Анализ ассоциативных тезаурусов
и возможность их применения в задачах машинного перевода

Выломова Екатерина, *evylomova@gmail.com*

АИСТ

Материал

Для исследования использовался Русский Ассоцитиавный Словарь объемом 102516 различных стимульно-реактивных пар. Данные представляли собой набор триплетов 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 соответствующей пары. Впоследствии данная частота была заменена на относительный вес $weight_{ij} = \frac{freq_{ij}}{\sum_{j=1}^{n} freq_{ij}}$.

Методы

1. На основе тезауруса был построен ассоциативный граф G=(V,E), вершинами которого являлись концепты («язык», «Великая Отечественная Война»), а ребрами – соответствующие ассоциации между ними;
2. Был проведен анализ графа;
3. С использованием методов LSA и MDS было произведено преобразование концептуального пространства в векторное;
4. Полученные вектора были кластеризованы алгоритмом k-means;
5. Центрами кластеров являлись базовые концепты такие, как «человек», «язык», «время»

Поддержка

Работа выполняется в рамках гранта РГНФ №12-04-12039в

Введение

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.

Ассоциативные эксперименты

Данные: <стимул, реакция, частота стимульно-реактивной пары>

Впервые проведен Френсисом Гальтоном в 1879 году.

К текущему моменту АЭ были проведены в США, Великобритании, России, Чехии, Югославии, Швеции, Бельгии, Нидерландах, Японии, Южной Корее, Израиле.

Результаты

В рамках работы были получены следующие результаты:

- Ассоциативные сети и, в частности, РАС относятся к классам "small-world" и "scale-free" сетей
- Ассоциативные сети содержат вершины-хабы, являющиеся базовыми концептами, общими для различных языков

Применение в машинном переводе

На основе полученных выводов предлагаются следующие направления применения в задачах машинного перевода:

- Проверка гипотез :

  читать лекции->read lectures vs give lectures

  старый мужчина -> gever zaken

  старый стол -> shulhan yashan

- Ontology mapping: построение отображения графа на русском языке в граф на английском или каком-либо другом языке. Раскрытие омонимии:

  английский язык(tongue, language)-> english language

«Small-World»

Milgram, 1967: «6 степеней разделения»

$L \propto \log(N)$, где L- средняя длина кратчайшего пути между парами вершин, N – количество вершин.

К ним относятся: система энергоснабжения США, международная сеть киноактеров, нервная система червя Caenorhabditis elegans,WWW, сети научных сотрудничеств, социальные сети.

«Scale-Free»

Amaral, Scala, Barthélémy, Stanley , 2000

$P(k) \approx k^{-\gamma}$, $\gamma \in (2..4)$, где P(k) – функция распределения степени вершин

К ним относятся: социальные сети, WWW, международная сеть киноактеров, семантические сети, системы аэропортов

Сравнение сетей

	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

N – общее количество вершин
L – средняя длина кратчайшего пути
D – диаметр сети
γ – показатель в функции распред-я
<k> – средняя степень вершины