OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822: 004.912

ПРИМЕНЕНИЕ КГГВ-АЛГОРИТМА ДЛЯ НАУЧНЫХ ТЕКСТОВ

Ландэ Д.В.[*], Снарский А.А.[*], Ягунова Е.В.[**]

Институт проблем регистрации информации НАН Украины,
Национальный технический университет Украины «Киевский политехнический институт»,
г. Киев, Украина
DWLande@gmail.com
ASnarskii@gmail.com

[**] *Санкт-Петербургский государственный университет,*
Санкт-Петербург, Россия
Iagounova.Elena@gmail.com

Описывается применение предложенной авторами методики построения компактифицированного графа горизонтальной видимости (КВГГ) для выявления в научных текстах тех слов, которые определяют не только их структурную связность, но и информационную структуру. Сравниваются результаты, полученные на основе применения КГГВ-алгоритма и алгоритма построения простой сети слов.

Ключевые слова: опорные слова; научные тексты; граф горизонтальной видимости; информационная структура.

Введение

Построение сетей слов (Language Networks), узлами которых являются термины – слова или словосочетания, фрагменты естественного языка, уже традиционно позволяет выявлять структурные элементы текста, без которых он теряет свою связность [Ferrer-i-Cancho, 2001], [Ландэ и др., 2013]. При этом некоторые из важных структурных элементов текста оказываются также информационно-значимыми, определяющими информационную структуру [Солганик, 1991], [Черняховская, 1983]. Такие элементы, (будем называть их «опорными словами») могут использоваться для построения информационных портретов [Ягунова, 2010], идентификации не достаточно четко определенных компонент текста, таких как коллокации, сверхфразовые единства [Giora ,1983], [Ягунова, 2012].

Авторами были проведены исследования применения алгоритма КГГВ для выявления опорных слов для художественных текстов [Lande etc, 2013], однако вопрос о возможности его применения для массивов научных текстов относительно небольшого объема (в частности, материалов конференций) до сих пор остается открытым.

Графы горизонтальной видимости

В рамках теорий цифровой обработки сигналов (Digital Signal Processing) и сложных сетей (Complex Network) [Albert, 2002], [Strogatz, 2001] предложено несколько методов построения сетей на основе числовых рядов, среди которых можно назвать семейство методов построения графов видимости, в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG) [Gutin, 2011], [Luque, 2009]. Эти подходы также позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом ставятся в соответствие некоторые начальные числовые весовые значения.

При построении сетей слов в данной работе использована дисперсионная оценка важности слов [Ortuño etc, 2002]. Пусть текст состоит из N слов ($n = 1,...N$, n – порядковый номер слова в тексте, позиция слова). Обозначим средний интервал (количество слов) между появлениями слова A в тексте через $<\Delta A>$, а средний квадрат значений этих интервалов через $<\Delta A^2>$. Дисперсионная оценка слова A – σ_A рассчитывается как

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

По сути, дисперсионная оценка позволяет отделить слова, встречающиеся в тексте относительно равномерно (для равномерно распределенных слов эта оценка равна нулю), от слов, распределенных неравномерно. Т.е. это оценка различительной, дискриминантной силы слов, в частности, для информационного поиска. Идея дисперсионной оценки очень близка к традиционной оценке TFIDF, однако более корректно применима к полным единичным текстам, а не к массивам текстов, как TFIDF.

Сеть слов на основе КГГВ-алгоритма

В отличие от обычных числовых рядов, изучаемых в рамках цифровой обработки сигналов, ряды из цифровых значений, соответствующих словам, преобразуются в графы, узлам которых соответствуют не только цифровые значения, но 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, см. рис.1).

На втором этапе строится традиционный граф горизонтальной видимости. При этом считается, что между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. Этот геометрический критерий можно записать следующим образом: два узла (слова) слова, например, B с весом σ_B и C с весом σ_C соединены ребром, если $\sigma_B, \sigma_C > \sigma_X$ для всех слов X с весом σ_X, расположенных между словами B и C.

На третьем этапе, полученная на предыдущем этапе сеть компактифицируется. Все узлы с данным словом, например словом A, объединяются в один узел. Все связи таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более одного ребра, кратные ребра изымаются. В результате получается новая сеть слов – *компактифицированный граф горизонтальной видимости* (см. рисунок 1).

На заключительном этапе формирования КГГВ также отфильтровываются слова, важные для согласованности текста, но не имеющие информационной значимости. Для этого использовался так называемый «стоп-словарь», сформированный на основе агрегации информации, доступной по адресам:

- *http://code.google.com/p/stop-words/source/browse/trunk/stop-words/stop-words/stop-words-russian.txt?spec=svn3&r=3*
- *https://github.com/punbb/langs/blob/master/Russian/stopwords.txt*
- *http://www.ranks.nl/stopwords/russian.html*
- *https://trac.mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.txt*

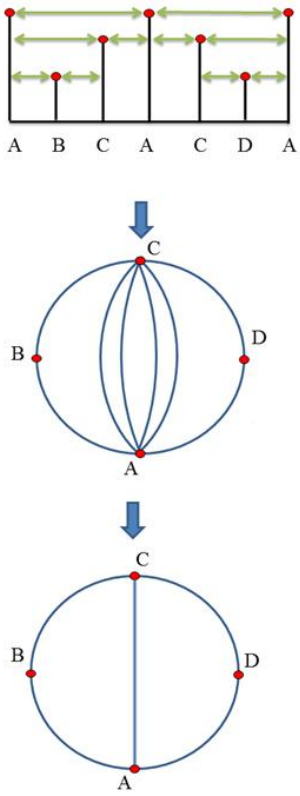

Рисунок 1 – Этапы построения компактификационного графа горизонтальной видимости

Следует отметить, что в рамках предложенного алгоритма не рассматривалась нормализация слов, т.е. в качестве разных узлов иногда использовались различные словоформы одного и того же слова.

Эксперимент

В качестве корпуса научных текстов авторами использовались труды международной научной конференции «Корпусная лингвистика – 2008», в частности тексты таких докладов:

- Е.А. Сидорова «Подход к построению предметных словарей по корпусу текстов» (длина текста – 10946 символов, включая пробелы);
- В.И. Шестопалова, Т.И. Петрова, М.А. Болгов «Региональный вариант живой русской речи как объект корпусной лингвистики» (12948 символов);
- В.Ф. Выдрин «На пути к электронному корпусу языка Бамана: обозначение тонов» (21633 символов);
- Е.В. Падучева «Прямая и косвенная диатеза ментального глагола: корпусное исследование» (24761 символов).

Для всех указанных произведений построены сети слов, центральные фрагменты которых визуально представлены на рисунках 2-5.

Для сравнения исследовано также поведение простейших сетей языка для научных текстов. Для построения этих сетей на первом этапе связываются ребрами соседние слова, а на втором происходит компактификация сети. Очевидно, вес узлов в этой сети соответствует частоте встречаемости слов, а их распределение – закону Ципфа. При этом очевидно, что самые большие степени имеют узлы, соответствующие словам с наибольшей частотой, зачастую имеющим большое значение для связности текста, но малоинтересным с точки зрения информационной структуры.

В таблицах 1-4 приведено сопоставление наиболее весомых узлов КГГВ и простых сетей языка (с учетом фильтрации по стоп-словарю), построенных для приведенного выше списка научных текстов. Выделенные жирным шрифтом слова соответствуют специальной лексике, информационно значимой для рассматриваемых произведений. Обычным шрифтом выделена общенаучная лексика, курсивом – слова, не несущие по мнению асессоров важного информационного смысла. При этом оказалось, что чем больше длина рассматриваемого текста, тем выше качество выбора опорных слов с помощью КГГВ-алгоритма, тем он более эффективен, чем метод выбора наиболее весомых слов по простой сети языка.

Небольшой по объему доклад Е.А. Сидоровой (таблица 1, рисунок 2) посвящен построению предметных словарей по корпусу текстов, поэтому вполне логично, в состав наиболее весомых узлов соответствующих сетей попали термины, такие как: ТЕРМИН, СЛОВАРЬ, ТЕКСТ, МОДУЛЬ, ТЕХНОЛОГИЯ РАЗМЕТКА, ОБУЧЕНИЕ, АВТОМАТИЧЕСКАЯ, КОНКОРДАНС, ИЕРАРХИЯ, КОРПУС. Вместе с тем, ввиду небольшого размера текста, в состав наиболее весомых узлов попала значительная часть общенаучных слов.

Таблица 1 – Наиболее весомые узлы КГГВ и простой сети языка по тексту доклада Е.А. Сидоровой

№	КГГВ	№	Простая сеть языка
1	**ТЕРМИН**	1	**ТЕКСТ**
2	**СЛОВАРЬ**	2	**СЛОВАРЬ**
3	**ТЕКСТ**	3	**ТЕРМИН**
4	**МОДУЛЬ**	4	**ОБУЧЕНИЕ**
5	**ТЕХНОЛОГИЯ**	5	**МОДУЛЬ**
6	**РАЗМЕТКА**	6	*СЛЕДУЮЩИЕ*
7	ОСНОВА	7	**КОРПУС**
8	**ОБУЧЕНИЕ**	8	**СТАТИСТИКА**
9	**АВТОМАТИЧЕСКАЯ**	9	**АНАЛИЗ**
10	РИС	10	СОЗДАНИЕ
11	**АНАЛИЗ**	11	РИС
12	**КОНКОРДАНС**	12	**РАЗМЕТКА**
13	КОЛИЧЕСТВО	13	*ЯВЛЯЕТСЯ*
14	**ИЕРАРХИЯ**	14	ПОЛЬЗОВАТЕЛЬ
15	**КОРПУС**	15	*НАЛИЧИЕ*
16	СОЗДАНИЕ	16	**КОНКОРДАНС**
17	*СЛЕДУЮЩАЯ*	17	**КЛАССИФИКАЦИЯ**
18	**СТАТИСТИКА**	18	ВСТРЕЧАЕМОСТЬ
19	**МЕТОД**	19	**СХЕМА**
20	**МЕХАНИЗМ**	20	**МЕХАНИЗМ**
21	**СООБЩЕНИЕ**	21	**ИЕРАРХИЯ**
22	*ЯВЛЯЕТСЯ*	22	**АВТОМАТИЧЕСКАЯ**
23	КОНФЕРЕНЦИЯ	23	**ВЫБОРКА**
24	**СХЕМА**	24	*ВКЛЮЧАЕТ*
25	**СИСТЕМА**	25	**ТЕМАТИЧЕСКАЯ**
26	**РАСПРЕДЕЛЕНИЕ**	26	**ТЕХНОЛОГИЯ**

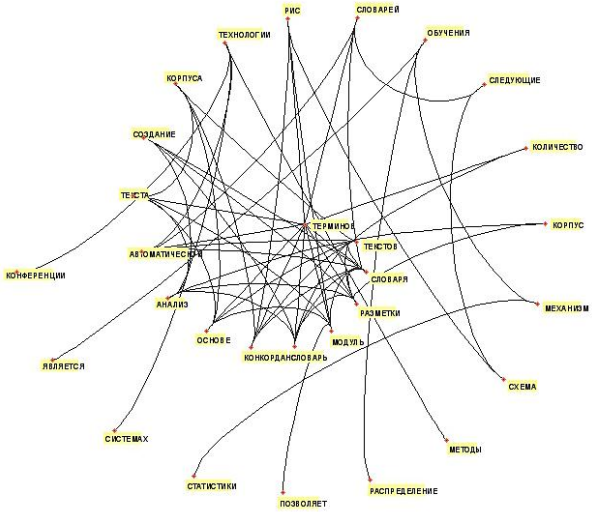

Рисунок 2 – центральный фрагмент КГГВ по докладу Е.А. Сидоровой

Также небольшой доклад В.И. Шестопаловой, Т.И. Петровой, М.А. Болгова (таблица 2, рисунок 3) посвящен региональному варианту живой русской речи как объекту корпусной лингвистики. В состав наиболее весомых узлов в этом случае попали такие термины: РЕЧЬ, ГОРОД, ТЕКСТ, ЯЗЫК, ИНФОРМАНТ, ЖИВОЙ, КОРПУС, РУССКИЙ. Вместе с тем, и в этом случае велика часть общенаучных слов.

Таблица 2 – Наиболее весомые узлы КГГВ и простой сети языка по тексту доклада В.И. Шестопаловой, Т.И. Петровой и М.А. Болгова

№	КГГВ	№	Простая сеть слов
1	**РЕЧЬ**	1	**РЕЧЬ**
2	**ГОРОД**	2	**ГОРОД**
3	МАТЕРИАЛ	3	МАТЕРИАЛ
4	**ТЕКСТ**	4	**ТЕКСТ**
5	**ЯЗЫК**	5	**ЖИВОЙ**
6	**ИНФОРМАНТ**	6	**ЯЗЫК**
7	**ЖИВОЙ**	7	ПРОБЛЕМА
8	**КОРПУС**	8	**ИНФОРМАНТ**
9	**РУССКИЙ**	9	**КОРПУС**
10	ВЫДЕЛЯЕМ	10	**УСТНЫЙ**
11	ПРОБЛЕМА	11	СПЕЦИФИКИ
12	**УСТНЫЙ**	12	*СОЗДАНИЕ*
13	**НТИ**	13	**РЕГИОН**
14	**РЕГИОН**	14	*ЯВЛЯЕТСЯ*
15	СОЦИАЛЬНЫЙ	15	ОТБОР
16	**СЕР**	16	*ОСОБЕННОСТИ*

17	РЕШЕНИЕ	17	**НКРЯ**
18	*ЯВЛЯЕТСЯ*	18	*ВОЗМОЖНОСТИ*
19	**НКРЯ**	19	СБАЛАНСИРОВАННОС
20	*ВОЗМОЖНОСТ*	20	ТЬ
21	*И*	21	ОСНОВА
22	РАЗМЕЩЕНИЕ	22	ИССЛЕДОВАНИЕ
23	**ДАННЫЕ**	23	ВЫДЕЛЯЕМ
24	*УЧЕТ*	24	**ТЕРРИТОРИАЛЬНЫЙ**
25	ПРИНЦИП	25	*СЧИТАЕМ*
26	**ПОДКОРПУС**	26	СОЦИАЛЬНЫЙ
	ИССЛЕДОВАН		*СИТУАЦИЯ*
	ИЕ		

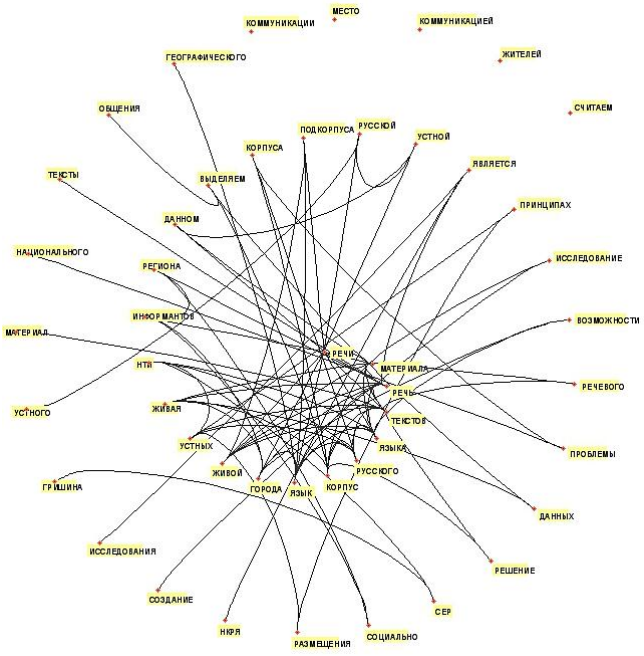

Рисунок 3 – центральный фрагмент КГГВ по докладу В.И. Шестопаловой, Т.И. Петровой и М.А. Болгова

Средний по размеру доклад В.Ф. Выдрина (таблица 3, рисунок 4) посвящен электронному корпусу языка Бамана. В состав наиболее весомых узлов в этом случае попали такие термины: БАМАНА, ТОН, АРТИКЛЬ, ЯЗЫК, ТЕКСТ, СЛОВО, СТЕПЕНЬ. Как можно видеть, в данном случае абсолютное большинство выбранных слов имеет специальный, информационно значимый для данного доклада характер.

Таблица 3 – Наиболее весомые узлы КГГВ и простой сети языка по тексту доклада В.Ф. Выдрина

№	КГГВ	№	Простая сеть слов
1	**БАМАНА**	1	**БАМАНА**
2	**ТОН**	2	**ТОН**
3	**АРТИКЛЬ**	3	**СЛОВО**
4	**ЯЗЫК**	4	*ОКАЗЫВАЕТСЯ*
5	**ТЕКСТ**	5	**АРТИКЛЬ**
6	**СЛОВО**	6	**ТЕКСТ**
7	**СТЕПЕНЬ**	7	*ФАКТИЧЕСКИ*
8	**DUMESTRE**	8	*ЯВЛЯЕТСЯ*
9	**BAMBARA**	9	*ОБОЗНАЧАТЬ*
10	*ОКАЗЫВАЕТСЯ*	10	ЯЗЫК
11	*ЯВЛЯЕТСЯ*	11	ПУБЛИКАЦИИ
12	**ФРАЗА**	12	*ОБРАЗОМ*
13	**ЭЛЕМЕНТ**	13	**СВЯЗИ**
14	*ИМЕЮТСЯ*	14	*СУЩЕСТВЕННО*
15	*НИЗКИЙ*	15	**СТЕПЕНЬ**
16	**СВЯЗИ**	16	ОБОЗНАЧЕНИЕ
17	КОМПОНЕНТА	17	*НИЗКИЙ*
18	**ПРИЛАГАТЕЛЬНЫХ**	18	**КОРПУСА**
19	ОБОЗНАЧЕНИЕ	19	КОЛИЧЕСТВО
20	*ФАКТИЧЕСКИ*	20	**BAMBARA**
21	**ТОНАЛЬНЫЙ**	21	**ТОНАЛЬНЫЙ**
22	*ОТСУТСТВИЕ*	22	КОМПОНЕНТА
23	**MALI**	23	*ИМЕЮТСЯ*
24	**КОРПУС**	24	*ДОМ*
25	**КЛАСС**	25	**ЭЛЕМЕНТА**
26	**ГЛОССАРИЙ**	26	*ВЫСОКИЙ*

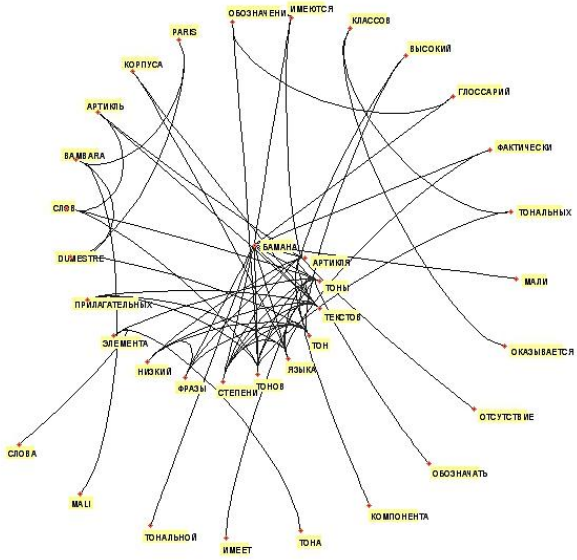

Рисунок 4 – центральный фрагмент КГГВ по докладу В.Ф. Выдрина

Самый большой из рассматриваемых по размеру доклад Е.В. Падучевой (таблица 4, рисунок 5) посвящен прямой и косвенной диатезе ментального глагола. В состав наиболее весомых узлов в данном случае попали многие из рассматриваемых глаголов, которые также являются информационно значимыми для данного доклада.

Таблица 4 – Наиболее весомые узлы КГГВ и простой сети языка по тексту доклада Е.В. Падучевой

№	КГГВ	№	Простая сеть слов
1	**ИВАН**	1	**МНЕНИЕ**
2	**АКЦЕНТ**	2	**ГЛАГОЛ**
3	**ПРЕДПОЛАГАТЬ**	3	ЗНАНИЯ
4	**МНЕНИЕ**	4	**ПРЕДПОЛАГАТЬ**
5	**ГЛАГОЛ**	5	**МАША**
6	**ПАДУЧЕВА**	6	**ПОДОЗРЕВАТЬ**
7	**ПОДОЗРЕВАТЬ**	7	**ИВАН**
8	**МАША**	8	ВОПРОС
9	**ДОГАДЫВАТЬСЯ**	9	**ПРЕСУППОЗИЦИЯ**
10	**ПОДЧИНЯТЬ**	10	**АКЦЕНТ**
11	ЗНАНИЯ	11	**ПАДУЧЕВА**
12	**ПРЕСУППОЗИЦИЯ**	12	**ПРОПОЗИЦИО-**
13	**АКТАНТ**		**НАЛЬНЫЙ**
14	**ДИАТЕЗ**	13	**КОНТЕКСТ**
15	ДОКАЗАТЬ	14	*ИМЕЕТ*
16	**КОНТЕКСТ**	15	**ГОВОРЯЩИЙ**

17	СИТУАЦИЯ	16	*ДУМАТЬ*
18	ВОПРОС	17	ДОКАЗАТЬ
19	**ГОВОРЯЩИЙ**	18	*ЯВЛЯЕТСЯ*
20	*ПРЯМОЙ*	19	**ПРОПОЗИЦИЯ**
21	**ДОГАДЫВАЕТСЯ**	20	**ДИАТЕЗ**
22	**ПРОПОЗИЦИЯ**	21	СИТУАЦИИ
23	УЕХАЛ	22	РЕМАТИЧЕСКИЙ
24	**СОЧЕТАЕМОСТЬ**	23	ДОГАДЫВАТЬСЯ
25	**ПРОПОЗИЦИОНАЛЬНЫЙ**	24	ВОСПРИЯТИЯ
		25	УПОТРЕБЛЕНИЯ
26	ЗНАЧЕНИЕ	26	**РЕЧЬ**

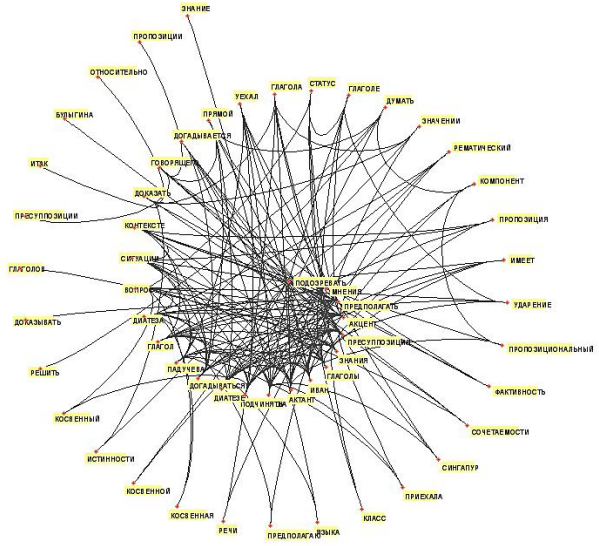

Рисунок 4 – центральный фрагмент КГГВ по докладу Е.В. Падучевой

Для всех построенных КГГВ-сетей слов было определено распределение степеней узлов, которое оказалось степенным ($p(x) = Cx^{\alpha}$), т.е. эти сети являются безмасштабными. Были проведены расчеты параметров сетей для всех рассмотренных научных текстов. В результате оказалось, что для всех из них коэффициент α изменялся в диапазоне от –2.2 до –2,4 при высокой точности аппроксимации степенного распределения, которая повышается при увеличении длины текста.

В состав узлов с наибольшими степенями в КГГВ, наряду со служебными словами, попали слова, определяющие их информационную структуру.

Следует, отметить, что авторами также проводились подобные исследования на базе десятков других произведений, в том числе, значительно более объемных. Анализировались также законодательные акты Украины и России.

Заключение

Представления об информационной значимости наборов слов, полученных с помощью КГГВ-алгоритма, их важности для понимания смысла научных текстов были подтверждены в ходе экспериментов с асессорами. Так, для всех текстов были проведены эксперименты со стандартной инструкцией «Прочитайте текст. Подумайте над его содержанием. Выпишите слова, наиболее важные для его содержания» (группа из 20 информантов) [Ягунова, 2012].

В результате проведенных исследований:

1. Предложен алгоритм построения компактифицированного графа горизонтальной видимости для научных текстов относительно небольших объемов.
2. На основе последовательности дисперсионных оценок слов с помощью предложенного метода построены сети языка для научных текстов различной длины.
3. Установлено, что построенные графы горизонтальной видимости характеризуются степенным распределением весов узлов, т.е. являются безмасштабными, в некотором смысле инвариантными к масштабу рассмотрения. Определены значения степенных коэффициентов α от -2,2 до -2,4.
4. Для научных текстов среди узлов соответствующих КГГВ с наибольшими степенями присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие их информационную структуру, отражающие семантику.
5. Качество выявления опорных слов с помощью КГГВ-алгоритма повышается при увеличении объемов текстов.

Библиографический список

[Ландэ и др., 2013] Ландэ Д.В. Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текста / Д.В. Ландэ, А.А. Снарский, Е.В. Ягунова // Труды 15-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г. – С. 315-321.

[Солганик, 1991] Солганик Г.Я. Синтаксическая стилистика. Сложное синтаксическое целое. – 2-е изд., испр. и доп. – М.: Высш. шк., 1991. – 182 с.

[Черняховская, 1983] Черняховская Л.А. Смысловая структура текста и ее единицы // Вопросы языкознания, 1983. – № 6. – С. 118–126.

[Ягунова, 2010] Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика / Отв. ред. В.Т. Юнгблюд. – Пермь, 2010. – Вып. 1. – С. 85-91.

[Ягунова, 2012] Ягунова Е.В. Вариативность структуры нарратива и разнообразие стратегий понимания // Человек и язык. К юбилею Тамары Ивановны Ерофеевой. Сб. научн. ст. Пермь: Изд-во "Прикамский социальный институт", 2012. – С. 65-83.

[Albert, 2002] Albert R., Barabasi A.-L. Statistical mechanics of complex networks // Reviews of Modern Physics, 2002. – 74. – P. 47.

[Ferrer-i-Cancho, 2001] Ferrer-i-Cancho R., Sole R. V. The small world of human language // Proc. R. Soc. Lond, 2001. – B 268, 2261.

[Giora ,1983] Giora R. Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam, 1983. – 3. – № 2. – P. 155-181.

[Gutin, 2011] Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatoris on words // Physica A, 2011, – 390 – P. 2421-2428.

[Lande etc, 2013] Lande D.V. Compactified HVG for the Language Network / D.V. Lande, A.A. Snarskii // // International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20-23 aug. 2013, Chi.in.u, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, 2013. - P. 108-113.

[Luque, 2009] Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs: Exact results for random time series // Physical Review E, 2009. – P. 046103-1–046103-11.

[Strogatz, 2001] Strogatz S.H. Exploring Complex Networks // Nature, 2001. – 410. – P. 268-276.

[Ortuño etc, 2002] Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, 2002, – 57(5). – P. 759-764.

APPLICATION OF THE CHVG-ALGORITHM FOR SCIENTIFIC TEXTS

Lande D.V. [*], *Snarskii A.A.*[*], *Yagunova E.V.*[**]

[*] *Institute for Information Recording NAS of Ukraine, NTUU "Kiev Polytechnic Institute" , Kiev , Ukraine*

DWLande@gmail.com

ASnarskii@gmail.com

[**] *Saint-Petersburg State University , St.-Petersburg, Russian Federation*

Iagounova.Elena@gmail.com

In work describes the application of the CHVG-algorithm for scientific texts. A CHVG-algorithm for identify the words that define the information structure of the text is proposed. It was found that the networks constructed in such way have a property that among the nodes with largest degrees there are words that determine not only a text structure communication, but also its informational structure.

INTRODUCTION

Along with "linear" text analysis, construction of a net with text elements such as words and word combinations as its nodes can help reveal the structural elements of a text which make it coherent. Finding those structural elements which also have informational significance and form informational structure of a text is an important problem.

The originality of the research is contained within the application of horizontal visibility graph used in digital signal processing to scientific texts. The proposed algorithm enables to extract the words which not only have informational significance but also are important for text coherence.

MAIN PART

According to the horizontal visibility algorithm, language network is built in three stages. First, a series of nodes are plotted on the horizontal axis with uniform spacing, each node corresponding to a word, in the order the words appear in the text. At the same time numeric weights (dispersion estimated value) corresponding to the words are plotted on the vertical axis. At the second stage a traditional horizontal visibility graph is built. Visibility is considered for the highest points of the nodes columns. An edge is put between the nodes, if there is a visible connection between them, e.g., if they can be connected by a horizontal line which does not cross any column.

At the third stage language network is compactified. All the nodes with the given word are merged into one node. All the edges of such nodes are also merged. It is important to note that multiple edges are removed, and there is no more than one edge left between every two nodes. As a result, we have a new language network – a compactified horizontal visibility graph.

For the sake of comparison we analyzed the simplest types of language networks where on the first stage of the algorithm neighboring words in the text are connected, and on the second one network compactification takes place. At the same time we have maximum degree nodes for the maximum frequency words which are of great importance for text coherence and of little interest for the informational structure of a text.

CONCLUSION

The following results were obtained from studying the language networks:

- An algorithm for constructing compactified horizontal visibility graph (CHVG) was proposed.

- Language networks for different scientific texts are built on the basis of dispersion estimated values and CHVG.

- For scientific texts the CHVG nodes with maximum degree correspond to the words which not only provide for text coherence but also determine its informational structure and the semantics of the pieces of literature.