УДК 81’322.2

О ПОСТРОЕНИИ МОДЕЛИ ЛОКАЛЬНО-ГЛОБАЛЬНОГО КОНТЕКСТА

Никитина С.А.

РГП на ПХВ «Казахский Национальный Педагогический Университет им. Абая» (КазНПУ), Алматы, e-mail: nikitina.svetlana@gmail.com

Методы структуризации и анализа больших объемов текстовых данных из-за экспоненциального роста объемов сети приобретают всё большую популярность. Применимость методов обусловлена широким кругом важных прикладных задач из следующих областей: мониторинг общественного мнения, маркетинг в социальных сетях, информационный поиск, визуализация текстовой информации. В исследовании делается попытка при помощи использования математического аппарата, теории вероятности, обработки естественных языков описать алгоритмы сбора данных из гетерогенных онлайн-сервисов в русскоязычном Интернете.

Ключевые слова: теория случайного графа, астротурфинг, искусственные нейронные сети, обработка естественных языков

ABOUT MODELLING LOCAL-GLOBAL CONTEXT

Nikitina S.A.

Kazakh National Pedagogical University Abai, Almaty, e-mail: nikitina.svetlana@gmail.com

The methods of structuring and analysis of big volume text data are getting more and more popular due to the exponential growth of network volume. The use of these methods has been caused by a wide range of important applied objectives of the following fields: monitoring of public opinion, marketing in social networking sites, information search, visualization of text information. In the research it is attempted to describe the algorithms of data collection from heterogeneous online-servers in the Russian-language Internet with the help of use of mathematical criticus, theory of relativity, natural language processing.

Keywords: the theory of random grath, astroturfing, artificial neural net, natural language processing

Информация, распространяемая в глобальной информационной среде через различные форумы, блоги и социальные сети оказывает определенное влияние на современные социальные процессов, происходящие в стране. Построение модели глобально-модального контекста поможет объединить две смежные области: маркетинг и мониторинг общественного мнения.

Научное направление компьютерной лингвистики возникло в 50-е годы 20 века. Присущие этой науке подходы делятся на статистические и основанные на правилах. Такой выбор исторически обусловлен тем фактом, что, начиная с 60-х годов, область теоретической лингвистики была значительно проработана в духе Ноама Хомского [5], автора общеизвестной иерархии формальных грамматик. Другим значимым вкладом в современную вычислительную науку о языке являются работы Игоря Мельчука [2], в частности, повсеместно заимствованный подход комбинаторного словаря. В русле его направления работают школа Н.Н. Леонтьевой [1], в центре которой находится «Русский обще семантический словарь» или, как пример западных проектов, – OpenGraph и WordNet [10].

Подходы, предлагаемые современной вычислительной лингвистикой, включают построение статистических моделей языка на основе ручной разметки огромного количества текстов. Разметка добавляет в текст метаинформацию о семантической, синтаксической или морфологической структуре предложения, что само по себе является довольно трудоемким процессом. Такой подход обработки естественных языков выделился в отдельное направление, а именно корпусную лингвистику, обучение на примерах [8]. Из существенных недостатков следует отметить тот факт, что реальный язык развивается достаточно быстро, из-за чего требуется постоянно обновлять аннотированные текстовые корпуса и морфологические словари. Лингвистикой, основанной на правилах, должен заниматься специалист по конкретному языку [8]. Область задач при этом ограничена мощностью набора правил и морфологического словаря.

Специалистами по машинному обучению выдвигается идея о возможности синтаксического разбора на основе статистических закономерностей, полученных на не размеченном корпусе [4, 6, 9].

Первые значимые результаты были получены для английского языка, синтаксические парсеры для большинства синтетических, полисинтетических языков недоступны. Причины этого могут быть обусловлены различными факторами, например, отсутствием сколько-нибудь значимых текстовых корпусов.

В некоторых работах используются грамматические модели в стиле Lucien Tesniere, в частности, Мельчук [2] вводит понятие грамматики зависимостей (Dependency Grammar, DG). Целью разбора зависимостей является конструирование дерева предложения, где все узлы представлены словами и грани – это связи между ними. Всего выделяется четыре типа связей: синтаксические, морфологические, семантические и просодические. Очевидным преимуществом разбора зависимостей является тот факт, что дерево синтаксических зависимостей является резонным приближением к семантической структуре предложения. Для одного предложения деревья всех четырех типов связей могут совпадать частично или полностью, в случае частичного совпадения, к примеру, совпали вершины графа, но не совпали направления связей. Корневым элементом дерева синтаксического разбора обычно является глагол.

Чисто статистический подход имеет определенные преимущества, для него достаточно иметь ограниченный неаннотированный корпус размером не более 1 миллиона слов, что может являться важным для исчезающих языков. Статистический подход к синтаксическому анализу предложения представлен следующими моделями:

1) языковая модель зависимостей (dependency language model, DLM) [6];

2) ориентированный на данные разбор зависимостей без учителя (Unsupervised Data-Oriented Parsing, U-DOP) [4];

3) метод общих покрывающих связей (Common cover links, CCL) [9].

U-DOP модель – одна из первых работ успешной апробации алгоритмов разбора без учителя для английского, немецкого и китайского языков. Главная идея заключается в том, чтобы поставить в соответствие все возможные бинарные деревья множеству предложений и потом использовать все поддеревья для того чтобы вычислить наиболее вероятное дерево [4]. Следует отметить, что вычислительная сложность алгоритма для одного среднестатистического предложения на русском языке предполагает 4862 варианта бинарного дерева.

CCL модель использует несколько универсальных свойств естественных языков, а именно схватывает асимметрию синтаксического дерева, последовательно ограничивает пространство поиска, обрабатывая последовательно высказывания и основываясь на законе Ципфа, принимает решения разбора. Используется элементарные методы самонастройки для выделения основных свойств обрабатываемого языка.

DLM модель позволяет выделить лингвистические ограничения через структуру зависимости – множество вероятностных зависимостей, выраженных между заглавным словом каждой фразы в предложении с помощью ацикличного, плоского, ненаправленного графа. Предложенный алгоритм поэтапного подхода к разбору предложения был развит Потемкиным С.Б. [3], в частности применен к разбору предложения на русском языке.

Для дальнейшего развития алгоритма необходимо решить проблемы анафорических ссылок, выделения направлений в графе, разделения терминологических связок и грамматических отношений, маркирования граней дерева разбора, эллипсиса, омонимичности.

Вышеупомянутый метод не лишен и общих недостатков DG моделей, а именно: экстрапозиции, перестановки слов, Wh-fronting, тематизации.

На наш взгляд, необходимо развить подход локальных связей (Model of local links, MLL), упомянутый в работе Потемкина С.Б. [3], а именно усилить модель глобальным контекстом[7].

Отношение близости [5] позволяет выделить несколько языковых парадигм:

1) лексическая парадигма слов, а именно близкие по контексту употребления «утренний – завтрак – ранний – кофе – восход»;

2) словообразовательная парадигма: делать, переделать, сделать, делающий,...; дело, деловой... .

Необходимо также улучшить выделение на основе [5] до списков синонимичности и антонимичности. Очевидным преимуществом подобного алгоритма является обучение без учителя. Степень синонимичности может быть представлена с помощью семейства алгоритмов иерархической кластеризации, где на каждом уровне иерархии выделяется отдельный уровень обобщённости синонимов. Возможно изложенная выше идея подойдет для большинства синтетических языков.

Одной из целей статистических языковых моделей является определение вероятности предложения W среди всех возможных предложений T, таких что $P(W) = \Sigma P(W,T)$, где P(W,T) – вероятность предложения W для структуры T.

Задача разбора является обратной, иными словами, $P(T\backslash W) = \Pi P((i,I)W)$, где P((i,I) W) – вероятность связи в конкретном предложении W.

Сложно определить напрямую условную вероятность, потому как из-за закона Ципфа частоты для большинства заглавных

слов не будут статистически значимыми, так как неаннотированный корпус размером около 18Гб содержит более 5*105 слов с количеством упоминаний меньше 100 Закон Ципфа хорошо аппроксимируется распределением Парето, поэтому традиционные статистические подходы плохо применимы для статистических языковых моделей.

Список литературы

1. Леонтьева Н.Н. О смысловой неполноте текста (в связи с семантическим анализом) // МП и ПЛ, вып. 11. – М., 1970.

2. Мельчук И.А. Опыт теории лингвистических моделей «Смысл <=> Текст». – М., 1974.

3. Потемкин С.Б. Неконтролируемый синтаксический анализ. МГУ им. М.В. Ломоносова// http://www.dialog-21.ru/digests/dialog2009/materials/html/63.htm (дата обращения: 24.01.2016).

4. Bod R. An all-subtrees approach to unsupervised parsing.: Proceedings of COLINGACL.

5. Chomsky Noam. Knowledge of language: its nature, origin, and use. NY: Praeger, 1986.

6. Gao J., Suzuki H. Unsupervised learning of dependency structure for language modeling. ACL, 2003. – C. 521–528.

7. Huang Eric. Improving Word Representations via Global Context. // http://nlp.stanford.edu/pubs/HuangACL12.pdf.

8. Klein D., Manning C.D. Corpus-based induction of syntactic structure. Models of dependency and constituency. In: Proc. Of ACL, 2004.

9. Seginer Y. Prague. Fast Unsupervised Incremental Parsing.: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007.

10. Word Net. A lexical database for English. // http://wordnet.princeton.edu/ (дата обращения: 23.01.2016).