THE BUILDING OF RELATIONS IN HYBRID ONTOLOGICAL NETWORK FOR SOLVING TESTING TASKS ON DIALOG'2015 FORUM EVALUATION

Ponomarev S. V. (serv@newmail.ru)

Sputnik LLC, Moscow, Russia

This paper describes the principles of building of relations such as "synonym", "hyperonym" and "hyponym" between words and word collocations by method of teaching by examples. As a knowledge base the system uses a number of ontological information of words and expressions from open-access sources and statistic information, collected by processing large text corpora.

The knowledge base is presented as a hybrid ontological network — an oriented graph, where vertices[1] are the words and expressions and edges are the links between words. In addition, each link between two words or expressions is oriented, typified and weighted. The link type characterizes the information source, from which this link and its type were extracted (for example, synonym from Wiktionary). Link weight is determined by reliable information source. All links, obtained from dictionaries and ontological bases, have the weight equals to one. The links, collected by processing text corpora, have the weight equals to frequency of relevant agreed bigrams (for example, a bigram adjective + noun).

The structure of the hybrid ontological network characterizes by a large number of links between the network vertices. Besides direct links connecting two particular network vertices, there could be used composite links, passes through intermediate vertices, which leads to cardinally increasing of number of possible ways between vertices.

Here's a training algorithm that allows setting in the hybrid ontological network the links between words and items in term of combinations of weighted paths between network vertices.

Key words: ontology, data mining, supervised learning, semantics

1. Смешанная онтологическая сеть

Находящиеся в открытом доступе онтологии на русском языке не предполагают обработки связей, имеющих вероятностную природу, поскольку в этих онтологиях не предусмотрено указание веса конкретного триплета. Соответственно, такие онтологии не могут быть расширены связями, накопленными при статистической обработке текстовых корпусов, и их слияние с другими источниками информации затруднено расхождениями между разными источниками информации. Добавление в триплет значения его достоверности (веса) позволяет решить эти проблемы.

---

[1] https://en.wikipedia.org/wiki/Vertex_(graph_theory)

Будем называть онтологию смешанной, если:

1. она составлена из нескольких независимых источников;
2. содержит триплеты (связи), накопленные статистической обработкой текстовых корпусов;
3. каждый триплет характеризуется типом и весом.

Основными свойствами такой онтологии являются избыточность и высокая связность. Избыточность возникает в связи с дублированием большей части онтологических связей в различных использованных источниках, а высокая связность возникает при включении связей, полученных при статистической обработке текстов.

Если представить такую онтологию в виде сети, с узлами-понятиями и рёбрами-связями, то смешанная онтологическая сеть характеризуется большим количеством возможных путей между узлами сети, в том числе — через несколько промежуточных узлов. Использование статистических данных гарантирует, что даже не представленные в использованных источниках информации понятия, к примеру, редкие слова или названия, связаны с другими узлами онтологической сети достаточным количеством связей.

Общее количество узлов в сети 1 355 135 и сводная информация по типам связей смешанной онтологической сети приведена в таблице 1.

Таблица 1. Структура смешанной онтологической сети

№	Связь	Кол-во связей	Тип связи	Источник
3	Идиоматические выражения	9334	Онтологическая	Викисловарь [1]
4	Эпитеты	49929		
5	Антонимы	24900		
6	Синонимы	739053		
7	Гиперонимы	29545		
8	Гипонимы	30871		
9	Вышестоящие категории	12332		
0	Устойчивые сочетания	16068		
13	Связанные слова	407895		
14	Холонимы	475		
15	Меронимы	667		
10	Категории	226800		
24	Примеры использования	16463		
2	Определяющие слова	4672480	Статистическая	Определяющие слова — это слова из статей толковых словарей, для которых отношение частоты слова в статье к частоте слова во всём корпусе — максимально.

№	Связь	Кол-во связей	Тип связи	Источник
12	Омни-связи	17 092	Статистическая	Омнимические связи устанавливаются между узлами путём сравнения всех возможных грамматических форм слова
11	Слова входят в одну фразу	231 416 665	Статистическая	Несогласованные N-граммы, полученные парсингом корпуса новостей
30	Слово примыкает слева	22 551 832		
17	N-грамма существительное + существительное «восход Солнца»	28 722 993	Статистическая	Сбор статистики при помощи SDK Грамматического словаря [2]
18	N-грамма наречие + глагол «много спит»	2 148 646		
19	N-грамма наречие + прилагательное «очень яркий»	1 722 124		
20	N-грамма предлог + существительное «на диване»	623 370		
21	N-грамма глагол + управляемый объект «видит мышку»	4 234 149		
22	N-грамма прилагательное + существительное «спиральная галактика»	10 249 513		
23	N-грамма существительное + глагол «кошка мышкует»	7 518 027		
25	Фраза состоит из	951 895	Внутренняя связь	Узлы сети, составленные из нескольких слов (словосочетания и фразы), имеют связи со словами, из которых состоят.
26	Первое слово фразы	310 817		
27	Второе слово фразы	310 817		
28	Количество вхождений слова во фразу	934 023		
29	Количество вхождений словосочетания во фразу	228 386		

Описанная структура смешанной онтологической сети позволяет устанавливать между узлами сети отношения разного рода, например — «синоним» или «значение атрибута». При этом, отношения характеризуются вычислимым уровнем достоверности, в диапазоне [0;1]. Другими словами, установленное отношение по сути является классификатором, оценивающим наличие между сущностями реального мира отношения на основании имеющейся в сети информации.

2. Простой алгоритм построения отношений

Автоматическое построение отношений основано на обучении на примерах. В качестве примеров использовалась представленная Организаторами тестирования обучающая выборка синонимов, гипернимов и гипонимов. Рассмотрим

построение отношений на примере синонимов. Часть слов, являющихся синонимами, связаны связью №6 «синонимы из Викисловаря». Однако, связью №6 связаны только 19 691 слово, и для остальных слов степень синонимичности необходимо вычислять. Для оценки степени синонимичности (взаимозаменяемости) двух слов будем использовать структуру связей обоих исследуемых слов. В качестве примера возьмём согласованные биграммы — синонимичные слова должны иметь схожую структуру связей с другими словами (таблица 2).

Таблица 2. Сравнение структуры связей слов в биграммах

	АПЕЛЬСИН (1026 связей всего)	**ПОМИДОР (747 связей всего)**
N-грамма существительное + существительное «восход Солнца»	0,0515 лимон 0,0327 сок **0,0302 долька** 0,0165 яблоко 0,0162 кожура 0,0151 цвет 0,0136 банан 0,0108 запах 0,0104 цедра	0,0672 огурец 0,0463 салат 0,0231 лук 0,0145 кг **0,0140 долька** 0,0131 яйцо 0,0127 перец 0,0104 чеснок 0,0100 сыр

Сравнение связей двух слов по одному типу связи даёт возможность вычислить метрику подобия, например, косинусную меру. Соответственно, для каждой пары слов доступен вектор из метрик подобия по каждому из типов связи в онтологической сети. Подстраивая коэффициенты, с которыми учитывается вес каждого из каналов, можно создать классификатор, дающий оценку степени наличия между двумя словами семантического отношения (синоним, гиперним, гипоним). Для отношения синоним часть вектора коэффициентов приведена в таблице 3. Обратные связи обозначаются знаком минус.

Таблица 3. Часть коэффициентов вектора, реализующего отношение «синоним»

Тип связи	**Весовой коэффициент**
(−15) Меронимы, (−14) Холонимы, (−10) Категории, (−9) Вышестоящие категории, (−0) Устойчивые сочетания	0,05806
(−29) Количество вхождений словосочетания во фразу	0,05723
(−25) Фраза состоит из	0,05375
(−8) Гипонимы	0,05278
(−6) Синонимы	0,05207
(29) Количество вхождений словосочетания во фразу	0,04982
(7) Гиперонимы	0,04946

Тип связи	Весовой коэффициент
(–13) Связанные слова	0,04110
(–2) Определяющие слова	0,03874
…	…
(–21) N–грамма глагол+управляемый объект «видит мышку»	0,00053
(23) N–грамма существительное+глагол «кошка мышкует»	0,00017
(–18) N–грамма наречие+глагол «много спит»	0,00011
(18) N–грамма наречие+глагол «много спит»	0,000097
(19) N–грамма наречие+прилагательное «очень яркий»	0,000052
(21) N–грамма прилагательное+существительное «спиральная галактика»	0,000003

Как видно из таблицы 3, основное решение по степени синонимичности (взаимозаменяемости) слов принимается на основе имеющейся онтологической информации. Статистическая информация оказывает на два порядка меньшее влияние на оценку. Однако, для редких слов, не представленных в Викисловаре и других источниках онтологической информации, статистические типы связей являются единственными доступными для таких слов.

Общий алгоритм построения классификатора, реализующего произвольный тип связи между словами:

1. Сформировать обучающую выборку в виде троек «слово1 — слово2 — значение», где значение — это сила связи между словами;
2. Для каждой тройки вычислить метрики подобия между словами для каждого типа связи;
3. Скорректировать коэффициенты каждого типа связи.

Можно привести такую аналогию — из слова1 мы последовательно переходим по конкретному типу связи ко всем узлам, к которым есть связь, после чего, по этому же типу связи возвращаемся обратно, но в обратном направлении (обратная связь). Если существуют такие узлы, у которых удельная доля связи к слову1 приблизительно равна удельной доле связи к слову2, то слова сильно связаны и вычисленная метрика будет иметь относительно высокое значение. Общее количество типов связей — 64, длина вектора коэффициентов — 64 элемента, соответственно, число настраиваемых параметров — тоже 64.

Описанный простой алгоритм оценки синонимичности/взаимозаменяемости слов называется в рамках смешанной онтологической сети отношением подобия. При программной реализации алгоритм обладает высокой скоростью вычисления.

3. Расширенный алгоритм построения отношений

Расширим алгоритм вычисления метрики подобия, допустив, что при переходе от слова1 по конкретному типу связи, мы сможем вернуться к слову2 по любому типу связи, а не только по обратной. Такое решение увеличивает число настраиваемых параметров до 64^2 = 4096 элементов.

Для принятия решения по наличию отношений между двумя словами использовалась библиотека на языке R, реализующая бинарную классификацию методом построения леса решающих деревьев. Обучающая выборка для классификатора готовилась следующим образом:

1. позитивные примеры по соответствующему отношению (синоним, гиперним и гипоним), использовались из обучающей выборки, представленной Организаторами;
2. Негативные примеры генерировались также по обучающей выборке Организаторов, методом выбора случайных пар слов;
3. Значение каждого из 4096 параметров вычислялось как сумма по всем связям произведений удельного веса связи от слова1 на удельный вес связи от слова2.

В результате были построены три классификатора — синонимы, гиперонимы и гипонимы.

4. Классификатор семантической близости

Классификатор семантической близости двух слов был подготовлен специально для участия в состязании «Семантическая близость — 2015», проходившего в рамках форума «Диалог-21». Классификатор построен как логистическая регрессия от значений факторов, приведённых в таблице 4.

Таблица 4. Факторы классификации семантической близости

Имя	Описание	Коэффициент регрессии
Синоним	Правило-классификатор синонимов (см. п. 3)	0,2599
Гипоним	Правило-классификатор гипонимов (см. п. 3)	0,0865
Гипероним	Правило-классификатор гиперонимов (см. п. 3)	0,0113
Подобие	Отношение подобия (см. п. 2)	0,1198
Word2Vec	Word2Vec в режиме skipgrams [3]	0,2583

Word2Vec обучался на относительно большой (16 Гб) выборке логов поисковых запросов пользователей к поисковому движку. Общее количество векторизованных слов — четыре миллиона, включая не только все словоформы всех распространённых слов русского языка, но и варианты слов с опечатками, а также многие слова из других языков.

Сами логи поисковых запросов не могут быть представлены в публичном доступе, но обученные базы Word2Vec с некоторыми утилитами можно получить по адресу http://servponomarev.livejournal.com/7667.html

5. Классификатор ассоциативных связей

Классификатор ассоциативных связей также был подготовлен для участия в состязании «Семантическая близость — 2015». Классификатор построен как лес деревьев решений, и классификация осуществляется по факторам, приведённым в таблице 5. Будем называть совместностью по первому слову долю словосочетаний, в которых присутствуют оба слова, в словосочетаниях, в которых присутствует только первое слово.

Таблица 5. Факторы классификации ассоциативной близости

Имя	Описание
Словосочетание	Признак, что в классифицируемой паре есть словосочетания (0,1)
Совместность по первому (второму) слову по смешанной онтологической сети	Используется смешанная онтологическая сеть (см. п. 1), совместность вычисляется по значениям узлов сети, большинство из которых являются названиями словарных статей толковых словарей.
Совместность по первому (второму) слову по названиям статей Википедии	Названия статей Википедии [4]
Косинусная мера по Word2Vec в режиме skipgrams	Word2Vec, обученный по логам поисковых запросов в режиме skipgrams
Косинусная мера по Word2Vec в режиме bag of words	Word2Vec, обученный по логам поисковых запросов в режиме bag of words
Отношение подобия	Отношение подобия, штатная функция смешанной семантической сети (см. п. 2)
Совместность по первому (второму) слову в логах поисковых запросов	Логи поисковых запросов — запросы пользователей к поисковым системам
Совместность по первому (второму) слову в логах поисковых запросов (с частотами запросов)	Логи поисковых запросов — запросы пользователей к поисковым системам с учётом частотности запросов
Пол	Признак несовпадения грамматического атрибута — пол
Число	Признак несовпадения грамматического атрибута — число
Лицо	Признак несовпадения грамматического атрибута — лицо

Имя	Описание
Время	Признак несовпадения грамматического атрибута — время
Падеж	Признак несовпадения грамматического атрибута — падеж
Синоним	Классификатор синонимов (см. п. 3)
Гипоним	Классификатор гипонимов (см. п. 3)
Гипероним	Классификатор гиперонимов (см. п. 3)
Классификатор семантической близости	Классификатор семантической близости (см. п. 4)

Полученный набор факторов был дополнен весовыми коэффициентами связей между парой понятий-узлов онтологической сети. Факторы использовались для генерации обучающей выборки на позитивных, и автоматически сгенерированных негативных примерах.

6. Анализ результатов тестирования

6.1. Тестирование на корреляцию с человеческими оценками

Результат 0,6641, 6-е место.

К сожалению, сильная ограниченность обучающей выборки по данной дорожке (65 позиций всего), не позволила использовать методы машинного обучения, а логистическая регрессия не показала приемлемых результатов. Другим фактором, снизившим результаты тестирования, является принцип построения системы только на связях между словами, без разбора внутренней структуры слова. В тестовой выборке присутствовали синтетические слова, например «киновидеотеатр», которые, как показали исследования, ни разу не встречались даже при статистической обработке больших текстовых корпусов.

6.2. Тестирование на степень семантической близости

Результат 0,9209, 3-е место.

Если проанализировать таблицу 4 видно, что наибольший вклад в оценку степени семантической близости вносит классификатор синонимов и Word2Vec. При этом, классификатор синонимов немного опережает Word2Vec по качеству работы. Относительно высокие результаты в тестировании получены благодаря комбинации двух методик — оценки синонимичности слов по онтологической сети — что имеет эффект на давно известных и распространённых словах и использования Word2Vec, который, будучи обученным за месяц до конкурса, содержал в себе новые слова и отношения между ними. Негативный

эффект от отсутствия в системе разбора состава слова сохранился, поскольку для Word2Vec такие слова как «адыгеец» и «адыгейка» являются совершенно разными, а частота их появления в поисковых запросах — близка к нулю, что не даёт возможности установить между словами связь.

6.3. Тестирование ассоциаций (Русский Ассоциативный Тезаурус)

Результат 0,9277, 3-е место.

При анализе обучающей выборки, представленной Организаторами, стало заметно, что многие ассоциации представлены в виде нескольких слов. Типичная ассоциация из нескольких слов — это продолжение поговорки («слово не воробей») или широко известного названия художественного произведения («белое солнце пустыни»). Таким образом, для предсказания ассоциаций необходимо иметь достаточно подробный перечень поговорок, афоризмов, названий и прочих культурных артефактов. Для этой цели использовались названия статей Википедии и логи поисковых запросов. В обоих источниках информации проверялось, существует или нет словосочетание, составленное из исследуемых слов/словосочетаний и если существует — вычислялась его удельная доля среди всех словосочетаний. Данный подход хорошо работает на названиях фильмов и книг, поскольку такие названия широко представлены в Википедии и логах поисковых запросов, но указанные источники бедны поговорками и афоризмами, что и привело к относительно невысоким результатам тестирования.

6.4. Тестирование ассоциаций (Sociation.org)

Результат 0,9849, 1-е место.

Структура тестовых данных из Sociation.org такова, что комбинация применявшихся подходов и источников информации оказалась эффективной. Метод вычисления подобия, описанный в пункте 2, по сути своей вычисляет степень ассоциированности пары слов, учитывая подобие структуры связей этих слов в онтологической сети. Word2Vec также вычисляет степень ассоциированности слов по их контекстам в поисковых запросах. При этом, Word2Vec обеспечил отработку новых слов, появившихся недавно, например «аватар» и «эйва».

References

1. Wiktionary, www.ru.wiktionary.org
2. Russian and English Morphology for Windows and Linux, http://solarix.ru/grammatical-dictionary-api-en.shtml
3. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
4. Wikipedia, www.ru.wikipedia.org