Міжнародна науково-технічна конференція

ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ ЛІНГВІСТИЧНОГО АНАЛІЗУ

22-23 жовтня 2013 року

ПЕРСПЕКТИВЫ МЕТОДА "ГРАФОВ ВИДИМОСТИ" В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

Д.В. ЛАНДЭ, д.т.н.,
А.А. СНАРСКИЙ, д.ф.-м.н., профессор,
Институт проблем регистрации информации НАН Украины, Киев

Киев, 22 октября 2013 г.

Аннотация

Предлагается методика создания компактифицированного графа горизонтальной видимости. Исследованы свойства таких сетей слов, показано, что они являются безмасштабными, а также, что среди узлов с наибольшими степенями имеются слова, определяющие не только структуру связности текста, но и его информационную структуру.

Актуальность

Наряду с последовательным, «линейным» анализом текстов, построение сетей, узлами которых являются их элементы – фрагменты естественного языка, позволяет выявлять структурные элементы текста, без которых он теряет свою связность. При этом актуальной является задача определения того, какие из важных структурных элементов оказываются также информационно-значимыми, определяющими информационную структуру текста. Такие элементы могут использоваться также для идентификации еще не достаточно четко теоретически определенных компонент текста, таких как коллокации, сверхфразовые единства, например, при поиске подобных фрагментов в различных текстах.

Сети слов

Первым шагом при применении теории сложных сетей к анализу текста является представление этого текста в виде совокупностиузлов и связей, построение сети языка (Language Network).

Существуют различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям сети языка.

Узлы могут быть соединенны меду собой, если соответствующие им слова стоят рядом в тексте, принадлежат одному предложению, соединены синтаксически или семантически.

Простейшие сети слов

- **L-пространство**. Связываются соседние слова, которые принадлежат одному предложению. Количество соседей для каждого слова (окно слова) определяется радиусом взаимодействия R, чаще всего рассматривается случай R = 1.

B-пространство. Рассматриваются узлы двух видов, соответствующие предложениям и словам, которые им принадлежат.

P-пространство. Все слова, которые принадлежат одному предложению, связываются между собой.

C-пространство. Предложения связываются между собой, если в них употреблены одинаковые слова.

Из рядов - графы

На стыке теорий цифровой обработки сигналов (Digital Signal Processing) и сложных сетей (Complex Network) предложено несколько методов построения сетей на основе временных рядов, среди которых можно назвать несколько методов построения графов видимости, в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG). Эти подходы позволяют строить сетевые структуры также и на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие некоторые весовые значения.

Весовые оценки слов

В качестве функции, ставящей в соответствие слову из текста число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова, «вес» слов в текстах, общепринятую оценку TFIDF или ее варианты, а также другие весовые оценки, в частности, статистические дисперсионные.

TFIDF

В качестве весовой оценки из полного текста, состоящего из слов, текст разбивается на фрагменты, содержащие заданное количество N слов M (например, M = 500). Затем для каждого слова i, входящего в текст, подсчитывается количество фрагментов df(i), в которые это слово входит, а также общее количество вхождений данного слова i в текст – n(i). После этого рассчитывается среднее значение весовой оценки каждого слова в тексте, близкое по идеологии к классическому TFIDF:

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{M \times df(i)}\right)$$

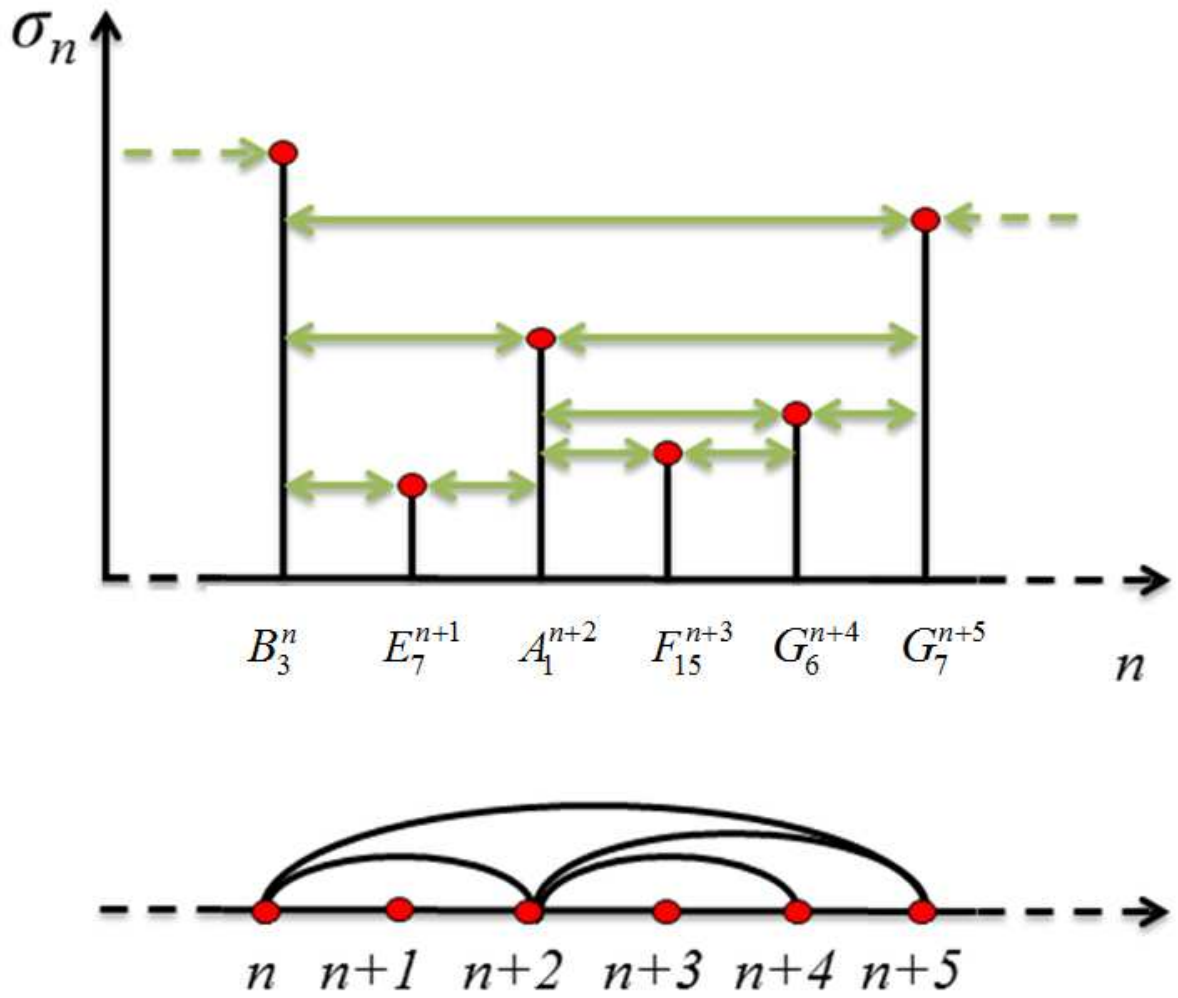
Дисперсионная оценка

Дисперсионная σ_A оценка для некоторого слова *A* из текста рассчитывается как

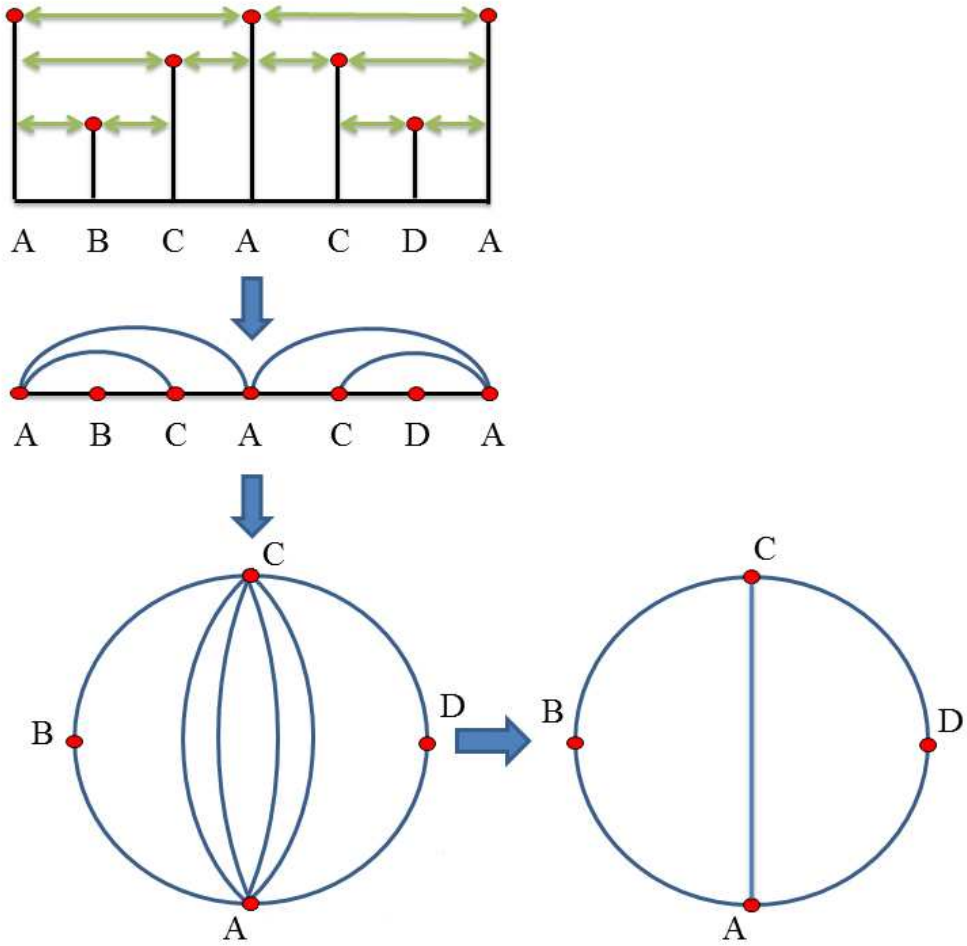
$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

где: $\langle \Delta A \rangle$ – среднее расстояние (в словах) между появлениями слова *A* в тесте; : $\langle \Delta A^2 \rangle$ – среднее квадрата расстояния между появлениями слова *A* в тексте.

Формирование графа горизонтальной видимости



Этапы построения КГГВ



Компактифицированный граф горизонтальной видимости

Основы оценки методов

Если обозначить Ψ – множество из N различных слов из «стоп-словаря» языка, а Λ -множество слов, соответствующих наиболее весомым узлам КГГВ, то множество $\Omega=\Lambda\setminus\Psi$ соответствует информативным словам, имеющим, кроме того, важное значение и для связности текста.

Наиболее весомые узлы КГГВ

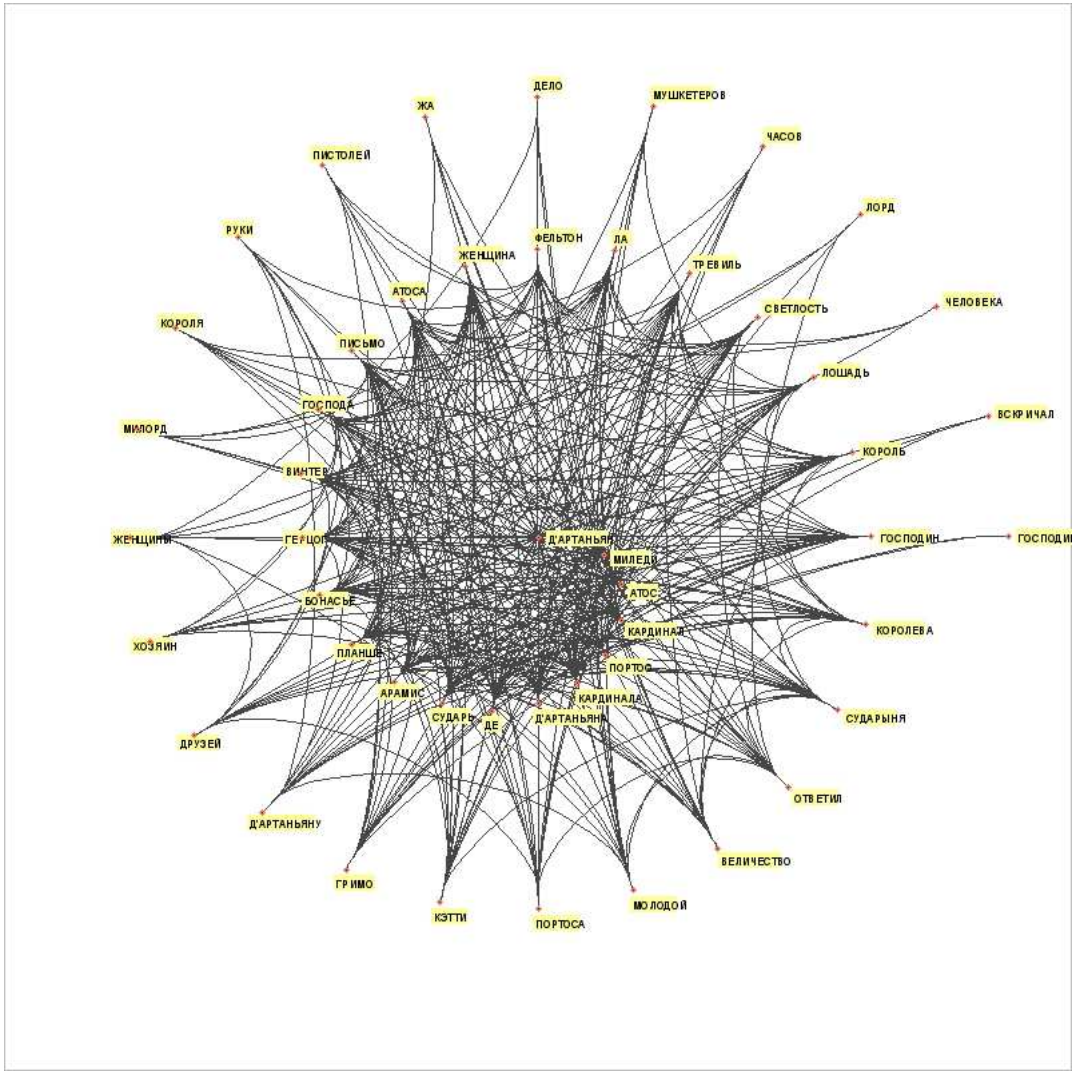
А. Дюма, «Три мушкетера»

14053 Д'АРТАНЬЯН
9347 МИЛЕДИ
9057 АТОС
4128 АРАМИС
3945 БОНАСЬЕ
3878 ПОРТОС
3755 ДЕ
3038 ФЕЛЬТОН
2599 КАРДИНАЛ
2384 СУДАРЬ
2248 ПЛАНШЕ
1940 ТРЕВИЛЬ
1689 ВЕЛИЧЕСТВО
1585 ПИСЬМО
1542 КОРОЛЬ
1400 ГЕРЦОГ
1342 ГОСПОДИН
1310 КЭТТИ

1301 СУДАРЫНЯ
1248 ЖЕНЩИНА
1229 ЛА
1128 КОРОЛЕВА
1033 ВИНТЕР
925 ГОСПОДА
898 ГРИМО
891 ЛОШАДЬ
888 МОЛОДОЙ
852 СВЕТЛОСТЬ
738 ДРУЗЕЙ
728 ЧЕЛОВЕКА
723 РУКИ
721 ХОЗЯИН
708 ДЕЛО
683 ЧАСОВ
669 ПИСТОЛЕЙ
657 МИЛОРД

Визуализация КГГВ

А. Дюма, «Три мушкетера»



Наиболее весомые узлы КГГВ

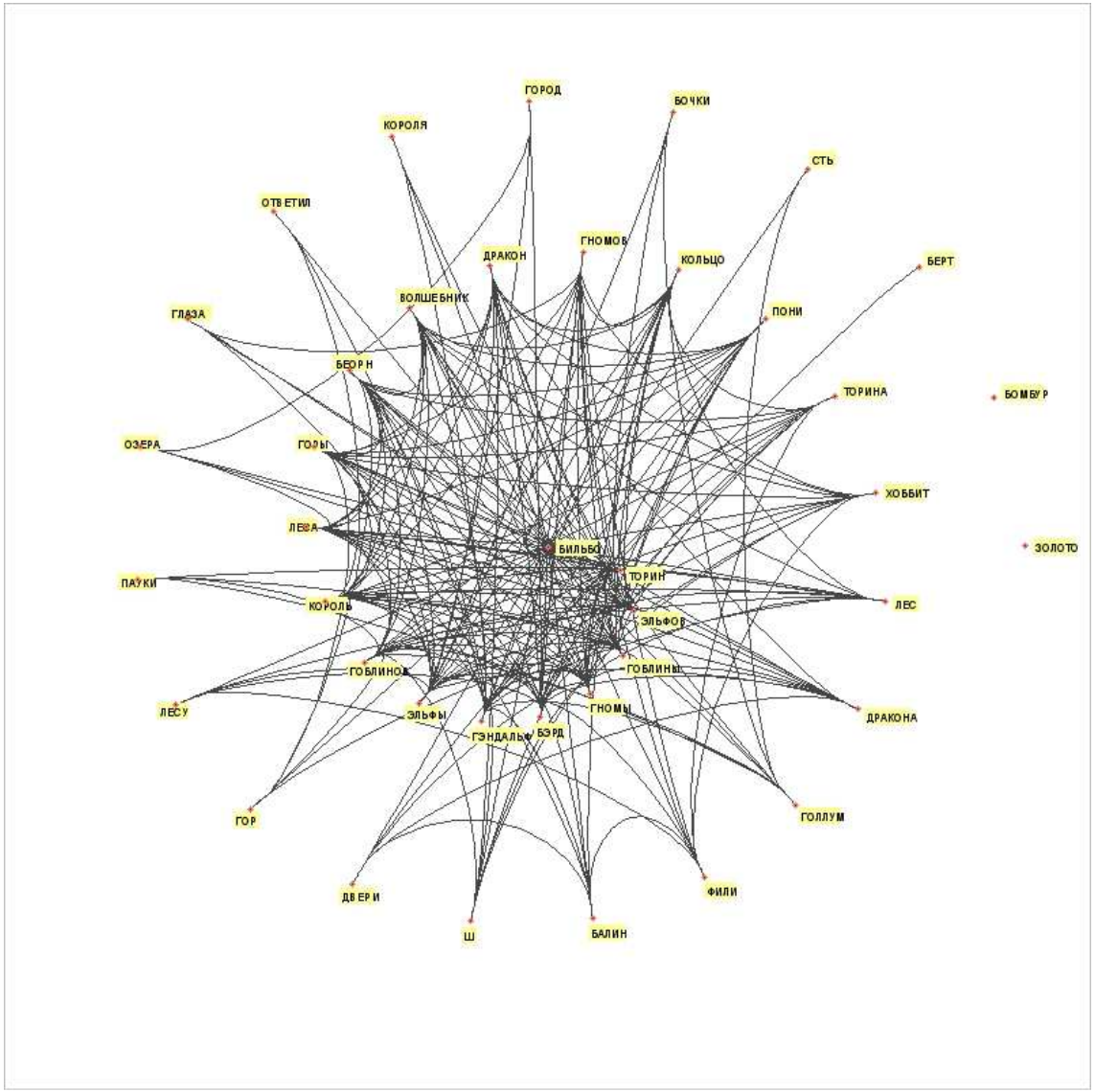
Дж. Толкиен, «Хоббит, или Туда и обратно »

5478 БИЛЬБО
2227 ГЭНДАЛЬФ
1879 ТОРИН
1476 ГНОМ
1221 ГОБЛИН
1009 ГОЛЛУМ
800 ЭЛЬФЫ
777 ДРАКОН
738 ПОНИ
707 БЭРД
702 ГОРЫ
617 КОРОЛЬ
530 ХОББИТ
509 БЕОРН
483 КОЛЬЦО
453 ВОЛШЕБНИК
432 ЛЕС
409 БАЛИН

382 ФИЛИ
365 ДВЕРИ
334 БЕРТ
287 ПАУКИ
284 ГЛАЗА
282 БОЧКИ
277 ОЗЕРА
272 ГОР
263 ГОРОД
256 КОРОЛЯ
254 БОМБУР
252 ЗОЛОТО
244 ПАУК
242 ГОРЕ
238 СМОГА
235 БУРГОМИСТР
228 АРКЕНСТОН
224 СВЕТ

Визуализация КГГВ

Дж. Толкиен, «Хоббит, или Туда и обратно »



Некоторые результаты

- Предложен алгоритм построения компактифицированного графа горизонтальной видимости (КГГВ).
- На основе последовательности дисперсионных оценок слов текста и КГГВ, построены сети слов различных текстов.
- Для литературных текстов среди узлов соответствующих КГГВ с наибольшими степенями присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражают семантику литературных произведений.
- Алгоритм определения веса слов, базирующийся на дисперсионной оценке оказался более эффективным для определения информационно-значимых слов, играющих важное значение для структурной связности в литературных текстах, чем алгоритм TFIDF.

Перспективы

Изучение свойства выявленных одновременно информационно и структурно важных лексических единиц как опорных слов для различных текстовых жанров, документов, представленных на разных языках. Это позволит:

- Формировать «более осмысленные» информационные портреты текстов;
- Более качественно выполнять автоматическое реферирование текстов;
- Формировать цепочки подобных документов, объединять тематические сюжеты, используя выявленные слова в качестве дескрипторов;
- Выявлять возможное содержательное дублирование документов, представленных на различных языках (необходимо дальнейшее исследование инвариантности опорных слов, для исходных текстов и их переводов);
- Составлять словари опорных слов, формировать тезаурусы и онтологии предметных областей.

Міжнародна науково-технічна конференція

ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ ЛІНГВІСТИЧНОГО АНАЛІЗУ

22-23 жовтня 2013 року

Спасибо за внимание!

Д.В. ЛАНДЭ

dwlande@gmail.com