УДК 681.3

А.А. МАРЧЕНКО

МЕТОД АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ. I. РАЗРАБОТКА СЕМАНТИКО-СИНТАКСИЧЕСКОЙ МОДЕЛИ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация. Представлена семантико-синтаксическая модель естественного языка. Применен тензорный подход к моделированию семантико-синтаксических связей между словами в предложениях. Использован аппарат управляющих пространств синтаксических структур естественного языка, позволяющий усовершенствовать тензорную семантико-синтаксическую модель, описывающую с помощью рекурсии и суперпозиции синтаксические структуры произвольной длины и сложности.

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

В данной работе описывается метод обогащения онтологических сетей новыми связями — семантическими отношениями между концептуальными узлами. Для этого применяется новая семантико-синтаксическая модель естественного языка, основанная на неотрицательной факторизации лингвистических тензоров — многомерных массивов лингвистических данных, полученных при частотном анализе больших корпусов текстов. В тензоре каждое измерение соответствует некоторому фиксированному члену предложения — подлежащему, сказуемому, дополнению, определению, обстоятельству. Данные N-мерные тензоры содержат оценки частоты употребления сочетаний различных наборов слов в предложениях. При этом учитываются синтаксические позиции слов. После обработки больших текстовых корпусов и накопления значительного объема данных в тензоре формируется N-мерный массив описания синтаксического поведения лексических единиц в предложениях, т.е. для множества слов, имеющихся в тензоре, задано, в какие синтаксические отношения, с какими словами и с какой частотой последние имеют свойство вступать. С помощью неотрицательной факторизации огромный и чрезмерно разреженный N-мерный тензор можно представить в более экономном и удобном виде.

Неотрицательная факторизация N-мерного тензора при ранге разложения k формирует N двумерных матриц, состоящих из k вектор-столбцов, представляющих отображение каждого измерения тензора на k фактор-измерений латентного семантического пространства. Это уникальное средство для моделирования и выявления взаимосвязей лингвистических переменных в массиве N-мерных данных. Описанная модель позволяет 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.

Цель настоящей работы — построение алгоритма решения приведенной далее задачи.

Даны: иерархическая таксономия (WordNet), содержащая в своих узлах синсеты — синонимические наборы слов, обозначающие в английском языке соответствующие данным узлам понятия, а также текстовый корпус большой электронной энциклопедии (English Wikipedia), в статьях которой имеются определения понятий и описания их основных свойств, семантических отношений с другими концептами и характеристик этих отношений.

Требуется с помощью частотного синтаксического анализа текстов собрать тензорную лингвистическую структуру, провести неотрицательную факторизацию полученных тензоров, из сгенерированных тензорных матриц выделить в явном виде семантические отношения между концептами и представить их в качестве ребер к соответствующим узлам графа таксономии.

Неотрицательная тензорная факторизация — весьма востребованная технология в компьютерной лингвистике [1–4]. Так, в работах [1, 2] описываются модели тензорного представления данных о частоте различных типов синтаксических сочетаний слов в предложениях, например трехмерных сочетаний subject – verb – object или четырехмерных сочетаний subject – verb – direct_object – indirect_object, или других синтаксических сочетаний длины, не превышающей размерности тензора N.

Очевидной проблемой этой перспективной модели является негибкое и ограниченное представление синтаксиса предложений. Размерность тензора определяет максимальную длину предложений либо словосочетаний, описываемых данной моделью. Каждой размерности соответствует конкретная синтаксическая позиция. В работе [1] рассмотрен трехмерный тензор для моделирования одного синтаксического сочетания типа «подлежащее – сказуемое – дополнение». В работе [2] представлены тензоры размерностью 9 и 12 для моделирования двух десятков различных типов синтаксических отношений-сочетаний. Только лишь увеличение размерности тензора для обработки большего количества типов синтаксических отношений расширенной арности не является очень убедительным методом совершенствования модели. Актуальное и востребованное направление исследований в этом контексте — поиск универсальных средств представления синтаксических структур предложений. Целесообразно использовать такую формальную модель представления синтаксиса, которая с помощью рекурсии может выразить синтаксические отношения в предложениях произвольной длины и любой степени сложности структуры. Такая модель позволит записать многомерные структурные связи между словами в предложениях произвольной длины с помощью ограниченного числа массивов фиксированной размерности.

В качестве модели представления синтаксиса предлагается использовать управляющее пространство синтаксических структур предложений [5]. Существует ряд классических проверенных временем формальных моделей представления синтаксиса языка. Выбор именно управляющих пространств обусловлен тем, что в этой модели с помощью рекурсии описываются произвольные сложные конструкции посредством суперпозиции двух базовых синтаксических отношений: предикативных и синтагматических. Предложенная лексико-синтаксическая тензорная модель состоит из одного трехмерного тензора для предикативных отношений и одной матрицы для синтагматических. Применение управляющих пространств оказалось эффективным средством редукции произвольных N-арных синтаксических отношений в суперпозиции бинарных и тернарных отношений.

Тензорные модели содержат данные о семантико-синтаксических коммуникативных свойствах только тех слов, которые имеются в обработанных корпусах текста, и только в тех предложениях и словосочетаниях, в которых данные слова встречались. Иначе говоря, тензорная модель воспроизводит лишь предложения и словосочетания, содержащиеся в обработанных текстах. В настоящей статье предложено использовать иерархические лексико-семантические базы типа WordNet [6] для обобщения описаний коммуникационных свойств слов с применением неявных механизмов наследования по дереву таксономии. Если слово A имеет определенное свойство, то существует большая вероятность того, что это свойство может быть у всех слов синсета $\mathcal{A}$, в котором содержится A, а также у слов из синсетов — сыновей $\mathcal{A}$ и, возможно, у слов из синсета — родителя $\mathcal{A}$. Именно эти предположения явились основой для реализации механизма обобщения описания коммуникативных семантико-синтаксических свойств слов по принципу таксономического наследования.

Общеизвестно, что для использования естественного языка нужны знания непосредственно о языке (лексике, морфологии, орфографии, синтаксисе и т.д.),

а также о некоторой предметной области (языковые реалии, семантика). Тензорные модели содержат данные, в которых интегрированы описания семантических и синтаксических коммуникативных характеристик слов. Применение лексико-семантических баз типа WordNet усиливает семантическую составляющую модели. В качестве обучающих вместе с корпусом The Wall Street Journal использованы тексты статей English Wikipedia и Simple English Wikipedia, содержащие определения понятий и основную информацию о них, для увеличения объема семантических данных в модели.

УПРАВЛЯЮЩЕЕ ПРОСТРАНСТВО СИНТАКСИЧЕСКИХ СТРУКТУР ЕСТЕСТВЕННОГО ЯЗЫКА

Базовые синтаксические структуры языка описываются в классических схемах грамматики еще с античных времен. Отношения управления между словами выражаются в лингвистических моделях деревьев подчинения и систем составляющих. Главным преимуществом данных моделей является их корректность — адекватное представление характеристик синтаксической структуры предложения, однако имеются и недостатки. Модель деревьев подчинения ориентирована на управляющие связи между словами, а модель систем составляющих учитывает иерархическое отношение вложенности синтаксических групп в линейной структуре текста. При этом модели приближенно описывают коммуникативные свойства синтаксических структур.

Попытки построения моделей, более удобных для машинной обработки, обобщающих и объединяющих преимущества деревьев подчинения и систем составляющих, привели к созданию моделей системы компонент [7] и синтаксических групп [8]. В этих моделях в синтаксических структурах внимание, сфокусированное на линейном порядке, обусловленном последовательностью записи текста, было переключено на сложное пространство, образованное синтаксически связанными группами объектов. В работе [5] предлагается перейти к пространству представления, не зависящему от порядка записи текста, а значит, и от естественного языка. Такое пространство отображает все предикативные и синтагматические отношения, содержащиеся в синтаксических структурах, и называется управляющим пространством синтаксической структуры предложения. В отличие от сугубо лингвистического подхода предложение рассматривается как некоторый динамический вычислительный рекурсивный процесс, развивающийся в управляющем пространстве и связывающий синтаксически сгруппированные части предложения информационными каналами. Структура управляющего пространства отображает семантику синтагматических и предикативных конструкций языка.

Язык имеет фундаментальное свойство выражать динамические отношения слов. Так, глагол связывает отношениями объекты, попадающие под действие этого глагола, прилагательное задает отношение объекта с самим собой. Синтаксическая модель должна содержать описание, какие части предложения связаны между собой с помощью отношений и какого типа последние. Как отмечалось ранее, существуют два типа синтаксических отношений: предикативные и синтагматические. Предикативное отношение с помощью сказуемого-глагола выражает зависимость между синтаксическими объектами через понятие, обозначающее действие. Синтагма — это сочетание двух синтаксических объектов, один из которых является определением другого. Поэтому в модели должны полностью отображаться именно эти типы отношений. Кроме того, в широком понимании синтагмы образуют синтаксические группы.

Адекватная модель синтаксической структуры должна соответствовать основному свойству рекурсивности языка. Для построения данной модели удобнее всего задавать синтаксические отношения связями генерации и передачи отношений. При этом достигается более точная характеристика управляющих связей.

Если объекты A и B вступают в отношение C, то выделяется объект A, который вызывает (порождает) это отношение C, и объект B, на который это отноше-

ние передается. Таким образом, существует два вида направленных связей: α-связь (генерирование) от объекта — генератора отношения, непосредственно к отношению и β-связь (распространение) от отношения к подчиненному объекту. Объекты A, B и отношение C размещаются в точках управляющего пространства, которые соединяются α- и β-связями: $A \xrightarrow{\alpha} C \xrightarrow{\beta} B$.

Глаголы определяют отношения между объектами. Поэтому в стандартной схеме простого предложения: «существительное – глагол – существительное», α-связь направлена от первого существительного к глаголу, а β-связь — от глагола к существительному-определению.

Рассмотрим предложение «Профессор прочитал лекцию». Здесь объект «профессор» генерирует отношение «прочитал» и направляет его на объект «лекция». Поэтому α–β-структура этого предложения выглядит так: «Профессор $\xrightarrow{\alpha}$ прочитал $\xrightarrow{\beta}$ лекцию».

В словосочетании «талантливый футболист» объект «футболист» генерирует унарное отношение «талантливый» и передает его себе. Таким образом, появляется кольцевая связь, характеризующая определения «футболист $\xrightarrow{\alpha}$ талантливый $\xrightarrow{\beta}$ футболист».

Аналогично предложению «Талантливый студент быстро решил задачу» соответствует структура, показанная на рис. 1.

В приведенных предложениях имеются два типа α–β-связей: строго линейная зависимость (линейная констукция, отображающая предикативные конструкции языка) и замкнутая кольцевая зависимость (отображает синтагматические конструкции языка).

Формальная модель, ориентированная на представление сложных структур необходимого вида в форме управляющих пространств, строится следующим образом.

Имеется класс базовых объектов U, с каждым из которых ассоциируется определенный тип. Всего различных типов конечное число. Данные типы можно выразить числами из интервала $[0, N]$. Предполагается неоднозначность при сопоставлении объектов типам, т.е. функция приписывания типов φ отражает U во множество подмножеств, образованных числами из интервала $[0, N]$. Конструкциями являются либо объекты из U, либо конструкции, полученные из других конструкций с помощью подстановки последних в точки линейного или кольцевого отношения.

Относительно синтаксических структур данное определение уточняется следующим образом.

Базовые объекты — это слова и словосочетания, относящиеся к определенным частям речи (имени существительному, прилагательному, глаголу, частицам и т.д.) с соответствующими морфологическими признаками, а также отношения и корреляторы, предназначенные для соединения придаточных предложений

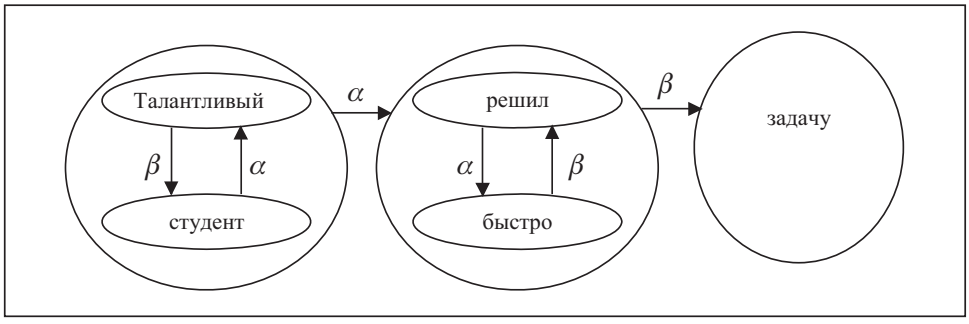

Рис. 1. Структура предложения «Талантливый студент быстро решил задачу»

с главными. Тип слова — это его полное грамматическое значение. Например, тип слова «книга»: имя существительное, неодушевленное, единственное число, именительный падеж. Подразумевается также расширение понятия типа слова добавлением некоторых семантических атрибутов. Неоднозначность в определении типа является следствием многозначности некоторых слов вне контекста. Например, слово «печь» может быть именем существительным или глаголом (инфинитивом).

Введем функцию $f(i, j, k)$, которая в точках управляющего пространства i, j и k задает типы соответствующего простого и сложного предложения. Булева функция $d(i, j)$ в точках управляющего пространства i и j задает согласование типов определяемого I и определяющего J объектов. Например, определяющими для имени существительного могут быть имена прилагательные, предложные группы или придаточные предложения, для глагола — наречие, деепричастие или придаточное предложение; глагол не может являться определюящим для имени существительного.

Для нетранзитивных (непереходных) глаголов для обозначения недостающего слова в предикативной конструкции используется специальное пустое слово $\varnothing$ (например, предложение «Тренер закурил» имеет структуру «тренер $\xrightarrow{\alpha}$ закурил $\xrightarrow{\beta} \varnothing$»), а для транзитивных (переходных) глаголов — параллельная редукция (например, предложение «Мама подарила сыну котенка» имеет структуру (Мама $\xrightarrow{\alpha}$ подарила $\xrightarrow{\beta}$ сыну) + (Мама $\xrightarrow{\alpha}$ подарила $\xrightarrow{\beta}$ котенка)).

В работе [5] показано, какими элементарными преобразованиями можно конвертировать управляющее пространство произвольного предложения как в дерево подчинения, так и в дерево вывода, т.е. структура управляющего пространства одновременно обобщает деревья подчинения и деревья вывода. Управляющие пространства могут отобразить синтаксическую структуру любой сложности и арности в виде набора бинарных и тернарных отношений, что легко позволяет точно описать все данные о семантико-синтаксических связях внутри предложения в одной матрице D и одном трехмерном тензоре F.

ПОСТРОЕНИЕ ЛЕКСИКО-СИНТАКСИЧЕСКОЙ МОДЕЛИ ЕСТЕСТВЕННОГО ЯЗЫКА

Для построения семантико-синтаксической модели естественного языка разработана система автоматического заполнения трехмерного тензора F и матрицы D в процессе синтаксического анализа и постобработки синтаксических структур предложений большого текстового корпуса. Система выполняет следующую последовательность действий:

— принимаются на вход предложения из большого текстового корпуса и выполняется их синтаксический анализ с помощью модуля грамматического разбора Stanford Parser, который генерирует синтаксические структуры предложений в виде деревьев подчинения и деревьев вывода [9, 10];

— анализируются дерево подчинения и дерево вывода текущего предложения с построением управляющего пространства его синтаксической структуры и с разбором связей между словами для выделения предикативных сочетаний длины $l = 3$ (подлежащее–сказуемое–дополнение), а также синтагматических сочетаний длины $l = 2$ (существительное–прилагательное, глагол–наречие и т.п.);

— после генерации управляющего пространства синтаксической структуры текущего предложения для каждой тройки точек (i, j, k), связанных линейной предикативной последовательностью α–β-связей, в тензоре F в ячейке $F[I, J, K]$ значение увеличивается на единицу: $F[I, J, K] = F[I, J, K] + 1$, где I, J, K — координаты элемента тензора, соответствующие парам (w_i, A_i), (w_j, A_j) и (w_k, A_k), при этом w_i, w_j, w_k — слова, которые являются лексическими значе-

ниями соответствующих точек (i, j, k), и A_i, A_j, A_k — закодированное описание грамматических значений этих лексем (часть речи, род, число, падеж и т.д.);

— аналогично в управляющем пространстве синтаксической структуры текущего предложения для каждой пары точек (i, j), связанных между собой кольцевой синтагматической α–β-связью, в матрице D в ячейке $D[I, J]$ значение увеличивается на единицу: $D[I, J] = D[I, J]+1$, где I, J — координаты, соответствующие парам (w_i, A_i) и (w_j, A_j), при этом w_i, w_j — слова, которые являются лексическими значениями соответствующих точек (i, j), и A_i, A_j — закодированное описание грамматических значений этих лексем.

После обработки большого объема текстов в матрице D и в трехмерном тензоре F накапливается достаточный объем данных о семантико-синтаксических коммуникативных свойствах множества слов для эффективной реализации лексико-синтаксической модели естественного языка.

Отметим, что полученные массивы F и D содержат синтаксическую составляющую описания языка: A_i представляют части речи и их грамматические значения (род, число, падеж и т.д.). Таким образом, описывается, какие последовательности частей речи и в какой форме могут образовывать определенный тип связи. Также массивы содержат и семантические ограничения на то, какие лексемы могут объединяться и вступать в отношения определенного типа.

Сверхбольшой размер и разреженность матрицы D и тензора F требуют трансформации структур данных в целях более экономного и удобного представления для хранения и обработки. Для оптимизации полученных огромных массивов данных лучше всего подходят методы неотрицательной матричной и тензорной факторизации.

ФАКТОРИЗАЦИЯ МАТРИЦЫ *D*

Для разложения матрицы D большого размера $(N \times M)$ в виде произведения двух матриц: $W(N \times k) \times H(k \times M)$, где $k << N, M$, целесообразно использовать алгоритм неотрицательной матричной факторизации NMF, предложенный Ли и Суном [11]. В целевой функции используется норма Фробениуса

$$\min_{W,H} || D - WH ||_F^2,$$

при этом элементы матриц W и H должны быть неотрицательными.

Для такой целевой функции и для начальных матриц W_0 и H_0 алгоритм NMF представляет собой итерационное выполнение двух шагов.

Шаг 1. Вычисляем

$$(H_k)_{i,j} = (H_{k-1})_{i,j} \times \frac{(W_{k-1}^{\mathrm{T}} D)_{i,j}}{(W_{k-1}^{\mathrm{T}} W_{k-1} H_{k-1})_{i,j}}.$$

Шаг 2. Вычисляем

$$(W_k)_{i,j} = (W_{k-1})_{i,j} \times \frac{(D H_{k-1}^{\mathrm{T}})_{i,j}}{(W_{k-1} H_{k-1} H_{k-1}^{\mathrm{T}})_{i,j}}.$$

На практике шаги алгоритма повторяются, пока не будет достигнута неподвижная точка или выполнено максимальное число итераций. Ли и Сун доказали два основных свойства этого метода: во-первых, целевая функция является монотонно убывающей при выполнении шагов алгоритма; во-вторых, матрицы W и H становятся константными только в случае достижения стационарной точки целевой функции.

ФАКТОРИЗАЦИЯ ТЕНЗОРА *F*

Для разложения тензора F используется неотрицательная тензорная факторизация [12]. Она является подобием параллельного факторного анализа с ограниче-

нием, что все данные должны быть неотрицательными. Параллельный факторный анализ — это мультилинейный аналог сингулярного разложения матриц, используемого в латентном семантическом анализе. Главная идея метода — минимизация суммы квадратов разностей между оригинальным тензором и его факторизированной моделью. Для трехмерного тензора $T \in R^{D_1 \times D_2 \times D_3}$ определяется целевая функция

$$\min_{x_i \in R^{D_1}, y_i \in R^{D_2}, z_i \in R^{D_3}} \left\| T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \right\|_F^2,$$

где k — размерность факторизированной модели, а $\circ$ — внешнее произведение (outer product).

Для неотрицательной факторизации добавляются ограничения неотрицательности значений элементов

$$\min_{x_i \in R_{\geq 0}^{D_1}, y_i \in R_{\geq 0}^{D_2}, z_i \in R_{\geq 0}^{D_3}} \left\| T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \right\|_F^2.$$

Результат работы алгоритма — представление тензора в виде трех матриц: X, Y и Z, описывающих отображение каждой размерности тензора на k фактор-измерений латентного семантического пространства. Данная NTF-модель подгоняется методом наименьших квадратов. На каждой итерации две из размерностей фиксируются, а третья подгоняется методом наименьших квадратов. Процесс продолжается до момента сходимости.

СВОЙСТВА ЛЕКСИКО-СИНТАКСИЧЕСКОЙ МОДЕЛИ ЕСТЕСТВЕННОГО ЯЗЫКА

С помощью факторизации матрицы D и тензора F система формирует мощную базу, которая содержит данные о строении синтаксических структур предложений, в которые интегрировано описание лексико-семантических отношений между словами. В отличие от грамматики, которая задает структуру предложения в общем абстрактном виде, база содержит лексико-семантические ограничения, определяющие, какие слова могут образовывать связь некоторого синтаксического типа. Для того чтобы определить, могут ли два слова: a и b, образовать кольцевую синтагматическую связь, необходимо выбрать из матрицы W вектор-строку W_a, соответствующую слову a, из матрицы H — вектор-столбик H_b, соответствующий слову b, и вычислить скалярное произведение (W_a, H_b^{T}). Если значение произведения превышает некоторый пороговый уровень, то данная связь является определенной. Для того чтобы решить, могут ли три слова: a, b и c, сформировать предикативную связь $(a \to b \to c)$, нужно из первой матрицы X разложенного тензора F выбрать вектор X_a, соответствующий слову a, из второй матрицы Y разложенного тензора F — вектор Y_b, соответствующий слову b, а из третьей матрицы Z разложенного тензора F — вектор Z_c, соответствующий слову c, и вычислить значение $S_{abc} = \sum_{i=1}^{k} X_a[i] * Y_b[i] * Z_c[i]$.

Если значение S_{abc} превышает некоторый пороговый уровень, то данная связь определена. Все связи, которые не являются определенными, считаются неопределенными.

Полученные матрицы в неявном виде задают множество определенных предложений языка, которое изначально задается текстами входного корпуса. Векторы слов из полученных матриц являются неявным описанием их «структурного поведения» — они определяют, в какие синтаксические отношения эти слова имеют свойство вступать и с какими словами они формируют эти отношения. Данные векторы назовем векторами семантико-синтаксической валентности

слов. С помощью векторов валентности можно выполнять синтаксический анализ предложений с построением управляющего пространства их синтаксических структур, используя восходящие алгоритмы синтаксического анализа, например Кока–Янгера–Касами [13].

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МОДЕЛИ

В качестве учебного текстового корпуса используются статьи English Wikipedia, Simple English Wikipedia, а также тексты корпуса The Wall Street Journal. Тексты последовательно обрабатываются парсером и блоком построения управляющих пространств синтаксических структур. Вначале предложение анализируется Стэнфордским парсером, в результате получается дерево вывода предложения (parse tree) и дерево подчинения (dependency tree). Для построения управляющих пространств предложений разработан алгоритм конвертации дерева подчинения и дерева вывода предложения в управляющее пространство синтаксической структуры предложения [14]. Алгоритм представляет собой рекурсивный обход дерева вывода предложения слева направо с порождением точек управляющего пространства в каждом узле этого дерева и с конвертацией соответствующих этим узлам связей из дерева подчинения в α–β-связи управляющего пространства — предикативные или синтагматические. За каждой точкой пространства закрепляется определенное лексико-семантическое значение (слово или словосочетания и признаки: часть речи, род, число, падеж и т.д.).

В начале работы алгоритма каждое слово — это несвязанная точка управляющего пространства. Когда точки A и B пространства соединяются и образуют новую точку C, представляющую α–β-связь между A и B, она получает собственное лексико-семантическое значение. Последнее может наследоваться от главного элемента пары (A, B). Например, в словосочетании «красный мяч» в паре (красный, мяч) главным является имя существительное, поэтому новая образованная точка наследует значение «мяч». Бывает, что в результате объединения двух точек их лексические значения образуют устойчивое словосочетание, которое можно найти в специальной базе — списке названий статей Википедии. Например, если объединяются точка A со значением «Тезис» и точка B со значением «Черча», образуется устойчивое словосочетание «Тезис Черча», которое становится лексико-семантическим значением новой образованной точки C.

После построения управляющего пространства синтаксической структуры предложения для всех кольцевых синтагматических α–β-связей в матрице кольцевых связей D наращиваются значения $d[I, J]$, где I и J — индексы первых и вторых слов соответственно, $d[I, J] = d[I, J]+1$. Для всех троек линейных предикативных связей $A \xrightarrow{\alpha} B \xrightarrow{\beta} C$ в трехмерном тензоре линейно-предикативных связей F наращиваются значения $f[I, J, K]$, где I, J и K — индексы слов A, B и C соответственно, $f[I, J, K] = f[I, J, K]+1$.

Обрабатывались 800 000 статей English Wikipedia и Simple English Wikipedia, а также корпус статей The Wall Street Journal. За счет того, что данный корпус размечен вручную и содержит корректные синтаксические структуры предложений, которые напрямую конвертировались в управляющие пространства, для учебной выборки было получено большое количество управляющих пространств синтаксических структур высокого качества (корректных почти на 100 %).

Таким образом, сгенерирована большая разреженная матрица кольцевых связей D размера $\approx 2{,}3$ млн слов $\times$ 2,3 млн слов с приблизительно 57 млн ненулевых элементов, а также большой трехмерный тензор линейно-предикативных связей F размера $\approx 2{,}3$ млн слов $\times$ 152 тыс. слов $\times$ 2,3 млн слов с приблизительно 78 млн ненулевых элементов. Данные массивы факторизованы с помощью алгоритма неотрицательной матричной факторизации Ли и Сунга и алгоритма параллельной факторизации трехмерного тензора PARAFAC [12]. Алгоритмы факторизации реализованы с применением методов параллельных вычислений на графических картах [15, 16].

Факторизованные массивы данных позволяют элементарно вычислять значения вероятности образования кольцевых синтагматических связей между двумя любыми словами с помощью скалярного произведения соответствующих им векторов. Аналогично можно вычислять значения вероятности образования линейных предикативных связей между тремя любыми словами.

На основе полученных массивов лексико-синтаксической сочетаемости реализован синтаксический анализатор, который разбирает предложения на английском языке и напрямую строит управляющие пространства их синтаксической структуры. В качестве базового метода для синтаксического анализатора использован алгоритм Кока–Янгера–Касами.

Предложенная модель содержит описание только тех связей между словами, которые фактически имели место в соответствующих предложениях обучающих текстов. Если для пары слов: A и B, кольцевая синтагматическая связь прописана, так как в обучающих текстах она имеется, то для пары: A_1 и B_1 (где A_1 — синоним A и B_1 — синоним B), такой связи может и не быть. Для тройки слов: A, B и C, которые связаны линейно-предикативной связью, это утверждение также имеет место. С применением словарей синонимов эта проблема достаточно легко решается. В разработанной системе в качестве такого словаря используется WordNet и его синсеты. Система предполагает, что если связь между A и B имеется, то она может существовать также между любой парой: A_i и B_i, где A_i и B_i — произвольные слова из синсетов, содержащих A и B соответственно. Здесь возникает проблема омонимии, когда одному слову в WordNet соответствует несколько синсетов, — каким образом определить пару или тройку корректных синсетов в каждом конкретном случае при синтаксическом анализе предложения.

Существует несколько подходов к решению этой классической проблемы неоднозначности слов (WSD). Наиболее подходящими в данном случае могут оказаться методы, которые разрабатывались специально для интеграции страниц Wikipedia в качестве новых узлов в WordNet [17–20].

Матрицы W и H, полученные в результате неотрицательной факторизации матрицы D, являются мощным инструментом для определения степени семантической близости между словами по методологии латентного семантического анализа.

Для решения проблемы неоднозначности слов в модели разработан следующий алгоритм.

Для определения наличия кольцевой синтагматической α–β-связи между словами a и b выполнить:

Шаг 1. Для того чтобы определить, могут ли слова a и b образовать кольцевую синтагматическую связь, нужно выбрать из матрицы W вектор-строку W_a, соответствующую слову a, из матрицы H — вектор-столбец H_b, соответствующий слову b, и вычислить скалярное произведение векторов (W_a, H_b^{T}). Если значение $(W_a, H_b^{\mathrm{T}}) > T$, где T — пороговый уровень, оптимальное значение которого определено экспериментально, то данная α–β-связь определена, в противном случае переходим к шагу 2.

Шаг 2. По словам a и b переходим к их синсетам в лексико-семантической базе WordNet. Получим наборы синсетов-узлов $\{A_i\}$ и $\{B_i\}$, на которые ссылаются соответственно слова a и b. Проверим попарно $\{A_i\}$ и $\{B_i\}$, существуют ли такие k и j, что в синсетах A_k и B_j содержатся соответственно слова a'_k и b'_j, для которых скалярное произведение векторов $(W_{a'_k}, H_{b'_j}^{\mathrm{T}}) > T$. Если такие k и j найдены, то данная связь между a и b определена, в противном случае переходим к шагу 3.

Шаг 3. Множество $\{A_i\}$ расширяется синсетами, соединяющимися с узлами $\{A_i\}$ отношениями гипонимии и гипернимии, аналогично расширяется множество $\{B_i\}$. После этого проверяется, существует ли для расширенных $\{A_i\}_{ext}$ и $\{B_i\}_{ext}$ такие значения k и j, что в A_k и B_j содержатся соответственно слова a'_k и b'_j, для кото-

Таблица 1

Шаг алгоритма	Оценки точности определения кольцевых синтагматических α–β-связей в корпусах текстов статей (%)		
	Simple Wikipedia	Wikipedia	WSJ corpus
1	95,17	91,23	93,71
2	91,29	89,91	91,05
3	89,17	83,06	85,07

Таблица 2

Шаг алгоритма	Оценки точности определения линейных предикативных α–β-связей в корпусах текстов статей (%)		
	Simple Wikipedia	Wikipedia	WSJ corpus
1	96,17	92,24	94,37
2	93,21	90,01	91,33
3	91,03	87,79	89,79

рых скалярное произведение векторов $(W_{a_k'}, H_{b_j'}^{\mathrm{T}}) > T$. Проверка выполняется только для тех пар синсетов, которые до этого не проверялись. Если такие k и j найдены, то данная связь между a и b определена, в противном случае выполняем еще раз расширение множеств $\{A_i\}$ и $\{B_i\}$ и поиск таких A_k и B_j, для которых $(W_{a_k'}, H_{b_j'}^{\mathrm{T}}) > T$.

Если за две итерации выполнения шага 3 условие $(W_{a_k'}, H_{b_j'}^{\mathrm{T}}) > T$ не выполняется, данной связи не существует.

При расширении $\{A_i\}$ и $\{B_i\}$ нужно избегать добавления синсетов из списка концептов наиболее общих значений из верхней части иерархии WordNet. При задействовании подобных синсетов быстро утрачивается смысловая близость между синсетами при наследовании свойств и отношений по связям гипонимии/гипернимии.

Для линейных предикативных α–β-связей данный алгоритм работает аналогично.

Таксономическая иерархия лексико-семантической базы WordNet и механизм неявного наследования, реализованный шагом 3 приведенного алгоритма, дают возможность обобщить описанную модель представления синтаксических связей и лексико-семантических отношений. Это делает построенную систему универсальным средством анализа синтаксиса и семантики естественного языка. В табл. 1 и 2 представлены оценки точности работы алгоритма анализа и построения управляющих пространств синтаксических структур предложений на английском языке, полученные в результате проведения экспериментов [21], которые подтверждают эффективность и адекватность предложенной рекурсивной тензорной модели семантико-синтаксического пространства слов естественного языка.

Таким образом, 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. Применен тензорный подход к моделированию семантико-синтаксических связей между словами в предложениях. Использование аппарата управляющих пространств синтаксических структур естественного языка позволило усовершенствовать тензорную семантико-синтаксическую модель. С помощью рекурсии и суперпозиции разра-

ботанная тензорная модель способна описывать синтаксические структуры произвольной длины и сложности.

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 — 2012. — P. 2703–2720.
3. Cohen S.B., Collins M. Tensor decomposition for fast parsing with latent-variable PCFGs // NIPS. — 2012. — P. 2528–2536.
4. Peng W., Li T. On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis // Applied Intelligence. — 2011. — **35**, N 2. — P. 285–295.
5. Anisimov A.V. Control space of syntactic structures of natural language // Cybernetics. — 1990. — N 3. — P. 11–17.
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. Нариньяни А.С. Формальная модель: общая схема и выбор адекватных средств. — Препринт № 400/ВЦ СО АН СССР. — Новосибирск, 1978. — 19 с.
8. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. — М.: Наука, 1985. — 144 с.
9. Klein D., Manning C.D. Accurate unlexicalized parsing // Proceedings of ACL-2003. — 2003. — P. 423–430.
10. De Marneffe M.C., MacCartney B.,Manning C.D. Generating typed dependency parses from phrase structure parses // Proceedings of LREC-2006. — 2006. — P. 449–454.
11. Lee D.D., Seung H.S. Algorithms for non-negative matrix factorization // Proceedings of NIPS-2000. — 2000. — P. 556–562.
12. Cichocki A., Zdunek R., Phan A.H., Amari S.I. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. — Chichester: J. Wiley & Sons, 2009. — 500 p.
13. Younger D.H. Recognition and parsing of context-free languages in time n3 // Information and Control. — 1967. — **10**, N 2. — P. 189–208.
14. Марченко О.О. Алгоритм конвертації дерева залежностей у керуючий простір синтаксичної структури речення // Вісник Київського національного університету імені Тараса Шевченка. Серія: фізико-математичні науки. — 2013. — № 4. — С. 146–151.
15. Antikainen J., Havel J., Josth R., Herout A., Zemcík P., Hauta-Kasari M. Nonnegative tensor factorization accelerated using GPGPU // IEEE Transactions on Parallel and Distributed Systems. — 2011. — **22**, N 7. — P. 1135–1141.
16. Kysenko V., Rupp K., Marchenko O., Selberherr S., Anisimov A. GPU-accelerated non-negative matrix factorization for text mining // Lecture Notes in Computer Science. — 2012. — **7337.** — P. 158–163.
17. Ponzetto S. P., Navigli R. Knowledge-rich word sense disambiguation rivaling supervised systems // Proceedings of ACL-2010. — 2010. — P. 1522–1531.
18. Ponzetto S.P., Navigli R. Large-scale taxonomy mapping for restructuring and integrating Wikipedia // Proceedings of IJCAI-2009. — 2009. — P. 2083–2088.
19. Ponzetto S. P., Navigli R. BabelNet: building a very large multilingual semantic network // Proceedings of ACL-2010. — 2010. — P. 216–225.
20. Ruiz-Casado M., Alfonseca E., Castells P. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets // Advances in Web Intelligence. — Heidelberg; Berlin: Springer, 2005. — P. 380–386.
21. Anisimov A., Marchenko O., Taranukha V., Vozniuk T. Semantic and syntactic model of natural language based on tensor factorization // Proceedings of NLDB-2014. Lecture Notes in Computer Science. — 2014. — **8455.** — P. 51–54.

Поступила 16.07.2015