Анализ ассоциативных тезаурусов и возможность их применения в задачах машинного перевода[1]

Е.А.Выломова

evylomova@gmail.com

МГТУ им. Н. Э. Баумана, каф. Системы Обработки Информации и Управления

Аннотация. В работе представлен анализ ассоциативных тезаурусов и, в частности, Русского Ассоциативного Тезауруса (РАС). Показано, что сеть, основанная на данных тезауруса, принадлежит к классам “small-world” и “scale-free”. Помимо этого, приведены результаты сравнения тезаурусов на различных языках. В работе также рассматривается вопрос о возможности применения данных тезаурусов для улучшения параметров машинного перевода.

Ключевые слова: ассоциативные тезаурусы; ассоциативные сети; языковые ресурсы; извлечение знаний.

Введение

В соответствии с предположением Фирта [1], что смысл понятия раскрывается в его взаимосвязи с соседними концептами, а также согласно представлению об ассоциации как базовом механизме сознания человека, описанному в работах Deese [2] and Cramer [3], вербальные ассоциации отражают структурные шаблоны

[1] Работа выполнена в рамках гранта РГНФ №12-04-12039в
Игнатов Д.И., Яворский Р.Э. (ред.): Анализ Изображений Сетей и Текстов, Екатеринбург, 16-18 марта, 2012

взаимоотношений между понятиями. Со времен Ф. Гальтона ассоциативные эксперименты начали активно использоваться как эмпирический метод наблюдения процессов мышления, запоминания и организации знаний.

Результаты ассоциативных экспериментов наиболее часто хранятся в виде ассоциативных словарей (тезаурусов), представляющих собой набор триплетов <*стимул, реакция, частота стимульно-реактивной пары*>. В качестве стимула и реакции выступают слова и словосочетания. Подобные ассоциативные тезаурусы на текущий момент существуют на английском [2,3,4,5], японском [6,7], шведском [8], русском [9, 10], чешском [11], корейском [12], голландском [13] языках и иврите [14].

Описанная выше структура данных позволяет на базе ассоциативного тезауруса создать ассоциативную сеть. Узлами в ассоциативной сети являются концепты (например, *«еда»*, *«Великая Отечественная Война»*). Концепты можно разделить на три вида: концепты-стимулы (S), концепты-реакции (R) и концепты-стимулы-реакции (SR). Два узла u и v сети соединены между собой дугой, если u является стимулом (S или SR), а v – реакцией (SR или R) и существует ассоциативная связь между ними. В основе большинства семантических сетей, к коим можно полноправно отнести и ассоциативную сеть, лежит граф $G = (V, E)$, где V – множество узлов (вершин), а E – множество ребер или дуг в случае ориентированного графа. В дальнейших рассуждениях понятия «граф» и «сеть» рассматриваются как синонимичные.

Основные понятия теории графов

В данном разделе хотелось бы кратко перечислить основные параметры, характеризующие граф, которые затем будут использованы для сравнения их между собой.

Две вершины, связанные ребром или дугой, являются *соседями*. В случае орграфа вершина характеризуется *степенью* k^{in} и k^{out}, то есть количеством входящих и исходящих дуг соответственно. Если же граф неориентированный, то параметру k соответствует количество ребер вершины.

В неориентированном графе под *путем* понимается последовательность ребер, соединяющая пару вершин. В ориентированном графе *путь* – это набор дуг, по которым можно перейти от одной вершины к другой. В случае конкретного *графа длина пути* определяется как количество ребер, или дуг, между вершинами.

Расстояние между вершинами u и v – длина кратчайшего пути между ними.

Диаметр сети D – это максимальное расстояние между вершинами для всех пар вершин. Поэтому диаметр сети D и средняя длина кратчайшего пути L взаимосвязаны.

Функция распределения степени $P(k)$ представляет собой вероятность, что случайно выбранная вершина будет иметь степень k. Для ориентированного графа количество входящих ребер считается более важным показателем и степень в случае такого графа вычисляется как количество входящих ребер.

Таблица 1. Основные характеристики графов

Обозначение	**Определение**
N	общее количество вершин
L	средняя длина кратчайшего пути между парами вершин
D	диаметр сети
k, k^{in}, k^{out}	степень вершины
$\langle k \rangle$	средняя степень вершины
γ	показатель в функции распределения степени

Сети “small-world” и “scale-free”

Впервые феномен “small-world” наблюдал Milgram [15] в экспериментах с социальными сетями. Он предположил, что любые два человека в США разделены сравнительно малым количеством знакомых или друзей ("6 степеней разделения"). Феномен прослеживается на случайных графах, где каждая пара вершин соединена ребром с вероятностью p. Когда p достаточно высока, вся сеть становится связной и средняя длина пути L между двумя случайными вершинами возрастает логарифмически по отношению к общему числу вершин: $L \propto \log(N)$[16].

Watts, Strogatz [17] в своих работах возобновили интерес к такому типу сетей и показали, что система энергоснабжения запада США, нервная система червя Caenorhabditis elegans, международная сеть киноактеров также к ним относятся.Watts и Strogatz обозначили понятие "small-world" структур, особенностью которых является малое значение среднего кратчайшего пути между парами вершин и сравнительно высокое значение среднего коэффициента кластеризации.

В дальнейших исследованиях было выявлено, что World Wide Web, сети научного сотрудничества, метаболические сети в биологии также обладают структурой “small-world”.

Amaral, Scala, Barthélémy, Stanley [18] изучали различные классы сетей со структурой "small-world", сравнивая функцию распределения степени *P(k)*. В результате были выделены два типа распределений: экспоненциальное и степенное. Во втором случае функция имеет вид $P(k) \approx k^{-\gamma}$, где γ∈(2..4). Данный тип сетей получил название “scale-free”, их основной особенностью является наличие вершин-хабов, через которые проходит множество путей, соединяющих остальные вершины.

Результаты

В рамках данного исследования преимущественно изучался Русский Ассоциативный Словарь [10] (РАС). Объем экспериментальной выборки РАС составил *102516* различных стимульно-реактивных пар вида $\langle c_i, r_j, freq_{ij}, \rangle$, где $c_i = \overline{1{,}6577}$ – стимулы, $r_j = \overline{1{,}21312}$ – реакции, $freq_{ij}$- частота стимульно-реактивных пар. Впоследствии частота была преобразована в относительную $weight_{ij} = \frac{freq_{ij}}{\sum_{j=1}^{n} freq_{ij}}$, где n – общее количество различных реакций на данный стимул. На основе ассоциативного тезауруса была построена сеть. Ниже приведено сравнение параметров полученной сети с американским аналогом, а также с сетью WordNet.

Таблица 2. Сравнение параметров американской, русской ассоциативных сетей и WordNet

	Ориент. (РАС)	**Неориент. (РАС)**	**Ориент. (амер.)**	**Неориент. (амер.)**	**WordNet**
N	23196	23196	5018	5018	122005
L	3.989	3.836	4.27	3.04	10.56
D	8	7	10	5	27
γ	2.12	2.103	1.79	3.01	3.11
<*k*>	4.423	8.236	12.7	22	1.6

Как видно, РАС можно также отнести к сетям типа “small-world” и “scale-free”. Свойства этих сетей, а именно малая величина среднего кратчайшего пути и наличие вершин-хабов, позволили разработать методику эффективного хранения ассоциативного графа. Методика рассчитана на частые запросы кратчайших путей между вершинами и в большинстве случаев позволяет, во-первых, не загружать полный граф в операционную память, а, во-вторых, обходить лишь необходимую часть графа.

Кроме того, было получено, что для сетей на различных языках множества концептов, соответствующих вершинам-хабам, во многом пересекаются и содержат такие базовые понятия как *«отец»*, *«дитя»*, *«язык»*, *«вода»*, *«хороший»* и другие. Данный факт послужил толчком для начала исследования возможности построения отображения сети на русском языке в аналоги на других. Эта технология основана на поиске близких концептуальных полей для двух и более языков с использованием словаря-переводчика. Предполагается, что подобный подход позволит улучшить машинный перевод по следующим причинам: 1) ассоциативные сети достаточно хорошо отображают взаимосвязи между понятиями; 2) ассоциативные сети имеются на многих языках и содержат схожий набор базовых стимулов.

Заключение

Целью данной статьи являлось не только представление результатов исследования, но также и привлечение внимания к ассоциативным тезаурусам как к дополнительным языковым ресурсам, на основе которых возможно проводить валидацию онтологий и семантических отношений, полученных автоматическими методами.

Основные направления дальнейшего исследования следующие: 1) построение комбинированной модели русского и английского ассоциативных тезаурусов, которая позволит улучшить параметры машинного перевода; 2) сравнение статистических характеристик данных ассоциативного тезауруса и результатов квантитативного анализа текстовых корпусов.

Список источников

1. Firth, J.R. Selected Papers of J. R. Firth 1952-1959. Palmer, F.R. (ed.), Longman. London. 1968

2. Deese, J. The Structure of Associations in Language and Thought. The John Hopkins Press. Baltimore. 1965

3. Cramer, P. Word association. NY: Academic Press, 1968

4. Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. Word Association, Rhyme and Fragment Norms.The University of South Florida, 1999 http://luna.cas.usf.edu/~nelson/

5. Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), The Computer and Literary Studies. Edinburgh: University Press, 1973

6. Okamoto and S. Ishizaki.. Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary (" 概念間距離の定式化と既存電子化辞書との比較 自然言語処理"), vol. 8, pages 37-54, 2001 (Japanese)

7. Joyce, Terry. Building a word association database for basic Japanese vocabulary. Poster session presented at The 3rd Annual Meeting of the Japanese Society for Cognitive Psychology, 28-29 May, Kanazawa University, Kanazawa, Japan, 2005

8. Lönngren, L. A Swedish Associative Thesaurus. Euralex '98 Proceedings, Vol. 2, pages 467-474, 1998

9. Леонтьев А.А. Словарь ассоциативных норм русского языка. Москва, 1977

10. Караулов Ю.Н., Тарасов Е.Ф., Сорокин Ю.А., Уфимцева Н.В., Черкасова Г.А. Ассоциативный тезаурус современного русского языка. РАН, 1999

11. Novák, Z. Volné slovní párové asociace v češtině. Praha. 1988 (Czech)

12. Jung J., Na L., Akama H. Network analysis of korean associations,Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, - Los Angeles, CA, 2010, pages 27-35, 2010

13. De Groot, A. M. B. Word association norms with response times(“Woordassociatienormen met reactietijden”). Nederlands Tijdschrift voor de Psychologie (Dutch Journal of Psychology), 43, pages 280-296, 1988

14. Rubinsten, O., Anaki, D., Henik, A., Drori, S., Faran, Y. Free association norms in the Hebrew language. In A. Henik, O. Rubinsten, & D. Anaki,

(Eds.). Word Norms in Hebrew, Ben-Gurion University of the Negev, pages 17-34, 2005 (Hebrew)

15. Milgram, S. The small-world problem. Psychology Today, 2, pages 60–67, 1967

16. Erdös, P., & Réyni, A. On the evolution of random graphs. Publications of the Mathematical Institute of the hungarian Academy of Sciences, 5, pages 17–61, 1960

17. Watts, D. J., Strogatz, S. H. Collective dynamics of 'small-world' networks. Nature, 393, pages 440–442, 1998

18. Amaral, L., Scala A., Barthelemy, M., Stanley, H. Classes of small-world networks.Proc. Natl. Acad. Sci. U. S. A. 97, pages 11149-11152, 2000