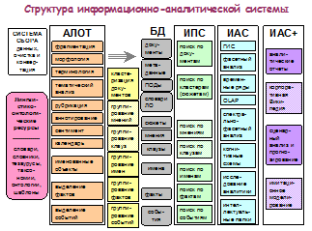


Современные тенденции в автоматической обработке текстов

Лукашевич Наталья Валентиновна, дтн

ведущий научный сотрудник НИВЦ МГУ имени М.В. Ломоносова, профессор МГТУ имени Н.Э. Баумана, профессор филологического факультета МГУ имени М.В. Ломоносова

louk_nat@mail.ru

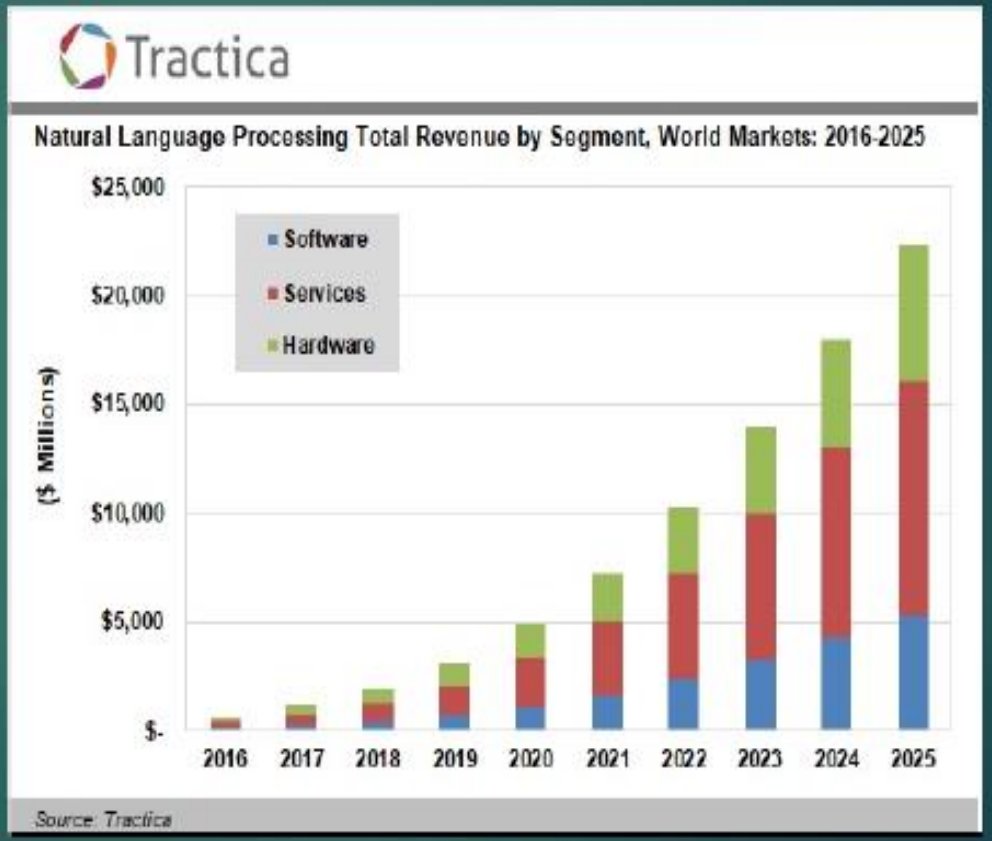
План

- Задачи и проблемы автоматической обработки текстов
- Данные в задачах АОТ
  - Размеченные данные vs. неразмеченные данные
- Особенности применения методов машинного обучения
  - Традиционные методы vs. глубокие нейронные сети
  - Пример подходов в области извлечения именованных сущностей
  - Достижения и проблемы методов на основе нейронных сетей

Рынок инструментов обработки естественного языка

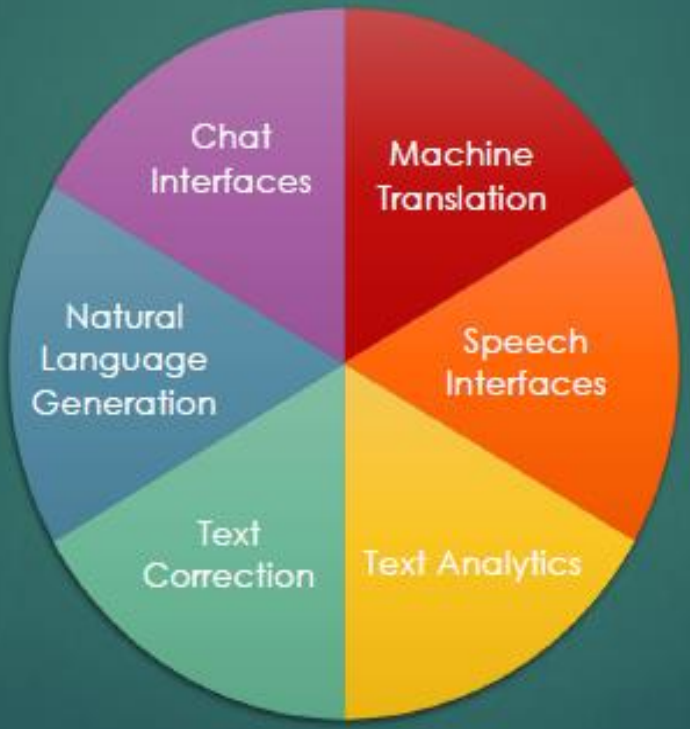
Natural Language Processing Market to Reach $22.3 Billion by 2025

Tractica



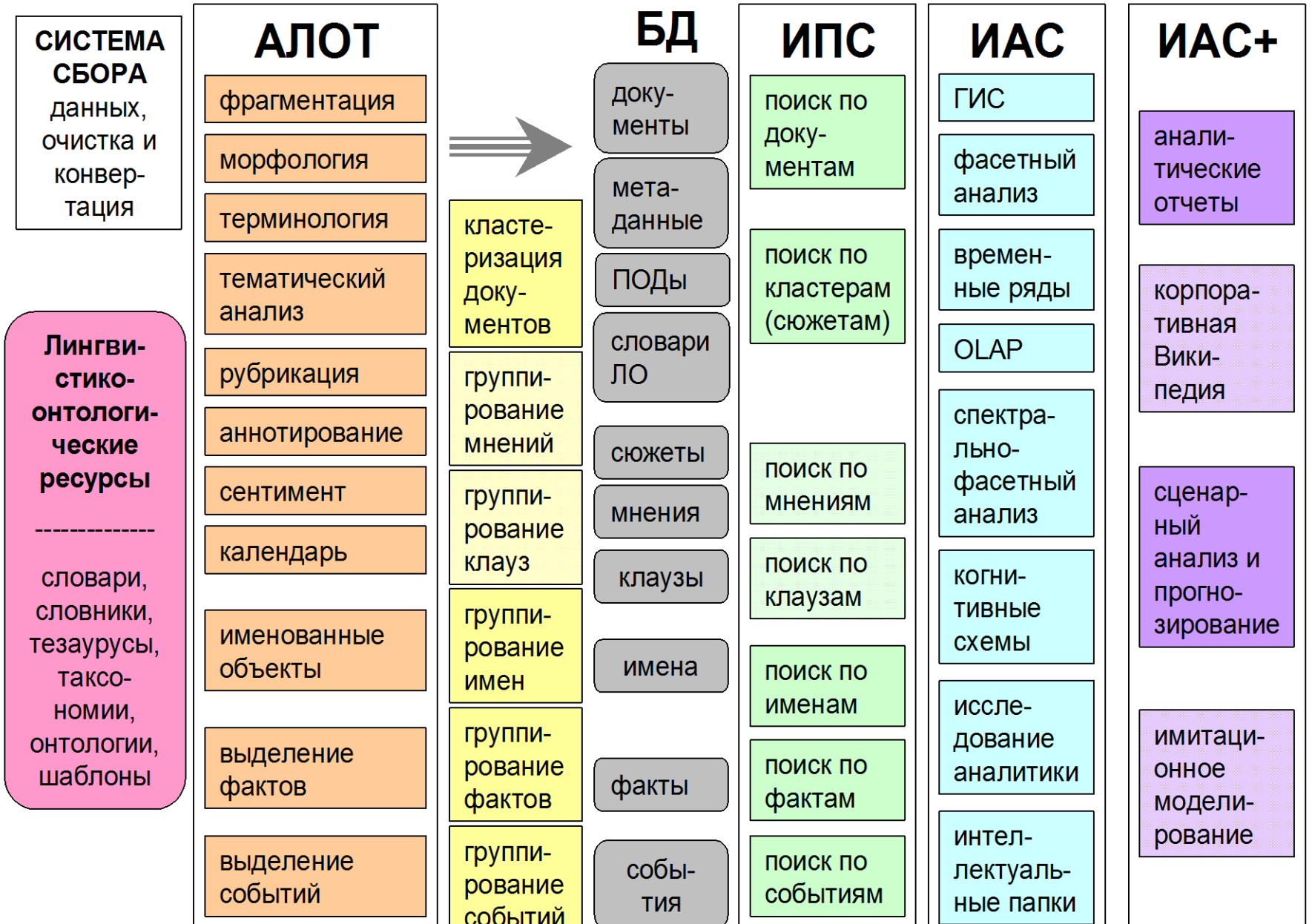
(slides by R.Dale 2017)

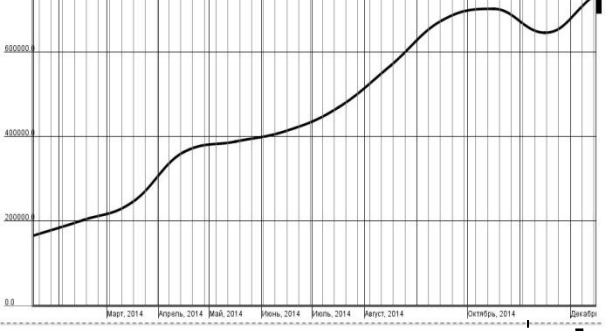
The Major Categories



(slides by R.Dale 2017)

Структура информационно-аналитической системы





Большие текстовые коллекции
(ср. понятие деловой прозы А.П. Ершова)

Новости

Твиттер

LJ

FB

VK

WWW

Активная зона

Архив

Сложность обработки текстов на ЕЯ

- Язык – многоуровневая знаковая система
- Описать всю информацию (словари, правила), необходимую для качественной обработки текстов, очень сложно
  - Многозначность на всех уровнях
  - Изменчивость, зависимость от контекста,
  - Изменение со временем
  - Необходимость использование больших объемов знаний о мире

Многозначность

- Фонетическая: *лук (луг* или *лук)*
- Морфологическая: *стали*
- Синтаксическая:
  - *Девочка шла по полю с цветами*
- Семантическая: *Возьми лук*

Свободные и устойчивые словосочетания: анекдоты о Штирлице

- Штирлиц выстрелил в упор, упор упал.
- Штирлиц шел по улице и поднял глаза. Это были голубые глаза пастора Шлага.
- Штирлиц стрелял из двух автоматов по очереди. Очередь заметно редела.
- Штирлица бил озноб. Озноб служил в Гестапо.
- Штирлиц топил печку. Через час печка утонула.
- Штирлиц порол чушь. Чушь отчаянно визжала.

Необходимость знаний о мире

– “Time flies like an arrow”
  - Сколько интерпретаций?

Необходимость знаний о мире

- “Time flies like an arrow”
  - Сколько интерпретаций?
    - 1. time passes quickly like an arrow?
    - 2. command: time the flies the way an arrow times the flies
    - 3. command: only time those flies which are like an arrow
    - 4. “time-flies” are fond of arrows

Подходы в компьютерной лингвистике

- До 90-х годов 20 века
  - Model-driven – модели, словари, правила
- 1995-2005
  - Переходный период
- После 2005
  - Data-driven – подходы, основанные на закономерностях, обнаруженных в данных
  - => машинное обучение
    - С 2013 года глубокие нейронные сети

Данные

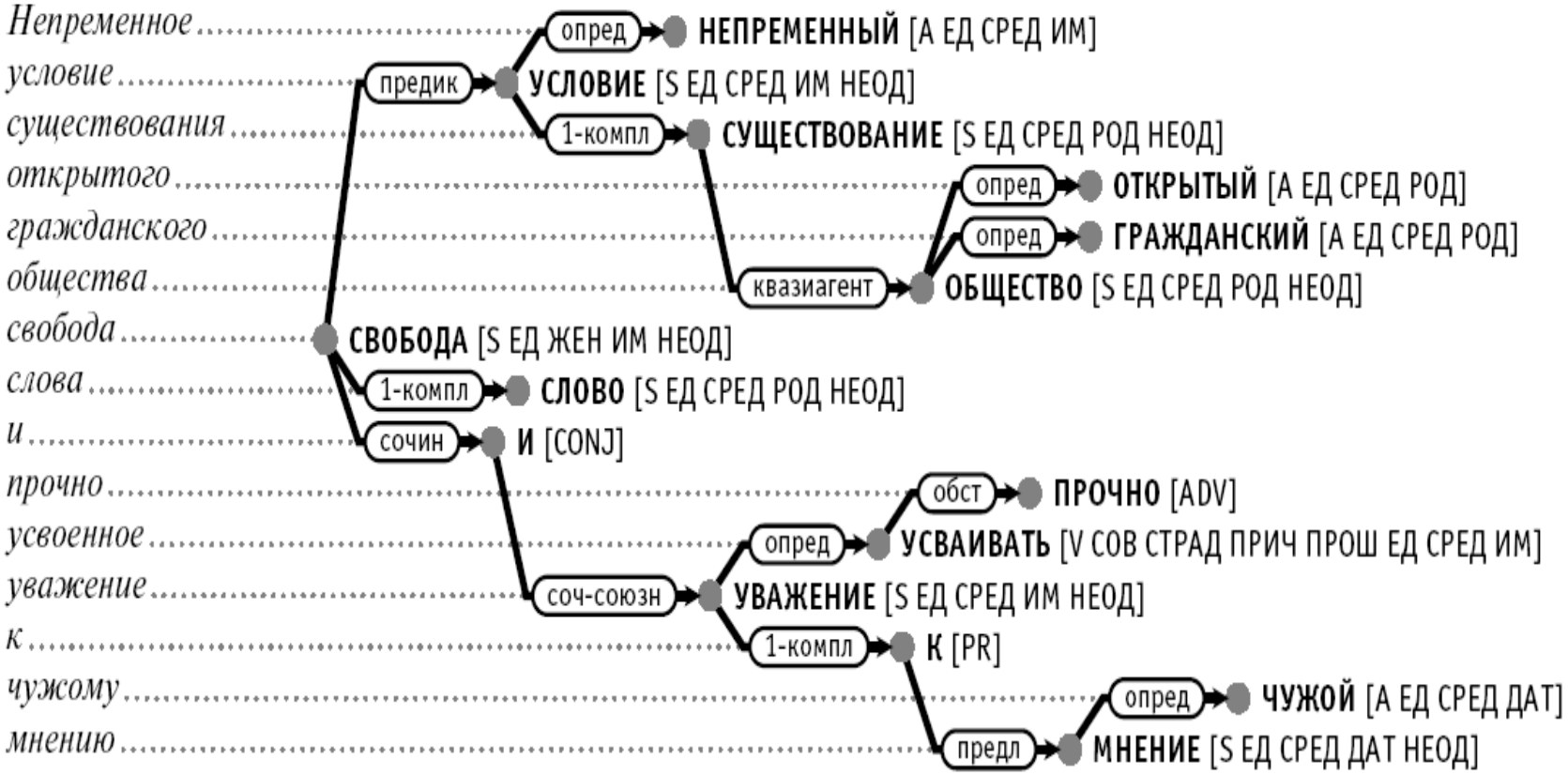
- Размеченные данные
  - Развитие алгоритмов, воспроизводящих ручную разметку
  - Машинное обучение с учителем (supervised machine learning)

- Неразмеченные данные
  - Набор данных (тексты, речь, диалоги)
  - Автоматический вывод закономерностей
  - Машинное обучение без учителя (unsupervised machine learning)

Разметка данных

- Экспертная разметка
  - Дорого
  - Может не отражать восприятие «обыкновенных» людей
- Краудсорсинг – набор неквалифицированных людей для подготовки данных
  - Подготовка заданий
  - Вознаграждение
  - Контроль качества
- Пользовательская разметка (user-generated content)
  - Хэштеги и смайлики для твитов
  - Ссылки в Википедии

Экспертная разметка:SyntagRus



Краудсорсинг

- Специальные платформы:
  - Mechanical turk (Mturk) – мало русскоязычных разметчиков
  - Yandex Toloka
- Индивидуальные проекты сбора ответов людей
- Игры со специальной целью: Game with a purpose (GWAP)
  - Как сделать так, чтобы многим людям было интересно играть и тем самым цель игры достигалась

Оценка семантической близости слов краудсорсингом

№	Первое слово	Второе слово	Насколько *связаны* эти слова?
1	чашка	напиток	не связаны
2	автомобиль	машина	не связаны
3	производство	поход	не связаны
4	Арафат	терроризм	не связаны
5	поездка	машина	не связаны
6	психология	наука	не связаны
7	перечисление	категория	не связаны
8	среда	новости	не связаны
9	любовь	секс	не связаны

GWAP (GAME WITH A PURPOSE)

- Players have fun, creators get data as by-product

foldit BETA
Solve Puzzles for Science

Phylo
A human computing framework for comparative genomics

eterna
Make Molecules Advance Science

GET READY TO TARGET TUBERCULOSIS

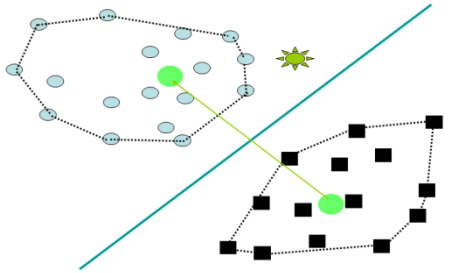
Пользовательский контент

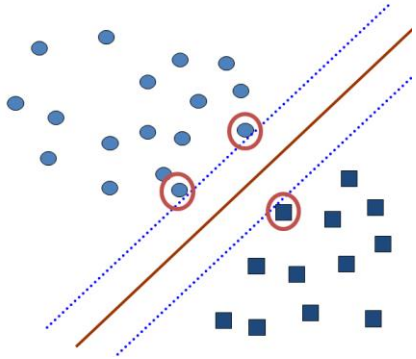
- Википедия
  - Энциклопедическая статья
  - Соответствия между языками
  - Привязка статей к категориям и др.

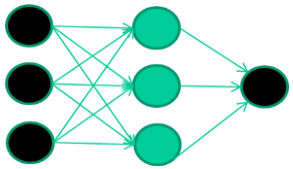
- Твиттер: разметка смайликами и хэштегам

Размеченные данные: машинное обучение с учителем

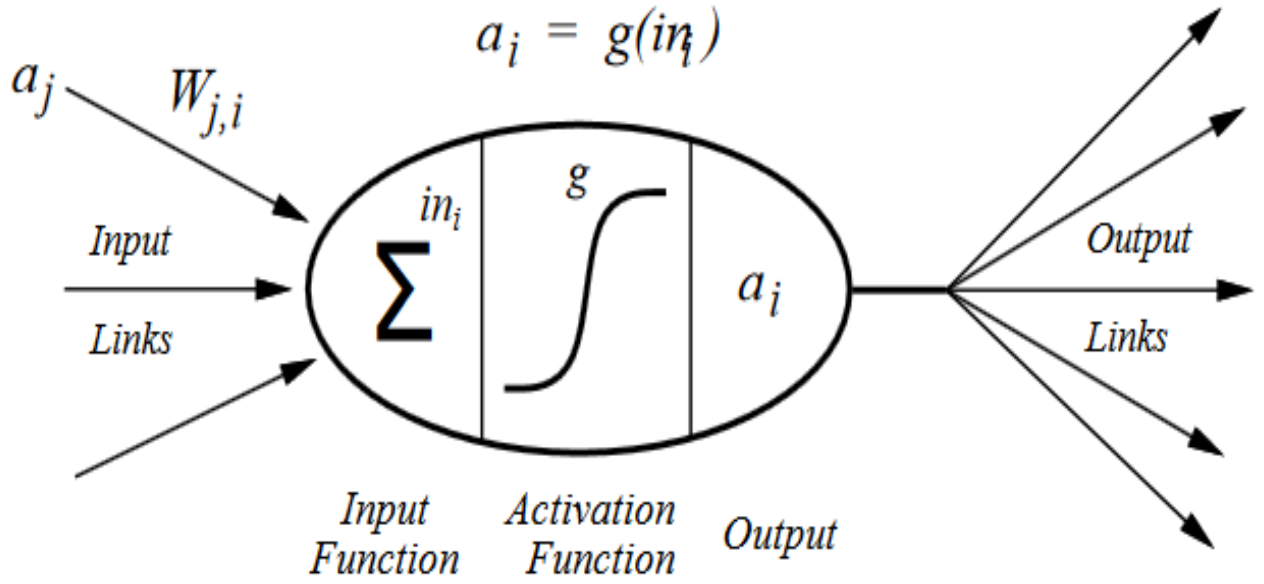
- Классические методы
  - Деревья решений
  - Наивный байесовский классификатор
  - Метод опорных векторов (SVM) и др.
- Методы анализа последовательносей
  - Скрытые марковские модели (HMM)
  - Случайные условные поля (CRF)
- Комбинация классификаторов
  - Gradient boosting
  - Random Forest
- Пакеты машинного обучения:
  - например http://scikit-learn.org/stable/
- Нейронные сети
  - Глубокое обучение





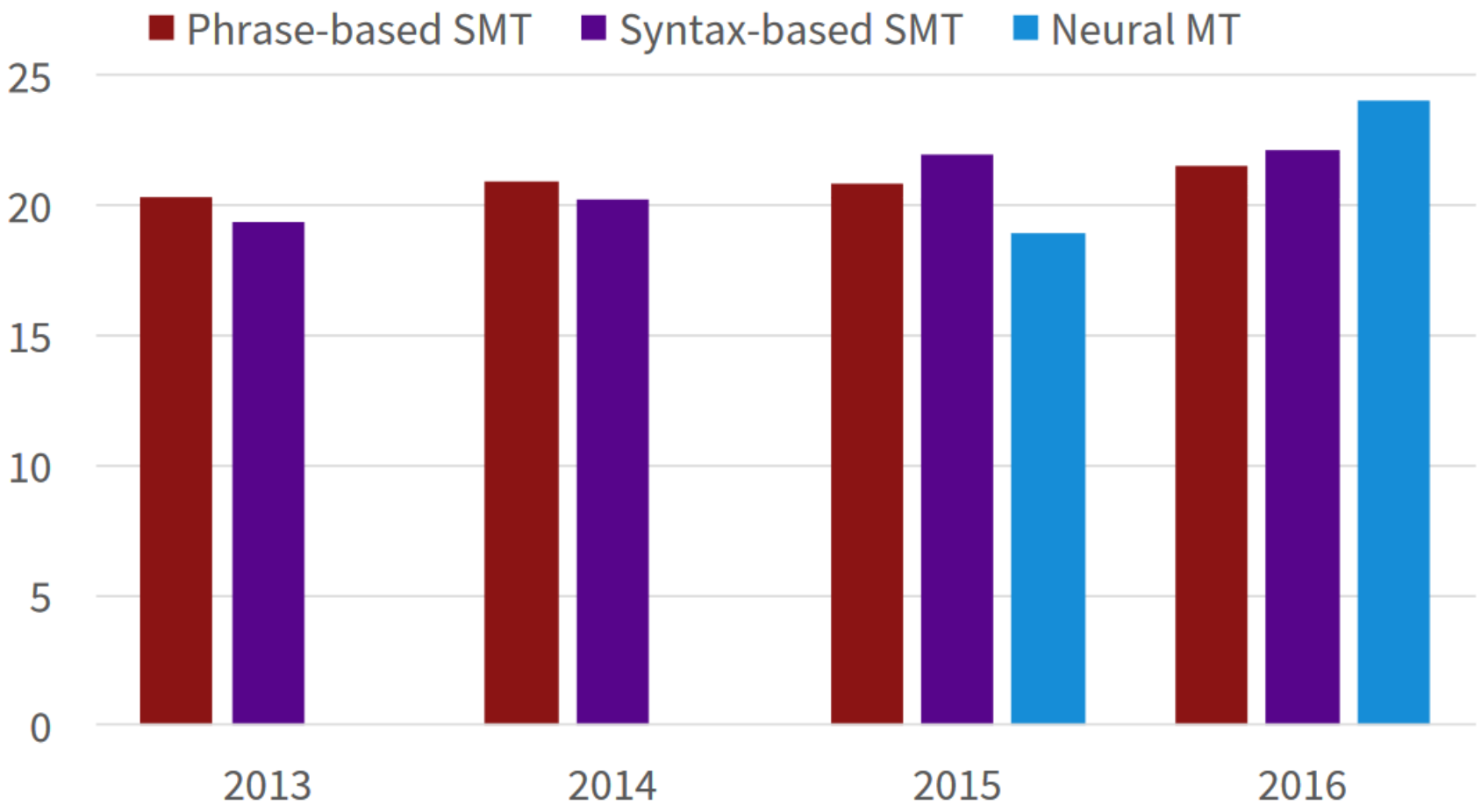


Структура искусственного нейрона



$$a_i = g(\sum_j W_{j,i} a_j)$$

Пример: Прогресс в машинном переводе



From [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

Неразмеченные данные

- Кластеризация – группирование сходных объектов
- Кластеризация документов
  - Кластеризация похожих новостей в новостных сервисах
  - Кластеризация похожих документов
  - Анализ текстовых коллекций
- Кластеризация слов
  - Статистическое выделение тем
  - Выделение похожих слов
    - Традиционные подходы дистрибутивной семантики
    - **Подходы на основе нейронных сетей:**

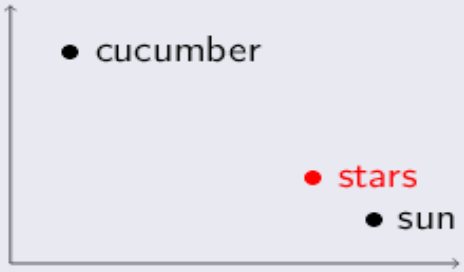
Кластеризация слов: традиционный дистрибутивный подход

he curtains open and the stars shining in on the barely
ars and the cold , close stars " . And neither of the w
rough the night with the stars shining so brightly , it
made in the light of the stars . It all boils down , wr
surely under the bright stars , thrilled by ice-white
sun , the seasons of the stars ? Home , alone , Jay pla
m is dazzling snow , the stars have risen full and cold
un and the temple of the stars , driving out of the hug
in the dark and now the stars rise , full and amber a
bird on the shape of the stars over the trees in front
But I could n't see the stars or the moon , only the
they love the sun , the stars and the stars . None of
r the light of the shiny stars . The plash of flowing w
man 's first look at the stars ; various exhibits , aer
rief information on both stars and constellations, inc

Construct vector representations

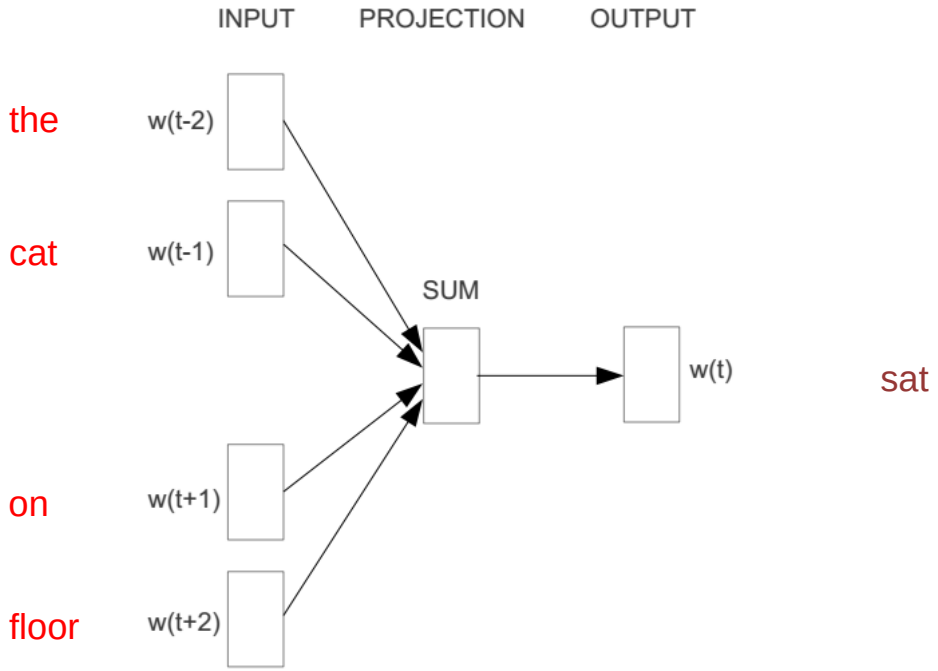
	shining	bright	trees	dark	look
stars	38	45	2	27	12

Similarity in meaning as vector similarity



Word2vec – Continuous Bag of Word

- "The cat sat on floor"
  - Window size = 2



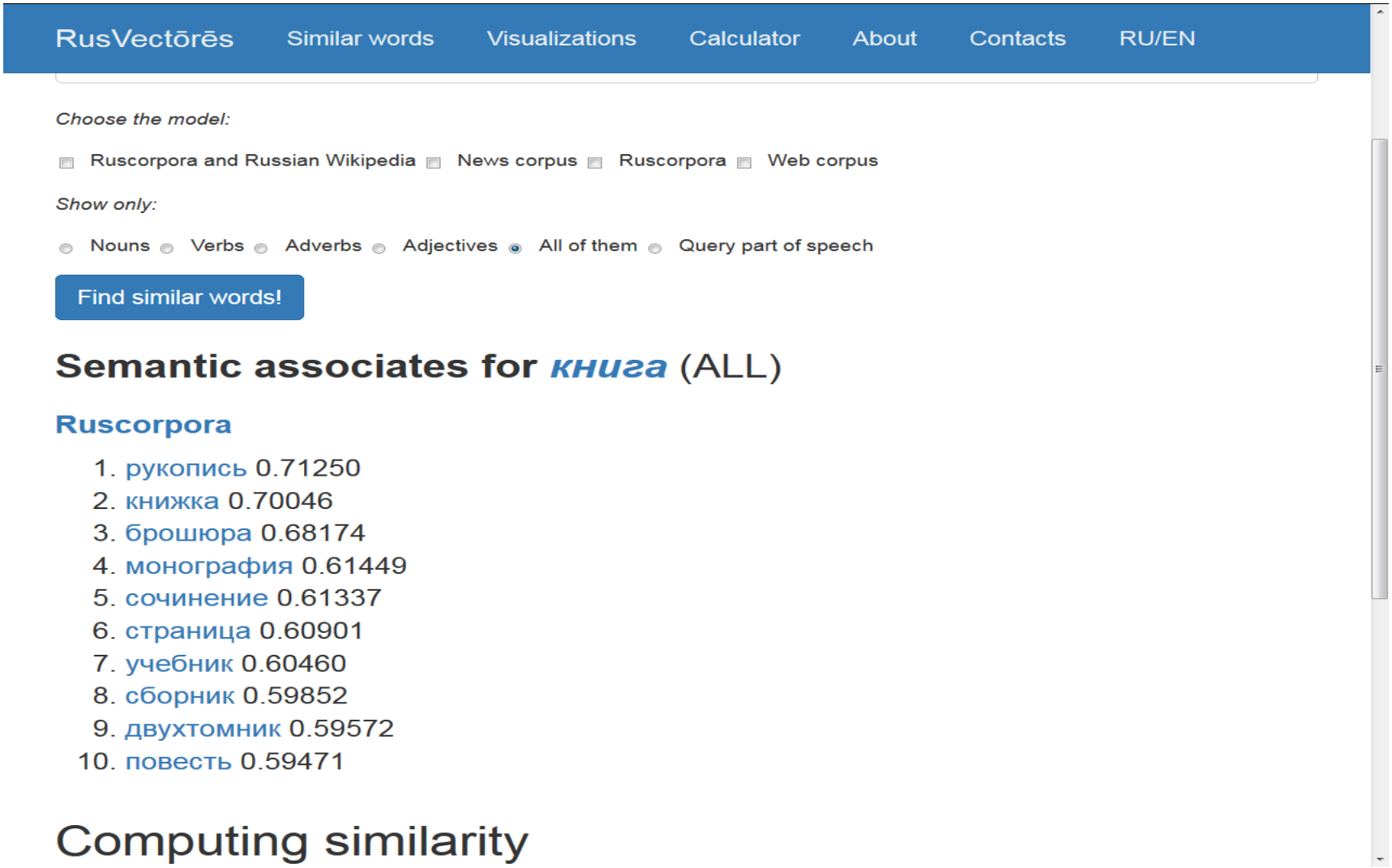
Новые подходы к представлению слов на основе векторов сокращенной размерности

- Word embeddings
- Размерность 300-1000 измерений
- Пакет - word2vec
- Высокая эффективность расчетов

- В последнее время еще более качественные и эффективные представления слов, конструкций, предложений

$$linguistics = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Проект Rusvectores: близкие слова по корпусу



Модели на сайте RusVectores

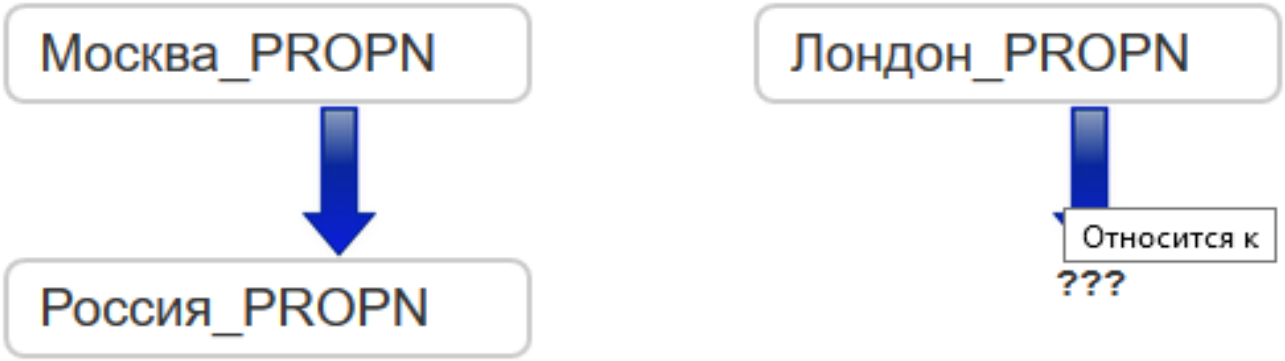
Модели

Все модели можно скачать и свободно использовать на условиях лицензии CC-BY (**жирным** выделены модели, доступные для использования в веб-интерфейсе).

Таблицу можно (и нужно) пролистывать по горизонтали!

Постоянный идентификатор	Скачать	Корпус	Размер корпуса	Объём словаря	Частотный порог	Тагсет	Алгоритм	Размерно вектора
ruscorpora_upos_cbow_300_20_2019	462 Мбайт	НКРЯ	270 миллионов слов	189 193	5 (потолок словаря 250K)	Universal Tags	Continuous Bag-of-Words	300
ruwikiruscorpora_upos_skipgram_300_2_2019	608 Мбайт	НКРЯ и Википедия за декабрь 2018	788 миллионов слов	248 978	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300
tayga_upos_skipgram_300_2_2019	610 Мбайт	Тайга	почти 5 миллиардов слов	249 565	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300
tayga_none_fasttextcbow_300_10_2019	839 Мбайт	Тайга	почти 5 миллиардов слов	192 415	5 (потолок словаря 250K)	Нет	fastText CBOW (3..5-граммы)	300
news_upos_skipgram_300_5_2019	611 Мбайт	Русскоязычные новости	2.6 миллиарда слов	249 318	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300
araneum_none_fasttextcbow_300_5_2018	854 Мбайта	Araneum	около 10 миллиардов слов	195 782	400	Нет	fastText CBOW (3..5-граммы)	300
ruscorpora_none_fasttextskipgram_300_2_2019	710 Мбайт	НКРЯ	270 миллионов слов	164 996	5 (потолок словаря 250K)	Нет	fastText Skipgram (3..5-граммы)	300
ruwikiruscorpora-func_upos_skipgram_300_5_2019	606 Мбайт	НКРЯ и Википедия за декабрь 2018 (с функциональными	788 миллионов слов	248 118	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300

Word2vec: Семантические закономерности



НКРЯ и Wikipedia

1. англия$_{PROPN}$ 0.58
2. европа$_{PROPN}$ 0.54
3. великобритания$_{PROPN}$ 0.52
4. страна$_{NOUN}$ 0.48
5. франция$_{PROPN}$ 0.47

Извлечение именованных сущностей

Пример задачи

Извлечение именованных сущностей

В штабе Зеленского заявили о принадлежности **Крыма Украине**

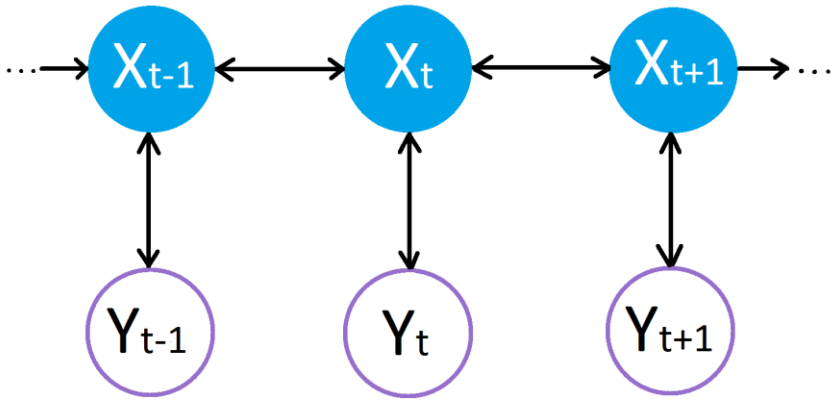
В **штабе кандидата в** президенты **Украины Владимира Зеленского заявили**, что **Крым**»? был, остается и будет частью **Украины**», сообщает телеканал «**112 Украина**».

NECRF
[PER] ЗЕЛЕНСКИЙ

«**Разговоров** о размене **Крыма** на **мир** на **Донбассе** быть не может ни при каких обстоятельствах. Так же, как не может **быть предметом переговоров** сепаратизация отдельных **районов** или **регионов Украины**», — **говорится** в **сообщении команды Зеленского**. **Отмечается**, что **Крым должен** вернуться в состав **Украины** с **компенсацией** от **России**.

В **штабе** добавили, что **представителям Киева придется вести переговоры с Москвой**, однако **делать** это необходимо при участии **западных партнеров** и **не на «оккупированных территориях»**.

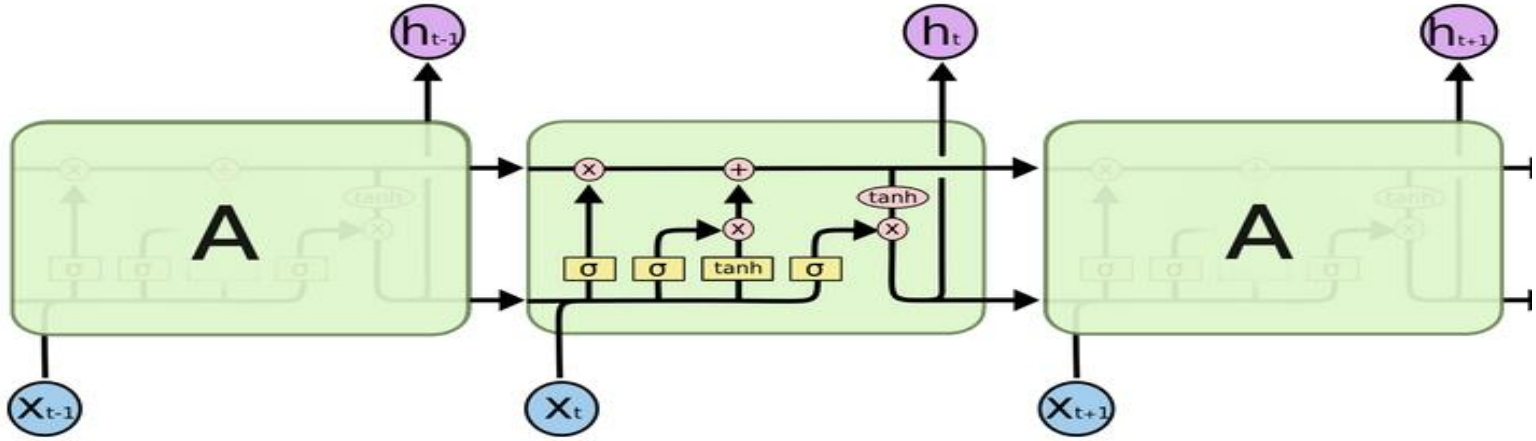
Лучший метод до появления нейронных сетей CRF



CRF: Представление слов в виде признаков

- Признаки токена
  - слово, часть речи, написание, суффиксы, знак препинания
- Такие же признаки для соседних слов (два влево – два вправо)
- Присутствие в словарях
  - Имена, фамилии, отчества
  - Географические названия
  - Названия должностей, типы организаций и т.п.
- Семантические кластеры слов

Сети LSTM: общая картина



Первый слой вычисляет, насколько на данном шаге ему нужно забыть предыдущую информацию — по сути множители к компонентам вектора памяти

Второй слой вычисляет, насколько ему интересна новая информация, пришедшая с сигналом — такой же множитель, но уже для наблюдения.

На третьем слое вычисляется линейная комбинация памяти и наблюдения с только вычисленными весами для каждой из компонент. Так получается новое состояние памяти, которое в таком же виде передаётся далее.

Все вместе

- Первый слой вычисляет, насколько на данном шаге ему нужно забыть предыдущую информацию.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

- Второй слой вычисляет, насколько ему интересна новая информация, пришедшая с сигналом

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

- На третьем слое вычисляется линейная комбинация памяти и наблюдения с только вычисленными весами для каждой из компонент

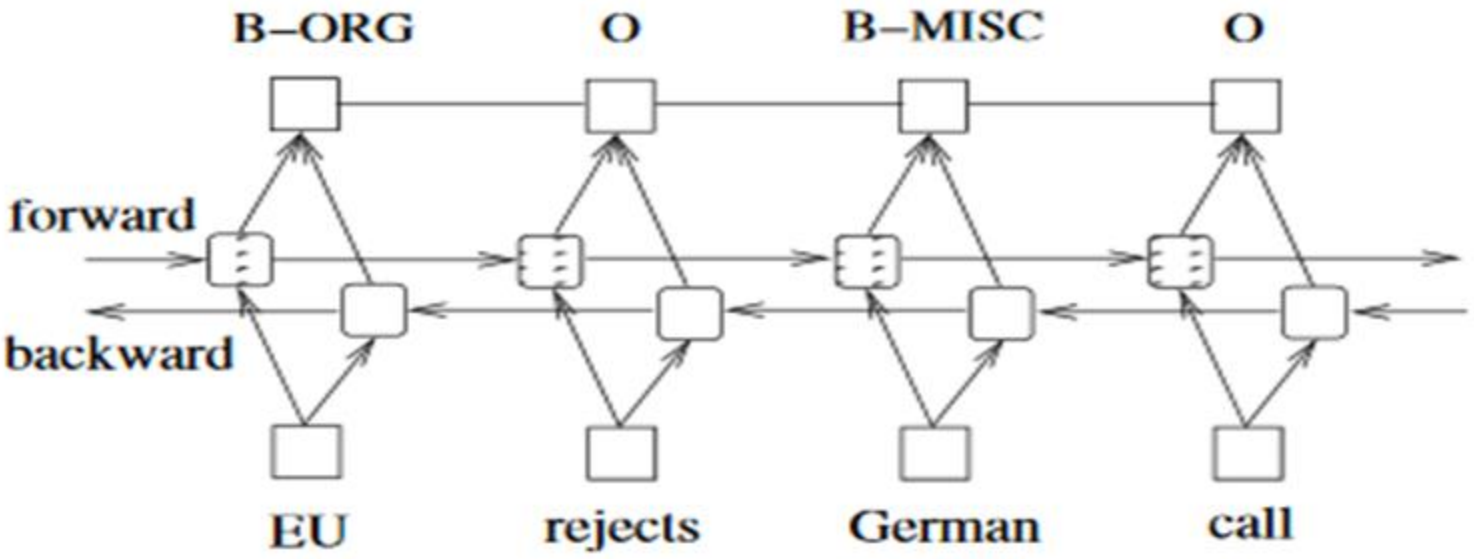
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- Выходной слой:

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

BI-LSTM-CRF



- На первом слое не слова, а векторные представления слов
- Можно делать посимвольные представления слов – дает возможность обрабатывать неизвестные слова
- Не нужно отдельное представление в виде признаков

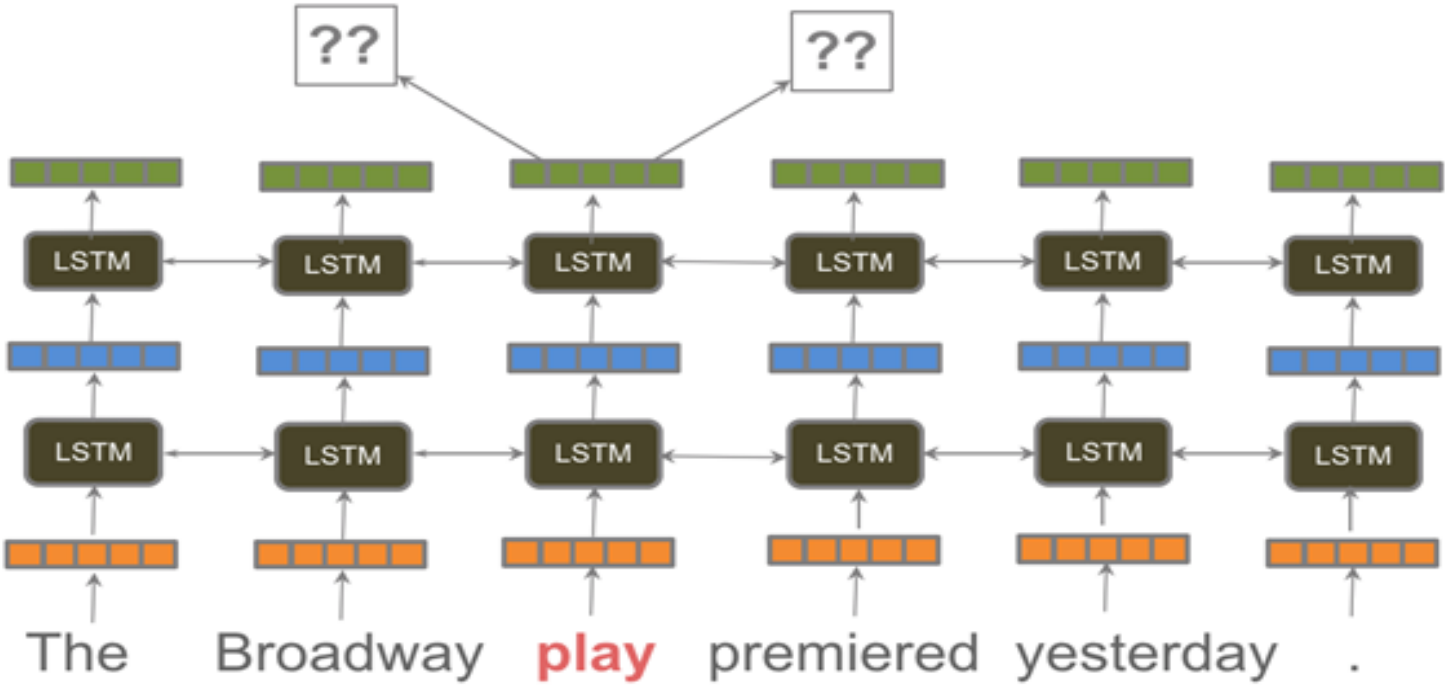
Сравнение моделей (Huang et al., 2015)

Pos – предсказание частей речи
CONLL 2000 - chunking, выделение именных групп
CONLL 2003 – извлечение именованных сущностей

		POS	CoNLL2000	CoNLL2003
Random	Conv-CRF (Collobert et al., 2011)	96.37	90.33	81.47
	LSTM	97.10	92.88	79.82
	BI-LSTM	97.30	93.64	81.11
	CRF	97.30	93.69	83.02
	LSTM-CRF	**97.45**	93.80	84.10
	BI-LSTM-CRF	97.43	**94.13**	**84.26**
Senna	Conv-CRF (Collobert et al., 2011)	97.29	94.32	88.67 (89.59)
	LSTM	97.29	92.99	83.74
	BI-LSTM	97.40	93.92	85.17
	CRF	97.45	93.83	86.13
	LSTM-CRF	97.54	94.27	88.36
	BI-LSTM-CRF	**97.55**	**94.46**	**88.83 (90.10)**

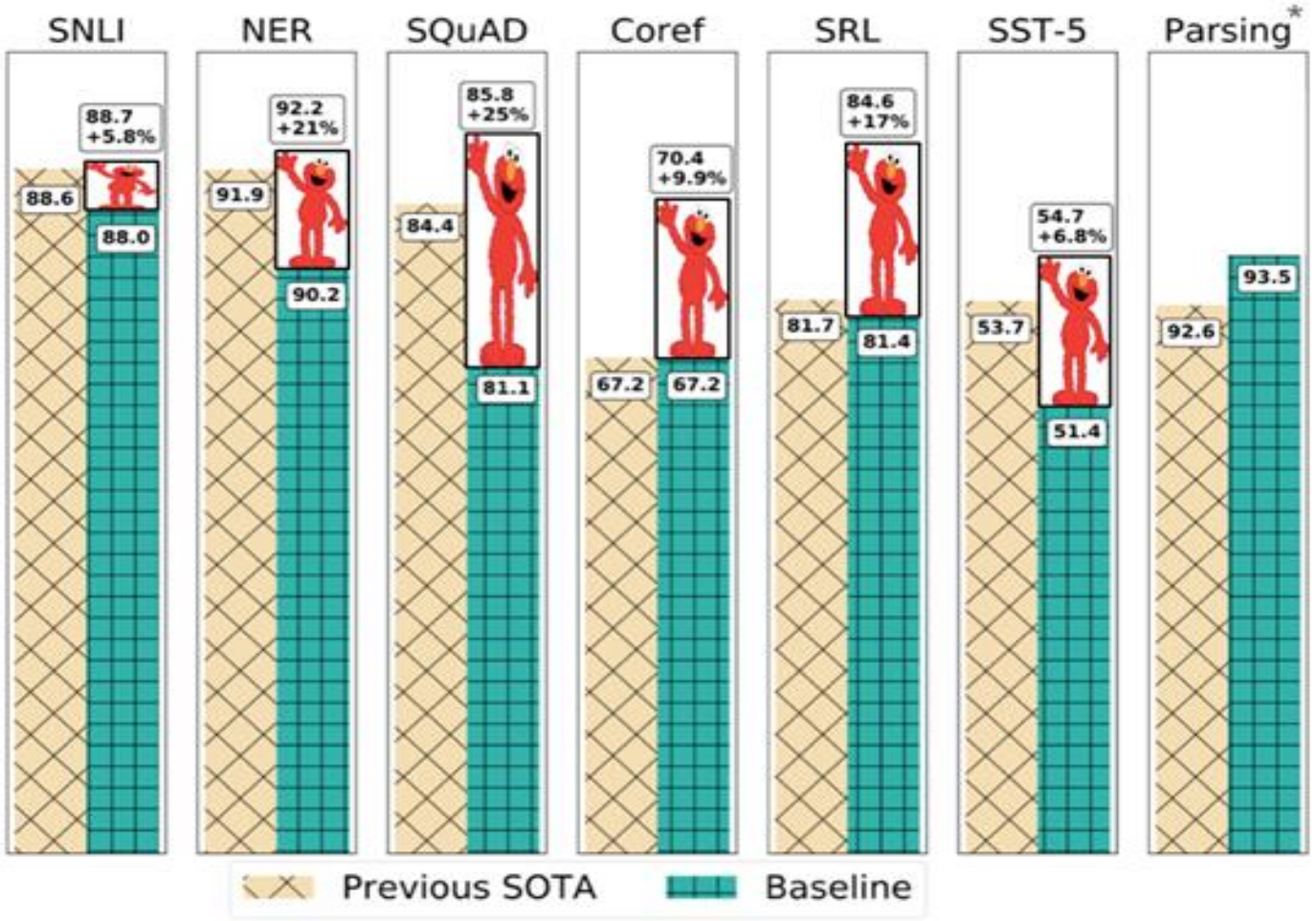
Random – векторные представления слов обучаются в процессе обучения Решения; Senna – заранее рассчитанные векторные представления на большой коллекции; Значения в скобках – добавлены признаки на словарях

2018: Контекстные представления слов: ELMO, BERT



Слово play получает разные векторные представления в разных контекстах

Применение контекстных представлений дает улучшения во многих задачах (ELMO)



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

Результаты на данных CONLL-2003 для новых способов подсчетов контекстных представлений слов: Bert 2018

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

sults. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

Однако

- Машинное обучение не решает все задачи
  - Полного искусственного интеллекта пока не создано
  - Нейронная сеть решает точно ту задачу, для которой она обучалась
  - Нужны хорошо подготовленные обучающие выборки
  - Обучающие выборки устаревают со временем
- Возможность намеренного искажения ввода для обхода системы построенной на глубоком обучении
  - Adversarial learning is a technique employed in the field of machine learning which attempts to fool models through malicious input

Добавление специального «шума» заставляет нейронную сеть ошибиться

In a paper by Goodfellow, Shlens, and Szegdy, modern machine learning methods are described as building “a Potemkin village that works well on naturally occuring data, but is exposed as a fake when one visits points in space that do not have high probability in the data distribution

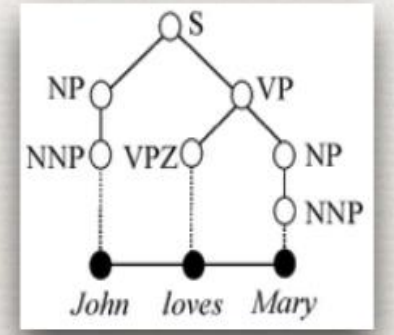
The Language Resource Spectrum: A Perspective from Google

Ryan McDonald

Google

Google NLU team
Google Linguistics team

2016

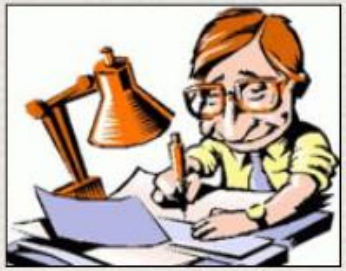


unsupervised

weakly supervised

fully supervised

α β γ δ



High quality annotations

Crowd-sourced

Pre-existing resources

Software Engineer

Auto resources

Models

Active Learning

Заключение

- Задачи автоматической обработки текстов
  - Необходимость данных (размеченных и неразмеченных)
  - Большинство лучших результатов на разных эталонных данных получено на основе нейронных сетей
  - На практике все сложнее:
    - меньше или нет данных,
    - возможность специального искажения данных (вопросы безопасности)

- Решение задач АОТ в настоящее время - это поиск подхода, в котором могут комбинироваться
  - Готовые ресурсы, сделанные вручную (словари и правила)
  - Ресурсы, построенные автоматически на основе больших объемов текстовых данных
  - Методы машинного обучения с учителем, работающие на основе специально подготовленной обучающей коллекции