ИСПОЛЬЗОВАНИЕ ГРАФОВ ГОРИЗОНТАЛЬНОЙ ВИДИМОСТИ ДЛЯ ВЫЯВЛЕНИЯ СЛОВ, ОПРЕДЕЛЯЮЩИХ ИНФОРМАЦИОННУЮ СТРУКТУРУ ТЕКСТОВ НА РАЗЛИЧНЫХ ЯЗЫКАХ

Ландэ Дмитрий Владимирович
Институт проблем регистрации информации НАН Украины, Украина, Киев
dwlande@gmail.com

Снарский Андрей Александрович
НТУУ «Киевский политехнический институт», Украина, Киев
asnarskii@gmail.com

Ягунова Елена Викторовна
Санкт-Петербургский гос. университет, Санкт-Петербург, Россия
iagounova.elena@gmail.com

Предлагается использование предложенного авторами компактифицированного графа горизонтальной видимости для создания сети слов и выявления информационно-значимых слов в текстах литературных произведений и в их переводах. Обнаружена большая степень совпадения значений таких слов для одних и тех же произведений, представленных на различных языках.

В [1] приведено описание алгоритма формирования компактифицированного графа горизонтальной видимости для создания сети слов. Данный подход позволяет определять, какие из важных структурных элементов текста оказываются также информационно-значимыми, определяющими его информационную структуру.

В рамках теорий цифровой обработки сигналов (Digital Signal Processing) и сложных сетей (Complex Network) [2] предложено несколько методов построения сетей на основе временных рядов, среди которых можно назвать несколько методов построения графов видимости (см. обзор [3]), в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG). Эти подходы также позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие числовые значения. В качестве функции, ставящей в соответствие слову число, можно рассматривать, например, порядковый номер уникального слова в тексте, «вес» слов в текстах, различные дисперсионные оценки, общепринятую оценку TFIDF или ее варианты [4] и т.д.

При построении сетей слов в данной работе также будет использована дисперсионная оценка важности слов [5], которая реализуется следующим образом:

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

где: $\langle \Delta A \rangle$ – среднее значение последовательности $\Delta A_1, \Delta A_2, ..., \Delta A_K$, $\langle \Delta A^2 \rangle$ – последовательности $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_K^2$, K –количество появления слова A в тексте.

По сути, дисперсионная оценка позволяет отделить слова, встречающиеся в тексте относительно равномерно, от слов, распределенных неравномерно.

В отличие от остальных рядов, изучаемых в рамках цифровой обработки сигналов, ряды из численных значений, соответствующих словам, преобразуются в графы горизонтальной видимости, в которых узлам соответствуют также сами слова, выражающие определенное смысловое значение.

Сеть слов с использованием алгоритма горизонтальной видимости строится в три этапа. На первом на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (визуально – набор вертикальных линий). На втором этапе строится традиционный граф горизонтальной видимости. В этом случае между узлами устанавливается связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. На третьем, заключительном этапе, полученный граф компактифицируется. Все узлы с данным словом объединяются в один узел. Все связи таких узлов также объединяются (кратные связи изымаются). В результате получается новая сеть слов – *компактифицированный граф горизонтальной видимости* (КГГВ) – рис. 1.

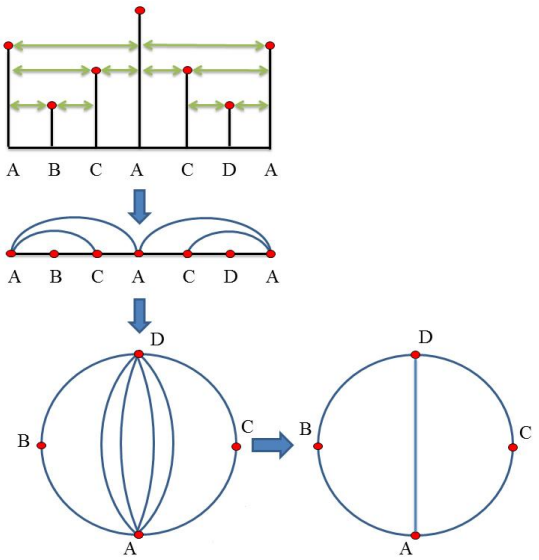


Рис. 1. Этапы построения компактификационного графа горизонтальной видимости

Для всех исследованных КГГВ-сетей слов было определено распределение степеней узлов (количества входящих связей), которое оказалось близким к степенному ($p(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Узлы КГГВ-сети с наибольшими степенями и рассматриваются в рамках данной модели как наиболее информационно-значимые, определяющие информационную структуру текста слова.

В качестве иллюстрации метода рассмотрим тексты повестей Джона Рональда Руэла Толкина (John Ronald Reuel Tolkien) «Хоббит, или Туда и обратно» (The Hobbit or There and Back Again) и Рэймонда Дугласа (Рэя) Брэдбери (Raymond Douglas (Ray) Bradbury) «Вино из одуванчиков» (Dandelion Wine) , представленных на оригинале – английском языке и в переводах на русский и украинский.

Следует, отметить, что авторами проводились подобные исследования на базе десятков других произведений самых разных объемов. Анализировались также законодательные акты Украины и России. Концептуальные результаты анализа при этом совпадали с приведенными ниже, поэтому остановимся на предложенных произведениях, как примерах. В состав узлов с наибольшими степенями в для КГГВ-сетей, наряду с личными

местоимениями и другими служебными словами (частицы, предлоги, союзы и т.д.), попали слова, определяющие информационную структуру текста [6].

Для сравнения исследовано поведение простейших сетей языка, когда не первом этапе построения сети связываются соседние слова, входящие в текст, а на втором происходит компактификация сети. В этом случае самые большие степени имеют узлы, соответствующие словам с наибольшей частотой – союзам, предлогами и т.п., имеющим большое значение для связности текста, но малоинтересным с точки зрения информационной структуры.

Если обозначить S – множество слов в стоп-словарях, соответствующих языкам исследуемых текстов, а Λ – множество из слов, соответствующих наиболее весомым узлам КГГВ, то множество $\Omega = \Lambda \setminus S$ соответствует информативным словам, имеющим, кроме того, важное значение и для связности текста. Ниже приведены сопоставления 20 наиболее весомых узлов для КГГВ-сетей слов (построенных на основе дисперсионных оценок слов) по указанным выше повестям. В частности, в КГГВ-сети по повести Дж. Толкина в список 12 наиболее весомых узлов (в порядке убывания весов) попали слова:

Английский: BILBO, GANDALF, THORIN, GOBLINS, DWARVES, MOUNTAIN, DOOR, DRAGON, FOREST, GOLLUM, ELVES, SMAUG.

Русский: БИЛЬБО, ГЭНДАЛЬФ, ТОРИН, ГОБЛИНЫ, ГНОМЫ, ГОЛЛУМ, ЭЛЬФЫ, ДРАКОН, ПОНИ, БЭРД, ГОРЫ, КОРОЛЬ.

*Украинский:*БІЛЬБО, ГАНДАЛЬФ, ГНОМИ, ТОРІН, ГОБЛІНИ, ГОРИ, ГАМ (GOLLUM в украинском переводе), ЕЛЬФИ, ГОБІТ, ДРАКОН, ДВЕРІ, ЧАРІВНИК.

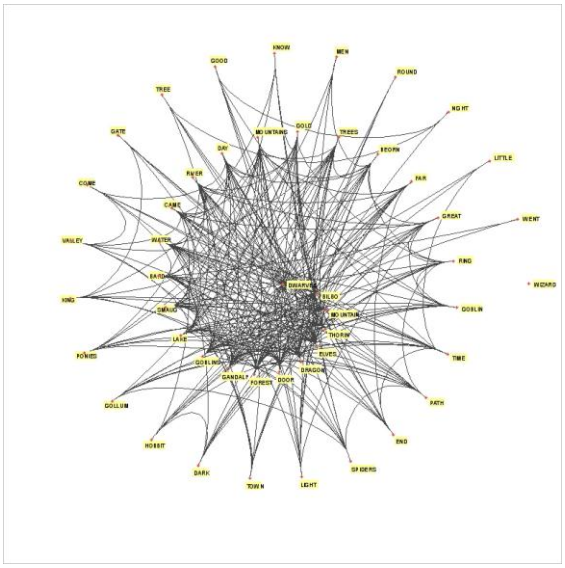
Соответственно, в КГГВ-сети по повести Рэя Брэдбери в список 12 наиболее весомых узлов попали слова:

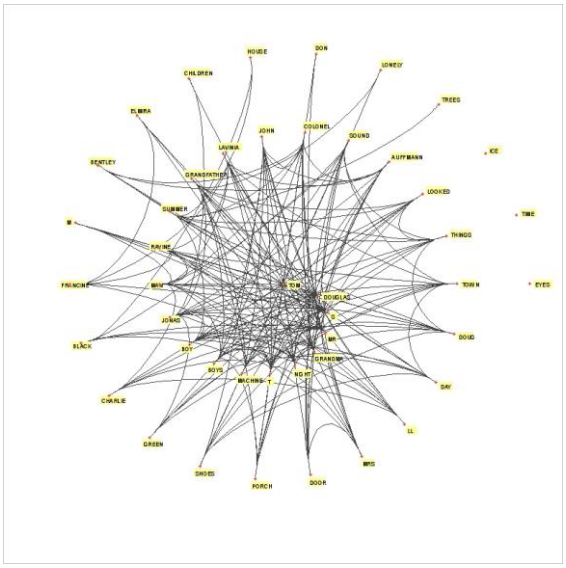
Английский: DOUGLAS, TOM, GRANDMA, LAVINIA, NIGHT, JOHN, MAN, RAVINE, MACHINE, MRS, AUFFMANN, ELMIRA.

Русский: ДУГЛАС, ТОМ, ЛАВИНИЯ, БАБУШКА, МИССИС, ЛЕО, ДЖОН, ДЕДУШКА, БЕНТЛИ, ПОЛКОВНИК, ЭЛЬМИРА, ГЛАЗА.

Украинский: ДУГЛАС, ТОМ, ЛЕО, МІСІС, ДІДУСЬ, ОЧІ, ДЖОН, БАБУСЯ, ЕЛМІРА, ЛАВІНІЯ, БЕНТЛІ, ЧАРЛІ.

На рис. 2 приведены примеры визуализации фрагментов КГГВ-сетей, соответствующих рассмотренным произведениям.





а) б)

Рис. 2. Фрагмент КГГВ-сети, соответствующей повестям «Хоббит, или Туда и обратно» (а) и «Вино из одуванчиков» (б)

В результате проведенных исследований сетей:

1. Реализован алгоритм построения компактифицированного графа горизонтальной видимости (КГГВ).
2. На основе последовательности дисперсионных оценок слов, с помощью метода КГГВ, построены сети слов различных текстов.
3. Для литературных текстов среди узлов соответствующих КГГВ-сетей с наибольшими степенями присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражают семантику литературных произведений.
4. Обнаружена большая степень совпадения значений наиболее весомых слов КГГВ-сетей, которые построены из одних и тех же произведений, представленных на различных языках.

Литература

1. Ландэ Д.В., Снарский А.А., Ягунова Е.В. Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текста // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г. – С. 67-76.
2. Strogatz S. H. Exploring Complex Networks // Nature. – 410. – P. 268-276 (2001).
3. Nunez A. M., Lacasa L., Gomez J. P., Luque B. Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6. – P. 119 – 152 (2012).
4. Salton G., McGill M. J. Introduction to Modern Information Retrieval. – New York: McGraw-Hill. – 448 p. (1983).
5. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, – 57(5). – P. 759-764 (2002).
6. Черняховская Л.А. Смысловая структура текста и ее единицы // Вопросы языкознания. – № 6. – С. 118–126. (1983).