Труды Карельского научного центра РАН
№ 8. 2016. С. 45–54
DOI: 10.17076/mat394

УДК 81.32

МОДЕЛЬ ГЕОМЕТРИЧЕСКОЙ СТРУКТУРЫ СИНСЕТА

А. Н. Кириллов, А. А. Крижановский

Институт прикладных математических исследований Карельского научного центра РАН

В статье поставлен вопрос формализации понятия синонимии. На основе векторного представления слов в работе предлагается геометрический подход для математического моделирования наборов синонимов (синсетов). Определен такой вычислимый атрибут синсетов как *внутренность синсета* (IntS). Введены понятия *ранг* и *центральность* слов в синсете, позволяющие определить более значимые, «центральные» слова в синсете. Для ранга и центральности даны математическая формулировка и предложена процедура их вычисления. Для вычислений использованы нейронные модели (Skip-gram, CBOW), созданные программой Т. Миколова word2vec. На примере синсетов Русского Викисловаря построены IntS по нейронным моделям корпусов проекта RusVectores. Результаты, полученные по двум корпусам (Национальный корпус русского языка и новостной корпус), в значительной степени совпадают. Это говорит о некоторой универсальности предлагаемой математической модели.

Ключевые слова: синоним; синсет; нейронная сеть; корпусная лингвистика; word2vec; RusVectores; gensim; Русский Викисловарь.

A. N. Kirillov, A. A. Krizhanovsky. SYNSET GEOMETRY STRUCTURE MODEL

The goal of formalization, proposed in this paper, is to bring together, as near as possible, the theoretic linguistic problem of synonym conception and the computer linguistic methods based generally on empirical intuitive unjustified factors. Using the word vector representation we have proposed the geometric approach to mathematical modeling of synonym set (synset). The word embedding is based on the neural networks (Skip-gram, CBOW), developed and realized as word2vec program by T. Mikolov. The standard cosine similarity is used as the distance between word-vectors. Several geometric characteristics of the synset words are introduced: the interior of synset, the synset word rank and centrality. These notions are intended to select the most significant synset words, i.e. the words which senses are the nearest to the sense of a synset. Some experiments with proposed notions, based on RusVectores resources, are represented.

Keywords: synonym; synset; neural network; corpus linguistics; word2vec; RusVectores; gensim; Russian Wiktionary.

Введение

Понятие синонима не имеет строгого определения, хотя на бытовом уровне оно прижилось и достаточно часто используется. Приведем описательное определение синонима из известного словаря синонимов русского языка Александровой З. Е. [1, с. 6]:

> Синонимами считаются слова, выражающие одно и то же понятие, тождественные или близкие по значению, отличающиеся друг от друга оттенками значений, принадлежностью к тому или иному стилистическому слою языка и экспрессивной окраской.

Это определение вызывает ряд вопросов: что такое понятие, значение и т. д.? В результате нет единого строгого определения синонимии. Имеются многочисленные научные работы, отражающие различные подходы в его понимании.

Таким образом, возникает необходимость введения некоторой формализации, которая позволила бы дать количественные характеристики для описания соотношений между словами, что особенно важно в задачах автоматической обработки языка (англ. *natural language processing*).

В настоящей работе предложен подход к математическому моделированию понятия синсета.

Понятие *синсет* (набор синонимов) обязано своим появление системе WordNet, в котором различные отношения (синонимия, антонимия и др.) указываются не между словами, а между *синсетами* (от англ. ***syn****onym* ***set***, группа синонимов) [15].

Для исследования были использованы синонимы Русского Викисловаря. Викисловарь — это свободно пополняемый многофункциональный многоязычный онлайн-словарь и тезаурус. Машиночитаемый Викисловарь, используемый в этой работе, регулярно обновляется и строится с помощью программы wikokit[1] на основе данных Викисловаря [7].

Авторы статьи ставят перед собой ряд задач, решение которых в большей или меньшей степени представлено в этой работе:

- автоматически упорядочивать синонимы внутри синсета по степени близости слов к тому смыслу, который представлен этим синсетом;
- предложить математический аппарат для анализа, характеристики и сравнения синсетов, проверить его экспериментально на данных онлайн-словаря (Русский Викисловарь);
- в перспективе с помощью предлагаемого математического аппарата найти «слабые» синсеты с целью повышения качества словаря;
- важное направление, занятие которым побудило авторов к этой работе, это разрешение лексической многозначности (word-sense disambiguation или WSD). Программа максимум заключается в том, чтобы использовать нейронные сети и предлагаемые методы для решения WSD-задачи на качественно новом уровне по сравнению с текущими методами [3].

Векторное представление слов: блеск и нищета построения нейронных сетей инструментом word2vec

Идея векторного представления слов с помощью нейронных сетей получила мощный толчок благодаря работам Томаса Миколова [12–14]. Главное достоинство работы Т. Миколова в том, что он разработал инструмент *word2vec* для создания моделей нейронных сетей (далее будем их называть *предсказательными моделями*, см. context-predicting models в работе [4]) на основе текстов корпусов. Забегая вперед, можно сказать, что, с нашей точки зрения, не меньший вклад сделали и отечественные ученые Андрей Кутузов и Елизавета Кузьменко, которые приготовили с помощью word2vec предсказательные модели для русского языка на основе ряда корпусов. Свой инструмент они назвали RusVectores [9].

Бедность подхода, предложенного Т. Миколовым в том, что поиск осмысленных пар семантических отношений работает только на некоторых ярких примерах, например ($queen - woman + man \approx king$). У нас есть обоснованные подозрения, что не на всем пространстве текстов слова будут подчиняться таким удивительно простым правилам. Слабость математической стороны работ Т. Миколова была подмечена в недавней работе Голдберга и Леви [5].

Работа И. Голдберга и О. Леви, посвященная обсуждению результатов Т. Миколова, заканчивается обращением к исследователям:

> *"Can we make this intuition more precise? We'd really like to see something more formal"* [5].
>
> Перевод: «Может ли интуитивный подход быть сделан более точным? Мы действительно хотели бы увидеть нечто более формальное.»

В какой-то мере настоящая статья является ответом на вызов этих известных исследователей в области компьютерной лингвистики.

Кратко осветим подход Т. Миколова.

[1] `https://github.com/componavt/wikokit`

Определение 1. Векторным словарем назовем множество $D = \{w_i \in \mathbb{R}^{|D|}\}$, где i-ая компонента вектора w_i равна 1, а остальные компоненты – нули.

Рассмотрим некоторый словарь и пронумеруем все слова, входящие в него. Пусть $|D|$ — количество слов в словаре, i — номер слова.

Задача векторного представления слов состоит в построении линейного отображения $L : D \to \mathbb{R}^N$, где $N \ll |D|$, а вектор $v = L(w),\ w \in D$, v имеет компоненты $v_j \in \mathbb{R}$. Результат отображения называется распределенным (distributed) векторным представлением слов. Цель его состоит в замене очень «тощего» (разреженного) множества $D \in \mathbb{R}^{|D|}$, в которое входят векторы с нулевым взаимным скалярным произведением, на некоторое подмножество из $\mathbb{R}^N$, векторы которого расположены таким образом, что их компоненты позволяют использовать скалярное произведение нормированных векторов в качестве меры их похожести (similarity), что принято в соответствующих задачах обработки языков. Полагая, что линейное отображение L реализуется с помощью матрицы W, получаем $v = Ww$, причем для нахождении матрицы W используют различные методы, в частности, основанные на нейронных сетях. Наибольшую популярность в самое последнее время приобрели CBOW (continuous bag of words) и Skip-gram методы, предложенные в работе [14] и являющиеся, по сути, модификацией метода максимального правдоподобия. При этом в методе Skip-gram матрица W максимизирует функцию $F(W)$ вида

$$F(W) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0} \ln p(w_{t+j}|w_t)$$

$$p(w_{t+j}|w_t) = \frac{\exp u_{t+j}}{\sum_{i=1}^{|D|}\exp u_i}, \qquad u_i = (Ww_i, Ww_t)$$

где $(\cdot,\cdot)$ — символ скалярного произведения, T — объем обучающего контекста. Здесь по слову w_t находится содержащий его контекст, составляющий «окно» размера $2c$ слов. В методе CBOW, наоборот, по контексту находится слово, входящее в него. Для максимизации $F(W)$ используется метод стохастического градиентного спуска.

В работах Т. Миколова при построении нейронных сетей учитывается только локальный контекст слов (упомянутое выше «окно»). Существуют попытки [6] учесть глобальный контекст (весь документ). Это полезно при разрешении лексической многозначности.

Геометрия синсета

Внутренность синсета IntS

Расстояние между векторами-словами (нормированными) измеряется их скалярным произведением, или углом между векторами, как в теории проективных пространств. Таким образом, увеличение скалярного произведения соответствует уменьшению расстояния между векторами-словами a, b, которое принято обозначать как $sim\{a, b\}$, что является сокращением термина *similarity* – «похожесть» или «сходство» слов[2]. Итак, $sim\{a, b\} = \frac{(a,b)}{||a||||b||}$ — это расстояние между векторами a и b.

Предлагаются и другие способы определения расстояния между словами-векторами, но в их основе также лежит скалярное произведение [10, 11, 18].

Введем обозначения для нормированных сумм векторов: $M((a_i), n) = \frac{\sum_{i=1}^{n} a_i}{||\sum_{i=1}^{n} a_i||}$. Расстояние между множествами векторов будем понимать как расстояния между средними векторов этих сумм. Таким образом, если даны два множества векторов $A = \{a_1, ..., a_n\}$ и $B = \{b_1, ..., b_m\}$, то расстояние между ними, $sim\{A, B\}$, определяется следующим образом $sim\{A, B\} = (M((a_i), n), (M((b_j), m)))$.

Рассмотрим синсет $S = \{v_k, k = 1, ..., |S|\}$. Удалим какое-либо слово v из синсета. Индекс слова опускаем для сокращения записи. Разобъем множество $S \setminus \{v\}$ на два непересекающихся подмножества: $S \setminus \{v\} = \{v_{i_s}\} \sqcup \{v_{j_p}\}$, $s = 1, ..., q,\ p = 1, ..., r,\ q + r = |S| - 1,\ i_s \neq j_p$. Обозначим $S_1 = \{v_{i_s}\}, S_2 = \{v_{j_p}\}$. Тогда введенное выше дизъюнктное разбиение запишется в виде $S \setminus \{v\} = S_1 \cup S_2$.

Определение 2. Внутренностью $IntS$ синсета S называется множество всех векторов $v \in S$, удовлетворяющих условию

$$IntS = \{v \in S : sim\{S_1, S_2\} < sim\{S_1 \cup v, S_2\} \bigwedge sim\{S_1, S_2\} < sim\{S_1, S_2 \cup v\}\} \qquad (1)$$

для всех дизъюнктных разбиений $S \setminus \{v\} = S_1 \cup S_2$, где $S_1 \neq \varnothing,\ S_2 \neq \varnothing$.

[2]Будем использовать фигурные скобки $sim\{a, b\}$, чтобы отличать запись от скалярного произведения $(\cdot,\cdot)$.

Смысл определения состоит в том, что добавление вектора $v \in IntS$ в любое из двух подмножеств множества $S \setminus \{v\}$, образующих его дизъюнктное разбиение, уменьшает расстояние между этими подмножествами.

Чтобы проиллюстрировать IntS и показать, какие слова в него входят, предположим, что вектора имеют размерность не 100 или 300, а всего два. На рисунке 1 представлена такая конфигурация синсета S, что вершина v не может не входить в $IntS$. То есть любые разбиения S будут «стягиваться», сближаться добавлением v к одному из разбиений (S_1 или S_2).

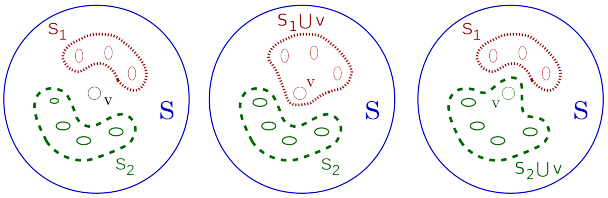

Рис. 1. Пример вершины v, сближающей любые непустые разбиения S, в частности — S_1 и S_2 (слева), а следовательно, $v \in IntS$. При добавлении вершины v к S_1 получаем множество $S_1 \cup v$, которое на рисунке (в центре) находится ближе к S_2, чем множество S_1. Подобным образом $S_2 \cup v$ ближе к S_1, чем множество S_2 (справа).

Ранг и центральность слов в синсете

Введём понятие **ранга синонима** $v \in S$. Дизъюнктное разбиение на два множества, элемента разбиения, будем называть разбиением. Пусть $P_v = \{p_i, i = 1, ..., 2^{n-2} - 1\}$ — множество всех пронумерованных каким-либо образом разбиений $(n-1)$-элементного множества $S \setminus \{v\}$, $n > 2$.

Рассмотрим какое-либо разбиение p_i множества $S \setminus \{v\}$ на подмножества S_1 и S_2, то есть $S \setminus \{v\} = S_1 \sqcup S_2$. Обозначим $sim_i = sim\{S_1, S_2\}$, $sim_i^1 = sim\{S_1 \cup v, S_2\}$, $sim_i^2 = sim\{S_1, S_2 \cup v\}$. При этом получаем более компактное определение внутренности $IntS$ синсета S

$$IntS = \{v \in S : sim_i < sim_i^1 \bigwedge sim_i < sim_i^2\} \quad (2)$$

Введем функцию $r_v : P_v \to \{-1, 0, 1\}$ следующего вида:

$$r_v(p_i) = \begin{cases} -1, & sim_i^1 < sim_i \bigwedge sim_i^2 < sim_i, \\ & v \text{ отдаляет } S_1 \text{ от } S_2 \\ 1, & sim_i^1 > sim_i \bigwedge sim_i^2 > sim_i, \\ & v \text{ сближает } S_1 \text{ и } S_2 \\ 0, & (sim_i^1 - sim_i) \cdot (sim_i^2 - sim_i) < 0. \\ & \text{сближение} - \text{отдаление} \end{cases} \quad (3)$$

Функция r_v определена для каждого разбиения и дает своего рода «кирпичики», из которых будет складываться ранг синонима.

Поясним краткую запись «сближение–удаление». Выражение $(sim_i^1 - sim_i) \cdot (sim_i^2 - sim_i) < 0$ эквивалентно и является компактной записью для $(sim_i^1 < sim_i \wedge sim_i^2 > sim_i) \bigvee (sim_i^1 > sim_i \wedge sim_i^2 < sim_i)$. Другими словами функция $r_v(p_i)$ дает значение 0, если добавление слова v одному из элементов разбиения p_i уменьшает (увеличивает) расстояние sim_i, а добавление ко второму элементу, наоборот, увеличивает (уменьшает) расстояние sim_i. То есть элемент v действует на множества в "противофазе". На рисунке 2 это разбиения 2 и 3.

Определение 3. Рангом синонима $v \in S$, где $|S| > 2$, называется целое число вида

$$rank\ (v) = \sum_{i=1}^{|P_v|} r_v(p_i). \quad (4)$$

Легко видеть, что если $v \in IntS$, то $rank\ (v) = 2^{|S|-2} - 1$ – это число всех непустых дизъюнктных разбиений $(|S|-1)$-элементного множества $S \setminus \{v\}$, т. е. $rank\ (v)$ максимален и совпадает с числом Стирлинга второго рода: $\left\{ {n \atop k} \right\} = \left\{ {|S| \atop 2} \right\}$, где n — мощность разбиваемого множества, а k — число подмножеств, здесь два [2, с. 24].

Взаимосвязь $IntS$ и ранга синонима в синсете S сформулируем в виде теоремы.

Теорема 1 (IntS theorem)**.** *Слово v принадлежит внутренности синсета S тогда и только тогда, когда это слово обладает максимально возможным рангом в данном синсете, этот ранг совпадает с числом Стирлинга второго рода.*

$$v \in IntS \Leftrightarrow rank\ (v) = 2^{|S|-2} - 1, \quad \text{где } |S| \geqslant 3,$$

При этом внутренность синсета $IntS$ определена для синсетов, содержащих три и более слов, поскольку для вычисления $IntS$ множество S нужно разбить на три части: $S \setminus \{v\} = S_1 \sqcup S_2$.

Доказательство.

$$v \in IntS \overset{(2)}{\Leftrightarrow} \forall p_i : IntS = \{v \in S : sim_i^1 > sim_i \wedge\ sim_i^2 > sim_i\}\ \ (v \text{ сближает } S_1 \text{ и } S_2) \overset{(3)}{\Leftrightarrow}$$

$$\forall p_i : r_v(p_i) = 1 \overset{(4)}{\Leftrightarrow}$$

$$rank\ (v) = \sum_{i=1}^{|P_v|} 1 = |P_v| = 2^{|S|-2} - 1, \quad (5)$$

поскольку $2^{|S|-2} - 1$ — это максимально возможное число непустых дизъюнктных разбиений, совпадающее с числом Стирлинга второго рода [2, с. 24]. □

Обратим внимание, что слова в $IntS$ имеют бóльший ранг и значение центральности относительно других слов синсета S.

Определение 4. Центральностью синонима $v \in S$ при разбиении p_i множества $S \setminus \{v\}$ называется величина

$$centrality(v, p_i) = (sim_i^1(v) - sim_i) + (sim_i^2(v) - sim_i) \quad (6)$$

Определение 5. Центральностью синонима $v \in S$ называется величина

$$centrality(v) = \sum_{i=1}^{|P_v|} centrality(v, p_i)$$

По-видимому, ранг и центральность указывают на значимость слова внутри синсета, то есть близость слова к тому значению, которое выражает синсет совокупностью слов.

Центральность дает более точную характеристику значимости слова v в синсете, чем ранг (см. таблицу 1). Это естественно следует из того, что ранг является целым ($\mathbb{Z}$), а степень центральности — вещественным числом ($\mathbb{R}$), при этом вычисляются они по одному и тому же алгоритму (см. далее).

Алгоритм вычисления ранга и центральности

Из определения центральности (см. выше) следует процедура её вычисления (алгоритмы 1 и 2)

Algorithm 1: Вычисление ранга и центральности вершины v для разбиения p_i синсета S

Data: разбиение p_i множества $S \setminus \{v\}$ на подмножества S_1 и S_2, то есть $S \setminus \{v\} = S_1 \sqcup S_2$.
Result: $rank(v, p_i)$, $centrality(v, p_i)$.

1. $sim_i = sim\{S_1, S_2\}$,
2. $sim_i^1(v) = sim\{S_1 \cup v, S_2\}$ // слово v добавляется к первому подмножеству S_1
3. $sim_i^2(v) = sim\{S_1, S_2 \cup v\}$ // слово v добавляется ко второму подмножеству S_2
4. $centrality(v, p_i) = (sim_i^1(v) - sim_i) + (sim_i^2(v) - sim_i)$
5. $rank(v, p_i) = \text{sgn}(sim_i^1(v) - sim_i) + \text{sgn}(sim_i^2(v) - sim_i)$,

   где $\text{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$

Algorithm 2: Вычисление ранга и центральности вершины v синсета S

Data: синсет S, вершина v.
Result: $rank(v)$, $centrality(v)$.

1. $centrality(v) = \sum_{i=1}^{|P_v|} centrality(v, p_i)$,
2. $rank(v) = \sum_{i=1}^{|P_v|} rank(v, p_i)$.

Гипотеза. Чем более многозначным является слово, тем меньше ранг ($\mathbb{Z}$) и степень центральности ($\mathbb{R}$) этого слова в разных синсетах.

Пример. Дан синсет S = *(баюкать, убаюкивать, укачивать, усыплять)*. Нужно найти $IntS$, вычислить ранг и центральность для каждого слова в синсете.

Пример вычисления ранга и степени центральности для слова «усыплять» в этом синсете показан на рисунке 2. Множество мощности $3 = |S \setminus \{v\}|$ можно разбить тремя способами на два непустых подмножества. Каждое такое разбиение добавляет в $rank(v)$ 1, 0 или -1 (рис. 2). Значение ранга получилось равным -1, степень центральности равна -0,071.

В таблице 1 указаны значения ранга, степени центральности и принадлежность $IntS$ для всех слов синсета.

В соответствии с изложенной выше Теоремой 1 ранг синонимов, принадлежащих внутренности синсета $IntS$, должен быть равен

$$2^{|S|-2} - 1 = 2^{|4|-2} - 1 = 3$$

В таблице 1 видно, что ранг 3 и наибольшие значения центральности у слов «баюкать», «убаюкивать». Итак, Int (баюкать, убаюкивать, укачивать, усыплять) = (баюкать, убаюкивать), то есть в $IntS$ вошли векторы, соответствующие словам «убаюкивать» и «баюкать». Это указывает на то, что эта пара наиболее близка по смыслу ко всем четырем словам синсета.

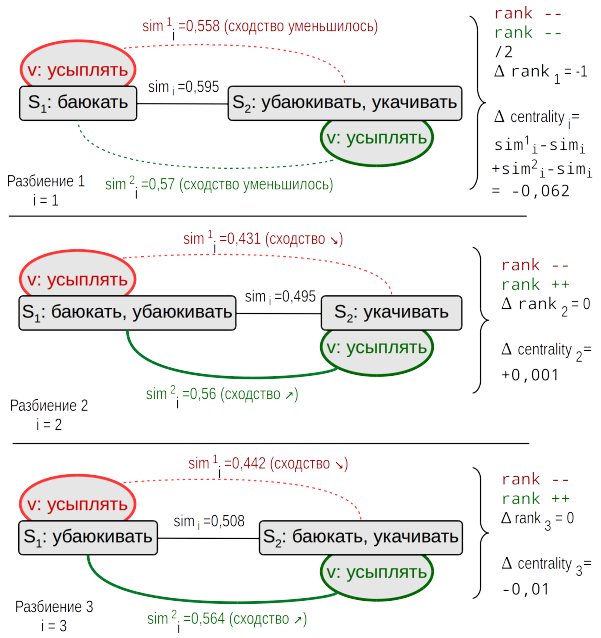

Рис. 2. Значение ранга и степени центральности для слова «усыплять» в синсете (баюкать, убаюкивать, укачивать, усыплять). Представлены три возможных разбиения множества (баюкать, убаюкивать, укачивать) на два непустых подмножества $S_1^i, S_2^i, i = 1, 2, 3$ без слова "усыплять"(вектор v). Значения $rank\ v$ и $centrality\ v$ вычисляются как сумма соответствующих $\Delta rank_i$ и $\Delta centrality_i$.

Эксперименты

В этой работе используются нейронные модели, созданные авторами проекта $RusVectores$ [9]. Первая модель построена по текстам Национального корпуса русского языка (НКРЯ или Ruscorpora), вторая модель — на основе текстов отечественных новостных сайтов (Новостной корпус или News corpus). Модели доступны на сайте проекта [16].

Авторы $RusVectores$ А. Кутузов и Е. Кузьменко обращают внимание читателя на такие особенности НКРЯ, как ручной отбор текстов для пополнения корпуса и регулирование соотношения объема текстов разных жанров, малый размер основного корпуса, порядка 107 млн слов (для сравнения Новостной корпус включает 2.4 млрд слов). В работе [8] вводится понятие *представительность корпуса* как способность отражать (указывать на) те ассоциации для слова, с которыми согласится большинство носителей. Ассоциации, порождаемые предсказательными моделями по данным НКРЯ и по данным веб-корпуса, как раз и используются для сравнения двух корпусов в этой работе. Задача сравнения свелась к поиску слов, значения которых в веб-корпусе существенно (или полностью) отличались бы от значений в НКРЯ. Если учесть, что для каждого слова в корпусе с помощью предсказательной модели можно получить список N ближайших слов (напомним, что слову соответствует вектор), то формулировка результата сравнения корпусов будет такой: более чем у половины слов (общих слов двух корпусов) совпадало три и более слов из 10 ближайших [8]. Это говорит о том, что в картине мира интеллектов, нейронных моделей, созданных на основе НКРЯ и на основе текстов Интернета, есть много общего. Однако необходима и обратная оценка — какова *степень различия* предсказательных моделей?

Отметим, что понятие *сбалансированность корпуса* приобретает новое значение в свете предсказательных моделей, создаваемых на основе корпуса. Несбалансированная выборка текстов приводит к перевесу в тематике корпусов, в итоге — к менее точной предсказательной модели.

Таблица 1. Ранг (rank) и степень центральности (centrality) для каждого слова в синсете, принадлежность синонима внутренности синсета (IntS)

synonym	усыплять	укачивать	убаюкивать	баюкать
centrality	-0.07	0.31	0.68	0.71
rank	-1	1	3	3
IntS	—	—	+	+

Таблица 2. Примеры синсетов, ряд которых имеют пустую внутренность ($IntS = \varnothing$). Синсеты взяты из словарных статей Русского Викисловаря, слова в синсете упорядочены по рангу и центральности. Указан корпус, по которому в проекте $RusVectōrēs$ построена предсказательная модель, использованная для вычислений $IntS$, здесь $OutS = S \setminus IntS$

словарная статья	синсет (из статьи), по умолчанию целиком входит в $OutS$	$\|\|S\|\|$	$\|\|IntS\|\|$	корпус
	существительные			
план	умысел, намерение, прожект, задумка, план, проект, замысел	7	0	НКРЯ
хвороба	нездоровье, хворость, хвороба, хворь, болезнь	5	0	НКРЯ
	наречия			
прекрасно	чудесно, замечательно, отлично, превосходно, прекрасно	5	0	НКРЯ
прекрасно	IntS(превосходно, замечательно), OutS(чудесно, прекрасно, отлично)	5	2	News
	прилагательные			
добрый	душевный, добросердечный, отзывчивый, сердечный, добрый	5	0	НКРЯ, News
каменный	каменный, бесчувственный, суровый, жестокий, безжалостный	5	0	НКРЯ
каменный	IntS(безжалостный), OutS(каменный, бесчувственный, суровый, жестокий)	5	1	News
	глаголы			
обличать	обличать, изобличать, обвинять, разоблачать, уличать	5	0	НКРЯ, News
казаться	сдаваться, представляться, думаться, казаться	4	0	НКРЯ, News
изготовлять	делать, создавать, производить, сооружать, мастерить, изготавливать, изготовлять	7	0	НКРЯ, News

Для последующих экспериментов важно следующее наблюдение работы [8]. Чем более слово является редким, чем меньше данных, контекстов с этим словом, тем более сомнительными, неточными будут ассоциативные слова, порождаемые предсказательной моделью.

Нами проведены эксперименты для апробации предложенной модели синсета. Были использованы две матрицы W (предсказательные модели), построенные авторами $RusVectores$ по корпусу НКРЯ и по Новостному корпусу.

Для работы с предсказательными моделями была выбрана программа gensim[3], поскольку она (помимо множества других алгоритмов) содержит реализацию word2vec на языке Python (программа gensim описана в работе [17]). Эта же программа gensim использовалась при создании предсказательных моделей авторами $RusVectores$ [16].

Авторами этой статьи разработан ряд скриптов на основе gensim для работы с предсказательными моделями, вычисления $IntS$, ранга, центральности. Скрипты доступны онлайн[4]. Для нескольких тысяч синсетов, извлеченных из Русского Викисловаря, вычислен ранг и определена внутренность синсета $IntS$. Эксперименты показали, что для редких в корпусе слов $IntS$ может оказаться пустым.

Обсудим данные таблицы 2. Очевидно, что одному и тому же слову в разных предсказательных моделях, построенных по разным корпусам, будут соответствовать разные вектора. И сами словари этих моделей будут отличаться, см. [8]. Именно по этой причине отрадно видеть, что результаты в таблице 2, полученные по разным корпусам, в значительной степени совпадают. Это говорит о некоторой универсальности предлагаемой математической модели.

[3] `http://radimrehurek.com/gensim/`

[4] `https://github.com/componavt/piwidict/tree/master/lib_ext/gensim_wsd`

Заключение

Мир современной лингвистики можно условно представить в виде двух тяготеющих друг к другу, но слабо связанных областей. Строгая формализация базовых понятий необходима для дальнейшего развития лингвистики как точной науки. Формулировка четкого определения для *значения слова*, *синонимии* и других позволит в должной мере опереться на методы и алгоритмы вычислительной лингвистики (корпусной лингвистики, нейронных сетей), дискретной математики, теории вероятностей.

В нашей работе предлагается формализация такого важного для машиночитаемых словарей и тезаурусов понятия, как набор синонимов (синсет). К этой формализации синсета предлагается ряд вычислимых атрибутов (IntS, rank, centrality), которые позволяют анализировать синсеты, сравнивать их, проводить количественный анализ.

Разработанный аппарат планируется применить к решению задачи разрешения лексической многозначности.

Работа поддержана грантом РГНФ (проект № 15-04-12006).

Литература

1. *Александрова З. Е.* Словарь синонимов русского языка. М.: Русский язык, 2001. 586 с.

2. *Баранов В. И., Стечкин Б. С.* Экстремальные комбинаторные задачи и их приложения. М.: Физматлит, 2004. 238 с.

3. *Каушинис Т. В., Кириллов А. Н., Коржицкий Н. И., Крижановский А. А., Пилинович А. В., Сихонина И. А., Спиркова А. М., Старкова В. Г., Степкина Т. В., Ткач С. С., Чиркова Ю. В., Чухарев А. Л., Шорец Д. С., Янкевич Д. Ю., Ярышкина Е. А.* Обзор методов и алгоритмов разрешения лексической многозначности: Введение // Труды КарНЦ РАН. 2015. № 10. С. 69–98. doi: 10.17076/mat135

4. *Baroni M., Dinu G., Kruszewski G.* Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors // Proceedings of the ACL '14, 2014. P. 238–247. URL: `http://anthology.aclweb.org/P/P14/P14-1023.pdf` (дата обращения: 9.05.2016)

5. *Goldberg Y., Levy O.* word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, 2014. P. 1–5.

6. *Huang E. H., Socher R., Manning C. D., Ng A. Y.* Improving word representations via global context and multiple word prototypes // Proceedings of the ACL '12, Jeju Island, Korea, 2012. P. 873–882. URL: `http://dl.acm.org/citation.cfm?id=2390524.2390645` (дата обращения: 9.05.2016).

7. *Krizhanovsky A. A., Smirnov A. V.* An approach to automated construction of a general-purpose lexical ontology based on Wiktionary // Journal of Computer and Systems Sciences International. 2013. № 2. P. 215–225. doi: 10.1134/S1064230713020068

8. *Kutuzov A., Kuzmenko E.* Comparing neural lexical models of a classic national corpus and a web corpus: the case for Russian // Computational Linguistics and Intelligent Text Processing. 2015. P. 47–58. doi: 10.1007/978-3-319-18111-0_4. URL: `https://www.academia.edu/11754162/Comparing_neural_lexical_models_of_a_classic_national_corpus_and_a_web_corpus_the_case_for_Russian` (дата обращения: 9.05.2016).

9. *Kutuzov A., Andreev I.* Texts in, meaning out: neural language models in semantic similarity task for Russian. arXiv preprint arXiv:1504.08183, 2015. URL: `http://www.dialog-21.ru/digests/dialog2015/materials/pdf/KutuzovAAndreevI.pdf` (дата обращения: 9.05.2016)

10. *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings // Transactions of the Association for Computational Linguistics. 2015. Vol. 3. P. 211–225.

11. *Mahadevan S., Chandar S.* Reasoning about linguistic regularities in word embeddings using matrix manifolds. arXiv preprint arXiv:1507.07636. 2015. P. 1–9.

12. *Mikolov T., Kombrink S., Burget L., Cernocky J., Khudanpur S.* Extensions of recurrent neural network language model // Proceedings of the 2011 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2011. doi: 10.1109/icassp.2011.5947611. URL: `http://dx.doi.org/10.1109/icassp.2011.5947611` (дата обращения: 9.05.2016).

13. *Mikolov T., Zweig G.* Context dependent recurrent neural network language model // Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), 2012. doi: 10.1109/slt.2012.6424228. URL: `http://dx.doi.org/10.1109/slt.2012.6424228` (дата обращения: 9.05.2016).

14. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. URL: `http://arxiv.org/abs/1301.3781` (дата обращения: 9.05.2016).

15. *Princeton University.* What is WordNet? URL: `http://wordnet.princeton.edu` (дата обращения: 9.05.2016).

16. *RusVectōrēs:* distributional semantic models for Russian. URL: http://ling.go.mail.ru/dsm/ru/ (дата обращения: 9.05.2016)

17. *Řehůřek R., Sojka P.* Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta: University of Malta, 2010. P. 45–50. URL: http://is.muni.cz/publication/884893/en (дата обращения: 9.05.2016).

18. *Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D.* Soft similarity and soft cosine measure: Similarity of features in vector space model // Computación y Sistemas. 2014. Vol. 18, no. 3. P. 491–504. URL: http://www.scielo.org.mx/pdf/cys/v18n3/v18n3a7.pdf (дата обращения: 9.05.2016).

Поступила в редакцию 26.05.2016

References

1. *Alexandrova Z. E.* Slovar' sinonimov russkogo jazyka [Dictionary of Russian Synonyms]. Moscow: Russkij jazyk, 2001. 586 p.

2. *Baranov V. I., Stechkin B. S.* Jekstremal'nye kombinatornye zadachi i ih prilozhenija [Extremal combinatorial problems and their applications]. Moscow: Fizmatlit, 2004. 238 p.

3. *Kaushinis T. V., Kirillov A. N., Korzhitsky N. I., Krizhanovsky A. A., Pilinovich A. V., Sikhonina I. A., Spirkova A. M., Starkova V. G., Stepkina T. V., Tkach S. S., Chirkova Ju. V., Chuharev A. L., Shorets D. S., Yankevich D. Yu., Yaryshkina E. A.* Obzor metodov i algoritmov razresheniya leksicheskoi mnogoznachnosti: Vvedenie [A review of word-sense disambiguation methods and algorithms: Introduction]. *Trudy KarNTs RAN* [*Transactions of KarRC of RAS*]. 2015. No. 10. P. 69–98. doi: 10.17076/mat135

4. *Baranov V. I., Stechkin B. S.* Jekstremal'nye kombinatornye zadachi i ih prilozhenija [Extreme combinatorial problems and their applications]. Moscow: Fizmatlit, 2004. 238 p.

5. *Goldberg Y., Levy O.* word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, 2014. P. 1–5.

6. *Huang E. H., Socher R., Manning C. D., Ng A. Y.* Improving word representations via global context and multiple word prototypes. *In Proceedings of the ACL '12,* Jeju Island, Korea, 2012. P. 873–882. URL: http://dl.acm.org/citation.cfm?id=2390524.2390645 (accessed: 9.05.2016).

7. *Krizhanovsky A. A., Smirnov A. V.* An approach to automated construction of a general-purpose lexical ontology based on Wiktionary. *Journal of Computer and Systems Sciences International.* 2013. No. 2. P. 215–225. doi: 10.1134/S1064230713020068

8. *Kutuzov A., Kuzmenko E.* Comparing neural lexical models of a classic national corpus and a web corpus: the case for Russian. *Computational Linguistics and Intelligent Text Processing.* 2015. P. 47–58. doi: 10.1007/978-3-319-18111-0_4. URL: https://www.academia.edu/11754162/Comparing_neural_lexical_models_of_a_classic_national_corpus_and_a_web_corpus_the_case_for_Russian (accessed: 9.05.2016).

9. *Kutuzov A., Andreev I.* Texts in, meaning out: neural language models in semantic similarity task for Russian. arXiv preprint arXiv:1504.08183, 2015. URL: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/KutuzovAAndreevI.pdf (accessed: 9.05.2016)

10. *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics.* 2015. Vol. 3. P. 211–225.

11. *Mahadevan S., Chandar S.* Reasoning about linguistic regularities in word embeddings using matrix manifolds. arXiv preprint arXiv:1507.07636. 2015. P. 1–9.

12. *Mikolov T., Kombrink S., Burget L., Cernocky J., Khudanpur S.* Extensions of recurrent neural network language model. In Proceedings of the 2011 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2011. doi: 10.1109/icassp.2011.5947611. URL: http://dx.doi.org/10.1109/icassp.2011.5947611 (accessed: 9.05.2016).

13. *Mikolov T., Zweig G.* Context dependent recurrent neural network language model. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT). 2012. doi: 10.1109/slt.2012.6424228. URL: http://dx.doi.org/10.1109/slt.2012.6424228 (accessed: 9.05.2016).

14. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. URL: http://arxiv.org/abs/1301.3781 (accessed: 9.05.2016).

15. *Princeton University.* What is WordNet? URL: http://wordnet.princeton.edu (accessed: 9.05.2016).

16. *Řehůřek R., Sojka P.* Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta: University

of Malta, 2010. P. 45–50. URL: http://is.muni.cz/publication/884893/en (accessed: 9.05.2016).

17. *RusVectōrēs:* distributional semantic models for Russian. URL: http://ling.go.mail.ru/dsm/en/ (accessed: 9.05.2016)

18. *Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D.* Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas, 2014. Vol. 18, no. 3. P. 491–504. URL: http://www.scielo.org.mx/pdf/cys/v18n3/v18n3a7.pdf (accessed: 9.05.2016).

Received May 26, 2015

СВЕДЕНИЯ ОБ АВТОРАХ:

Кириллов Александр Николаевич
ведущий научный сотрудник, д. ф.-м. н.
Институт прикладных математических исследований
Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск,
Республика Карелия, Россия, 185910
эл. почта: kirillov@krc.karelia.ru
тел.: (8142) 766312

Крижановский Андрей Анатольевич
рук. лаб. информационных компьютерных технологий, к. т. н.
Институт прикладных математических исследований
Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск,
Республика Карелия, Россия, 185910
эл. почта: andew.krizhanovsky@gmail.com
тел.: (8142) 766312

CONTRIBUTORS:

Kirillov, Alexander
Institute of Applied Mathematical Research,
Karelian Research Centre,
Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk,
Karelia, Russia
e-mail: kirillov@krc.karelia.ru
tel.: (8142) 766312

Krizhanovsky, Andrew
Institute of Applied Mathematical Research,
Karelian Research Centre,
Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk,
Karelia, Russia
e-mail: andew.krizhanovsky@gmail.com
tel.: (8142) 766312