Квантитативная семантика в проектировании систем обработки больших текстовых данных

Елена Борисовна Козеренко
ФИЦ ИУ РАН
kozerenko@mail.ru

Квантитативная семантика

- Квантитативная семантика – это актуальное направление прикладной и компьютерной лингвистики для автоматизации семантической обработки очень больших объемов неструктурированных данных, представленных в текстах на различных естественных языках.

Квантитативная семантика

- Термин «квантитативная семантика» в нашем случае употребляется в значении количественного подхода к исследованию семантики естественного языка и семантическому моделированию для построения информационных систем различных классов.

Содержание доклада

- Способы построения семантических представлений на основе векторных моделей (Vector Space Model – VSM), матриц;
- соотношение частотных характеристик языковых объектов и их значений;
- «вынесение значения»

Количественные и качественные описания

- В отличие от методов представления смысла в виде качественных описаний плана содержания языковых объектов, традиционно используемых в системах искусственного интеллекта (symbolic methods), количественные методы позволяют определять и сопоставлять значения слов и языковых структур по «численным образам» их контекстных окружений.

Содержание доклада

- Онтологическая семантика;
- дистрибутивная семантика;
- семантика синтаксиса;
- частотные словари, размеченные и неразмеченные текстовые корпуса;
- решения на основе гибридного использования логико-лингвистических и статистических подходов.

Векторные модели

- **Векторное** (или **линейное**) **простра́нство** – математическая структура, которая представляет собой набор элементов, называемых векторами, для которых определены операции сложения друг с другом и умножения на число — скаляр.

Модели лексической семантики

- Информация о дистрибуции лингвистических единиц представляется в виде многоразрядных векторов, которые образуют словесное векторное пространство. Векторы соответствуют лингвистическим единицам (словам или словосочетаниям), а измерения соответствуют контекстам. Координаты векторов представляют собой числа, показывающие, сколько раз данное слово или словосочетание встретилось в данном контексте.

Онтологическая семантика

Тезаурусы и онтологии

- Семантические базы данных
- Wordnet
- FrameNet
- Тезаурус РуТез
- Framebank

Дистрибутивная семантика

- Область лингвистики, которая занимается вычислением степени семантической близости между лингвистическими единицами на основании их распределения (дистрибуции) в больших массивах лингвистических данных (текстовых корпусах).
- Каждому слову присваивается свой контекстный вектор. Множество векторов формирует словесное векторное пространство.
- Семантическое расстояние между понятиями, выраженными словами естественного языка, обычно вычисляется как косинусное расстояние между векторами словесного пространства.

Дистрибутивный анализ

- был предложен Л. Блумфилдом в 20-х гг. XX века и применялся, главным образом, в фонологии и морфологии.
- З. Харрис и другие представители дескриптивной лингвистики развивали данный метод в своих работах в 30 — 50-х гг. XX века.
- Близкие идеи выдвигали основоположники структурной лингвистики Ф. де Соссюр и Л. Витгенштейн.
- Идея **контекстных векторов** была предложена психологом Ч. Осгудом в рамках работ по представлению значений слов, Контексты, в которых встречались слова, выступали в качестве измерений многоразрядных векторов.

Пример словесного векторного пространства, описывающего дистрибутивные характеристики слов tea и coffee, в котором контекстом выступает соседнее слово:

$$A_{m,n} = \begin{array}{c} \\ \textit{coffee} \\ w_2 \\ \textit{tea} \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & \textit{drink} & w_3 & \dots & w_m \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & \dots & 1 \\ 1 & 0 & 2 & \dots & 0 \\ 0 & 2 & 0 & \dots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \dots & 0 \end{bmatrix} \end{array}$$

Размер контекстного окна определяется целями исследования:

- установление синтагматических связей –1-2 слова;
- установление парадигматических связей — 5-10 слов;
- установление тематических связей — 50 слов и больше.
- Семантическая близость между лингвистическими единицами вычисляется как расстояние между векторами.

Семантическая близость между лингвистическими единицами вычисляется как расстояние между векторами: косинусная мера

$$\frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Дистрибутивный анализ

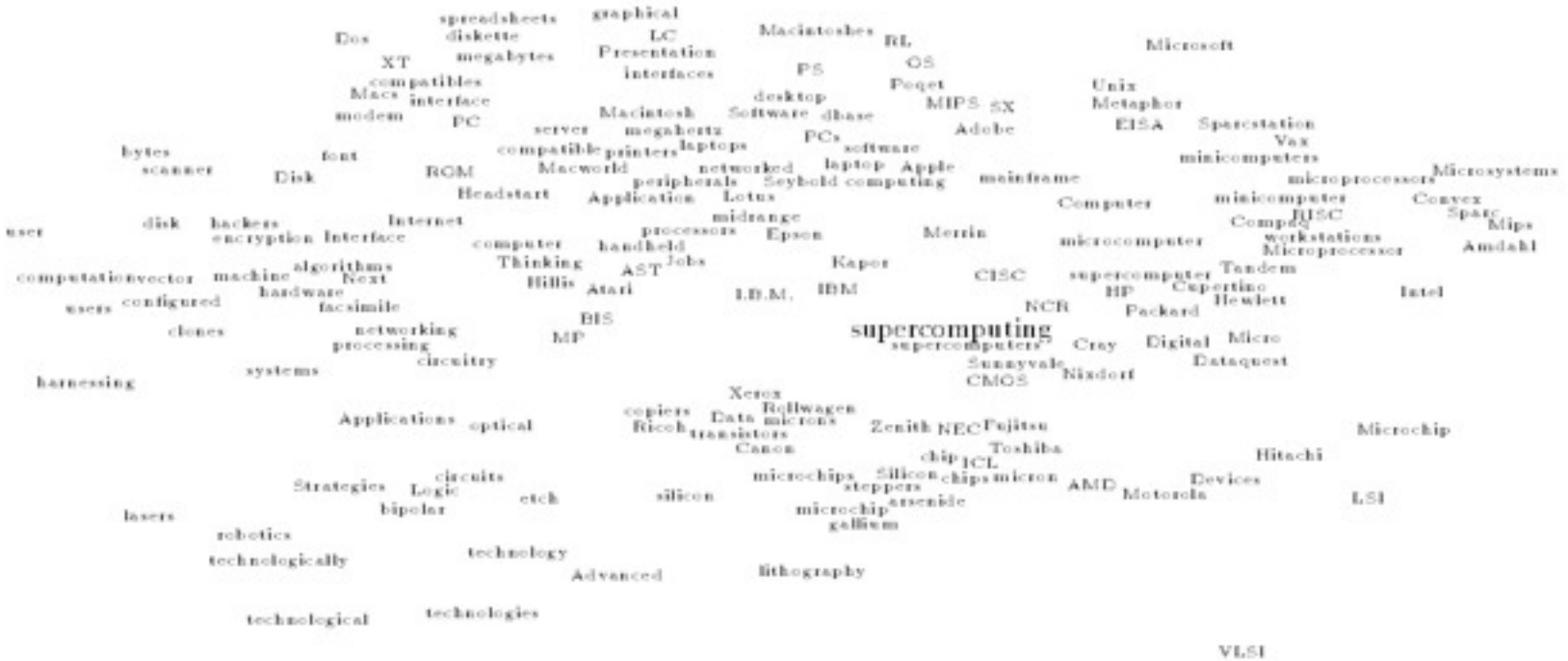
- После проведения дистрибутивного анализа становится возможным выявить наиболее близкие по смыслу слова по отношению к изучаемому слову.
- Пример наиболее близких слов к слову кошка (список получен на основании данных веб-корпуса русского языка, обработка корпуса выполнена системой Sketch Engine

Lemma	Score	Freq
кот	0.3	9252
собака	0.288	24102
птица	0.219	13889
зверь	0.215	7270
пес	0.214	4820
животное	0.21	16108
волк	0.199	6071
мальчик	0.198	28828
девочка	0.197	27136
медведь	0.196	5286
крыса	0.186	4021
парень	0.174	25625
мама	0.172	42197
корова	0.168	5466
папа	0.164	22231
лошадь	0.164	12582
мышь	0.164	4887

Графическое представление

- В графическом виде слова могут быть представлены как точки на плоскости, при этом точки, соответствующие близким по смыслу словам, расположены близко друг к другу.

Пример словесного пространства, описывающего предметную область



Программные средства с открытым кодом

для исследований по дистрибутивной семантике:

- S-Space
- Semantic Vectors
- Gensim
- word2vec
- WebVectors

Применение

- выявление семантической близости слов и словосочетаний;
- автоматическая кластеризация слов по степени их семантической близости;
- автоматическая генерация тезаурусов и двуязычных словарей;
- разрешение лексической неоднозначности;
- расширение запросов за счет ассоциативных связей;
- определение тематики документа;
- кластеризация документов для информационного поиска;
- извлечение знаний из текстов;
- построение семантических карт различных предметных областей;
- моделирование перифраз;
- определение тональности высказывания;
- моделирование сочетаемостных ограничений слов

Модели дистрибутивной семантики различаются по следующим

- тип контекста: размер контекста, правый или левый контекст, ранжирование;
- количественная оценка частоты встречаемости слова в данном контексте: абсолютная частота, TF-IDF, энтропия, совместная информация и пр.;
- мера расстояния между векторами: косинус, скалярное произведение, расстояние Минковского и пр.;
- метод уменьшения размерности матрицы: случайная проекция, сингулярное разложение, случайное индексирование и пр.

Наиболее широко известны следующие дистрибутивно-семантические модели:

- Модель векторных пространств
- Латентно-семантический анализ
- Тематическое моделирование
- Предсказательные модели

Уменьшение размерности векторных пространств

- удаление определенных измерений векторов в соответствии с лингвистическими или статистическими критериями;
- сингулярное разложение;
- метод главных компонент (PCA);
- случайное индексирование.

Проблема слишком большой размерности векторов

- В реальных приложениях возникает проблема слишком большой размерности векторов, соответствующей огромному числу контекстов, представленных в текстовом корпусе. Возникает необходимость в применении специальных методов, которые позволяют уменьшить размерность и разреженность векторного пространства и при этом сохранить как можно больше информации из исходного векторного пространства. Получающиеся в результате сжатые векторные представления слов в англоязычной терминологии носят название word embeddings.

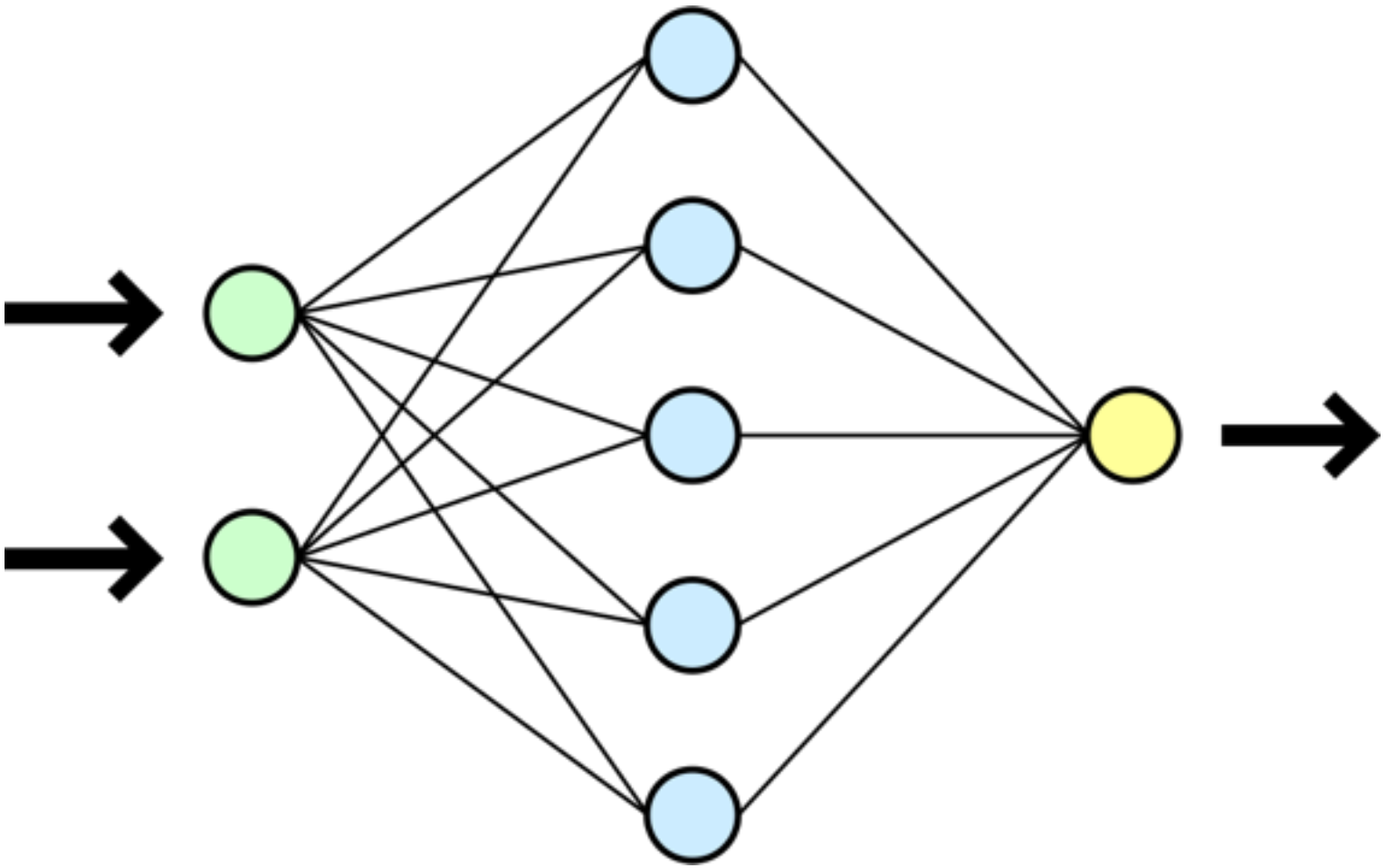
Предсказательные модели дистрибутивной семантики

- машинное обучение, в частности искусственные нейронные сети. При обучении таких предсказательных моделей (англ. predictive models) целевым представлением каждого слова также является сжатый вектор относительно небольшого размера (англ. embedding), для которого в ходе множественных проходов по обучающему корпусу максимизируется сходство с векторами соседей и минимизируется сходство с векторами слов, его соседями не являющихся.

Предсказательные модели дистрибутивной семантики

- Подобные предсказательные модели представляют семантику естественного языка более точно, чем счётные модели, не использующие машинное обучение.
- Наиболее известные представители подобного подхода — алгоритмы Continuous Bag-of-Words (CBOW) и Continuous Skipgram, впервые реализованные в утилите word2vec, представленной в 2013 году. Пример применения подобных моделей к русскому языку представлен на веб-сервисе RusVectōrēs.

Искусственные нейронные сети



Карен Спарк Джонс

- Внесла значительный вклад в две отдельные области: информационный поиск(ИП) и обработка естественного языка (ОЕЯ). Занималась интеграцией этих направлений в основные блок-схемы искусственного интеллекта (ИИ). Её наиболее важным вкладом является создание концепции учёта весов слов обратной частоты документа (IDF). IDF используется во многих поисковых системах, как правило, в составе схемы TF-IDF.

Karen Spark Jones

Лингвистические объекты как векторные модели

- Marco Baroni: https://www.aclweb.org/anthology/P14-1023.pdf
- Roberto Zamparelli
- Nouns are vectors, adjectives are matrices:

http://www.coli.uni-saarland.de/courses/comsem-13/material/Min_Baroni.pdf

- Роберто Навильи, Babelnet: https://babelnet.org/

Nouns are vectors, adjectives are matrices

- Adjective interpreted as a linear mapping of the noun vector
- p=Bv
- model-generated ANs should approximate corpus-observed ANs estimate the values of the weight matrix by solving linear regression problems
- Experiments show that . . . ANs in the corpus generally conform with our semantic intuitions and can be used as a goal of approximation (1st study)
- B&Z's alm method provides the best approximation (2nd study), followed by the additive model
- under the functional view, adjectives can still be meaningfully represented and compared even though the adjectives do not have an independently collected vector.

Вероятностная нейросетевая модель языка

- Модель NPLM в процессе предсказания слова u по предшествующим словам $v1{:}n$ обучает матрицу векторных представлений Θ размерности $T \times W$. Предсказания осуществляются по формуле:
- $p(u|v1{:}n) = \text{softmax}(b + Wx + U\text{th}(d + Hx))$, где
- x – это вектор размерности $nT \times 1$, составленный из векторных представ-лений контекстов $\theta vi, i = 1 \ldots n$. Все остальные вектора и матрицы b, W, U, d, H – это параметры нейронной сети.

Преобразование softmax

- Преобразование softmax переводит произвольный вещественный вектор в нормированный неотрицательный вектор той же размерности ($W \times 1$):

$\text{softmax}(z) = \exp z_k -> k \exp z_k$.

- Недостатком модели является огромное число параметров и долгое обучение.

Семантика синтаксиса

- Понятие синтаксемы
- Коммуникативная грамматика
- Категориальные грамматики
- Древоприсоединительные грамматики
- Унификация; атрибуты и значения
- CFG; GPSG; RGPSG; HPSG; LFG
- Грамматики зависимостей

Языковые ресурсы

- Интеллектуальный статистический вербалайзер
- Частотные словари, и неразмеченные текстовые корпуса
- НКРЯ
- BNC

Моделирование грамматических преобразований

- на основе векторных пространств и тензоров:
- Тензор (от лат. tensus, напряженный) – объект линейной алгебры, преобразующий элементы одного линейного пространства в элементы другого. Часто тензор представляют как многомерную таблицу, заполненную числами – компонентами тензора $d \times d \times \ldots \times d$, где d – размерность, над которым задан тензор, а число сомножителей совпадает с т. н. валентностью, или рангом тензора.

Моделирование грамматических преобразований

- Важно, что такое представление (кроме скаляров, т. е. тензоров валентности ноль) возможно только после выбора базиса (или системы координат): при смене базиса компоненты тензора меняются определенным образом. Сам тензор как «геометрическая сущность» от выбора базиса не зависит, компоненты вектора меняются при смене координатных осей, но сам вектор – образом которого может быть просто нарисованная стрелка – от этого не изменяется.

Тензор как сущность любой системы

- Тензор обычно обозначают некоторой буквой с совокупностью верхних (контрвариантных) и нижних (ковариантных) индексов: . При смене базиса ковариантные компоненты меняются так же, как и базис (с помощью того же преобразования), а контрвариантные – обратно изменению базиса (обратным преобразованием). Тензор является сущностью любой системы реального мира и сохраняется, несмотря на происходящие изменения в этой системе

John drinks strong beer quickly

$$\textit{drinks} \otimes \text{subj} \otimes \textit{John} \otimes \text{obj} \otimes (\textit{beer} \otimes \text{adj} \otimes \textit{strong}) \otimes \text{adv} \otimes \textit{quickly}$$

Решения на основе гибридного подхода

- Использование логико-лингвистических и статистических подходов

Основные методы решения

- Вероятностный грамматический разбор: правило Байеса.
- Значения вероятностей для каждого возможного варианта грамматического разбора (т.е. развертывания нетерминального узла) вычисляются на основе частот встречаемости таких вариантов разбора в существующих текстовых корпусах с синтаксической разметкой (treebanks).
- Значения вероятностей для вариантов разбора могут быть также получены и в виде лингвистических экспертных оценок.

Машинное обучение

- Основано на стохастической исследовательской парадигме. Алгоритмы обучения могут быть двух типов: неуправляемые и управляемые.
- Неуправляемый алгоритм должен вывести модель, пригодную для обобщения новых данных, которые ему ранее не предъявлялись, и этот вывод должен быть основан только на данных.
- Управляемый алгоритм обучается на множестве правильных ответов на данные из обучающей выборки.

Схема многовариантного англо-русского трансфера

- to form water :
- [Category: VerbInf] -> {to form water }
- OR {[Category:ParticipleAdv] [вес 1]; /*образуя воду/
- [Category: VerbFinit] [вес 2]; /*образуют воду/
- [Category: VerbNounIng] [вес 3]; } /*с образованием воды/

Машинный перевод на основе вероятностного трансфера

- Языковые структуры представлены в виде иерархии правил когнитивной трансферной грамматики, которая является разновидностью унификационно-порождающей грамматики.
- Отношения зависимости реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов.
- Вероятностные оценки возможных вариантов разбора и перевода предложений в виде весов вводятся в правила трансфера.

Литература

- Kozerenko E. B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications. – Las Vegas, USA: CSREA Press, 2003. P. 49–55.
- Kozerenko E. B. Parallel Texts Alignment Strategies: the Semantic Aspects // Informatics and Applications, 2013. Vol. 7. No. 1. P. 82–89. Козеренко Е. Б. Стратегии выравнивания параллельных текстов: семантические аспекты // Информатика и ее применения, 2013. Т. 7. Вып. 1. С. 82–89.
- Kuznetsov I. P., Kozerenko E. B., Matskevich A. G. Intelligent extraction of knowledge structures from natural language texts // Proceedings of 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology – Workshops WI-IAT 2011 (Campus Scientifique de la Doua, Lyon, France, August 22–27, 2011). – IEEE Computer Society, 2011. P. 269–272.

Литература

- Clark S., Pulman S. Combining symbolic and distributional models of meaning // Proceedings of AAAI Spring Symposium on Quantum Interaction. – AAAI Press, 2007. http://www.cl.cam.ac.uk/~sc609/pubs/aaai07.pdf
- Enhancing syntactic models in the set-phrase machine translation Khoroshilov, A., Kozerenko, E. 2012. Proceedings of the 2012 International Conference on Artificial Intelligence, ICAI 2012
- Intelligent tools for the semantic internet navigator design
- Kuznetsov, I., Charnine, M., Kozerenko, E., Nikolaev, V., Matskevich, A. 2012. CEUR Workshop Proceedings.
- Curran J. R. From Distributional to Semantic Similarity: PhD thesis. – Edinburgh: University of Edinburgh, 2004.
- Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behavior Research Methods, Instruments & Computers, 1996. Vol. 28. No. 2. P. 203–208.

Литература

- McCarthy D., Koeling R., Weeds J., Carroll J. Finding predominant senses in untagged text // Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. – Barcelona, Spain: ACL, 2004. P. 280–287.
- Danielson D. A. Vectors and Tensors in Engineering and Physics. 2nd ed. – Boulder, CO: Westview (Perseus), 2003. ISBN 978-0-8133-4080-7.
- Pang B., Knight K., Marcu D. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences // NAACL'03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. – Stroudsburg, PA, USA: ACL, 2003. Vol. 1. P. 102–109.
- Malkov K. V., Tunitsky D. V. On Extreme Principles of Machine Learning in Anomaly and Vulnerability Assessment // MLMTA'06: Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications. – Las Vegas, USA: CSREA Press, 2006. P. 24–29.
- Lambek J. From Word to Sentence: a Computational Algebraic Approach to Grammar. – Monza, Italy: Polimetrica Publisher, 2008.
- Hermann K. M., Blunsom P. The Role of Syntax in Vector Space Models of Compositional Semantics // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. – Stroudsburg, PA, USA: ACL, 2013. P. 894–904.

Литература

- Moortgat M. Symmetric Categorial Grammar // Journal of Philosophical Logic, 2009. Vol. 38. No. 6. P. 681–710.
- 26. Baroni M., Zamparelli R. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Stroudsburg, PA, USA: ACL, 2010. P. 1183–1193.
- Clark S., Curran J. R. Wide-coverage efficient statistical parsing with CCG and log-linear models // Computational Linguistics, 2007. Vol. 33. No. 4. P. 493–552.
- Grefenstette E., Sadrzadeh M. Experimental support for a categorical compositional distributional model of meaning // Proceedings of the Conference on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK, July 27–31, 2011). – Stroudsburg, PA, USA: ACL, 2011. P. 1394–1404.

Литература

Лукашевич Наталья ВалентиновнаТезаурусы в задачах информационного поиска –М., 2010. –396с.:

- https://nsu.ru/xmlui/bitstream/handle/nsu/9086/louk_book.pdf
- Hermann K. M., Blunsom P. The Role of Syntax in Vector Space Models of Compositional Semantics // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. – Stroudsburg, PA, USA: ACL, 2013. P. 894–904.
- Church K., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics, 1996. Vol. 16. No. 1. P. 22–29.
-

Литература

- Elena Kozerenko. Parallel texts alignment strategies. Inform. Primen., 2013, Volume 7, Issue 1, Pages 82–89 (Mi ia247)http://www.mathnet.ru/links/ceabc3c7ec0dbdb3c299f68b18d35495/ia247.pdf
- Yu. I. Morozova, E. B. Kozerenko, M. M. Sharnin, Method for extracting single-word translation correspondences from parallel texts using distributional semantics models, Sistemy i Sredstva Inform., 2014, Volume 24, Issue 2, 131–142
- http://www.mathnet.ru/links/6da76492c120c647bd1b758611e7d918/ssi349.pdf
- Elena Kozerenko, Alexander Khoroshilov, Alexei A. Khoroshilov
- Syntactic Parameters in the Phrasal Machine Translation.Proceeding of the International Conference on Artificial Intelligence (ICAI'13) 2013 World Congress on Computer Science Research, Educarion and Advanced Technologies, CSREA, 2013, Las Vegas, USA
- http://worldcomp-proceedings.com/proc/p2013/ICA2118.pdf
- Koehn P. Statistical machine translation. — Cambridge: University Press, 2009.
- Schütze, H. 1998. Automatic word sense discrimination. Computational Linguistics 24(1): 97–123.

- Спасибо!