Отождествление синтагматических структур в процессе формирования тематических знаний

© М.С. Шибут, В.С. Яковишин

Академия управления при Президенте Республики Беларусь
m_shi@tut.by

Аннотация

Предлагаемый метод формирования знаний основан на использовании специального формального языка, в котором обычные текстовые предложения выражаются в форме множеств синтагм. Теоретико-множественная форма представления синтагматических структур позволила свести процесс семантического отождествления текстовой информации к установлению отношения включения множеств. В процессе формирования знаний входные предложения сначала преобразуются в теоретико-множественную форму, а затем полученные формальные синтагматические структуры отбираются и объединяются в растущие представления знаний. Любое объединение предложений, в которых выражается заданная пользователем тема, рассматривается как представление тематических знаний, а любое множество таких представлений, полученных в процессе формирования знаний, – как ориентированное на конкретного пользователя описание предметной области.

Формирование тематических знаний может служить основой для автоматического создания адаптированных (ориентированных на конкретного пользователя) текстовых документов, таких как информационно-аналитические обзоры, индивидуальные электронные учебники, а также любые другие вторичные тестовые материалы.

1. Введение

В развитии электронных библиотек все более актуальным становится создание сервисных средств, осуществляющих извлечение знаний на основе семантической обработки получаемой при поиске электронной литературы. В библиотеках, располагающих этими средствами, может осуществляться не только поиск источников, но и их смысловая обработка и формирование тематических знаний, необходимых для порождения специальных адаптированных документов, ориентированных на заданную читателем предметную область. Такими вторичными документами могут быть, в частности, информационно-аналитические обзоры, обновленные или адаптированные (создаваемые по заданным компетентностям) учебные курсы, всевозможные виды изложения содержания источника (конспекты, «шпаргалки» и т.д.).

Заметим, что переход от полученного представления тематических знаний к создаваемому текстовому документу не представляет принципиальных трудностей: этот переход может осуществляться с помощью обычных лингвистических средств перевода (синтеза) текста или путем отображения полученного представления знаний на тексты обрабатываемых источников и извлечения из них соответствующих фрагментов, необходимых для создаваемого адаптированного документа. В последнем случае переход от знаний к создаваемому документу напоминает процесс конспектирования с использованием извлекаемых из источника текстовых выражений (цитат).

В данной работе сосредоточено внимание на реализации анализа и семантической обработки (смыслового отождествления) текстовой информации, осуществляемых в процессе формирования *тематических знаний* (subject knowledge). В соответствии с предлагаемым методом все предложения, полученные в результате документального (полнотекстового) поиска, сначала преобразуются в формальную запись, затем осуществляется отбор интенсионально более информативных предложений и их агрегация в представлениях знаний, формируемых по заданным темам [1]. Предполагается, что тема в отождествляемых предложениях выражается как номинативное словосочетание, которое представляет тему-подлежащее (в формальном членении «подлежащее-сказуемое», «subject-predicate») или тему-детерминант (в актуальном членении «тема–рема», «topic-comment»). Любое множество предложений, в которых выявлена одна

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

и та же тема, может рассматриваться как представление тематических знаний (subject knowledge representation), а любое множество представлений, полученных в процессе формирования тематических знаний, – как ориентированное на конкретного пользователя описание предметной области.

Ниже рассматриваются языковые и алгоритмические средства, позволяющие реализовать процесс формирования тематических знаний на основе автоматической обработки электронных информационных ресурсов. К этим средствам относятся

- язык формальных синтагматических структур;
- алгоритмы перевода текстовой информации на формальный язык;
- правила отождествления синтагматических структур.

2. Формальные синтагматические структуры

Используемый формальный язык представляет двухосновную алгебру с множеством слов и множеством предложений. Множество слов, как обычно, представляет свободную полугруппу (моноид) над некоторым фиксированным алфавитом, а множество предложений – кольцевидную алгебру с одной унарной операцией и парой бинарных операций - координацией и детерминацией. Координация («сложение») ассоциативна и коммутативна, а детерминация («умножение») неассоциативна, некоммутативна и (односторонне) дистрибутивна относительно координации [3, 4].

Следовательно, в формальном представлении предложений могут использоваться как слова, цепочки над некоторым *базовым* алфавитом, так и специальные, *вспомогательные*, символы, обозначающие различные алгебраические операции. Слова и символы операций являются *синтаксической* основой для выражения *семантических* значений: слова служат для выражения лексических значений, символы бинарных операций (как обозначения сочинительной и подчинительной синтаксической связи) - для выражения значений членов предложения, а символ унарной операции - для выражения общих сентенциальных значений (таких, как модальность, отрицание и др.). Таким образом, в формальном языке отчетливо различаются два уровня описания предложений: чисто абстрактный, *синтаксический,* уровень, отвечающий постулируемым свойствам алгебраических операций, и *семантический* уровень, представляющий смысловую интерпретацию абстрактных синтаксических структур.

Синтаксис формального языка, отвечающий указанным выше свойствам алгебраических операций, может быть задан следующим множеством правил:

$$S \rightarrow \Diamond S \,|(X \nabla X) \,|\, (X \Delta S)| \, X,$$
$$X \rightarrow X \nabla X \,|\, (X \Delta S),$$

где S (“предложение”) – начальный символ грамматики; X (“слово”) – начальный символ вывода слов; $\Diamond$ (“модальность”) – символ унарной операции; ∇ - символ координации; Δ - символ детерминации.

В предлагаемых правилах коммутативная и ассоциативная координация задается как элементарная бесскобочная формула $X \nabla X$, в которой один и тот же символ используется для обозначения обоих членов (операндов). Скобочное представление координации в правиле $S \rightarrow (X \nabla X)$ предусматривается для выражения односторонней (левой) дистрибутивности детерминации относительно координации, т.е. в связи с необходимостью порождения цепочек вида $X\Delta(X \nabla X\ldots)$. А представление детерминации требует и разных обозначений операндов (для выражения некоммутативности) и обязательных скобок (для выражения неассоциативности). При этом в правилах $S \rightarrow (X \Delta S)$, $X \rightarrow (X \Delta S)$ одна и та же формула $(X \Delta S)$ задается дважды: в первом случае операция «вычисляется» справа налево, а во втором – слева направо, т.е, порождаются предложения как с левым ветвлением $(X\Delta(X\Delta S\ldots))$, так и с правым $((\ldots X\Delta S)\Delta S)$.

Синтагматические структуры в порождаемых предложениях выражаются в эксплицитной форме как различные соединения элементарных формул – *синтагм,* двухсловных цепочек вида $X_1 \Delta X_2$, где первое слово X_1 - *определяемый* (главный, независимый) член синтагмы; второе слово ΔX_2 – *определяющий* (зависимый) член; X_1, X_2 (лексемы) - лексические значения; Δ (детерминация) - значение члена предложения. Заметим, что на семантическом этапе порождения слова X_1,X_2 могут принимать значение пустой цепочки (X_1,$X_2 \in A^*$), и тогда исходная синтагма $(X_1 \Delta X_2)$ сокращается: она становится либо *свернутой* (сокращение до главного члена X_1), либо *эллиптической* (сокращение до зависимого члена ΔX_2).

Таким образом, порождаемые синтагматические структуры выражаются в стандартной алгебраической форме, использующей обычную скобочную запись и фиксированный порядок. Но возможна и бесскобочная форма представления синтагматических структур. В предлагаемом методе используется теоретико-множественная форма, в которой каждое предложение выражается в виде множества синтагм [1].

Попытаемся убедиться, что теоретико-множественная (дискретная) форма представления является вполне достаточной для эксплицитного выражения всех существующих типов соединения синтагм (см. [4]). Ниже приводятся примеры типов синтагматических соединений, где с порождаемыми скобочными структурами соотносятся их

синтаксические эквиваленты, представленные в теоретико-множественной форме.

Соподчинение и последовательное подчинение. В первом типе соединения несколько зависимых членов соединяются с одним общим главным членом (*новая книга автора*):

$$((X_1\Delta_1X_2)\Delta_2X_3) = \{X_1\Delta_1X_2, X_1\Delta_2X_3\},$$

где X_1 - главный член; Δ_1X_2, Δ_2X_3 – зависимые члены. Во втором типе зависимый член предыдущей синтагмы служит главным членом последующей синтагмы (*книга нового автора*):

$$(X_1\Delta_1(X_2\Delta_2X_3))=\{X_1\Delta_1X_2, X_2\Delta X_3\},$$

где X_2 является и зависимым членом синтагмы $X_1\Delta_1X_2$ и главным членом синтагмы $X_2\Delta_2X_3$.

Структуры с однородными членами предложения. Символ детерминации в этих структурах представляет значение, общее для нескольких членов предложения, в то время как символ координации служит лишь для выражения наличия (сочинительной) связи между этими членами (*новые и старые книги*):

$$(X_1\Delta(X_2\nabla X_3))=\{X_1\Delta X_2, X_1\Delta X_3\},$$

где ΔX_2 и ΔX_3 – однородные члены; Δ (детерминация) – обозначение значения члена предложения; ∇ (координация) – выражение сочинительной связи.

Абсолютно определяемый член (подлежащее). В любой формальной синтагматической структуре выражается единственный немаркированный член (известный как грамматический субъект, подлежащее), занимающий в структуре начальную позицию. Этот член синтагматической структуры не имеет своего определяемого члена и, следовательно, в нем не выражается значение члена предложения. В теоретико-множественном представлении абсолютно определяемый член будет выделяться (для удобства отождествления структур) как отдельный элемент множества, например:

$$((X_1\Delta_1X_2)\Delta_2X_3) = \{X_1, X_1\Delta_1X_2, X_1\Delta_2X_3\},$$

где абсолютно определяемый член X_1 повторяется как свернутая синтагма.

Абсолютно определяющий член (детерминант). В формальной синтагматической структуре может выражаться общий определяющий член, или детерминант, относящийся ко всему предложению в целом. В теории актуального членения предложения этот определяющий член известен как тема (theme, topic) - элемент двухчленной структуры тема-рема (topic-comment). В теоретико-множественном представлении детерминант может выделяться (наряду с подлежащим) как отдельный элемент множества. Но в отличие от выделяемого подлежащего детерминант содержит не только лексическую часть, но и символ детерминации, т.е. он представляется как эллиптическая синтагма:

$$((X_1\Delta_1(X_2\Delta_2X_3))\Delta_3X_4) = \{\Delta_3X_4, X_1, X_1\Delta_1X_2, X_2\Delta_2X_3\},$$

где Δ_3X_4 выступает в позиции определяющего члена всего предложения – как, например, в предложении *По вечерам он гулял в саду*; ср. предложение *Он гулял в саду по вечерам,* где тот же определяющий член относится не к целому предложению, а лишь к отдельному определяемому члену (сказуемому).

В семантическом представлении предложений, получаемом на втором, семантическом, этапе порождения языка, лексические значения словоформ могут выражаться обычными словами (их основами или грамматически нейтральными словоформами), а значения членов предложения – специальными обозначениями. Слова в получаемых семантических синтагмах будем разделять, ради наглядности, знаком подчеркивания, а специальные обозначения и основы слов – точкой; например:

человек_а.молод ‘молодой человек’,
человек_р.чита ‘человек читает’,
чита_о.книг ‘чита[ет] книгу’,

где *а* (атрибут), *р* (предикат), *о* (объект) – значения членов предложения.

3. Перевод предложений на формальный язык

Перевод входных предложений на формальный язык сводится к распознаванию в тексте синтаксических связей и выявлению синтагм. Алгоритм перевода состоит из следующих этапов.

Морфологический анализ и кодирование грамматических значений словоформ

В каждой анализируемой словоформе определяется (с помощью словаря основ) окончание и тип парадигмы, что позволяет осуществлять замену окончаний соответствующими грамматическими значениями, представленными в виде цифровых грамматических кодов. Так, по правилу -*а*(f_1-f_4)→1.113 окончание -*а*, используемое в типах женского склонения f_1-f_4, заменяется значениями «существительное», «именительный падеж», «единственное число», «женский род» (*карта, книга*); по правилу *а*(m_1, m_6)→1.211/120 окончание -*а*, используемое в типах мужского склонения m_1, m_6, заменяется значениями «родительный падеж», «единственное число», «мужской род» (*дома, мастера, сторожа*) или значениями «именительный падеж», «множественное число» (*домб, мастерб, сторожб*) и т. д. [2].

Выявление и свертывание синтагм, основанных на согласовании

Полученные грамматические коды учитываются при выявлении синтаксической связи. Считается, что смежные в тексте словоформы синтаксически связаны, если их основные значения (первые части кодов) представляют сочетаемые части речи (например, существительное и прилагательное), а все остальные значения (мантиссы кодов) совпадают; например: 3.120.X_1 и 1.120.X_2 (*большие дома*), 3.51.X_1 и 1.511.X_2 (*большим домом*) и т. д.

В выявленных синтагмах осуществляется свертывание - удаление определяющих членов: $3.120.X_1 1.120.X_2 \rightarrow 1.120.X_2$, $3.51.X_1 1.511.X_2 \rightarrow 1.511.X_2$.

Выделение в оставшихся словосочетаниях гипотетических синтагм

После свертывания синтагм, основанных на согласовании, в анализируемом предложении могут оставаться словосочетания с неопознанными синтагмами, основанными на других видах подчинительной связи - управлении и примыкании. Предполагаемые синтагмы в таких словосочетаниях выделяются с учетом возможности существования в синтагматических структурах двух типов подчинительной связи – соподчинения и последовательного подчинения. Так, если в анализируемом предложении осталось неопознанным словосочетание *ABCD*, то необходимо рассмотреть, учитывая все альтернативные варианты соединений, следующие гипотетические синтагмы: *AB, AC, AD, BC, BD, CD*.

Верификация гипотетических синтагм

Верификация выявляемых синтагм осуществляется по создаваемому списку, который автоматически пополняется достоверными синтагмами, выявляемыми в безальтернативных ситуациях. Безальтернативные ситуации могут быть обнаружены в результате удаления несовместимых альтернативных вариантов; при этом учитывается свойство синтагматических структур, согласно которому определяющая словоформа любой синтагмы может быть связана только с одной единственной определяемой словоформой. Так, если известно, что в анализируемом словосочетании *ABCD* синтагмы *AC* и *CD* являются достоверными, то из всего множества гипотетических синтагм *AB, AC, AD, BC, BD, CD* можно удалить синтагмы *AD, BC* и *BD* (поскольку словоформы *C* и *D* уже используются как определяющие члены в достоверных синтагмах). Таким образом, в результате синтаксического анализа в словосочетании *ABCD* выявляются синтагмы *AB, AC* и *CD*, где единственная новая синтагма *AB* (представляющая безальтернативную ситуацию) может рассматриваться как достоверная.

Работу алгоритма рассмотрим на примере перевода предложения: *Писатель подарил ученику новую книгу с дарственной надписью.* В процессе данного перевода выполняются следующие действия.

(1) Замена словоформ основами, дополненными грамматическими кодами:

1.111.писател- 2.111.подари- 1.320.ученик-
3.413.нов- 1.413.книг- с 3.513.дарственн-
1.513.надпис- .

(2) В группе сказуемого выявляются (свертываются и заносятся в формируемое множество) синтагмы, образованные по способу согласования:

1.111.писател- 2.111.подари- 1.320.ученик-
1.413.книг- с 1.513.надпис-
{книг-_а. нов-, надпис-_а.дарственн-}.

(3) В группе сказуемого происходит выявление всех остальных синтагм, и в анализируемом предложении остается только предикативная синтагма:

1.111.писател- 2.111.подари-
{книг-_а. нов-, надпис-_а.дарственн-,
подари-_ adr.ученик-, подари-_о.книг-,
книг-_cmt.надпис-}.

(4) Свертывание предикативной синтагмы. В анализируемом предложении остается только абсолютно определяемый член:

1.111.писател-и
{писател-_ р.подари-, книг-_а. нов-,
надпис-_а.дарственн-, подари-_ adr.ученик-,
подари-_о.книг-, книг-_cmt.надпис-}.

(5) Занесение абсолютно определяемого члена в формируемое представление как отдельного элемента множества (завершение перевода):

{писател-, писател-_ р.подари-, книг-_а. нов-,
надпис-_а.дарственн-, подари-_ adr.ученик-,
подари-_о.книг-, книг-_cmt.надпис-}.

Здесь *a*, *p*, *o* - значения атрибута, предиката и объекта; *adr* – значение адресата; *cmt* (комитатив) - значение наличия признака (см. [3]).

4. Представление и формирование тематических знаний

Формальное выделение подлежащего (в виде свернутой синтагмы) и детерминанта (в виде эллиптической синтагмы) позволяет отделять *тему* предложения (как номинативную фразу, в которой содержится подлежащее или детерминант) от остальной его части (представляющей предикат или рему) и объединять тематически тождественные предложения в единое представление знаний. Таким образом, в результате анализа текстовой информации можно получить *представление тематических знаний* (subject knowledge representation) - как множество предложений, содержащих одну и ту же тему, а также *представление предметной области* (subject field representation) - как множество представлений тематических знаний, т. е.

$$\sigma(N)=\{S \supseteq N \mid S \text{ - предложение}\},$$
$$\sigma(N_1, N_2, \ldots) = \{\sigma(N_1), \sigma(N_2), \ldots\},$$

где N, N_1, N_2, … номинативные фразы, представляющие темы; $\sigma(N)$, $\sigma(N_1)$, $\sigma(N_2)$, … – представления тематических знаний («синопсисы тем»); $\sigma(N_1, N_2, \ldots)$ – представление предметной области («синопсис тематики»).

Согласно определению, в тематическом представлении допускается свободная

корректировка (обобщение или конкретизация) темы путем уменьшения или увеличения множества N, что позволяет получать разные объемные варианты представления знаний, отвечающие потребностям пользователя в расширении или сужении представляемой предметной области. Так, в предельном случае, когда в заданной номинативной части указано лишь одно номинативное слово ($|N| = 1$) можно получить максимальное множество $\sigma(N)$, т.е. представление, наиболее широко охватывающее заданную тему.

Среди фрагментов текста, извлекаемых в процессе поиска из большого количества различных документов, неизбежны многочисленные повторения одной и той же темы и предложений с одним и тем же содержанием. В связи с этим возникает задача архивации извлекаемых знаний – устранения в них семантической избыточности. Решение этой задачи может быть получено путем удаления из анализируемого текста содержательно менее информативных предложений и формирования сложных представлений знаний из предложений, в которых используется общая тематическая часть.

Теоретико-множественная форма представления знаний позволяет рассматривать интенсиональные различия анализируемых предложений как отношения включения множеств. При этом может происходить удаление менее содержательных предложений. Так, если в анализируемом тексте представлены предложения S_1 (*Поступила новая книга автора*) и S_2 (*Поступила новая книга*), то может быть отсеяно S_2 как менее информативное предложение (поскольку $S \supseteq S_2$). Таким образом, отбор предложений с учетом степени их интенсиональности реализуется с помощью правила

$\{S, S'\} \rightarrow \{S,$ если $S \supseteq S'\}$ – правило селекции.

Учитывая отношение включения множеств, можно осуществлять также объединение предложений в одном тематическом представлении знаний: любые предложения S_1, S_2, S_3 и т. д. могут объединяться в одном представлении $\sigma(N)$, если в них содержится общая номинативная часть N, т.е.

$\{ S_1, S_2, \ldots\} \rightarrow \{S \supseteq N | S -$ предложение$\}$ – правило агрегации.

Применение указанных правил покажем на следующем примере. Пусть имеются предложения S_1, S_2, S_3, S_4, S_5, представленные следующими множествами синтагм:

S_1 = {человек-, человек-_а.молод-, человек-_р.чита-, чита-_о.книг-}

‘Молодой человек читает книгу’

S_2 = {человек-, человек-_а.молод-, человек-_р.чита-, чита-_о.книг-, чита-_in.библиотек-}

‘Молодой человек читает книгу в библиотеке’

S_3 = {человек-, человек-_а.молод-, человек-_pt.гуля-, гуля-_in.парк-}

‘Молодой человек гулял в парке’

S_4 = {библиотек-, библиотек-_pPs. расположен-, расположен-_in.улиц-, улиц-_а.красив-}

‘Библиотека расположена на красивой улице’

S_5 = { человек-, человек-_а.молод-, человек-_ pt.удари-, удари-_о.мяч-}

‘Молодой человек ударил мяч’,

где *a*, *o, in* – значения второстепенных членов предложения; *p, pt, pPs* – значения сказуемого.

В соответствии с правилом селекции первое из этих предложений отсеивается, т.к. $S_1 \subseteq S_2$. А предложения S_2, S_3, S_4, S_5 могут по-разному объединяться по правилу агрегации. Так, если в запросе пользователя указаны темы “человек”, “молодой человек” и “библиотека”, то в результате применения правила агрегации получим следующие представления заданных тем:

$\sigma(\{$человек-$\}) = \{ S_2, S_3, S_5 \}$
$\sigma(\{$человек-, человек- _а.молод-$\}) = \{ S_2, S_5 \}$
$\sigma(\{$библиотек-$\}) = \{ S_2, S_4 \}$.

Заметим, что в предложении S_2 происходит актуализация: номинативная фраза *в библиотеке* становится темой, т.е. переходит в позицию детерминанта – путем изменения порядка слов (*В библиотеке молодой человек читает книгу*).

Литература

[1] Шибут М.С., Яковишин В.С. Метод создания адаптированных учебных материалов на основе обработки электронных информационных ресурсов. *В материалах научно-практической конференции «Прикладная лингвистика в науке и образовании» памяти Р.Г. Пиотровского*. Санкт-Петербург, Изд-во «Лема», 2010. С. 339 – 345.

[2] Шибут, М.С., Яковишин В.С. Метод представления знаний в интеллектуальной обучающей системе. *В материалах IV международной конференции «Дистанционное обучение – образовательная среда XXI века»*. Минск, БГУИР, 2004. С. 347–348.

[3] Яковишин В.С. Формальный язык: Теория. Грамматика. Применение. Минск, ИТК НАН Беларуси, 2000. – 152с.

[4] Yakovishin V.S. Algebraic representation of syntagmatic structures.In *Web Journal of Formal, Computational & Cognitive Linguistics*, Issue 11, 2009 [Electronic resource], http//fccl.ksu.ru/issue11.

Identification of Syntagmatic Structures in the Process of Subject Knowledge Formation

© M.S.Shibut, V.S.Yakovishin

The presented method is based on the use of the special formal language, in which all sentence syntagmatic structures are expressed as sets of their syntactic elements – syntagmes. The set-theoretical form allows to reduce the semantic identification of sentences to the use of set inclusion. In the knowledge formation process, the input text sentences are at first transformed into the set-theoretical form, then the resulting formal syntagmatic structures are selected and united into growing knowledge representations. The integration of the sentences that have one and the same subject (a noun phrase contained in user's request) can be considered as a subject knowledge representation; then any collection of the subject knowledge representations produced in the knowledge formation process is a user-oriented (highly tailored) representation of subject field.

The subject knowledge formation can be used as a basis for automatic creation (compiling) of various adapted (user-oriented) text documents, such as information-analytical reviews, individual electronic textbooks, as well as any other producible text materials.