Обнаружение понятий в графе синонимов

Д. А. Усталов

Институт математики и механики им. Н.Н. Красовского УрО РАН, Екатеринбург, Россия
Уральский федеральный университет им. Б.Н. Ельцина, Екатеринбург, Россия
Контактный e-mail: dau@imm.uran.ru

Рассмотрена проблема автоматического группирования семантически близких слов в понятия по материалам словарей синонимов. Представлен метод обнаружения понятий в графе синонимов Watset, основанный на кластеризации связанных значений слов в графе синонимов. Выполнено исследование вычислительной сложности предложенного метода. Проведено его сравнение с аналогичными методами. Эксперименты показывают высокую эффективность предложенного метода на основании попарных информационно-поисковых критериев по материалам двух золотых стандартов для русского языка.

Ключевые слова: онтология, вычислительная семантика, вывод значений слова, кластеризация графа, понятие, синсет.

Введение

Лексические онтологии (тезаурусы) представляют собой словари, в которых слова и словосочетания с близкими значениями сгруппированы в единицы, называемые *понятиями* или *синсетами*, и в которых явно указываются семантические *отношения* между этими понятиями [1]. Тезаурусы широко применяются в компьютерной лингвистике и информационном поиске для разрешения семантической многозначности слов, расширения поисковых запросов и т. п. С другой стороны, тезаурусы позволяют формализовать экспертные знания при построении интеллектуальных систем [2] и составляют основу более сложных формальных онтологий для различных задач инженерии знаний (подробнее проблематика формальных онтологий рассмотрена в [3, гл. 3]).

В зарубежных работах большой популярностью пользуются электронные лексические онтологии WordNet [4] и BabelNet [5]. Анализ современного состояния электронных тезаурусов русского языка подтверждает высокую актуальность исследований, посвященных связыванию и развитию существующих ресурсов [6]. Кроме того, разработка автоматических методов построения тезаурусов позволяет упростить создание и развитие лексических ресурсов для языков с меньшим количеством доступных языковых данных, например церковнославянского языка [7].

Настоящая работа посвящена задаче формирования понятий на основе существующих слабоструктурированных словарей синонимов неизвестного качества. В отличие от традиционного лексикографического подхода, предполагающего привлечение коллектива специалистов-лексикографов для построения такого ресурса [8], предложенный метод позволяет повторно использовать существующие словари.

1. Обзор литературы

Задача автоматического обнаружения понятий упоминается в англоязычной литературе как *concept discovery*, она предполагает использование методов обучения без учителя для преобразования некоторого языкового ресурса, например корпуса текстов, во множество понятий, объединяющих близкие по значению слова. В классических работах Шутце [9], а также Лина и Пантеля [10] для этого применяется агломеративная кластеризация и используется большой корпус текстов для извлечения статистической информации о совместной встречаемости слов.

Применяются и методы кластеризации графов общего назначения. Например, алгоритм испорченного телефона (*Chinese Whispers*), предложенный Биманном [11] и являющийся вариацией марковского алгоритма кластеризации (*Markov clustering — MCL*) [12], широко используется в задачах вычислительной семантики. В результате выполнения алгоритма каждая вершина попадает только в один кластер, это называется жесткой кластеризацией. Несмотря на эту особенность, алгоритм легко адаптируется под задачу вывода значений слов при помощи подхода, предложенного Дороу и Виддоусом [13], состоящего в кластеризации определенных подграфов исходного графа с исключением из них вершины, соответствующей требуемому слову.

Известны и специализированные алгоритмы вывода и группировки отдельных значений многозначных слов. Алгоритм MaxMax, предложенный Хоупом и Келлером, производит мягкую кластеризацию графа на несколько пересекающихся подграфов, каждый из которых представляет собой отдельное понятие [14]. Данный алгоритм работает в два этапа. Сначала исходный неориентированный граф преобразуется в ориентированный таким образом, что между смежными вершинами с наибольшим значением близости строятся направленные ребра. Затем каждая вершина помечается как корневая, а все следующие от нее вершины как некорневые. Благодаря этому образуются пересекающиеся кластеры с общими некорневыми вершинами.

Гончало Оливейра и Гомес при построении WordNet-подобного тезауруса португальского языка Onto.PT [15] предложили подход “Извлечение, кластеризация, онтологизация” (сокр. ECO от англ. Extraction, Clustering, Ontologisation). В частности, этап кластеризации предполагает использование марковского алгоритма кластеризации [12] для обнаружения понятий в графе синонимов. Поскольку данный алгоритм является алгоритмом жесткой кластеризации, перед его запуском в веса ребер исходного графа добавляется стохастический шум. Такая процедура повторяется тридцать раз. После этого оценивается вероятность попадания пары слов в один кластер, которая сравнивается с заданным пороговым значением.

Кроме того, в области анализа сложных сетей применяются алгоритмы поиска пересекающихся сообществ в неориентированных графах. Например, метод перколяции клик (*clique percolation method — CPM*), осуществляющий поиск k-клик в заданном неориентированном графе [16]. Каждая обнаруженная k-клика рассматривается как отдельное сообщество, т. е. кластер.

Теоретико-графовые методы используются и для расширения лексических ресурсов. Например, формирование понятий в многоязычной лексической онтологии BabelNet осуществляется путем интеграции материалов Википедии с иерархией понятий тезауруса WordNet [5]. В этом случае применяются не алгоритмы кластеризации, а различные вариации алгоритма обхода графа для разрешения многозначности представленных понятий.

2. Метод обнаружения понятий WATSET

Известно, что множество синонимов, выражающих одно и то же понятие, естественным образом формирует клику в графе, причем не обязательно полную клику [17]. Значит, задача обнаружения понятий может быть сведена к задаче поиска клик в графе, чему препятствуют два существенных ограничения. Во-первых, задача поиска клик в графе является NP-полной [18], что требует искать более эффективные методы с точки зрения вычислительной сложности. Во-вторых, в структуре графа синонимов никак не отражается явление полисемии, когда слово может иметь несколько различных значений.

Наличие специализированных методов кластеризации [11, 12, 14] позволяет перейти от задачи поиска клик к задаче кластеризации графа, обладающей меньшей вычислительной сложностью. Например, вычислительная сложность алгоритма испорченного телефона составляет $O(|E|)$, где $|E|$ — количество ребер в исходном графе [11]. В свою очередь, проблема группировки отдельных значений слов остается нерешенной.

В данной работе используется следующая постановка задачи обнаружения понятий в графе синонимов. Пусть имеется неориентированный граф $G = (V, E)$, множество вершин V которого образуется множеством известных слов, а множество ребер E сформировано так, что $(v, u) \in E \iff$ слова $v \in V$ и $u \in V$ являются синонимами. Необходимо построить такое множество понятий S, что составляющие его элементы, именуемые понятиями, содержат близкие по лексическому значению слова.

2.1. Общая схема метода

На основе предположения о структуре графов многозначных слов [13] и допущения о кликах [17] предлагается метод WATSET для выделения значений слов в графе G и объединения близких по значению слов во множество понятий S. Общая схема метода представлена на рис. 1. Метод включает четыре этапа, в том числе предварительный этап построения графа синонимов на основе исходных словарей синонимов. В таком графе осуществляется вывод значений каждого слова. После этого производятся связывание значений слов друг с другом и формирование графа значений слов. Кластеризация графа значений слов группирует близкие значения слов в понятия и является финальным этапом работы метода. Концептуально метод WATSET близок к двум первым этапам подхода ECO [15], но на этапе кластеризации явным образом осуществляется вывод значений слов.

При обозначении слова и идентификатора его отдельного значения используется нотация, принятая в BabelNet [5]. Например, запись *лук*^{1} и *лук*^{2} имеет два различных значения одной и той же лексемы “лук”. В частности, *лук*^{1} может означать род травянистых растений, а *лук*^{2} — разновидность оружия дальнего боя.

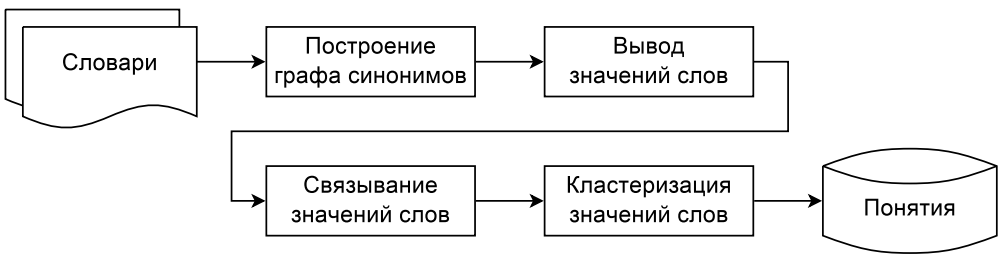

Рис. 1. Общая схема метода WATSET

2.2. Построение графа синонимов

Словари включают имена существительные, имена прилагательные, глаголы и другие части речи без их явного указания. Исходный неориентированный граф синонимов $G = (V, E)$ формируется следующим образом. Множество вершин V образуется множеством всех слов. Множество ребер E формируется так, что $(v, u) \in E \iff$ слова $v \in V$ и $u \in V$ являются синонимами хотя бы в одном словаре. Допускается использование дополнительной информации для назначения весов ребрам.

2.3. Вывод значений слов

Для решения проблемы полисемии предлагается воспользоваться методом выделения значений слов на основе эго-сетей, предложенным в работах [11, 13]. Эго-сети используются в социологии, они отражают совокупность социальных связей отдельной социальной единицы [19], например человека и его родственников. Таким образом, эго-сеть вершины u в графе G — это граф $\textsc{Ego}(u) = (V_u, E_u)$, где $V_u \subseteq V$ — множество вершин, смежных с u, включающее u; $E_u \subseteq E$ — множество ребер, связывающих эти вершины:

$$V_u = \{u\} \cup \{v : v \in V \wedge (u, v) \in E\},$$
$$E_u = E \cap (V_u \times V_u).$$

Многозначные слова связывают семантически не связанные слова. Их исключение из эго-сети приводит к разбиению ее на несколько компонентов связности [13]. Поэтому для каждой вершины $u \in V$ выполняется следующая процедура: формируется эго-сеть $\textsc{Ego}(u)$, вершина u исключается из нее, затем производится кластеризация эго-сети при помощи метода жесткой кластеризации графа [11, 12]. Каждый кластер из полученного множества кластеров C представляет собой контекст i-го значения слова u, т. е. множество синонимов слова в данном значении u^i: $\mathrm{ctx}(u^i) = C_i$, $1 \leq i \leq |C|$.

На рис. 2 представлен пример эго-сети слова “билет”. При кластеризации слово исключается из сети, что приводит к появлению трех соответствующих отдельным значениям этого слова кластеров: “бакс, банкнота, купюра”, “патент, свидетельство, аттестат” и “билетик”.

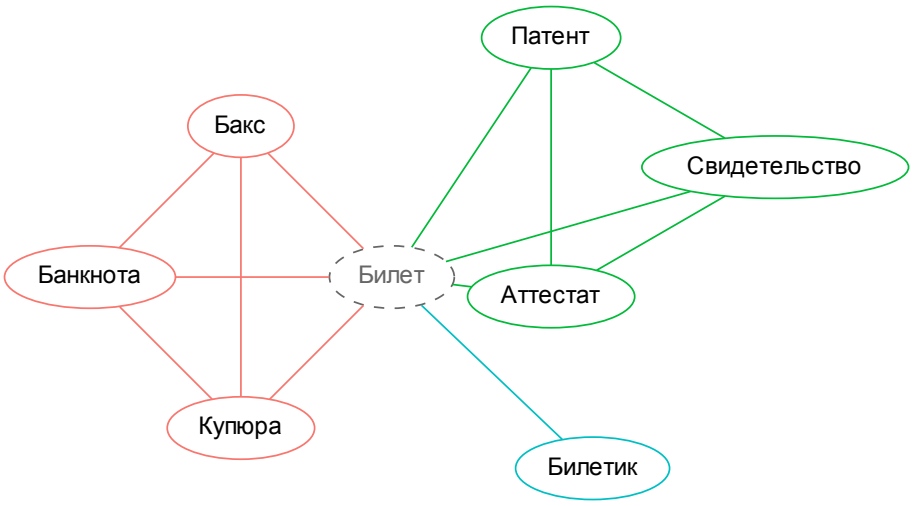

Рис. 2. Кластеризация эго-сети без слова “билет”

2.4. Связывание значений слов

Вывод значений слов позволяет извлечь из графа G отдельные значения слов и связанные с ними синонимы в виде контекстов $\mathrm{ctx}(s)$, где s — некоторое значение слова. В табл. 1 содержатся результаты вывода значений слова "билет" из примера на рис. 2: в колонке "Значение" представлены обнаруженные значения слова, в колонке "Контекст" перечислены контексты для каждого из значений. Видно, что при этом не указываются номера значений слов, образующих контексты. Для построения графа значений слов требуется провести снятие неоднозначности каждого слова в каждом контексте.

Аналогичная проблема возникает при связывании семантических сетей [20], она решается путем подбора наиболее близкого значения $\hat{u}$ для каждого слова u в контексте заданного значения слова s, т. е.

$$\hat{u} = \underset{u' \in \mathrm{senses}(u)}{\arg\max}\ \mathrm{sim}(\mathrm{ctx}(s), \mathrm{ctx}(u')),$$

где $\mathrm{senses}(u)$ — список всех значений слова u. В качестве меры близости используется косинусная мера близости между векторами, представляющими контексты в векторно-пространственной модели: $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \cos(\mathbf{a}, \mathbf{b}) = \dfrac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

На основе данной процедуры исходный контекст $\mathrm{ctx}(s)$ преобразуется в контекст со снятой неоднозначностью $\widehat{\mathrm{ctx}}(s)$ для заданного значения слова s:

$$\widehat{\mathrm{ctx}}(s) = \{\hat{u} : u \in \mathrm{ctx}(s)\}.$$

Разрешение многозначности слов в контекстах позволяет сформировать граф значений слов $G' = (V', E')$, где V' — множество вершин, т. е. значений слов; E' — множество ребер, связывающих эти вершины:

$$V' = \{s : u \in V \wedge s \in \mathrm{senses}(u)\},$$
$$E' = \{(s, \hat{u}) : s \in V' \wedge \hat{u} \in \widehat{\mathrm{ctx}}(s)\}.$$

Полученный граф G' является графом синонимов, но отличается от исходного графа G тем, что его вершинами являются не слова, а отдельные лексические значения этих слов. При этом ребра представляют отношение синономии относительно значений слов.

2.5. Кластеризация значений слов

На основании допущения о кликах [17] предполагается, что граф значений слов G' также содержит клики, причем не обязательно полные клики, соответствующие понятиям. Поскольку многозначность слов была разрешена на предыдущих этапах, в качестве заключительного этапа производится кластеризация графа значений слов при помощи

Т а б л и ц а 1. Пример контекстов слова "билет"

Значение	Контекст
билет^1	{бакс, банкнота, купюра}
билет^2	{патент, свидетельство, аттестат}
билет^3	{билетик}

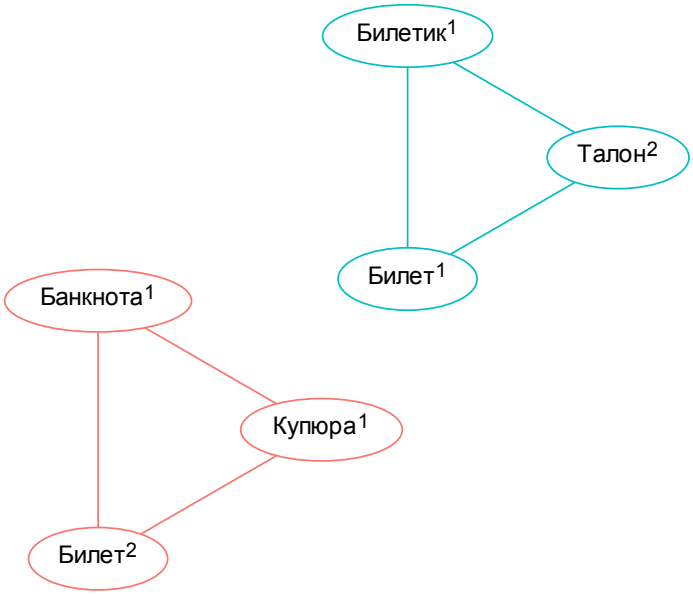

Рис. 3. Кластеризация графа значений слов

алгоритма жесткой кластеризации графов, например при помощи алгоритма испорченного телефона [11]. Полученное в результате кластеризации множество кластеров S является искомым множеством понятий.

На рис. 3 представлен пример кластеризации графа значений слов, полученного путем вывода значений слов в исходном графе синонимов (см. рис. 2) и разрешения неоднозначности в контекстах (табл. 1). В данном примере слово “билет” участвует в образовании двух понятий, связанных с различными значениями этого слова: “билет, талон, билетик” и “билет, банкнота, купюра”.

3. Реализация метода Watset

Метод Watset представлен в виде псевдокода в алгоритме 1. Исходный код эталонной реализации метода на языке программирования Python с инструкциями по запуску доступен в репозитории на GitHub: `https://github.com/dustalov/watset`. В целях обеспечения компактности записи Ego(u) означает построение эго-сети слова u с исключением этого слова из сети (см. рис. 2).

Алгоритм 1 состоит из трех основных этапов: сначала производится вывод значений слов, затем осуществляется связывание слов и значений, после чего выполняется кластеризация графа значений слов. На этапе вывода значений слов строятся эго-сети $|V|$ вершин исходного графа G, что делается за $O(|E|)$ шагов. Алгоритм испорченного телефона является линейным по количеству ребер в графе [11], при этом максимальное количество ребер в эго-сети не превышает $\frac{\Delta(G)(\Delta(G)-1)}{2}$, где $\Delta(G)$ — максимальная степень вершины в графе G. Значит, этап вывода значений выполняется за $O\left(|E| + |V|\frac{\Delta(G)(\Delta(G)-1)}{2}\right)$ шагов. Этап связывания значений состоит в разрешении многозначности всех контекстов каждого значения каждого слова при помощи

Алгоритм 1

WATSET$(V, E) \to S$

for all $u \in V$ **do** ▷ Вывод значений слов.
  $C \leftarrow$ CLUSTER(EGO(u))
  for $i \leftarrow 1 \ldots |C|$ **do**
    $\text{ctx}(u^i) \leftarrow C_i$
    $\text{senses}(u) \leftarrow \text{senses}(u) \cup \{u^i\}$
  end for
end for
$V' \leftarrow \{s : u \in V \wedge s \in \text{senses}(u)\}$ ▷ Связывание значений слов.
$E' \leftarrow \{(s, \hat{u}) : s \in V' \wedge \hat{u} \in \widehat{\text{ctx}}(s)\}$
return CLUSTER(V', E') ▷ Кластеризация значений слов.

косинусной меры близости векторов, т. е. выполняется за $O(|V| \cdot \max(|\text{senses}|) \cdot \max(|\text{ctx}|) \cdot \max(|\text{senses}|) \cdot \max(|\text{ctx}|))$ шагов, где $\max(|\text{senses}|)$ — максимальное количество значений слова, $\max(|\text{ctx}|)$ — максимальный размер контекста. Количество значений слова и размер контекста ограничены сверху степенью вершины, соответствующей слову, т. е. $\max(|\text{senses}|) \leq \Delta(G)$ и $\max(|\text{ctx}|) \leq \Delta(G)$. Следовательно, этап связывания значений выполняется за $O(|V| \cdot \Delta^4(G))$ шагов. При формировании графа значений слов не создаются новые ребра, т. е. $|E| = |E'|$, поэтому финальный этап кластеризации графа значений слов выполняется за $O(|E|)$ шагов. Таким образом, общая сложность алгоритма составляет $O(|V| \cdot \Delta^4(G))$, так как $|V| \cdot \Delta^4(G) \gg |E|$.

4. Эксперименты

Для экспериментальной оценки метода WATSET производится сравнение с аналогичными методами по материалам двух различных золотых стандартов. Как и в работе [14], при оценке использованы попарные значения точности, полноты и F_1-меры, принятые в информационном поиске [21]. В сравнении приняли участие четыре алгоритма:

- алгоритм Chinese Whispers[1] [11], реализация которого на языке программирования Java предоставлена создателями алгоритма;
- алгоритм MaxMax[2] [14], реализация которого на языке программирования Java выполнена автором данной работы самостоятельно и корректно проходит набор тестов из [14];
- алгоритм кластеризации ECO[3] [15], реализация которого на языке программирования Python выполнена автором данной работы самостоятельно, при этом из-за нехватки подробностей в описании данного метода вероятность попадания слов i и j в один кластер сравнивается с заданным пороговым значением и оценивается как

$$p_{ij} = \frac{f(i,j)}{f(i) + f(j) - f(i,j)},$$

где $f(\cdot)$ — частота появления;

[1] `https://github.com/tudarmstadt-lt/chinese-whispers`
[2] `https://github.com/dustalov/maxmax`
[3] `https://github.com/dustalov/watset/tree/db4a78/impl/onto-pt`

- метод перколяции клик CPM[4] [16], реализация которого входит в библиотеку NetworkX для работы с сетевыми структурами при помощи языка программирования Python.

В качестве золотого стандарта используются материалы двух различных семантических ресурсов для русского языка:

- тезаурус PyTez [1], построенный коллективом экспертов-лексикографов для решения задач информационного поиска, доступный на условиях лицензии Attribution-NonCommercial-ShareAlike 3.0 Unported и включающий в себя около 32 тыс. понятий и 112 тыс. лексем (слов и словосочетаний);
- тезаурус Yet Another RussNet [22] (YARN), построенный при помощи краудсорсинга, доступный на условиях лицензии Attribution-ShareAlike 3.0 и включающий около 2 тыс. понятий и 9 тыс. лексем (слов и словосочетаний) после процедуры фильтрации, состоящей в удалении всех понятий, имеющих менее восьми правок от участников проекта.

4.1. Подготовка данных

Набор данных для оценки сформирован на основе материалов трех различных открытых русскоязычных ресурсов и включает списки синонимов словаря Абрамова [23], Универсального словаря концептов [24] и русского Викисловаря [25]. Использовано три различных подхода к взвешиванию ребер графа: `ones` — все ребра получают вес, равный единице, `count` — каждое ребро получает вес, равный количеству появлений соответствующей пары слов в словарях, и `w2v` — каждое ребро получает вес, равный значению семантической близости по материалам дистрибутивного тезауруса русского языка [26]. В результате выполнения такой операции получено $|V| = 74\,386$, $|E| = 202\,316$.

Поскольку лексическое покрытие используемых словарей отличается от лексического покрытия золотого стандарта, при вычислении информационно-поисковых оценок использовались только те пары, оба слова которых входят в пересечение словника золотого стандарта и объединенного словника сравниваемых методов. Из понятия, содержащего n лексических входов, создается $\frac{n(n-1)}{2}$ пар синонимов для оценки.

Таким образом, использовано 474 517 пар синонимов из тезауруса PyTez и 56 831 пар синонимов из тезауруса Yet Another RussNet. Важно отметить, что при определенных условиях методы MaxMax и CPM генерировали понятия, содержащие 150 и более слов. Анализ результатов показал нерелевантность таких понятий, что привело к их исключению из процесса попарной оценки. Другие методы не демонстрировали подобное поведение.

4.2. Сравнение методов

В процессе сравнения тестировалось несколько конфигураций метода Watset, различающихся алгоритмом кластеризации эго-сетей и кластеризации графа значений слов: mcl — марковский алгоритм кластеризации, top — оригинальный вариант алгоритма испорченного телефона, nolog — вариант алгоритма испорченного телефона с использованием степени вершины для назначения кластеров, log — то же, что и предыдущий вариант, но с использованием натурального логарифма степени вершины. Запись

[4] `http://networkx.readthedocs.io/en/networkx-1.11/reference/algorithms.community.html`

Т а б л и ц а 2. Сравнение методов по материалам РуТез

Метод	# понятий	# пар	Точность	Полнота	F_1-мера
Chinese Whispers	16 738	645 093	0.070	0.212	0.105
MaxMax	24 250	560 096	0.147	0.180	**0.162**
ECO	43 639	68 938	**0.329**	0.070	0.115
CPM[$k = 2$]	10 932	35 618	**0.498**	0.028	0.053
CPM[$k = 3$]	3 729	46 007	**0.189**	0.045	0.073
CPM[$k = 4$]	1 971	192 756	0.056	0.072	0.063
Watset[top, log]	48 443	326 497	0.098	0.219	**0.135**
Watset[log, top]	48 441	325 931	0.097	0.218	**0.134**
Watset[nolog, log]	48 441	327 019	0.097	0.219	**0.134**
Watset[mcl, nolog]	31 135	431 287	0.084	**0.225**	0.122
Watset[mcl, log]	31 118	439 208	0.083	**0.226**	0.121
Watset[mcl, top]	31 102	434 102	0.083	**0.225**	0.121

Т а б л и ц а 3. Сравнение методов по материалам Yet Another RussNet

Метод	# понятий	# пар	Точность	Полнота	F_1-мера
Chinese Whispers	16 738	645 093	0.381	0.413	0.396
MaxMax	24 250	560 096	**0.566**	0.214	0.310
ECO	43 639	68 938	**0.846**	0.029	0.056
CPM[$k = 2$]	10 932	35 618	**0.932**	0.008	0.016
CPM[$k = 3$]	3 729	46 007	0.561	0.057	0.103
CPM[$k = 4$]	1 971	192 756	0.313	0.209	0.251
Watset[top, log]	48 443	326 497	0.381	0.440	**0.408**
Watset[top, nolog]	48 430	328 541	0.378	0.443	**0.408**
Watset[nolog, log]	48 441	327 019	0.379	0.441	**0.408**
Watset[mcl, nolog]	31 135	431 287	0.359	**0.454**	0.401
Watset[mcl, log]	31 118	439 208	0.352	**0.458**	0.398
Watset[mcl, top]	31 102	434 102	0.349	**0.460**	0.397

Watset[mcl, top] означает, что для вывода значений слов использован марковский алгоритм кластеризации, а для кластеризации графа значений слов — оригинальный вариант алгоритма испорченного телефона. Кроме того, исследовалось поведение алгоритма CPM на значениях параметра $k \in \{2, 3, 4\}$.

Результаты сравнения методов на материалах двух золотых стандартов приведены в табл. 2 и 3. В обеих таблицах, колонка "# понятий" означает количество понятий, выделенных методом; колонка "# пар" означает общее количество пар синонимов, образованных понятиями. Таблицы содержат оценки, полученные при использовании подхода `w2v`; на подходах `count` и `ones` все методы проявили себя хуже (рис. 4). Полужирным шрифтом в таблицах выделены наибольшие значения соответствующих критериев, причем лучший результат дополнительно выделен подчеркиванием.

4.3. Анализ результатов

На рис. 5 представлены совмещенные гистограммы распределения количества слов в понятиях как в использованных методах, так и в золотых стандартах.

Алгоритм испорченного телефона (Chinese Whispers) корректно сгенерировал большое количество понятий, образованных моносемичными словами и именами собствен-

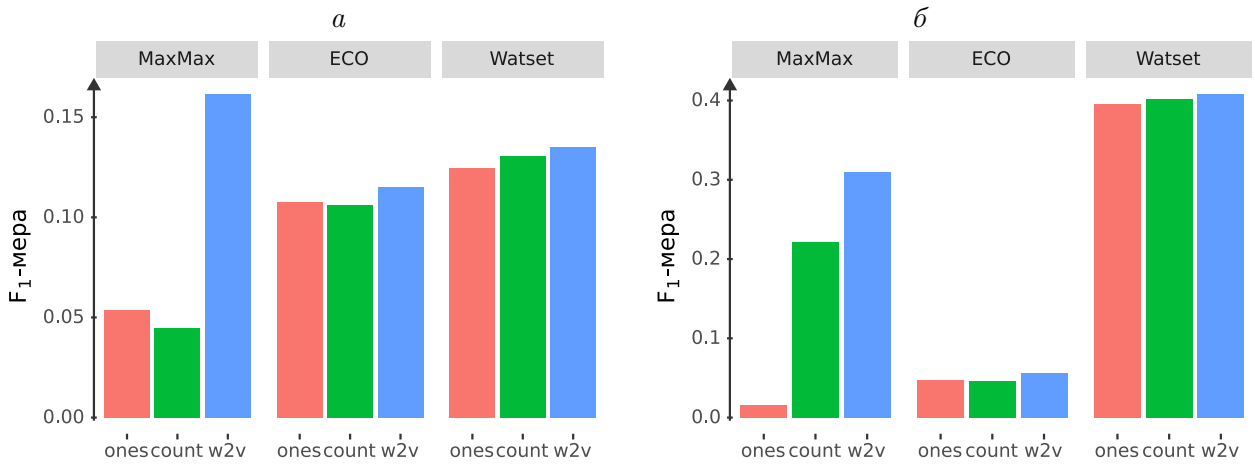

Рис. 4. Сравнение подходов к взвешиванию ребер `ones`, `count` и `w2v`: *а* — РуТез; *б* — Yet Another RussNet

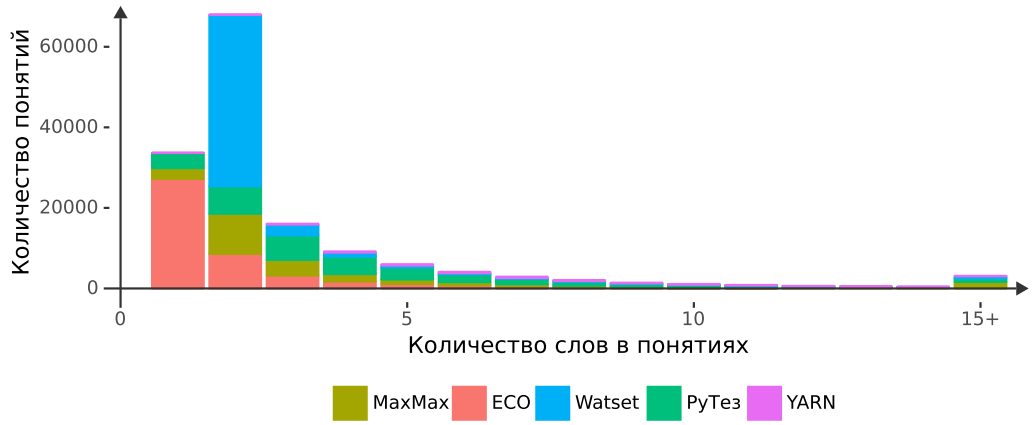

Рис. 5. Гистограммы распределения количества слов

ными, например {*туфля*[1], *полуботинок*[1], ...}. В свою очередь, при проявлении полисемии возникают заметные проблемы, состоящие в объединении никак не связанных друг с другом слов, например {*лук*[1], *порей*[1], *налучник*[1], *налучь*[1], ...}.

Несмотря на то что алгоритм MaxMax хорошо проявил себя с точки зрения используемых критериев оценки на материалах двух золотых стандартов, обнаружены две трудности в его практическом применении. Во-первых, данный алгоритм очень чувствителен к однородности весов ребер на этапе преобразования графа, что выражается в появлении крупных кластеров, связывающих семантически никак не связанные слова, например {*прайс*[1], *бином Ньютона*[1], *программный пакет*[1], ...}. Во-вторых, алгоритм не имеет механизма контроля гранулярности понятий и при определенных условиях генерирует понятия, состоящие из большого количества семантически связанных слов, не являющихся синонимами, например {*Афродита*[1], *Мефистофель*[1], *Самаэль*[1], ...}. Это существенно затрудняет практическое использование алгоритма MaxMax.

Результаты выполнения метода ECO не согласуются с отчетами об его успешном применении [15]. Понятия, состоящие из двух или более слов, образованы только моносемичными словами. При этом для многозначных слов вероятность попасть в понятия

с другими словами не превысила пороговое значение, поэтому были образованы некорректные однословные понятия: {*колонна*^{1}}, {*шатун*^{1}} и др. С одной стороны, структура графа в данной работе может отличаться от структуры графа, использованной в исследовании словарей португальского языка. С другой стороны, вероятность попадания слов в кластер может оцениваться иным образом, но в описании метода ECO не хватает подробностей. В оригинальной работе [15] заявлено пороговое значение $\theta = 0.2$, но это приводило к худшим результатам, чем использованное в данной работе $\theta = 0.05$.

Метод перколяции клик показал неудовлетворительные результаты. Вероятно, потому, что структура k-клик отличается от фактической структуры графа синонимов. Это приводит к появлению таких понятий, как {*МП*^{1}, *медицинский пункт*^{1}, *Московская Патриархия*^{1}, ...} и {*заливное*^{1}, *студень*^{1}, *студенческий билет*^{1}, ...}. Как и в случае алгоритма испорченного телефона и метода ECO, моносемичные слова оказались сгруппированы корректно.

Предложенный в данной работе метод Watset при оценке на тезаурусе RuTez по критерию F_1-меры уступил только алгоритму MaxMax, а при оценке на тезаурусе YARN получил максимальные значения как F_1-меры, так и полноты. Обнаруженные понятия корректно отражают явление полисемии, например {*пустота*^{1}, *бессодержательность*^{1}, *бессмысленность*^{1}, ...} и {*вакуум*^{1}, *пустота*^{2}, *ничто*^{1}, ...}. На этапе кластеризации значений слов также замечается ранее выявленная в методе MaxMax тенденция к связыванию семантически близких слов, не являющихся синонимами. Это приводит к снижению точности, что особенно заметно при использовании алгоритма MCL для вывода значений слов, генерирующего более крупные по размеру кластеры. Обработка таких ситуаций составляет предмет дальнейших исследований. Другим направлением развития метода Watset является обеспечение лучшей связности синонимов при помощи методов вычисления семантической близости слов по корпусу текстов [26].

Важно отметить, что тезаурусы RuTez и Yet Another RussNet, использованные в качестве двух золотых стандартов при сравнении методов, имеют различную природу и созданы для решения разных задач. Тезаурус RuTez предназначен для решения задач информационного поиска [1] и содержит понятия, состоящие из слов разных частей речи, например {*движение*, *двигаться*, *ход*, *двигательный*, ...}. Концептуально, тезаурус Yet Another RussNet более близок к оригинальному тезаурусу WordNet [22]: понятия могут быть образованы только словами одной части речи. Это объясняет разницу в результатах, полученных алгоритмом MaxMax и методом Watset, обусловленную принятыми в этих методах допущениями о структуре исходного графа синонимов.

Таким образом, представленный в данной работе метод обнаружения понятий Watset, основанный на кластеризации связанных значений слов в графе синонимов. Экспериментальное исследование метода подтверждает его высокую эффективность по итогам оценки с использованием попарных информационно-поисковых критериев. Полученные в результате работы наборы данных также доступны[5] для изучения и использования на условиях лицензии Creative Commons Attribution-ShareAlike 3.0.

Благодарности. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а, а также при финансовой поддержке РГНФ в рамках научного проекта № 16-04-12019 “Интеграция тезаурусов RussNet и YARN”.

[5] `http://ustalov.imm.uran.ru/pub/watset.tar.gz`

Автор благодарит Александра Панченко и Михаила Черноскутова за плодотворное обсуждение данного исследования, Наталью Лукашевич за предоставленный тезаурус РуТез в машиночитаемом виде, Андрея Крижановского за предоставленные материалы Викисловаря в машиночитаемом виде, а также двух анонимных рецензентов за ценные замечания по настоящей работе.

Автор также благодарит компанию Microsoft Research за предоставленные вычислительные ресурсы в облачной среде Microsoft Azure в рамках программы Azure for Research.

Список литературы / References

[1] **Лукашевич Н.В.** Тезаурусы в задачах информационного поиска. М.: Изд-во Московского ун-та, 2011. 512 с.
Lukashevich, N.V. Thesauri in information retrieval tasks. Moscow: Izd-vo Moskovskogo Un-ta, 2011. 512 p. (In Russ.)

[2] **Загорулько Ю.А.** Семантическая технология разработки интеллектуальных систем, ориентированная на экспертов предметной области // Онтология проектирования. 2015. Т. 5, № 1(15). С. 30–46.
Zagorulko, Yu.A. Semantic technology for development of intelligent systems oriented on experts in subject domain // Ontology of Designing. 2015. Vol. 5, No. 1(15). P. 30–46. (In Russ.)

[3] **Николаев И.С., Митренина О.В., Ландо Т.М.** Прикладная и компьютерная лингвистика. М.: URSS, 2016. 320 с.
Nikolaev, I.S., Mitrenina, O.V., Lando, T.M. Applied and computational linguistics. Moscow: URSS, 2016. 320 p. (In Russ.)

[4] **Fellbaum, C.** WordNet: An electronic database. Cambridge, MA, USA: MIT Press, 1998. 449 p.

[5] **Navigli, R., Ponzetto, S.P.** BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network // Artificial Intelligence. 2012. Vol. 193. P. 217–250.

[6] **Киселев Ю., Поршнев С., Мухин М.Ю.** Современное состояние электронных тезаурусов русского языка: качество, полнота и доступность // Программная инженерия. 2015. № 6. С. 34–40.
Kiselev, Yu., Porshnev, S., Muhkin, M.Yu. Current status of russian electronic thesauri: Quality, completeness and availability // Software Engineering. 2016. No. 6. P. 34–40. (In Russ.)

[7] **Shokina, N.Yu., Mocken S.** A text mining system for creating electronic glossaries with application to research of Church Slavonic language // Comput. Technologies. 2016. Т. 21, № 4. С. 3–15.

[8] **Константинова Н.С., Митрофанова О.А.** Онтологии как системы хранения знаний. Адрес доступа: http://www.ict.edu.ru/lib/index.php?id_res=5706 (дата обращения 04.11.2016).
Konstantinova, N.S., Mitrofanova, O.A. Ontologies as knowledge storage systems. Available at: http://www.ict.edu.ru/lib/index.php?id_res=5706 (accessed 04.11.2016). (In Russ.)

[9] **Schütze, H.** Automatic word sense discrimination // J. of Comput. Linguistics. 1998. Vol. 24. P. 97–123.

[10] **Lin, D., Pantel, P.** Concept discovery from text // Proc. of the 19th Intern. Conf. on Comput. Linguistics (COLING '02). Vol. 1. Taipei, Taiwan: Association for Comput. Linguistics, 2002. P. 1–7.

[11] **Biemann, C.** Chinese whispers — an efficient graph clustering algorithm and its application to natural language processing problems // Proc. of the First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1). New York, USA: Association for Comput. Linguistics, 2006. P. 73–80.

[12] **Van Dongen, S.** Graph clustering by flow simulation: Ph.D. thesis. Utrecht, Netherlands: Univ. of Utrecht, 2000. 169 p.

[13] **Dorow, B., Widdow,s D.** Discovering corpus-specific word senses // Proc. of the Tenth Conf. on Europ. Chapter of the Association for Comput. Linguistics (EACL '03). Vol. 2. Budapest: Association for Comput. Linguistics, 2003. P. 79–82.

[14] **Hope, D., Keller, B.** MaxMax: a graph-based soft clustering algorithm applied to word sense induction // Proc. of the 14th Intern. Conf. on Comput. Linguistics and Intelligent Text Processing (CICLing 2013). Pt I. Berlin; Heidelberg: Springer, 2013. P. 368–381.

[15] **Oliveira, H.G., Gomes, P.** ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically // Language Resources and Evaluation. 2014. Vol. 48, No. 2. P. 373–393.

[16] **Palla, G., Derenyi, I., Farkas, I., Vicsek, T.** Uncovering the overlapping community structure of complex networks in nature and society // Nature. 2005. Vol. 435. P. 814–818.

[17] **Kamps, J., Marx, M., Mokken, R.J., de Rijke, M.** Using WordNet to measure semantic orientations of adjectives // Proc. of LREC'2004. Paris: Europ. Language Resources Association, 2004. P. 1115–1118.

[18] **Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M.** The maximum clique problem // Handbook of Combinatorial Optimization. Springer, 1999. P. 1–74.

[19] **Freeman, L.C.** Centered graphs and the structure of ego networks // Mathematical Social Sciences. 1982. Vol 3, No. 3. P. 291–304.

[20] **Faralli, S., Panchenko, A., Biemann, C., Ponzetto, S.P.** Linked disambiguated distributional semantic networks // The Semantic Web — ISWC 2016: 15th Intern. Semantic Web Conf., Pt II. Springer Intern. Publ., 2016. P. 56–64.

[21] **Powers, D.M.W.** Evaluation: From precision, recall and F-Measure to ROC, informedness, markedness & correlation // J. of Machine Learning Technologies. 2011. Vol. 2, No. 1. P. 37–63.

[22] **Braslavski, P., Ustalov, D., Mukhin, M., Kiselev, Y.** YARN: Spinning-in-progress // Proc. of the 8th Global WordNet Conf. (GWC 2016). Global WordNet Association, 2016. P. 58–65.

[23] **Абрамов Н.** Словарь русских синонимов и сходных по смыслу выражений (8-е издание). М.: АСТ, 2007. 672 с.
Abramov, N. The dictionary of Russian synonyms and semantically related expressions, 8th edition. Moscow: AST, 2007. 672 p. (In Russ.)

[24] **Dikonov, V.G.** Development of lexical basis for the Universal Dictionary of UNL Concepts // Comput. Linguistics and Intellectual Technologies: Papers from the Annual Intern. Conf. "Dialogue". Moscow: RGGU, 2013. P. 212–221.

[25] **Krizhanovsky, A.A., Smirnov, A.V.** An approach to automated construction of a general-purpose lexical ontology based on Wiktionary // J. of Computer and Systems Sci. Intern. 2013. Vol. 52, No. 2. P. 215–225.

[26] **Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., Biemann, C.** Human and machine judgements for Russian semantic relatedness // Analysis of Images, Social Networks and Texts: Intern. Conf., AIST 2016, Revised Selected Papers. Springer Intern. Publ., 2017. P. 303–317.

Поступила в редакцию 20 января 2017 г.

Concept discovery from synonymy graphs

Ustalov, Dmitry A.

Krasovskii Institute of Mathematics and Mechanics UrB RAS, Ekaterinburg, 620990, Russia

Ural Federal University, Ekaterinburg, 620002, Russia

Corresponding author: Ustalov, Dmitry A., e-mail: dau@imm.uran.ru

This paper addresses the problem of automatic concept discovery from synonymy graphs. The purpose of the present study is to reuse the widely available semi-structured synonymy dictionaries for discovering the concepts. For that, Watset, a novel concept discovery method, based on graph clustering, has been proposed.

The method is designed under the assumption that the concept structures form cliques in the input synonymy graph. Watset has three primary steps. Firstly, it uses word sense induction to deal with ambiguous words. Secondly, it produces a disambiguated version of the input synonymy graph representing the synonymy relations between the particular word senses. Finally, it clusters the latter graph to produce a set of clusters corresponding to the concepts. The overall time complexity of this method has been assessed and found to be proportional to the number of the input words multiplied by the biquadratic maximum degree of the input graph.

A series of experiments has also been conducted to evaluate the performance of the proposed method. Watset outperformed four analogous state-of-the-art methods in terms of pairwise recall while being comparable in terms of pairwise precision and pairwise F-score on two datasets derived from the different Russian golden standards.

The software implementing the proposed approach has been made publicly available for further use.

Keywords: ontology, computational semantics, word sense induction, graph clustering, concept, synset.

Acknowledgements. The reported study was funded by RFBR according to the research project no. 16-37-00354 mol_a. This work is supported by the Russian Foundation for the Humanities project No. 16-04-12019 “RussNet and YARN thesauri integration”.

The author is grateful to Alexander Panchenko and Mikhail Chernoskutov for fruitful discussions on the present study, to Natalia Loukachevitch for the RuThes thesaurus provided in a machine-readable form, to Andrew Krizhanovsky for the Russian Wiktionary provided in a machine-readable form, and to two anonymous referrees who offered useful comments on the present paper.

The author is also grateful to Microsoft Research for providing free access to computational resources of the Microsoft Azure cloud under the Azure for Research Award program.

Received 20 January 2017