УДК 681.3

А.А. МАРЧЕНКО

МЕТОД АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ. II. АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ В ОНТОЛОГИЧЕСКОЙ СЕТИ[1]

Аннотация. Представлена семантико-синтаксическая модель естественного языка. После факторизации построенных тензоров модели генерируются векторы семантико-синтаксической валентности слов, описывающие коммутативное поведение слов в предложении. Разработан метод вычисления векторов семантико-синтаксической валентности концептов онтологии, которые являются неявным описанием их семантических отношений. Приведен алгоритм выделения явных семантических отношений между концептами онтологии из векторов их семантико-синтаксической валентности.

Ключевые слова: автоматическое извлечение знаний, корпусная лингвистика, онтология, неотрицательная факторизация тензоров.

ВВЕДЕНИЕ

Факторизованные тензорные лингвистические модели позволяют весьма успешно автоматически выделять из корпусов текстов такие лингвистические структуры, как предпочтения сочетаемости в предложениях (selectional preferences) [1] и субкатегориальные фреймы глаголов (verb subcategorization frame) [2], которые включают данные о семантических и синтаксических свойствах связей между глаголами и их аргументами — существительными в предложениях. Из этого следует возможность автоматического выделения из полученного латентного семантического пространства семантических отношений типа «ролевых» падежей Филмора–Грубера [3]. Семантические ролевые падежи в онтологии представляют систему различных связей предикатно-аргументного типа между концептами-узлами, описывающими некоторые процессы или действия, и концептами, которые используются в схемах реализации данных процессов в той или иной роли, например, в роли субъекта или объекта, или реципиента и т.д. Таким образом, ролевые падежи описывают предикатно-аргументную структуру семантики концептов-глаголов. Определение семантических связей между концептами онтологий в процессе обработки и анализа разложенных тензоров текстовых корпусов позволяет автоматизировать наполнение контентом онтологических баз знаний.

В работе [4] описано построение семантико-синтаксической модели естественного языка на основе неотрицательной факторизации лингвистических тензоров, собранных частотным анализом синтаксических структур предложений из больших текстовых корпусов. Векторы из матриц факторизованных тензоров модели определяют коммуникативные свойства слов: в какой тип отношений вступают эти слова и с какими словами данные отношения устанавливаются.

Вектор из матрицы разложенного лингвистического тензора, соответствующий некоторому слову, будем называть вектором семантико-синтаксической валентности данного слова в синтаксической позиции соответствующей матрицы.

Основная сложность состоит в том, что при построении тензоров семантико-синтаксической сочетаемости слов объектами изучения и анализа являются лексемы — слова, которые имеют разные значения. Векторное представление семантико-синтаксической валентности любого слова w, определяемое соответствующим ему вектором из матрицы разложенного тензора, — это, по сути, сумма составляющих слагаемых векторов отдельных различных семантических значений данного

[1] Начало см. в № 1, 2016

слова w — концептов $Sw_1, Sw_2, \ldots, Sw_t$ в некоторой онтологии. Необходимо по вектору валентности $(x_1, x_2, \ldots, x_k)$ слова w получить составляющие слагаемые векторы валентностей $(x_{11}, x_{12}, \ldots, x_{1k}), (x_{21}, x_{22}, \ldots, x_{2k}), \ldots, (x_{t1}, x_{t2}, \ldots, x_{tk})$ для каждого из его t значений. Вектор валентностей фиксированного значения — концепта онтологии, является неявным описанием его семантических отношений с другими концептами онтологической базы знаний. Простейшая операция вычисления скалярного произведения векторов валентностей двух концептов онтологии может подтвердить наличие семантической связи между ними.

Разработан метод определения семантических отношений между концептами-синсетами WordNet [5]. Он реализуется посредством анализа матриц разложенного тензора, сформированного при обработке корпусов статей English Wikipedia и Simple English Wikipedia, с расщеплением векторов семантической валентности слов на составляющие векторы семантической валентности их различных значений и с конкретной привязкой расщепленных векторов к соответствующим концептуальным узлам сети WordNet [6]. Предложенный метод протестирован на точность разделения векторов семантической валентности слов на составные слагаемые векторы семантической валентности концептов-значений данных слов, а также на точность их привязки к синсетам WordNet [5]. Основным преимуществом данного метода есть полная автоматизация определения новых семантических отношений между концептами онтологической базы знаний в процессе анализа матриц факторизованного тензора больших текстовых корпусов. Несмотря на то, что отношения задаются неявно — через векторы семантических валентностей, именно эта форма представления дает возможность решать такие классические задачи компьютерной лингвистики, как разрешение неоднозначности слов (word sense disambiguation), измерение семантической близости слов, семантический анализ текстов и др.

РАСЩЕПЛЕНИЕ ВЕКТОРОВ СЕМАНТИКО-СИНТАКСИЧЕСКОЙ ВАЛЕНТНОСТИ СЛОВ НА СОСТАВЛЯЮЩИЕ СЛАГАЕМЫЕ ВЕКТОРЫ ВАЛЕНТНОСТЕЙ ИХ РАЗЛИЧНЫХ ЗНАЧЕНИЙ

После факторизации матрицы D и трехмерного тензора F получены соответственно две матрицы: W и H, для кольцевых синтагматических связей и три матрицы: X, Y и Z, для предикативных линейных связей, которые состоят из векторов размерности k — степени факторизации. Каждый вектор из этих матриц соответствует некоторому слову или словосочетанию. Векторы описывают семантико-синтаксическое поведение лексемы, а именно в каких синтаксических позициях, с какими словами и какого типа связи образует данная лексема. Как уже отмечалось, слова неоднозначны, т.е. они, как правило, имеют несколько значений. Вектор слова является суммой векторов всех его значений. Одно слово может иметь несколько векторов из различных матриц, соответствующих разным синтаксическим позициям. Задача расщепления каждого из этих векторов решается отдельно. Разработанный алгоритм расщепления векторов семантико-синтаксической валентности слова на множество векторов семантико-синтаксических валентностей всех его значений — синсетов WordNet, решает следующую задачу.

Дан вектор семантической валентности X_w размерности k, который соответствует некоторому слову w в матрице X (или в любой другой из пяти матриц, когда метод работает аналогично). Положим, что существительному w соответствует t значений — синсетов в WordNet. Требуется разделить X_w на составляющие слагаемые $X_{w_1}, X_{w_2}, \ldots, X_{w_t}$, соответствующие данным t синсетам.

Алгоритм расщепления векторов семантико-синтаксической валентности слов на составляющие слагаемые векторы валентностей их различных значений — концептов-синсетов WordNet, описан в [5]. Он представляет собой анализ каждого из k значений вектора X_w с последующим определением, которому из t значений слова w и соответственно из t синсетов WordNet принадлежит каждое конкретное ненулевое

значение вектора. Таким образом, алгоритм расщепляет вектор семантико-синтаксической валентности слова w на t векторов его отдельных значений. При этом полученные векторы привязываются алгоритмом к конкретным синсетам WordNet.

ЭКСПЕРИМЕНТЫ С АЛГОРИТМОМ РАСЩЕПЛЕНИЯ И ПРИВЯЗКИ

Вычисление оценок точности работы алгоритма расщепления и привязки на векторах из матриц X, Y, Z, W и H, задающих линейные предикативные и кольцевые синтагматические связи в предложениях, проводились согласно методике, описанной в [5]. В результате получены следующие оценки точности работы алгоритма расщепления векторов валентности слов на составные слагаемые векторы их различных значений и привязки их к синсетам WordNet: X — 93,17 %; Y — 85,93 %; Z — 87,81 %; W — 92,89 %; H — 90,03 %. Эти оценки свидетельствуют о высокой эффективности алгоритма и перспективах его использования.

Отметим, что значительное преимущество предложенного метода состоит в высокой степени автоматизации каждого этапа его работы: парсинг статей, сборка тензора, неотрицательная тензорная факторизация, расщепление векторов семантической валентности слов на векторы их значений и привязка к соответствующим синсетам WordNet.

Задание семантических отношений между концептами-узлами онтологического графа в неявном виде с помощью k-мерных векторов семантических валентностей также имеет преимущество универсальности представления семантических связей. При обнаружении $(n+1)$-го типа связи между существующими концептами система фиксирует в базе его наличие в векторном представлении, не выдвигая немедленного требования пополнения списка отношений новым типом, его полного описания в онтологии и указания для него соответствующего синтаксического шаблона.

В данном представлении имеется и существенный недостаток. В полученной модели можно легко проверить наличие линейной предикативной α–β-связи между конкретными тремя концептами-узлами WordNet: a, b и c, или кольцевой синтагматической α–β-связи между концептами a и b, что является достаточным для ряда алгоритмов анализа предложений. Однако алгоритмы, использующие поиск кратчайших путей — цепочек между узлами онтологий, не могут напрямую применять данной модели без явного описания отношений между концептами-узлами в виде инцидентных ребер. Кроме того, сложно рассматривать пополнение семантической сети новыми отношениями как процесс обогащения онтологий в общепризнанном понимании без генерации явного описания семантических связей между понятийными узлами.

Из этого следует необходимость построения метода автоматического выделения явных семантических связей из векторов семантико-синтаксической валентности концептов онтологии.

АЛГОРИТМ ВЫДЕЛЕНИЯ ЯВНЫХ СЕМАНТИЧЕСКИХ СВЯЗЕЙ ИЗ ВЕКТОРОВ СЕМАНТИКО-СИНТАКСИЧЕСКОЙ ВАЛЕНТНОСТИ КОНЦЕПТОВ ОНТОЛОГИИ

Идея метода выделения явных семантических связей предикативно-аргументного или кольцевого синтагматического типа из векторной модели представления отношений между узлами-концептами онтологии состоит в повторном чтении обработанных текстов корпуса посредством процедуры синтаксического анализа, реализованного на основе алгоритма Кока–Янгера–Касами [7], работающего на базе векторов семантико-синтаксической валентности концептов, привязанных к синсетам WordNet.

Схематически можно представить этот процесс следующим образом.

Алгоритм последовательно анализирует предложения текстов из обучающего корпуса и строит управляющие пространства синтаксических структур.

Когда алгоритму нужно узнать, возможно ли построение кольцевой синтагматической связи между словами a и b, переходим к синсетам-узлам WordNet $\{A_i\}$, на которые ссылается слово a, и синсетам-узлам $\{B_i\}$, на которые ссылается слово b.

Вычисляем значения $(W_{a'}, H^{\mathrm{T}}_{b'})$ для слов $a' \in \{A_i\}$, $b' \in \{B_i\}$, если среди них существуют такие a'' и b'', для которых $(W_{a''}, H^{\mathrm{T}}_{b''}) > T1$ (где $T1$ — некоторый пороговый уровень, определяемый эмпирически, например $T1 \geq 1$), то образуется кольцевая синтагматическая связь и устанавливается семантическое отношение типа «объект–свойство» между концептом-синсетом $A_k : a'' \in A_k$ и концептом-синсетом $B_j : b'' \in B_j$.

Существование хотя бы одной такой пары (а именно a'' и b'') гарантирует тот факт, что данные для матрицы D получены из тех же текстов и предложений, которые обрабатываются в настоящий момент. Соответствующее синтаксическое отношение между словами a и b занесено в матрицу D и попало в векторы ее факторизованных матриц W и H. Во время расщепления этих векторов семантико-синтаксических валентностей соответствующие значения привязывались к некоторым двум синсетам, которые и будут найдены описанным алгоритмом — между ними установится данное отношение.

Алгоритм гарантированно корректно находит и определяет семантическое отношение между синсетами при выполнении двух следующих условий:

— при обучении соответствующая синтаксическая связь правильно записана в дерево подчинения и дерево вывода предложения Стенфордским парсером, в результате соответствующее управляющее пространство, на котором обучалась система, составлено корректно и данная связь правильно записана в матрице D и, следовательно, гарантированно имеется в векторах факторизированных матриц W и H;

— алгоритм расщепления векторов семантико-синтаксической валентности правильно расщепил соответствующие данной связи векторы из W и H и корректно выполнил привязку к правильно установленным узлам-синсетам WordNet.

Проведенные эксперименты свидетельствуют о высокой точности работы используемых алгоритмов, а значит, о высокой надежности данного метода определения семантических отношений между синсетами WordNet.

Для увеличения степени надежности алгоритма можно ввести дополнительную проверку наличия подобных связей у сыновей и/или отцов данных синсетов A_k и B_j. Если эти связи имеются, т.е. $\exists A'_k$ — отцовский/сыновий синсет A_k и $\exists B'_j$ — отцовский/сыновий синсет B_j: $\exists a' \in A'_k$ и $\exists b' \in B'_j$ такие, что $(W_{a'}, H^{\mathrm{T}}_{b'}) \geq T1$, то вероятность построения корректной семантической связи между синсетами A_k и B_j значительно увеличивается.

Отметим, что связи между синсетами A_k и B_j будут найдены сразу, если приведенные ранее условия выполнены. Если при обучении связь не найдена или найдена, но неправильно привязана к некорректным синсетам, то дополнительные проверки увеличивают надежность работы метода определения семантической связи между синсетами.

По сути, для повторного чтения текстов обучающего корпуса применяется парсер-построитель управляющих пространств (см. [4]) с той лишь разницей, что в [4] для определения наличия связей между словами использовались векторы семантико-синтаксической валентности слов, полученные после факторизации матрицы D и тензора F, а в настоящей статье осуществляется переход от слов к их значениям — синсетам WordNet, так как в данной модели векторы семантико-синтаксической валентности привязаны исключительно к синсетам. Таким образом, парсер «вынужденно» переходит от слов и словосочетаний к их семантическим значениям и данная фаза семантического анализа осуществляется параллельно с синтаксическим.

Рассмотрим случай, когда алгоритм выявляет более чем одну пару синсетов A_k и B_j $(a'' \in A_k,\ b'' \in B_j) : (W_{a''}, H^{\mathrm{T}}_{b''}) > T1$.

Алгоритм Кока–Янгера–Касами представляет собой процесс динамической сборки всех возможных вариантов синтаксической структуры предложения. На каждом уровне процесса построения управляющего пространства синтаксической структуры происходит слияние двух структур (двух точек управляющих пространств) в одну более крупную структуру — точку, которая наследует тем или иным образом лексическое значение у своих образующих точек. Дальнейшее присоединение на более высоком структурном уровне произойдет согласно векторам семантико-синтаксической валентности тех синсетов, на которые сошлется это новое лексическое значение образованной структуры — точки управляющего пространства (УП). Поэтому можно сохранять в ячейках таблицы объединения всех возможных пар A_k и B_j $(a'' \in A_k, b'' \in B_j) : (W_{a''}, H^{\mathrm{T}}_{b''}) > T1$. В процессе генерации общей структуры предложения алгоритмом Кока–Янгера–Касами некорректные варианты будут исключены вследствие невозможности установления семантико-синтаксической связи на верхних уровнях таблицы и построения целостной структуры. В случае успешного завершения работы полностью построенная структура УП в точках будет содержать объединения соответствующих корректных значений — синсетов WordNet, с использованием правильных семантико-синтаксических отношений между ними. Следовательно, можно внести эти отношения в семантическую базу WordNet путем добавления соответствующих семантических связей между синсетами уже после окончания построения полной и целостной структуры УП для данного предложения.

Рассмотрим более формально описание алгоритма.

Вход. Лексико-семантическая база WordNet с векторами семантико-синтаксической валентности, привязанными к ее синсетам-узлам, и входная цепочка слов предложения $\omega = a_1 a_2 \ldots a_n \in \Sigma^+$.

Выход. Таблица семантико-синтаксического анализа T, описывающая управляющее пространство синтаксической структуры входного предложения, содержащее в своих точках значения слов — синсеты WordNet, соединенные семантическими α–β-связями.

Алгоритм

Шаг 1. Положим $t_{i1} = \{A_i \mid A_i$ — синсеты, на которые ссылаются a_i $\forall i = 1, \ldots, n\}$. Если таковых не существует (например, для предлогов, союзов и т.д.), то вместо синсетов t_{i1} содержит только лексемы, также имеющие собственные векторы семантико-синтаксической валентности, с помощью которых они на следующем этапе вступают в связь с семантически значащими лексемами.

Шаг 2. Допустим, что уже вычислены t_{ij} для всех $1 \le i \le n$ и всех $1 \le j' < j$. Положим $t_{ij} = \{A_i$ для некоторого $1 < k \le j$, $\{B_i\} \in t_{ik}$ и $\{C_i\} \in t_{i+k,j-k}$, $\xi(B_i, C_i) = A_i\}$. Поскольку $1 < k \le j$, то k и $j-k$ меньше j. Таким образом, t_{ik} и $t_{i+k,j-k}$ вычисляются раньше, чем $A_i \Rightarrow^+ a_i a_{i+1} \ldots a_{i+j-1}$.

После этого шага из $A_i \in t_{ij}$ следует, что

$$A_i \Rightarrow (B_i, C_i) \Rightarrow^+ (a_i \ldots a_{i+k-1}, C) \Rightarrow \ldots \Rightarrow a_i \ldots a_{i+k-1} a_{i+k} \ldots a_{i+j-1}.$$

Шаг 3. Повторять шаг 2 до тех пор, пока не будут известны t_{ij} для всех $1 \le i \le n$ и $1 \le j \le n-i+1$.

Рассмотрим работу функции $\xi(B, C)$. Она проверяет, могут ли точки B и C установить между собой кольцевую синтагматическую α–β-связь либо α-связь линейного предикативного отношения, либо β-связь линейного предикативного отношения.

Для определения возможности установления кольцевой синтагматической α–β-связи между точками B и C проверка выполняется вычислением скалярного произведения (W_B, H^{T}_C), где W_B — вектор концепта-синсета, являющегося се-

мантическим значением точки B, а H_C — вектор концепта-синсета, являющегося семантическим значением точки C. Если $(W_B, H_C^{\mathrm{T}}) \geq T_{\alpha\beta}$, то функция ξ устанавливает данную кольцевую синтагматическую связь, размещает ее в новой точке A и вычисляет ее новое лексическое и семантическое значение. Оно либо наследуется из точки B как главной точки пары, либо в результате объединения образуется новое значение («Черная» + «дыра» = «Черная дыра»). Точка A получает найденное функцией ξ лексическое значение и соответствующий синсет (или набор синсетов в случае неоднозначности) в качестве семантического значения. К синсету (или синсетам) привязаны его (их) векторы семантико-синтаксической валентности из различных матриц тензорной модели языка.

Для определения возможности установления β-связи линейного предикативного отношения между точками проверка выполняется вычислением скалярного произведения (Y_B, Z_C), где Y_B — вектор Y концепта-синсета, являющегося семантическим значением точки B, и Z_C — вектор Z концепта-синсета, являющегося семантическим значением точки C. Если $(Y_B, Z_C) \geq T_\beta$, то функция ξ устанавливает данную β-связь линейного предикативного отношения, размещает ее в новой точке $A1$ и вычисляет ее лексическое и семантическое значения. Как правило, в $A1$ наследуются оба лексических и оба семантических значения из точек B и C.

Для определения возможности установления α-связи линейного предикативного отношения между точками D и $A1$ проверка выполняется вычислением $\sum_{i=1}^{k} X_D[i] * Y_B[i] * Z_C[i]$, где X_D — вектор концепта-синсета, являющегося семантическим значением точки D; Y_B — вектор концепта-синсета, являющегося семантическим значением точки B из $A1$; Z_C — вектор концепта-синсета, являющегося семантическим значением точки C из $A1$. Если $\sum_{i=1}^{k} X_D[i] * Y_B[i] * Z_C[i] \geq T_P$, то функция ξ устанавливает α-связь линейного предикативного отношения и завершает тем самым полную линейную предикативную последовательность α–β-связей УП данного предложения.

После того как алгоритм успешно завершает сборку целостного управляющего пространства предложения полностью построенная структура УП в точках содержит объединения соответствующих корректных значений — синсетов WordNet, с использованием правильных семантико-синтаксических отношений между ними.

Далее следует этап обхода построенной структуры УП предложения с перенесением найденных между синсетами семантических связей в базу данных WordNet в виде соответствующих семантических отношений между узлами-концептами онтологии.

Если между точками УП, содержащими синсеты A и B в качестве их семантических значений, установлена кольцевая синтагматическая α–β-связь, то между A и B в WordNet записывается явное отношение «объект–свойство» либо «действие–свойство».

Если между точками УП, содержащими синсеты A и B в качестве их семантических значений, установлена линейная предикативная α-связь, то между A и B в WordNet записывается явное отношение «субъект–действие» либо «актант–действие».

Если между точками УП, содержащими синсеты B и C в качестве их семантических значений, установлена линейная предикативная β-связь, то между B и C в WordNet записывается явное отношение «действие–объект» либо «действие–реципиент», либо «действие–инструмент» и т.д. в зависимости от использования того или иного синтаксического шаблона в предложении для выражения данной β-связи. Семантические ролевые падежи Филмора–Грубера [3] (табл. 1) описывают соответствие синтаксических шаблонов типам семантических отношений. Одновременно с добавлением в онтологию набора семантических связей из одного предложения фиксируются метки об их совместном использовании в рамках одной функциональной предикатно-аргументной схемы глагола.

Таблица 1. Семантические модально-ролевые падежи Филмора–Грубера, описывающие соответствие синтаксических шаблонов типам семантических отношений

Модально-ролевое отношение	Семантическое значение	Синтаксический шаблон	Пример
Актант или субъект	Инициатор действия, концепт типа «объект»	Соответствует группе существительного NG	**Jonn** writes a letter
Тема (объект)	Объект, над которым проводится действие, концепт типа «объект»	Соответствует группе существительного NG, но в другой синтаксической позиции	Jonn writes **a letter**
Реципиент (подвид объекта)	Объект, в направлении которого проводится действие	Соответствует предложной группе PG	Jonn writes a letter **to Mary**
Инструмент	Объект, с помощью которого проводится действие	Соответствует предложной группе PG с предлогами “with” и “by”	Jonn writes a letter **with his Parker pen**
Стиль	Способ выполнения действия	Выражается с помощью наречий ADV	Jonn writes **easily**

После повторной обработки обучающих текстовых корпусов все неявные семантические связи синсетов, выраженные векторами семантико-синтаксических валентностей, записываются в WordNet явным образом добавлением соответствующих ребер в граф сети онтологии.

ЗАКЛЮЧЕНИЕ

В статье описана семантико-синтаксическая модель естественного языка, реализованная с помощью неотрицательной факторизации лингвистических тензоров, построенных в результате частотного анализа синтаксических структур предложений из обучающих текстовых корпусов. На основе построенной модели разработан алгоритм пополнения онтологий новыми семантическими отношениями между узлами-концептами. Метод заключается в вычислении векторов семантико-синтаксической валентности концептов онтологии, после чего выполняется повторный анализ текстов обучающих корпусов с использованием структур данных обученной модели и занесением найденных семантико-синтаксических отношений предикатно-аргументного типа в базу данных онтологии.

СПИСОК ЛИТЕРАТУРЫ

1. Van de Cruys T. A non-negative tensor factorization model for selectional preference induction // Journal of Natural Language Engineering. — 2010. — **16**, N 4. — P. 417–437.
2. Van de Cruys T., Rimell L., Poibeau T., Korhonen A. Multi-way tensor factorization for unsupervised lexical acquisition // Proceedings of COLING 2012. — Mumbai, India — P. 2703–2720.
3. Fillmore C.J. The case for case // Universals in linguistic theory / E. Bach, R.T. Harms (Eds.). — New York: Holt, Rinehart, and Winston, 1968. — P. 88.
4. Марченко А.А. Метод автоматического построения онтологических баз знаний I. Разработка семантико-синтаксической модели естественного языка // Кибернетика и системный анализ. — 2016. — **52**, № 1. — С. 23–33.
5. Анисимов А.В., Марченко А.А., Вознюк Т.Г. Определение семантических валентностей концептов онтологий с помощью неотрицательной факторизации тензоров больших текстовых корпусов // Кибернетика и системный анализ. — 2014. — **50**, № 3. — С. 3–16.
6. Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Introduction to WordNet: An on-line lexical database. — http://wordnetcode.princeton.edu/5papers.pdf.
7. Younger D.H. Recognition and parsing of context-free languages in time n3 // Information and Control. — 1967. — **10**, N 2. — P. 189–208.

Поступила 16.07.2015