующий тот факт, что объект $o \in O$ обладает свойством $p \in P$. Этот предикат принимает значения $\{+, -, 0, \tau\}$, что означает следующее:

POSSESS(o, p) = + означает, что объект *o* обладает свойством *p*.

POSSESS(o, p) = - означает, что объект *o* не обладает свойством *p*.

POSSESS(o, p) = 0 означает, что существует противоречие, и можно сказать, что объект *o* как обладает свойством *p*, так и не обладает им.

POSSESS(o, p) = τ означает неизвестность относительно того, обладает объект *o* свойством *p*, или нет.

Множество $O^+_p = \{o \mid o \in O \wedge p \in P \wedge \text{POSSESS}(o, p) = +\}$ назовем множеством положительных примеров для свойства *p*.

Множество $O^-_p = \{o \mid o \in O \wedge p \in P \wedge \text{POSSESS}(o, p) = -\}$ назовем множеством отрицательных примеров для свойства *p*.

Множество $O^0_p = \{o \mid o \in O \wedge p \in P \wedge \text{POSSESS}(o, p) = 0\}$ назовем множеством противоречивых примеров для свойства *p*.

Множество $O^\tau_p = \{o \mid o \in O \wedge p \in P \wedge \text{POSSESS}(o, p) = \tau\}$ назовем множеством недоопределенных примеров для свойства *p*.

Задача состоит в том, чтобы выяснить, обладают недоопределенные объекты-примеры свойством *p* или нет. В зависимости от особенностей предметной области и характера обучающей выборки возможно принятие различных моделей причинно-следственных зависимостей. Опишем простую симметричную модель без блокировок и запретов на контрпример.

Пусть h – характеристика сходства множества из более одного объекта.

Множество $H^{+}_{Op} = \{h \mid (h \in \prod O^{+}_{p}) \wedge (h \notin \prod O^{-}_{p}) \wedge (h \notin \prod O^{0}_{p})\}$ назовём множеством положительных гипотез для свойства *p*. Это множество гипотез о причинах наличия свойства у объектов.

Множество $H^{-}_{Op} = \{h \mid (h \in \prod O^{-}_{p}) \wedge (h \notin \prod O^{+}_{p}) \wedge (h \notin \prod O^{0}_{p})\}$ назовём множеством отрицательных гипотез для свойства *p*. Это множество гипотез о причинах отсутствия свойства у объектов.

Множество $H^{0}_{Op} = \{h \mid ((h \in \prod O^{-}_{p}) \wedge (h \in \prod O^{+}_{p})) \vee (h \in \prod O^{0}_{p})\}$ назовём множеством противоречивых гипотез для свойства *p*. Это множество гипотез, на основании которых можно говорить как о наличии свойства, так и о его отсутствии.

Правила формирования перечисленных множеств гипотез соответствуют правилам первого рода в терминологии ДСМ-метода:

- Характеристика сходства *h* является гипотезой о наличии свойства *p*, если она является результатом операции вычисления сходства положительных примеров для свойства *p*, и не является результатом операции вычисления сходства отрицательных или противоречивых примеров для свойства *p*.
- Характеристика сходства *h* является гипотезой об отсутствии свойства *p*, если она является результатом операции вычисления сходства отрицательных примеров для свойства *p*, и не является результатом операции вычисления сходства положительных или противоречивых примеров для свойства *p*.
- Характеристика сходства *h* является противоречивой гипотезой о наличии свойства *p*, если она является результатом операции вычисления сходства одновременно отрицательных и положительных примеров, или результатом операции вычисления сходства противоречивых примеров для свойства *p*.

Определим правила для установления значения предиката POSSESS'(o, p), $o \in O$, $p \in P$, который выражает предположения о наличии свойства *p* у объекта *o*. POSSESS'(o, p) = POSSESS(o, p), если POSSESS(o, p) $\neq \tau$. Для тех объектов и свойств, для которых POSSESS(o, p) = τ:

$$\mathrm{POSSESS'}(o,p) = \begin{cases} +, \text{если } (\exists s: s \le o \wedge s \in H^{+}_{Op}) \wedge (\neg\exists s': s' \le o \wedge s' \in H^{-}_{Op}) \wedge (\neg\exists s'': s'' \le o \wedge s'' \in H^{0}_{Op}), \\ -, \text{если } (\exists s: s \le o \wedge s \in H^{-}_{Op}) \wedge (\neg\exists s': s' \le o \wedge s' \in H^{+}_{Op}) \wedge (\neg\exists s'': s'' \le o \wedge s'' \in H^{0}_{Op}), \\ 0, \text{если } ((\exists s: s \le o \wedge s \in H^{+}_{Op}) \wedge (\exists s': s' \le o \wedge s' \in H^{-}_{Op})) \vee (\exists s'': s'' \le o \wedge s'' \in H^{0}_{Op}), \\ \tau, \text{если } (\neg\exists s: s \le o \wedge s \in H^{+}_{Op}) \wedge (\neg\exists s': s' \le o \wedge s \in H^{-}_{Op}) \wedge (\neg\exists s'': s'' \le o \wedge s'' \in H^{0}_{Op}). \end{cases} \quad (1)$$

Правила определения значения предиката POSSESS' позволяют предположить, обладает объект свойством или нет, и соответствуют правилам второго рода в терминологии ДСМ-метода:

- Объект *o* обладает свойством *p*, если в него вложена гипотеза о наличии свойства *p*, не вложена ни одна гипотеза об отсутствии свойства *p* и ни одна противоречивая гипотеза;
- Объект *o* не обладает свойством *p*, если в него вложена гипотеза об отсутствии свойства *p*, не вложена ни одна гипотеза о наличии свойства *p* и ни одна противоречивая гипотеза.

2.1.2. Операция вычисления сходства для объектов с признаками произвольной природы

Обычно при использовании ДСМ-метода объекты представляются в виде кортежа бинарных признаков, и операция вычисления сходства при этом рассматривается как пересечение множества признаков.

Расширим операцию вычисления сходства так, чтобы можно было оперировать объектами с признаками любой природы. Предположим, что на множестве признаков существует *отношение выводимости* ‹ , которое можно интерпретировать так, что признак а выводится из признака а*, а ‹ а*, если при наблюдении признака а*, можно говорить также о наблюдении признака а, при этом верно, что а ‹ а. *Характеристикой сходства* двух признаков назовем признак, выводимый одновременно из каждого из двух данных признаков. Понятие характеристики сходства распространяется и на большее число признаков.

Определение 5. *Операцией вычисления сходства на множестве признаков А называется отображение* $\cap_A : A \times A \to A$, *такое, что сопоставляет двум признакам их характеристику сходства.*

Операция вычисления сходства для двух объектов о' ∈ О и о" ∈ О будет выполняться по следующей формуле:

$$o' \cap o'' = \bigcup_{a' \in o', a'' \in o''} (a' \cap_A a'') \qquad (2)$$

В связи с появлением операции вычисления сходства на множестве признаков, необходимо переопределить понятие характеристики сходства объектов.

Определение 6. *Множество признаков называется характеристикой сходства множества объектов О', О'* ⊂ *О, если каждый его элемент является характеристикой сходства признаков каждого объекта из О'.*

Поскольку переопределенная характеристика сходства является множеством признаков, то для неё остаются действительными отношения вложенности и максимальности, и, следовательно, остается в силе определение операции вычисления сходства множества объектов (определение 4.).

Верно следующее утверждение:

Утверждение 1. *Для того, чтобы операция вычисления сходства* ∩ *объектов о'* ∈ *О и о''* ∈ *О, обладала свойствами коммутативности и ассоциативности, достаточно, чтобы теми же свойствами обладала операция вычисления сходства* $\cap_A$ *на признаках.*

2.1.3. Составные признаки

Для того, чтобы оперировать понятием «синтаксема» будем рассматривать совокупности признаков, в которых элементы логически не могут существовать друг без друга. Такая совокупность является неделимым признаком, который мы назовем *составным признаком*. Например, синтаксема представляет собой тройку простых составляющих – предлога, падежа и категориально-семантического класса.

Составной признак состоит из нескольких признаков, но рассматривается не как множество независимых признаков, а как один признак. Для составного признака определяется своя операция вычисления сходства, результатом которой является составной признак, но не его составляющие по отдельности. Составной признак может сам быть в составе другого составного признака.

Простейшим примером операции вычисления сходства двух составных признаков является объединение результатов операции вычисления сходства для всех составляющих их признаков.

Пусть са' и са" – составные признаки. $ca' = \{a'_i : a'_i \in A\}$, $i = 1 \ldots N_{ca'}$, $ca'' = \{a''_i : a''_i \in A\}$, $i = 1 \ldots N_{ca''}$. Тогда

$$ca' \cap_A ca'' = \bigcup_{i=1}^{Nca'} \bigcup_{j=1}^{Nca''} (a'_i \cap_A a''_j) \qquad (3)$$

Но составные признаки специально введены для того, чтобы операцию вычисления сходства на них можно было определить особенным образом, при соблюдении свойств коммутативности и ассоциативности этой операции. Например, можно потребовать, чтобы число признаков результата операции вычисления сходства двух составных признаков равнялось числу признаков одного из операндов.

$$ca' \cap_A ca'' = \begin{cases} \bigcup_{i=1}^{Nca'} \bigcup_{j=1}^{Nca''} (a'_i \cap_A a''_j), \text{ если } \left| \bigcup_{i=1}^{Nca'} \bigcup_{j=1}^{Nca''} (a'_i \cap_A a''_j) \right| = \min(N_{ca'}, N_{ca''}), \\ a_\varnothing, \text{ иначе.} \end{cases} \quad (4)$$

Верно следующее утверждение:

Утверждение 2. *Для того, чтобы операция вычисления сходства составных признаков обладала свойствами коммутативности и ассоциативности, достаточно, чтобы этими же свойствами обладала операция вычисления сходства для компонентов составного признака.*

2.1.4. Объекты в контексте

В некоторых предметных областях на обладание объектом свойством может влиять окружение, контекст, в котором находится объект[8]. Например, если объектом является синтаксема, то на её синтаксическое значение, как уже говорилось, могут влиять другие элементы предложения – соседние синтаксемы, слова, принадлежащие некоторым частям речи, например, сравнительные прилагательные или наречия.

Пусть контекст *c* объекта $o \in O$ описывается множеством признаков из универсального множества A. Тогда можно говорить о множестве C *контекстов объектов*. Множество контекстов образовано так же, как множество объектов, поэтому все свойства множества объектов выполняются и для множества контекстов. Например, операцию вычисления сходства двух контекстов естественно определить так же, как операцию вычисления сходства двух объектов с признаками произвольной природы.

Определение 7. *Объектом в контексте* o_c *назовем пару* $<o, c>$, *где* $o \in O$, $c \in C$.

Определение операции вычисления сходства для двух объектов в контексте зависит от особенностей (аксиом) предметной области. Может оказаться так, что наличие свойства у объекта зависит только от его признаков, а на отсутствие свойства влияет только контекст. Возможно влияние признаков только самого объекта, или только контекста. Рассмотрим случай, когда на наличие свойства влияют характеристики как объекта, так и контекста, но контекст в отдельности не может влиять на появление свойства. Операция вычисления сходства двух объектов в контексте будет выглядеть так:

$$o'_c \cap o''_c = \begin{cases} < o' \cap o'', c' \cap c'' >, \text{ если } o' \cap o'' \neq o_\varnothing, \\ < o_\varnothing, c_\varnothing >, \text{ иначе.} \end{cases} \quad (5)$$

Верно следующее утверждение:

Утверждение 3. *Чтобы операция вычисления сходства объектов в контексте была коммутативна и ассоциативна, достаточно, чтобы этими же свойствами обладала операция вычисления сходства над элементами множеств C и O.*

Отношение вложенности объектов в контексте и операция вычисления сходства для множества объектов в контексте определяются так же, как и для обычных объектов:

Объект контексте o'_c *вложен в объект* o''_c, *если* o'_c *является фрагментом* o''_c, *т.е.* $o'_c \le o''_c \Leftrightarrow o'_c \cap o''_c = o'_c$.

В связи с этим все рассуждения, касающиеся операции вычисления сходства для множества объектов, остаются в силе при замене множества объектов O на множество объектов в контексте O_C.

2.2. Метод порождения правил установления значений синтаксем

2.2.1. Объекты, признаки и правила

Пусть A – множество характеристик, каждая из которых описывает некоторый объект с какой-либо одной стороны. Пусть на этом множестве некоторым образом определена операция вычисления сходства $\cap_A$: $A \times A \to A$, которая обладает свойствами коммутативности и ассоциативности. Пару $<A, \cap_A>$ назовём *типом признака*. Тип признака – это множество его всевозможных значе-

ний с заданной операцией вычисления сходства на нём. В общем случае операция вычисления сходства признаков может определяться произвольно, однако она должна удовлетворять свойствам коммутативности и ассоциативности. *Признаком*, соответственно, называется элемент множества A. Обычно имеется несколько множеств, характеризующих объект с разных сторон, каждое со своей операцией вычисления сходства, поэтому объект описывается признаками нескольких разных типов.

Пусть PPS – множество предлогов русского языка, PPS = {в, над, под, …} (всего 102 предлога). CAS – множество падежей русского языка, CAS = {именительный, родительный, дательный, винительный, творительный, предложный}. CAS* - множество непустых подмножеств множества CAS[1]. Множество категориально-семантических классов KSC содержит следующие элементы: KSC = {личное, предметное, признаковое, пространственное, темпоративное, параметр_измерения, единица_измерения}. Множество PST = {до, после, не важно} – множество позиций, множество POS = {числительное, прилагательное в сравнительной степени} – множество частей речи.

Для каждого множества элементарных признаков определим операцию вычисления сходства.

$$1.\ \forall pps' \in PPS,\ \forall pps'' \in PPS,\ pps' \cap_A pps'' = \begin{cases} pps', \text{если } pps' = pps'' \\ a_\emptyset, \text{иначе} \end{cases} \quad (6)$$

Результатом операции вычисления сходства двух признаков типа «предлог» является признак типа предлог со значением одного из операндов, если значения операндов совпадают, и пустой признак иначе.

$$2.\ \forall ksc' \in KSC,\ \forall ksc'' \in KSC,\ ksc' \cap_A ksc'' = \begin{cases} ksc', \text{если } ksc' = ksc'' \\ a_\emptyset, \text{иначе} \end{cases} \quad (7)$$

Результатом операции вычисления сходства двух признаков типа «категориально-семантический класс» является признак типа «категориально-семантический класс» со значением одного из операндов, если значения операндов совпадают, и пустой признак иначе.

$$3.\ \forall cas*' \in CAS*,\ \forall cas*'' \in CAS*,\ cas*' \cap_A cas*'' = \begin{cases} cas*' \cap cas*'', \text{если } cas*' \cap cas*'' \neq \emptyset \\ a_\emptyset, \text{иначе} \end{cases} \quad (8)$$

Результатом операции вычисления сходства двух признаков типа «падеж» является признак типа «падеж» со значением равным пересечению значений признаков операндов, если оно не пусто, и пустой признак иначе.

$$4.\ \forall pst' \in PST,\ \forall pst'' \in PST,\ pst' \cap_A pst'' = \begin{cases} pst', \text{если } pst' = pst'', \\ \text{"не важно"}, \text{иначе} \end{cases} \quad (9)$$

Результатом операции вычисления сходства двух признаков типа «позиция» является признак типа «позиция» со значением равным значению операндов при их совпадении, и равным «не важно» иначе.

$$5.\ \forall pos' \in POS,\ \forall pos'' \in POS,\ pos' \cap_A pos'' = \begin{cases} pos', \text{если } pos' = pos'' \\ a_\emptyset, \text{иначе} \end{cases} \quad (10)$$

Результатом операции вычисления сходства двух признаков типа «часть речи» является признак типа «часть речи» со значением одного из операндов, если значения операндов совпадают, и пустой признак иначе.

Множество синтаксем SYN содержит тройки <pps, cas*, ksc>, где pps ∈ PPS, cas* ∈ CAS*, ksc ∈ KSC. Множество синтаксем в позиции SYNPST состоит из пар <syn, pst>, где syn ∈ SYN,

[1] Лингвистический анализатор не всегда точно разрешает падежную омонимию, поэтому приходится учитывать все выдаваемые для слова падежи

pst ∈ PST. Множество частей речи в позиции POSPST состоит из пар <pos, pst>, где pos ∈ POS, pst ∈ PST. Зададим операции вычисления сходства:

1.

$$\forall syn' \in SYN,\ \forall syn'' \in SYN,\ syn' \cap_A syn'' = \begin{cases} < pps' \cap_A pps'', cas^{*}' \cap_A cas^{*}'', ksc' \cap_A ksc'' >, \text{если} \\ \quad (pps' \cap_A pps'' \neq a_\emptyset) \wedge (cas^{*}' \cap_A cas^{*}'' \neq a_\emptyset) \wedge (ksc' \cap_A ksc'' \neq a_\emptyset) \\ a_\emptyset, \text{иначе} \end{cases} \quad (11)$$

2.

$$\forall synpst' \in SYNPST,\ \forall synpst'' \in SYNPST,\ synpst' \cap_A synpst'' = \begin{cases} < syn' \cap_A syn'', pst' \cap_A pst'' >, \text{если } syn' \cap_A syn'' \neq a_\emptyset \\ a_\emptyset, \text{иначе} \end{cases} \quad (12)$$

3.

$$\forall pospst' \in POSPST,\ \forall pospst'' \in POSPST,\ pospst' \cap_A pospst'' = \begin{cases} < pos' \cap_A pos'', pst' \cap_A pst'' >, \text{если } pos' \cap_A pos'' \neq a_\emptyset \\ a_\emptyset, \text{иначе} \end{cases} \quad (13)$$

Определение 8. *Множеством синтаксем в контексте называется пара O_C = <O, C>, где O = <SYN>, C = <SYNPST, POSPST>.*

Сходство двух синтаксем в контексте вычисляется как поэлементное выполнение операции вычисления сходства для однотипных признаков. Результатом операции вычисления сходства двух синтаксем в контексте является характеристика сходства синтаксем в контексте, т.е. множество признаков, каждый их которых выводится из признаков каждой из синтаксем в контексте. Сходство множества синтаксем в контексте определяется как поэлементное выполнение операции вычисления сходства для всех элементов множества.

Определим операцию вычисления сходства для синтаксем в контексте:

$$o'_C \cap o''_C = \begin{cases} < o' \cap o'', c' \cap c'' >, \text{если } o' \cap o'' \neq o_\emptyset \\ o_{C\emptyset}, \text{иначе} \end{cases} =$$

$$= \begin{cases} < syn' \cap syn'',\ synpst' \cap synpst'',\ pospst' \cap pospst'' >,\ \text{если } syn' \cap syn'' \neq a_\emptyset \\ o_{C\emptyset}, \text{иначе} \end{cases} \quad (14)$$

Множество свойств объектов в рассматриваемой предметной области соответствует множеству синтаксических значений синтаксем.

В качестве модели причинно-следственных связей в предметной области выбрана несимметричная положительная модель, предполагающая существование только положительных причин для установления значений синтаксем. Это связано с тем, что отрицательных примеров в обучающем множестве нет. В обучающем множестве нет также противоречивых примеров. В связи с этим, далее оперировать будем только положительными примерами.

Под правилом обычно подразумевается логическая связка «если *предпосылка*, то *заключение*». Введем формальное определение:

Определение 9. *Множеством правил установления значения синтаксем $\boldsymbol{R_p}$ называется пара <H_p, p>, где $H_p \subset O_C$ – множество предпосылок, получаемых в результате выполнения операции вычисления сходства для множества синтаксем, имеющих значение **p**, а **p** – заключение правил, являющееся значением синтаксем.*

Определение 10. *Множеством конфликтных правил* $\boldsymbol{R_{CONF}}$ *называется множество правил с одинаковыми предпосылками, но с разными заключениями.*

Опишем эвристики поиска предпосылок правил установления значений синтаксем.

Пусть P – множество синтаксических значений, P = {аблатив, абстинатив,…} (всего 81 элемент). Тот факт, что синтаксема в контексте $o_c \in O_C$ имеет или не имеет значение ***p*** будем выражать предикатом POSSESS(o_c, p). Этот предикат в несимметричной положительной модели принимает два значения {+ (истина), τ (неизвестно)}.

Множество синтаксем в контексте, для которых предикат POSSESS(o_c, p) истинен, обозначим как O_C^p.

$$O_C^p = \{o_c \mid o_c \in O_C \wedge p \in P \wedge \mathrm{POSSESS}(o_c, p) = +\}.$$

Характеристика сходства ***h*** *считается предпосылкой правила установления значения синтаксемы* ***p****, если она является результатом операции вычисления сходства множества синтаксем* O_C^p*, имеющих данное значение, и не существует других значений, таких, что* ***h*** *является результатом операции вычисления сходства множества синтаксем, имеющих это значение, т.е.*

$$h \in H_p \Leftrightarrow (h \in \prod O_C^p) \wedge (\neg \exists p' : h \in \prod O_C^{p'}) \qquad (15)$$

Характеристика сходства ***h*** *считается предпосылкой конфликтного правила установления значения синтаксемы* ***p****, если она является результатом операции вычисления сходства как множества синтаксем, имеющих значение* ***p****, так и множества синтаксем, имеющих другое значение.*

$$h \in H_{CONF} \Leftrightarrow (h \in \prod O_C^p) \wedge (\exists p' : h \in \prod O_C^{p'}) \qquad (16)$$

В приведенных определениях ничего не говорится о вложенности, хотя возможны случаи, когда характеристика сходства множества синтаксем, имеющих одно значение, вложена в одну из характеристик сходства синтаксем, имеющих другое значение. Эти случаи здесь не будут рассмотрены.

Применение правил установления значений синтаксем:

Синтаксема в контексте имеет значение ***p****, если в неё вложен фрагмент* ***f****, являющийся предпосылкой правила установления значения* ***p****, и не вложен ни один фрагмент, являющийся предпосылкой конфликтного правила*.

$$\mathrm{POSSESS}'(o_c, p) = \begin{cases} +, \text{если } (\exists r_p = < h_p, p >: h_p \le o_c) \wedge (\neg \exists r_{conf} = < h_{conf}, p >: h_{conf} \le o_c) \\ \tau, \text{иначе.} \end{cases} \qquad (17)$$

2.2.2. Алгоритм порождения правил установления значений синтаксем

Для порождения предпосылок правил установления значений синтаксем находятся максимальные характеристики сходств множества из двух и более синтаксем в контексте, обладающих данным свойством, которые вычисляются как поочередное применение операции вычисления сходства для двух объектов.

Общая схема алгоритма порождения правил основывается на алгоритме нахождения минимальных пересечений множества объектов [9]:

Шаг А. Предварительно все синтаксемы разбиваются на множества синтаксем, имеющих одинаковые значения. Далее для каждой синтаксемы определяются её морфологические признаки и устанавливаются признаки контекста. Строятся объекты «синтаксема в контексте» Далее каждое множество объектов-синтаксем в контексте обрабатывается по следующему алгоритму:

Шаг А.1. Выбирается первый не пустой объект, он считается текущим.

Шаг А.2. Берется любой другой объект и находится характеристика сходства его с текущим объектом. Если характеристика сходства - пустой объект, то происходит переход к дру-

гому объекту, если характеристика сходства не пуста, применяется операция сходства для неё и следующего объекта и так далее, пока не просмотрено все множество объектов.

Шаг А.3. Запоминается найденная характеристика сходства и объекты первоначального множества, в которые она вложена. Из объектов, в которые вложена характеристика сходства, вычитаются признаки, формирующие эту характеристику сходства. Процедура повторяется с шага 1, пока находятся непустые характеристики сходства.

Шаг А.4. Для каждой полученной характеристики сходства выполняется операция сходства для объектов из первоначального множества, в которые вложена данная характеристика сходства. Если полученный результат совпадает с данной характеристикой сходства, то данная характеристика сходства добавляется в правило в качестве предпосылки. Заключением этого правила является значение синтаксем текущего множества.

Шаг Б. После того, как обработаны все множества объектов-синтаксем в контексте с одинаковыми значениями, для каждого полученного правила проверяется, не вложена ли его предпосылка в какую-либо предпосылку правил для установления другого значения. Если вложенности нет, то правило помещается в конечное множество правил, если вложенность есть, то правило помещается в множество конфликтных правил. Каждому правилу назначается вес, который вычисляется как отношение числа обучающих примеров, из которых построено правило, к числу всех обучающих примеров для синтаксемы.

Шаг **А.4** необходим для исключения локальных сходств объектов с удаленными на шаге **А.3** признаками.

2.3. Алгоритмы установления значений и снятия семантической многозначности синтаксем на основе правил установления значений синтаксем

Алгоритм предсказания значений синтаксем на основе правил установления значений синтаксем должен понижать влияние конфликтных правил. Он учитывает веса полученных правил установления значений синтаксем.

Алгоритм установления значений синтаксем

Шаг 1. Для синтаксемы в предложении создаются объекты-синтаксемы в контексте. Для одной синтаксемы может быть создано несколько объектов, например с правым и левым контекстом, причем для этих объектов могут быть предсказаны разные значения.

Шаг 2. Для каждого объекта, построенного для синтаксемы, поочередно просматриваются все правила. Вычисляется характеристика сходства объекта и предпосылки очередного правила – операция вычисления сходства та же, что и при обучении. Если характеристика сходства не пуста, то результат операции вычисления сходства сравнивается с самой предпосылкой правила. Их равенство означает, что предпосылка полностью вложена в объект и правило является кандидатом для правильного предсказания. Правило-кандидат добавляется в специальное множество правил.

Шаг 3. После того, как выбраны правила-кандидаты, предпосылки которых вложены во все объекты, построенные для синтаксемы, происходит перерасчет весов правил с одинаковым заключением: веса правил с одинаковым заключением складываются и умножаются на некоторый коэффициент, а полученное значение умножается ещё и на число этих правил. В итоге всем правилам с одинаковым заключением приписывается новое значение веса. Таким образом, чем большее количество раз для синтаксемы предсказано одно и то же значение, тем выше будет вес этого значения.

Шаг 4. Далее проверяется равенство предпосылки каждого правила с объектом. Равенство предпосылки правила и объекта прибавляет весу каждого правила ещё некоторый коэффициент; это означает, что объект «является» предпосылкой правила, а не содержит в себе предпосылку (объект сам является причиной свойства, а не содержит в себе причину), что также должно повышать вес правила.

Шаг 5. Выбирается наиболее весомое правило.

Описанный выше алгоритм подходит как для множеств правил с конфликтами, так и без них.

Напомним, что проблема смысловой многозначности синтаксем заключается в том, что, имея лишь морфологические характеристики и категориально-семантический класс синтаксемы, невозможно однозначно определить её семантическое значение. С использованием словаря предикатов каждой синтаксеме без учета её контекста можно приписать в среднем четыре значения, в то время как верным является лишь одно. Предлагаемый алгоритм снятия смысловой многозначности использует алгоритм установления значений синтаксем, описанный выше.

Алгоритм снятия семантической многозначности синтаксем

Для синтаксемы:

Шаг 1: Определить с помощью словаря предикатов все возможные значения синтаксемы.

Шаг 2: Установить значения синтаксемы согласно описанному выше алгоритму установления значений синтаксем на основании правил установления значений синтаксем. Результатом этого является множество возможных значений синтаксемы, упорядоченных по убыванию их веса.

Шаг 3: Выполнить пересечение значений, полученных на шаге 1 и шаге 2.

Шаг 4: Если пересечение значений, полученных на шаге 3, пусто, то случайным образом выбрать одно значение из множества, полученного на шаге 1. Если пересечение, полученное на шаге 3 не пусто, то выбрать из полученного пересечения значение с наибольшим весом.

Предложенный алгоритм снижает воздействие полученных на шаге 2 конфликтных правил тем, что отдаёт предпочтение значениям, внесенным в словарь предикатов вручную экспертом-лингвистом. Алгоритм помогает снять многозначность двумя способами: во-первых, он снижает количество возможных значений пересечением первоначального множества значений с множеством значений, полученных в результате применения правил; если пересечение всё же содержит более одного значения, то из него выбирается единственное значение с наибольшим весом, что будет соответствовать выбору наиболее типичного (частого) значения для данной синтаксемы (вернее значения синтаксемы, наиболее типичного для текстов, на которых проводилось обучение). Такой уточняющий алгоритм позволяет максимально снизить противоречивость конфликтных правил.

3. Компьютерные эксперименты и их результаты

3.1. Порождение правил установления значений синтаксем и оценивание их предсказательной силы

Материалом для построения обучающих примеров послужила электронная версия синтаксического словаря Г.А. Золотовой, предоставленная сотрудниками Машинного фонда русского языка Института русского языка РАН[2]. В электронной версии словаря границы синтаксем выделены с помощью знаков подчеркивания «_». Это дает возможность автоматически выделить фрагменты текста, содержащие примеры синтаксем, и построить обучающие примеры.

Обучающие примеры – синтаксемы в контекстах - строились для всех синтаксем словаря, кроме синтаксем именительного падежа.

Для полученного множества обучающих примеров выполнялся метод порождения правил установления значений синтаксем. Всего было порождено более тысячи правил.

Для каждого правила сохранялись примеры, из которых оно было получено. Каждый пример, помимо признаков, содержит тексты, из которых были созданы целевая и соседняя синтаксемы, а также обрабатываемое предложение целиком. Таким образом, каждое правило хранит своё обоснование, которое может быть полезным как для оценивания адекватности реализации метода, так и при анализе лингвистом результатов предсказания на новых примерах.

Для более доступного восприятия была реализована специальная процедура формирования словесной формулировки правил. Словесная формулировка представляет собой описание правила

[2] http://cfrl.ru

на естественном языке и предназначена для экспертов-лингвистов, обычно затрудняющихся понимать запись в виде математических формул.

Правила вносятся в размеченный текстовый файл, из которого их можно впоследствии загрузить и использовать для предсказания значений новых синтаксем.

Приведем пример работы процедуры вывода полной текстовой информации для правила установления значения «дестинатив» (назначение предмета или действия) для синтаксемы родительного падежа с предлогом «для»:

Правило: *Если встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <дестинатив - назначение предмета или действия >*

Обоснование:

Пример 1:

ЗНАЧЕНИЕ = дестинатив

ЦЕЛЕВАЯ СИНТАКСЕМА = для тебя; КСК: личное

СОСЕДНЯЯ СИНТАКСЕМА = Все; ПРЕДЛОГ: ;ПАДЕЖ: им.вин.; КСК: предметное; ПОЗИЦИЯ: до

===КОНТЕКСТ: и песни, и силы - Все для тебя.

Пример 2:

ЗНАЧЕНИЕ = дестинатив

ЦЕЛЕВАЯ СИНТАКСЕМА= для различных рачков; КСК: личное

СОСЕДНЯЯ СИНТАКСЕМА = пища; ПРЕДЛОГ: ;ПАДЕЖ: им.; КСК: предметное; ПОЗИЦИЯ: до

===КОНТЕКСТ: Эти растения - пища для различных рачков

В примерах поле «КОНТЕКСТ» содержит предложение, из которого был построен пример.

Автоматически получаемые словесные формулировки правил имеют тот же вид, что и правила, формулируемые экспертом-лингвистом. Это позволяет сравнивать их между собой.

Для проверки корректности работы метода порождения правил установления значений синтаксем порожденные правила сравнивались с лингвистическими правилами, сформулированными до этого экспертом-лингвистом для определенного числа синтаксем.

Всего для проверки использовались 33 лингвистических правила, каждое из которых сопоставлялось с правилами, порожденными для той же целевой синтаксемы. Для 19-и из 33-х нашлись правильно порожденные правила, для 4-х правила были порождены неверно. Для остальных десяти правила были конфликтными или не порождены. Полученные результаты можно интерпретировать как предсказательную точность и полноту правил, которые составляют 0,83 и 0,58 соответственно.

3.2. Снижение ошибок семантического анализа на основе правил установления значений синтаксем

Как уже было сказано, значение синтаксемы не всегда однозначно определяется по морфологическим признакам и категориально-семантическому классу слова, реализующего синтаксему в тексте. Например, для синтаксемы родительного падежа с предлогом «от», принадлежащей категориальному классу «предметные», возможны следующие варианты смысловых значений:

- дестинатив;
- абстинатив;
- сурсив;
- деструктив.

Это означает, что, например, в предложении «от рассвета до заката» словосочетание «от рассвета» может играть роль или назначения объекта (дестинатив), или источника действия или отправителя нежелательного объекта, от которого следует отстраняться (абстинатив) или источника восприятия (сурсив), или объект разрушающего воздействия (деструктив). На самом деле, это словосочетание выражает начало отсчёта времени, и приписывание ему других смыслов ошибочно.

Следующее правило помогает установить это автоматически:

Правило*: Если встречается синтаксема в падеже <родительный> с предлогом <от>, имеющая категориальный класс <предметное>, а после неё встречается синтаксема в падеже <родительный> с предлогом <до>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <темпоратив>*

Обоснование:

Пример 1:

ЗНАЧЕНИЕ=темпоратив

ЦЕЛЕВАЯ СИНТАКСЕМА:От первых проталин; КСК: предметное

СОСЕДНЯЯ СИНТАКСЕМА: СОДЕРЖАНИЕ = грозы;ПРЕДЛОГ: до;ПАДЕЖ: им.род.вин.;КСК: предметное;ПОЗИЦИЯ: после

===КОНТЕКСТ:От первых проталин до первой грозы .

Пример 2:

ЗНАЧЕНИЕ=темпоратив

ЦЕЛЕВАЯ СИНТАКСЕМА:От сева; КСК: предметное

СОСЕДНЯЯ СИНТАКСЕМА: СОДЕРЖАНИЕ = жатвы;ПРЕДЛОГ: до;ПАДЕЖ: им.род.вин.;КСК: предметное;ПОЗИЦИЯ: после

===КОНТЕКСТ:От сева до жатвы.

Именно наличие синтаксемы родительного падежа с предлогом «до», стоящей после рассматриваемой синтаксемы, указывает на то, что ей необходимо установить значение «темпоратив».

В среднем каждая синтаксема может иметь 4 возможных значения, одно из которых верно, поэтому можно сказать, что применение правил установления значений синтаксем, в результате которого остается одно значение для синтаксемы, уменьшает число ошибок семантического анализа текста в среднем в 4 раза, тем самым значительно повышая точность семантического поиска по предложениям определенного типа (в первую очередь по безглагольным предложениям).

Заключение

Предложенная операция вычисления сходства на признаках произвольной природы позволяет выполнять ДСМ-метод для объектов с признаками, обладающими сложной структурой, которые не легко представить бинарными векторами. Введенное понятие составного признака позволяет оперировать объектами с признаками ещё большей сложности, например, вложенными друг в друга. Предложенная операция вычисления сходства для составных признаков, однако, позволяет однообразно оперировать с элементарными и составными признаками, что делает операцию вычисления сходства объектов универсальной и легко реализуемой на практике.

Предложенные модификации ДСМ-метода позволяют работать со сложными объектами, такими как, например, «синтаксема в контексте», с сохранением структуры этих объектов и наглядности их представления. На основе модифицированного ДСМ-метода разработан метод порождения правил установления семантических значений синтаксем.

Проведенные компьютерные эксперименты показали корректность разработанного метода порождения правил установления значений синтаксем. Предсказательная точность полученных правил установления значений синтаксем составила 0.83 при полноте 0.56, что соответствует средним показателям результативности применения логических методов анализа данных в задачах установления смысловых значений лексических единиц.

Предложенный алгоритм снятия смысловой многозначности синтаксем на основе правил установления значений синтаксем позволяет выбрать одно значение для синтаксемы из всех возможных, что уменьшает ошибки семантического анализа текста в среднем в 4 раза.

Реализованные алгоритмы установления значений и снятия семантической многозначности синтаксем на основе порожденных правил установления значений синтаксем внедрены в семантическую поисковую машину Exactus, демонстрационная версия которой доступна в Интернет по адресу http://www.exactus.ru.

Литература

1. Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. – М. 2004. – 544 с.
2. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997. – 112 с.
3. Завьялова О.С. О принципах построения словаря глаголов для задач автоматического анализа текста.// Труды международной конференции Диалог'2004.
4. Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М.: Эдиториал УРСС, 2001. – 440 с.
5. Финн В.К. ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных системах. // НТИ* Сер. 2. - 2000 - №11. – с.1-5.
6. Финн В.К. Об особенностях ДСМ-метода как средства интеллектуального анализа данных. // НТИ Сер. 2. – 2001. - №5. – с. 1-4.
7. Аншаков О.М. Об одной интерпретации ДСМ-метода автоматического порождения гипотез. // НТИ Сер. 2. – 1999. - №1. – с. 45-53.
8. Финн В.К., Михеенкова М.А. О ситуационном расширении ДСМ-метода автоматического порождения гипотез. // НТИ Сер. 2. - 2000 - №11. – с.20-30.
9. Объедков С.А. Алгоритмические аспекты ДСМ-метода автоматического порождения гипотез. // НТИ. Сер. 2. – 1999. - №1. – с. 64-75.

Смирнов Иван Валентинович. Инженер-исследователь лаборатории интеллектуальных динамических систем ИСА РАН. Окончил Российский университет дружбы народов (РУДН) в 2003 году. Имеет 12 публикаций. Область научных интересов: искусственный интеллект, компьютерная лингвистика.

* ВИНИТИ, Ежемесячный научно-технический сборник «Научно-техническая информация», Сер. 2, Информационные процессы и системы