УДК 004.7

СЕТЬ ЕСТЕСТВЕННЫХ ИЕРАРХИЙ ТЕРМИНОВ НА ПРИМЕРЕ АНАЛИЗА НАУЧНЫХ ТЕКСТОВ

Д.В. Ландэ (*dwlande@gmail.com*)
Институт проблем регистрации информации НАН Украины, Киев
А.А. Снарский (*asnarskii@gmail.com*)
НТУУ «Киевский политехнический институт», Украина, Киев
Е.В. Ягунова (*iagounova.elena@gmail.com*)
Санкт-Петербургский гос. университет, Санкт-Петербург, Россия

Описывается методика построения сетей естественных иерархий терминов на основе анализа текстовых корпусов. Методика базируется на применении компактифицированных графов горизонтальной видимости для терминов, а также установлении связей включения между ними. Построена и исследована сеть языка, сформированная на основе обработки полных текстов конференции OSTIS-2014.

Введение

Построение большой отраслевой онтологии – сложная проблема, которая требует больших ресурсных затрат. При этом определенным этапом построения общих онтологий является построение словарных номенклатур, тезаурусов, терминологических онтологий. Эффективный автоматический отбор отдельных терминов для таких конструкций – не решенная окончательно задача, есть отдельные попытки ее решения методами «взвешивания» слов и устойчивых словосочетаний, выполняемых с помощью различных алгоритмов [Лукашевич и др., 2007]. Проблема установления связей при построении сетей из таких терминов также остается открытой.

Ниже описана методика построения сети естественной иерархии терминов (СЕИТ), которую можно рассматривать как терминологическую основу для формирования соответствующей терминологической онтологии. Сеть естественной иерархии терминов базируется на информационно-значимых элементах текста, опорных терминах [Yagunova и др., 2012], методология выявления которых приведена в [Lande и др., 2013-1, 2, 3]. Использование таких эле-

ментов позволяет формировать терминологические сети, охватывать целые области знаний в качестве основ для дальнейшего построения общих онтологий. Опорные термины при этом обычно выбираются с учетом такого их свойства, как «дискриминантная сила» [Salton, 1983]. Вместе с тем одного этого свойства часто оказывается недостаточно для построения терминологических сетей. Иногда слова с низкой дискриминантной силой, в частности, наиболее частотные из выбранной предметной области (например, слова «Информация», «Семантика», «Интеллект» в корпусе текстов по семантическим технологиям) являются важными для рассматриваемой задачи.

1. Алгоритм формирования СЕИТ

Формирование сети естественных иерархий терминов (СЕИТ) базируется на контенте текстовых корпусов выбранной для анализа направленности. «Естественность» иерархий терминов в этом случае понимается как отказ при формировании сети от специальных методов семантического анализа; все связи в такой сети определяются естественным взаимным положением слов и словосочетаний, которые экстрагируются из текстов статистически значимых объемов (как показывает практика, минимальный объем для такого анализа – 20 КБ). СЕИТ, создаваемая полностью автоматически, может рассматриваться как основа для формирования онтологии с участием экспертов. Алгоритм формирования СЕИТ предусматривает реализацию последовательности шагов [Lande, 2014]:

1. На первом этапе выбирается исходный текстовый корпус, в качестве которого рассматриваются тексты научных докладов, вошедшие в сборник трудов научно-технической конференции OSTIS-2014 (Open Semantic Technologies for Intelligent Systems – OSTIS), прошедшей в г. Минске в феврале 2014 года [OSTIS, 2014]. В сборник включено около 100 докладов общим объемом свыше 2 млн. символов. Предварительная обработка такого текстового корпуса предусматривает выделение фрагментов текстов (докладов, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний слов.

2. На втором этапе каждому отдельному термину из текста – слову, двух- или трехсловному сочетанию (биграмме или триграмме) ставится в соответствие оценка его «дискриминантная сила», а

именно TFIDF, которая в каноническом виде определяется как произведение частоты соответствующего термина (Term Frequency) из фрагмента текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот терм встретился (Inverse Document Frequency) .

3. Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) и выполняется повторное определение весовых значений слов уже по этому алгоритму. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики текстового корпуса [Lande и др., 2013-1]. В качестве весовых оценок терминов используются степени соответствующих им узлов в CHVG. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов CHVG. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Используемый в рамках данной работы стоп-словарь был построен на основе различных стоп-словарей в доступном виде на веб-ресурсах (http://code.google.com/p/stop-words/source/browse/trunk/stop-words/ stop-words-russian.txt; http://www.ranks.nl/stopwords/; https://trac. mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.txt).

4. Экспертным методом определяется необходимый размер СЕИТ (число N), после чего выбирается соответствующее количество единичных слов, биграмм и триграмм (всего $N+N+N$ элементов) с наибольшими весовыми значениями по CHVG.

5. Из отобранных терминов строятся СЕИТ, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождениям одних терминов в другие. На рис. 1 проиллюстрирован принцип построения связей СЕИТ. Различные геометрические фигуры на рис. 1 соответствуют различным словам. Первой строке соответствует выбранное множество единичных слов, второму – множество биграмм, а третьему – множество триграмм. Если единичное слово входит в биграмму или триграмму, или биграмма входит в триграмму, образуется связь. Множество узлов и связей образует трехуровневую сеть естественной иерархии терминов.

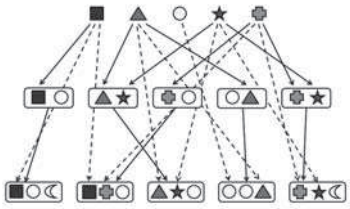

Рис. 1. Трехуровневая сеть естественной иерархии терминов

После формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки СЕИТ в базы данных формируется матрица инцидентности в формате csv. В табл. 1 приведены списки 20 наиболее весомых терминов (слов, биграмм и триграмм) из сборника трудов конференции.

Таблица 1

№	Слова	Биграммы	Триграммы
1	СИСТЕМА	ПРЕДМЕТНАЯ ОБЛАСТЬ	ПОДДЕРЖКА ПРИНЯТИЯ РЕШЕНИЙ
2	ЗНАНИЯ	БАЗА ЗНАНИЙ	ОНТОЛОГИЯ ПРЕДМЕТНОЙ ОБЛАСТИ
3	ОНТОЛОГИЯ	СЕМАНТИЧЕСКАЯ СЕТЬ	ПРОЕКТИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ
4	МОДЕЛЬ	ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА	УНИВЕРСАЛЬНЫЙ СЕМАНТИЧЕСКИЙ КОД
5	ДАННЫЕ	ЕСТЕСТВЕННЫЙ ЯЗЫК	ХРАНИЛИЩЕ СВЯЗАННЫХ ДАННЫХ
6	СЕМАНТИЧЕСКИЙ	РЕШЕНИЕ ЗАДАЧ	ПОНЯТИЯ ПРЕДМЕТНОЙ ОБЛАСТИ
7	ЯЗЫК	ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ	ЯЗЫК ПРЕДСТАВЛЕНИЯ ЗНАНИЙ
8	ЗАДАЧА	ПРИНЯТИЕ РЕШЕНИЙ	ЗНАНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ
9	ОБЛАСТЬ	ПРЕДСТАВЛЕНИЕ ЗНАНИЙ	ОСНОВАННАЯ НА ЗНАНИЯХ СИСТЕМА
10	ОБЪЕКТ	БАЗА ДАННЫХ	МАШИННАЯ ОБРАБОТКА ЗНАНИЙ
11	РЕШЕНИЕ	НЕЙРОННАЯ СЕТЬ	БАЗА КОНТЕКСТНЫХ ПРАВИЛ
12	СЕТЬ	СИСТЕМА УПРАВЛЕНИЯ	КЛАСС СЕМАНТИЧЕСКИХ СЕТЕЙ
13	ТЕКСТ	ИНФОРМАЦИОННАЯ СИСТЕМА	РЕШЕНИЕ ПОСТАВЛЕННЫХ ЗАДАЧ

№	Слова	Биграммы	Триграммы
14	ПОНЯТИЕ	СИТУАЦИОННОЕ УПРАВЛЕНИЕ	ИНФОРМАЦИОННЫЕ ПОТРЕБНОСТИ ПОЛЬЗОВАТЕЛЕЙ
15	МНОЖЕСТВО	ЖИЗНЕННЫЙ ЦИКЛ	УПРАВЛЕНИЕ ПОЛЕТАМИ МКС
16	ПРЕДСТАВЛЕНИЕ	СЕМАНТИЧЕСКИЙ АНАЛИЗ	СИСТЕМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
17	ОТНОШЕНИЕ	УПРАВЛЕНИЕ ПРОЕКТАМИ	ГРАФ ГОРИЗОНТАЛЬНОЙ ВИДИМОСТИ
18	ОСНОВА	ИНФОРМАЦИОННЫЙ РЕСУРС	ЭЛЕКТРОННОЕ УЧЕБНОЕ ИЗДАНИЕ
19	ИНФОРМАЦИОННОЕ	РУССКИЙ ЯЗЫК	СЕМИОТИЧЕСКАЯ КОГНИТИВНАЯ КАРТА
20	УПРАВЛЕНИЕ	ОБЛАСТЬ ЗНАНИЙ	ОТКРЫТЫЕ СЕМАНТИЧЕСКИЕ ТЕХНОЛОГИИ
21	ПОЛЬЗОВАТЕЛЬ	ОБУЧАЮЩАЯ ВЫБОРКА	МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
22	ОПИСАНИЕ	КОГНИТИВНАЯ КАРТА	ДАННЫЕ ПРЕДМЕТНОЙ ОБЛАСТИ
23	БАЗА	ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС	ПОРТАЛ НАУЧНЫХ ЗНАНИЙ

Рис. 2. Вид СЕИТ размером 20+20+20

На рис. 2 представлена небольшая сеть естественной иерархии терминов размером 20+20+20 (20 слов, 20 биграмм, 20 триграмм), которая визуализирована средствами системы Gephi (https://gephi.org/).

2. Исследование СЕИТ

Для построенных сетей естественных иерархий терминов различных размеров по выбранному тексту было определено распределение исходящих степеней узлов, которое оказалось близким к степенному ($p(k) = Ck^{a}$), т.е. эти сети являются безмасштабными. Оказалось, что коэффициент a для сетей различных размеров (от 20+20+20 до 200+200+200) составляет от 2,1 до 2,3, что вполне соответствует сетям языка (Language Networks) [Большакова и др., 2011].

Очевидно, что в соответствии с предложенным алгоритмом максимальное количество входных связей для узлов данной сети составляет 5 (для узлов из одного слова – 0 входящих связей, для узлов из 2 слов – максимально 2 связи, для узлов из 3 слов – максимально 5 связей – три связи от отдельных слов и две от пар слов).

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с максимальным количеством входных связей, среди которых можно выделить такие словосочетания: «язык представления знаний»; «система искусственного интеллекта»; «поддержка принятия решений»; «онтология предметной области»; «универсальный семантический код».

Из материалов конференции было выбрано 5 докладов из сборника трудов конференции, в которые входили данные словосочетания, а именно доклады:

- *Ефименко И.В., Хорошевский В.Ф.* УСК Мартынова – тридцать лет спустя;
- *Тарасов В.Б.* От семантического кода к когнитивной лингвистике и информатике: наследие В.В. Мартынова;
- *Массель Л.В., Массель А.Г.* Ситуационное управление и семантическое моделирование в энергетике;
- *Стефанюк В.Л.* Условно-рефлекторная основа запоминания знаний;
- *Кулинич А.А.* Когнитивное моделирование в условиях неопределенности (семиотический подход).

По отдельным докладам также были рассчитаны значения CHVG для слов, биграмм и триграмм, построены сети естественных иерархий терминов. Взаимосвязь терминов из выбранных докладов приведена на рис. 3, на котором можно видеть, что каждому докладу (узлы, идентифицированные фамилиями авторов) соответствуют термины. При этом в центральной части сети располагаются термины, общие для нескольких докладов (О-зона), а «гребешки» на периферии соответствуют специальным терминам, отражающим специфику конкретных докладов (С-зоны).

О-зона вовсе не обязательно включает термины из абсолютно всех докладов, достаточно, чтобы термины соответствовали лишь их определенной части (порогу), например половине. Чем больше в докладе терминов, попадающих в центральную часть, тем он лучше вписывается в тематику конференции, тем он точнее попадает в ее тренд. В данном случае (рис. 3) именно доклады Хорошевского и Тарасова в большей мере соответствуют тематическому направлению конференции. В соответствии с этим наблюдением можно предложить такой лингвистический критерий «релевантности» доклада тематике конференции: чем большая часть топ-лексики из него попадает в О-зону, тем он более релевантен.

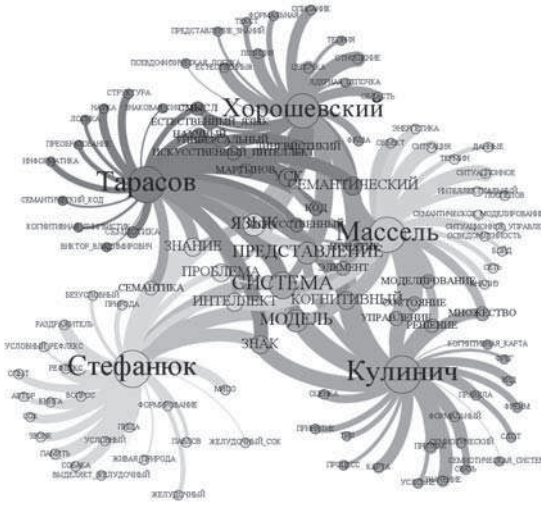

Рис. 3. Сеть связи терминов выбранных докладов

Заключение

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла научного текста были подтверждены в ходе экспериментов с информантами. Так, для всех текстов были проведены эксперименты со стандартной инструкцией «Прочитайте текст. Подумайте над его содержанием. Выпишите 10–15 слов, наиболее важных для его содержания» (более 20 информантов для текста каждого из исследуемых докладов, а также всего сборника) [Ягунова, 2010]. Полнота отражения информационно-значимых слов, биграмм и триграмм для сборника в этом случае превысила 50%.

Таким образом, в результате проведенных исследований:

- Описан алгоритм построения сетей естественных иерархий терминов на основе анализа текстов.
- На основании этого алгоритма по текстам докладов научно-технической конференции построена сеть естественной иерархии терминов.
- Сеть естественных иерархий терминов оказалась скейл-фри по исходящим связям.
- Выбраны программные средства визуализации сети естественных иерархий терминов.
- Предложен критерий релевантности доклада по тематике конференции.

Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построения общей онтологии (в рассмотренном примере – по тематике семантических технологий), использовать на практике в качестве готового к применению средства навигации в базах данных, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

СПИСОК ЛИТЕРАТУРЫ

[Лукашевич и др., 2007] Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог-2008». М., 2008. С. 339–344.

[Yagunova и др., 2012] Yagunova E., D. Lande D. Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects

// CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference «Digital libraries: Advanced Methods and Technologies, Digital Collections» Pereslavl-Zalessky, Russia, October 15-18, 2012. – P. 150–159.

[Lande и др., 2013-1] Lande D.V., Snarskii A.A. Compactified HVG for the Language Network // 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, Chisinau, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, 2013. – P. 108–113.

[Lande и др., 2013-2] Lande D., Snarskii A., Yagunova E. The Use Of Horizontal Visibility Graphs To Identify The Words That Define The Information Structure Of The Text // CEUR Workshop Proceedings. Vol-1108 urn:nbn:de:0074-1108-1. ISSN 1613-0073. Selected Papers of the 15th All-Russian Scientific Conference «Digital libraries: Advanced Methods and Technologies, Digital Collections» Yaroslavl, Russia, October 14–17, 2013. – P. 158–164.

[Lande и др., 2013-3] Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209–215.

[OSTIS, 2014] Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2014): материалы IV междунар. науч.-техн. конф. (Минск 20–22 февраля 2014 года) / – Минск: БГУИР, 2014. – 576 с.

[Lande, 2014] Lande D.V. Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint viXra 1404.0069.

[Salton, 1983] Salton G., McGill M.J. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.

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

[Большакова и др., 2011] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.

[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. – Пермь, 2010. – Вып. 1. – С. 85–91.