УДК 004.912
DOI: 10.17586/0021-3454-2019-62-1-50-55

ОБЪЕДИНЕНИЕ СЕМАНТИЧЕСКИХ СЕТЕЙ НА ОСНОВЕ ЭКВИВАЛЕНТНОСТИ ТОПОЛОГИЙ

А. Е. Письмак, С. В. Клименков, Е. А. Цопа,
А. Ю. Слободкин, В. В. Николаев

Университет ИТМО, 197101, Санкт-Петербург, Россия
E-mail: alexey.pismak@cs.ifmo.ru

Представлен метод, реализующий алгоритм слияния семантических графов на основе эквивалентности их топологий. Результатом применения метода является семантическая сеть, сформированная из двух разнородных источников и имеющая высокую связность.

Ключевые слова: *семантические сети, графы, тезаурусы, топология семантических сетей, Wiktionary, RuThes*

Введение. Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а ребра задают отношения между ними [1, 2]. В настоящее время семантические сети широко используются при решении множества различных задач, в частности при построении баз знаний, в задачах машинного перевода и обработки текста на естественном языке. Вследствие широкого спектра использования подобных графов возникает необходимость в их доработке — увеличении числа узлов и повышении связности между ними.

При традиционном способе построения семантической сети ее формирование осуществляется вручную, что требует значительных трудозатрат. Такие сети содержат небольшое количество узлов, тем не менее они обладают важным преимуществом — их узлы и связи проверены вручную и являются корректными. Альтернативный подход — автоматическое построение семантической сети на базе внешнего источника, формируемого пользователями сети Интернет [3]. Ярким примером такого источника является словарь Wiktionary [4].

Однако автоматическое формирование семантической сети на основе Wiktionary имеет существенный недостаток: этот источник сам по себе не является семантической сетью, т.е. при работе с ним семантические отношения между узлами приходится восстанавливать по ряду косвенных признаков. Это приводит к тому, что восстанавливаются далеко не все семантические отношения. Особенно критичной низкая связность является для семантического ядра сети. Описанная ситуация наглядно демонстрирует необходимость в слиянии нескольких семантических сетей для получения более качественного артефакта.

В настоящей статье рассматриваются два разнородных источника словарных данных — Jackalope и RuThes [5]. Первый из них — это тезаурус, построенный автоматически из данных, содержащихся в словаре Wiktionary [6]. Его особенности — наличие большого числа смысловых значений и множество небольших компонентов связности на периферии семантического графа. RuThes — это экспертный тезаурус, содержащий достаточно связное семантическое ядро. Именно на примере этих двух словарных источников предлагается метод слияния семантических сетей с использованием особенностей топологии объединяемых графов.

Алгоритм слияния графов. Различные семантические сети могут иметь специфичные особенности структуры, тем не менее в целом они представлены множеством узлов и ребер между ними, где каждый узел содержит смысловое значение, а ребра обозначают семантические отношения между понятиями. Подобное структурное сходство позволяет проанализиро-

вать эквивалентность небольших компонентов связности [7] двух тезаурусов. Если ряд критериев эквивалентности окажется приемлемым, то недостающие связи и смысловые понятия, содержащиеся в одном компоненте, могут быть импортированы в компонент другой сети.

На первом этапе анализа графов для каждого понятия или текстового входа C (лексемы, определяющей смысловое значение в RuThes) из RuThes определяется соответствующая ему лексема L в Jackalope (рис. 1). Далее необходимо определить, какому смысловому значению $\{S_1, S_2, S_3\}$ лексемы L соответствует концепт C (рис. 2; здесь и на последующих рисунках выделены узлы, анализируемые на конкретном шаге алгоритма). Так как на предыдущем этапе было установлено соответствие между лексемой L и концептом C, то одно из смысловых значений лексемы L с большой долей вероятности соответствует этому концепту.

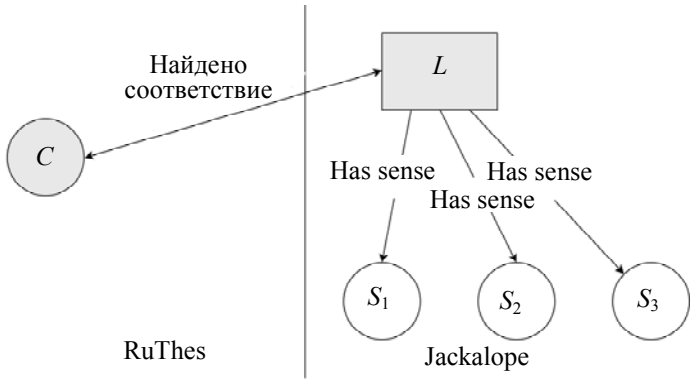

Рис. 1

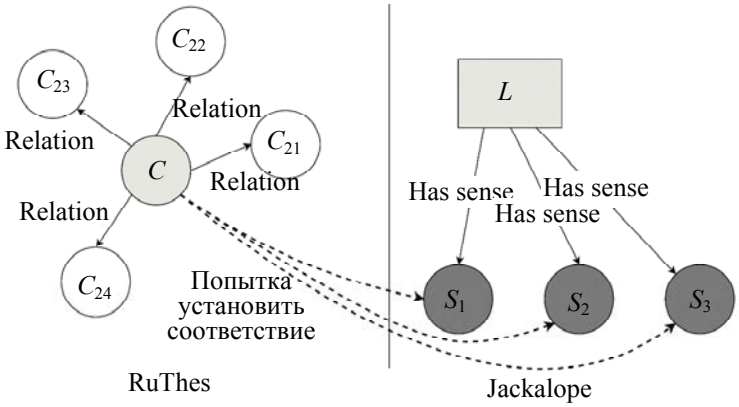

Рис. 2

Далее определяется соответствие между концептом C и каждым смысловым значением S_n лексемы L, полученным на предыдущем этапе (рис. 3). В данном случае возможны три варианта:

1) лексема L не имеет смысловых значений, тогда осуществляется возврат к началу алгоритма;

2) лексема L имеет только одно смысловое значение, тогда это искомое соответствие;

3) перебор всех смысловых значений S_n и сравнение их связей с семантическими связями концепта C.

Сравнение связей концепта C и смыслового значения S происходит путем сопоставления всех смежных понятий и типов семантических связей к ним. На этом шаге выбирается наиболее подходящее смысловое значение S_x, имеющее хотя бы одну аналогичную связь. Если ни одной связи не найдено, то осуществляется возврат к началу алгоритма. Следует отметить, что на данном этапе S_x имеет связи, направленные непосредственно к другим

семантическим узлам, а также невосстановленные связи к лексемам, на рис. 3 эти связи отмечены как „link“ и „option“ соответственно.

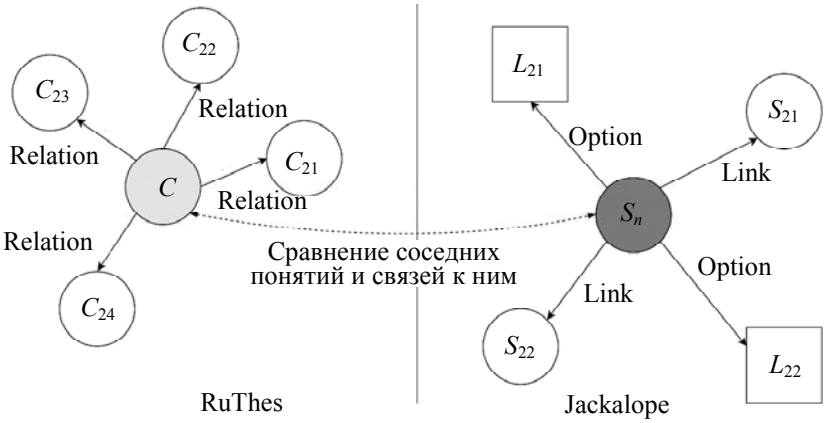

Рис. 3

После того как найдено соответствие между S и C, можно утверждать, что они означают одно и то же понятие. Следовательно, эти понятия должны иметь эквивалентные семантические связи (рис. 4).

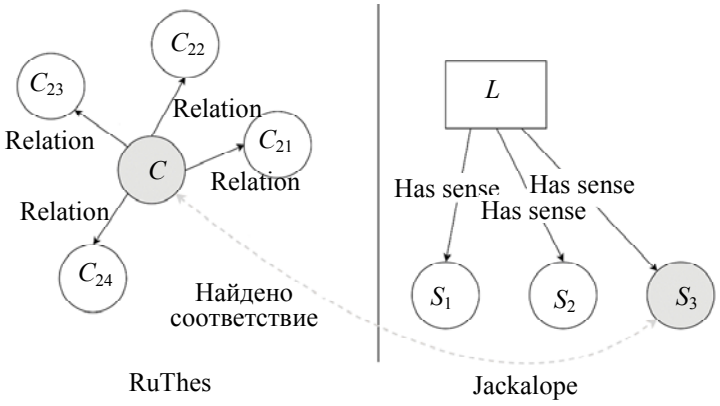

Рис. 4

Далее происходит поиск узла, к которому направлено отношение „sense option“ значения S (рис. 5). На предыдущем шаге получено соответствие между концептом C и смыслом S, поэтому можно предположить, что неуточненные связи смыслового значения S (направленные к лексемам, а не к другим понятиям) также есть у концепта C. Неуточненными связями семантической сети Jackalope принято считать связи, которые не были разрешены автоматически при построении сети, но содержат данные, косвенно указывающие на гипотетические семантические отношения между узлами.

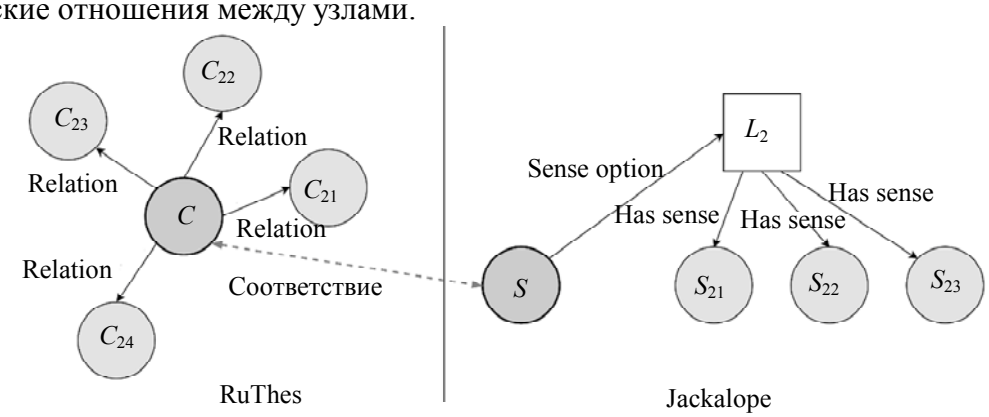

Рис. 5

Для каждой лексемы L_2, к которой есть невосстановленная ссылка, отмеченная на рис. 5 как „sense option“, рассматривается список смысловых узлов S_2 этой лексемы. Для каждого

узла S_2 лексемы L_2 производится сравнение связей со смежным набором понятий C_2 концепта C (рис. 6), которое аналогично сравнению на этапе поиска соответствия (см. рис. 2).

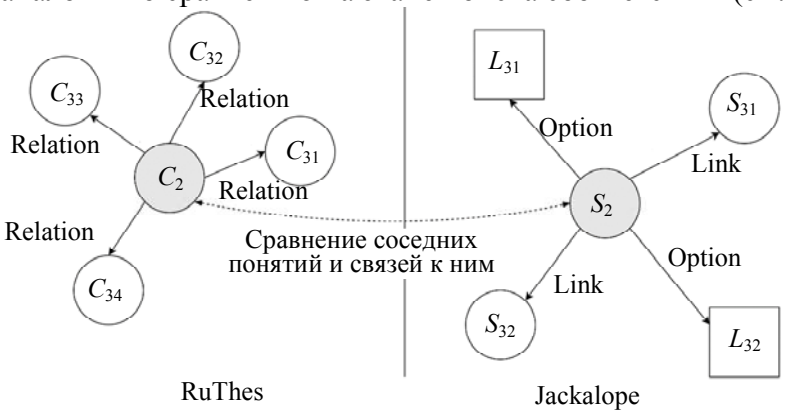

Рис. 6

Следует отметить, что на этом этапе типы связей также сравниваются на соответствие: отношение „sense option“ от смыслового узла S к лексеме L_2 сравнивается с каждым отношением „relation“ между концептом C и концептом из набора C_2, поэтому для сравнения со смысловыми значениями из набора S_2 выбираются только те концепты C_2, которые имеют связь „relation“, сходную со связью „sense option“.

Таким образом, на предыдущем этапе выбирается смысловое значение из набора S_2 (см. рис. 5), наиболее соответствующее концепту C. Если искомый смысловой узел из набора S_2 не был найден, то происходит возврат к началу алгоритма. При наличии такого смыслового узла становится возможным восстановить связь „sense option“ (рис. 7).

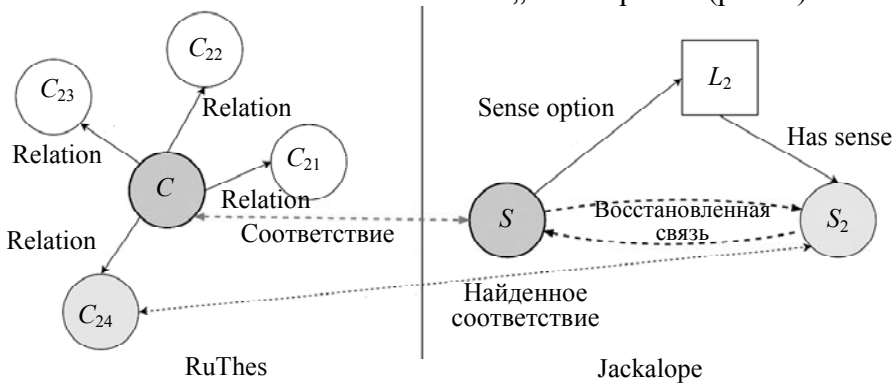

Рис. 7

Так, выполняя ряд шагов, рассмотренных выше, можно найти эквивалентные по топологии фрагменты семантических сетей и синхронизировать имеющиеся связи, что повысит связность графов.

Реализация и тестирование предложенного метода. Предложенный метод слияния семантических графов был реализован в виде модульного программного обеспечения:

— модуля преобразования исходных данных словарей RuThes и Jackalope в формат единого представления графов;

— модуля поиска эквивалентных семантических узлов;

— модуля синхронизации связей.

Применение метода слияния для указанных выше тезаурусов дало следующие результаты: количество восстановленных связей составило 5500; количество связей, импортированных из источника RuThes, — 25 500.

Заключение. Автоматическое построение семантических сетей — это сложная, комплексная задача. Открытые источники для формирования такой сети имеют ряд недостатков, связанных с низкой связностью узлов, либо имеют небольшое число этих узлов. Для

повышения связности семантических графов предложен алгоритм, базирующийся на схожести фрагментов их топологии. Применение алгоритма для слияния двух семантических сетей — Jackalope и RuThes — показало значительное увеличение количества семантических связей в первом из указанных источников.

Разработанный метод применим для различных источников семантических данных. Для слияния с другими графами необходимо лишь привести все источники к единому формату представления.

СПИСОК ЛИТЕРАТУРЫ

1. *Лату М. Н.* Принципы построения терминологических сетей: типы вершин и отношений // Вопросы когнитивной лингвистики. 2016. № 4. С. 142—149.

2. *Митрофанова О. А., Константинова Н. С.* Онтологии как системы хранения знаний // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению „Информационно-телекоммуникационные системы“. 2008 [Электронный ресурс]: <https://nsu.ru/xmlui/handle/nsu/8979>.

3. *Osika V. P., Klimenkov S., Tsopa E., Pismak A., Nikolaev V., Yarkeev A.* Method of reconstruction of semantic relations using translingual information // Proc. of the 9th Intern. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR). 2017. Vol. 2. P. 239—245.

4. *Письмак А. Е., Харитонова А. Е., Цопа Е. А., Клименков С. В.* Метод автоматического формирования семантической сети из слабоструктурированных источников // Программные продукты и системы. 2016. № 3. С. 74—78.

5. *Лукашевич Н. В., Добров Б. В., Четверкин И. И.* РуТез-Lite (опубликованная версия тезауруса русского языка РуТез) // Материалы Междунар. конф. по компьютерной лингвистике „Диалог-2014“. С. 340—349 [Электронный ресурс]: <https://pdfs.semanticscholar.org/270d/68ded1bfc3bc36f8e21724e1a992374f79a0.pdf>.

6. *Письмак А. Е., Харитонова А. Е., Цопа Е. А., Клименков С. В.* Оценка семантической близости предложений на естественном языке методами математической статистики // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16, № 2(102). С. 324—330.

7. *Райгородский А.* Модели случайных графов. ЛитРес, 2017.

Сведения об авторах

Алексей Евгеньевич Письмак — Университет ИТМО; кафедра вычислительной техники; ассистент; E-mail: alexey.pismak@cs.ifmo.ru

Сергей Викторович Клименков — Университет ИТМО; кафедра вычислительной техники; ассистент; E-mail: serge.klimenkov@cs.ifmo.ru

Евгений Алексеевич Цопа — Университет ИТМО; кафедра вычислительной техники; ассистент; E-mail: evgenij.tsopa@cs.ifmo.ru

Артем Юрьевич Слободкин — студент; Университет ИТМО; кафедра вычислительной техники; E-mail: artslob@yandex.ru

Владимир Вячеславович Николаев — Университет ИТМО; кафедра вычислительной техники; ассистент; E-mail: vladimir.nikolaev@cs.ifmo.ru

Поступила в редакцию
30.06.18 г.

Ссылка для цитирования: *Письмак А. Е., Клименков С. В., Цопа Е. А., Слободкин А. Ю., Николаев В. В.* Объединение семантических сетей на основе эквивалентности топологий // Изв. вузов. Приборостроение. 2019. Т. 62, № 1. С. 50—55.

MERGING OF SEMANTIC NETWORKS BASED ON EQUIVALENCE OF TOPOLOGIES

A. E. Pismak, S. V. Klimenkov, E. A. Tsopa,
A. Yu. Slobodkin, V. V. Nikolaev

ITMO University, 197101, St. Petersburg, Russia
E-mail: alexey.pismak@cs.ifmo.ru

A method realizing of semantic graphs merging algorithm based on features of their topologies is presented. The method application results in creation of a semantic network of high connectedness formed from two heterogeneous sources.

Keywords: semantic networks, graphs, thesaurus, semantic network topology, Wiktionary, RuThes

REFERENCES

1. Latu M.N. *Issues of cognitive linguistics*, 2016, no. 4, pp. 142–149. (in Russ.)
2. Mitrofanova O.A., Konstantinova N.S., 2008, 54 p. https://nsu.ru/xmlui/handle/nsu/8979. (in Russ.)
3. Osika V.P., Klimenkov S., Tsopa E., Pismak A., Nikolaev V., Yarkeev A. *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)*, 2017, vol. 2, pp. 239–245.
4. Pis'mak A.E., Kharitonova A.E., Tsopa E.A., Klimenkov S.V. *Programmnye produkty i sistemy*, 2016, no. 3, pp. 74–78. (in Russ.)
5. Lukashevich N.V., Dobrov B.V., Chetverkin I.I. *Mezhdunarodnaya konferentsiya po komp'yuternoy lingvistike Dialog-2014* (International Conference on Computational Linguistics Dialogue-2014), 2014, pp. 340–349. https://pdfs.semanticscholar.org/270d/68ded1bfc3bc36f8e21724e1a992374f79a0.pdf. (in Russ.)
6. Pis'mak A.E., Kharitonova A.E., Tsopa E.A., Klimenkov S.V. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, no. 2(16), pp. 324–330. (in Russ.)
7. Raygorodskiy A.M. *Modeli sluchaynykh grafov* (Random Graph Models), Moscow, 2017. (in Russ.)

Data on authors

Alexey E. Pismak	—	ITMO University; Department of Computer Science; Assistant; E-mail: alexey.pismak@cs.ifmo.ru
Sergey V. Klimenkov	—	ITMO University; Department of Computer Science; Assistant; E-mail: serge.klimenkov@cs.ifmo.ru
Evgeny A. Tsopa	—	ITMO University; Department of Computer Science; Assistant; E-mail: evgenij.tsopa@cs.ifmo.ru
Artem Yu. Slobodkin	—	ITMO University; Department of Computer Science; Assistant; E-mail: artslob@yandex.ru
Vladimir V. Nikolaev	—	ITMO University; Department of Computer Science; Assistant; E-mail: vladimir.nikolaev@cs.ifmo.ru

For citation: Pismak A. E., Klimenkov S. V., Tsopa E. A., Slobodkin A. Yu., Nikolaev V. V. Merging of semantic networks based on equivalence of topologies. *Journal of Instrument Engineering*. 2019. Vol. 62, N 1. P. 50—55 (in Russian).

DOI: 10.17586/0021-3454-2019-62-1-50-55