NLP && production

Эволюция задач и алгоритмов на текстах

Штех Геннадий @ NAUMEN • 07.04.2018

Задачи бизнеса

Введение

1. **Введение:** что такое НЛП, какие задачи решает?
2. **Куда идёт индустрия** и какие неочевидные задачи из этого возникают?

Секция реального опыта

1. **Сфера информационной доступности:** польза для бизнеса, науки, законодательства

Алгоритмы НЛП

Что происходит в мире

1. **Актуальные классические подходы:** обработка текста, topic modeling, полнотекстовый поиск
2. **Развитие методов:** дистрибутивная семантика, *2vec, нейронные сети

Что исследуем мы

1. **Информационный** поиск
2. **Семантика** документов
3. **Группировка** документов
4. **Рекомендательные** системы

Инструментарий

Требования к инструментам

1. **Классические проблемы** NLP in production
2. Какие решения **применяли мы**

Очень полезные ссылки

1. **Тезисный обзор открытых проектов,** о которых надо знать NLP-инженеру

Задачи

Историческая ретроспектива

Список

1. **Поиск**
2. **Машинный** перевод
3. **Извлечение информации:** заполнение форм
4. **Распознавание** речи
5. **Кластеризация**, каталогизация, автотеггирование
6. **Классификация**, сентимент

Поиск

1. Запросная строка с полнотекстовым (Wiki)
2. Четкие фильтры (по дате, количеству цитат, исключающие какие-то слова)

Машинный перевод

1. Помощь в составлении словарей
2. Статистический перевод фраз
3. Перевод документов/интерфейсов

Извлечение информации

1. Извлечение событий из потока новостей по шаблону
2. Автоматизированное заполнение форм, описаний, номенклатуры
3. Составление “карты знаний” с отношениями на объектах или понятиях “объект1-отношение-объект2”

Распознавание речи

1. Речевой ввод aka “Диктограф”
2. Автоматизация колл-центров
3. Субтитры и синхронный перевод

Задачи на текстах

Кластеризация

1. Маркетинговые исследования отзывов
2. Информационная разведка

Каталогизация

1. Составление банков знаний
2. Структуризация отчетности

Задачи на текстах

Классификация

1. Спам
2. Маршрутизация обращений
3. Оценка тональности отзыва

Автотегирование

1. Сортировка потока новостей
2. Краткая аннотация темы содержания

Куда двигается индустрия

1. Улучшение существующих техник методом “всё становится лучше с нейросетями”
2. End-to-End подходы (благодаря нейросетям, опять же)
3. Улучшение интерфейсов: взрыв “чат-ботов”
4. Q&A

Куда двигается индустрия

1. Стремление “спрямить” потоки информации
2. Комбинации методов
3. Больше контекста в поисковых сессиях: “диалоговые системы”

Почему я так думаю

Или список решенных Naumen задач

1. Семантический поиск документов
2. Группировка и каталогизация научно-технических библиотек
3. Поиск заимствований и повторов
4. Диалоговые интерфейсы

Алгоритмы

Историческая ретроспектива и справка

О типичном NLP-конвейере

Предобработка текста (нормализация)

- Токенизация, стемминг/лемматизация
- “Красненькая шапочка” -> [‘красный’, ‘шапка’]

Представление

- one-hot, Bag of Words, tf-idf, nPMI, hashing trick

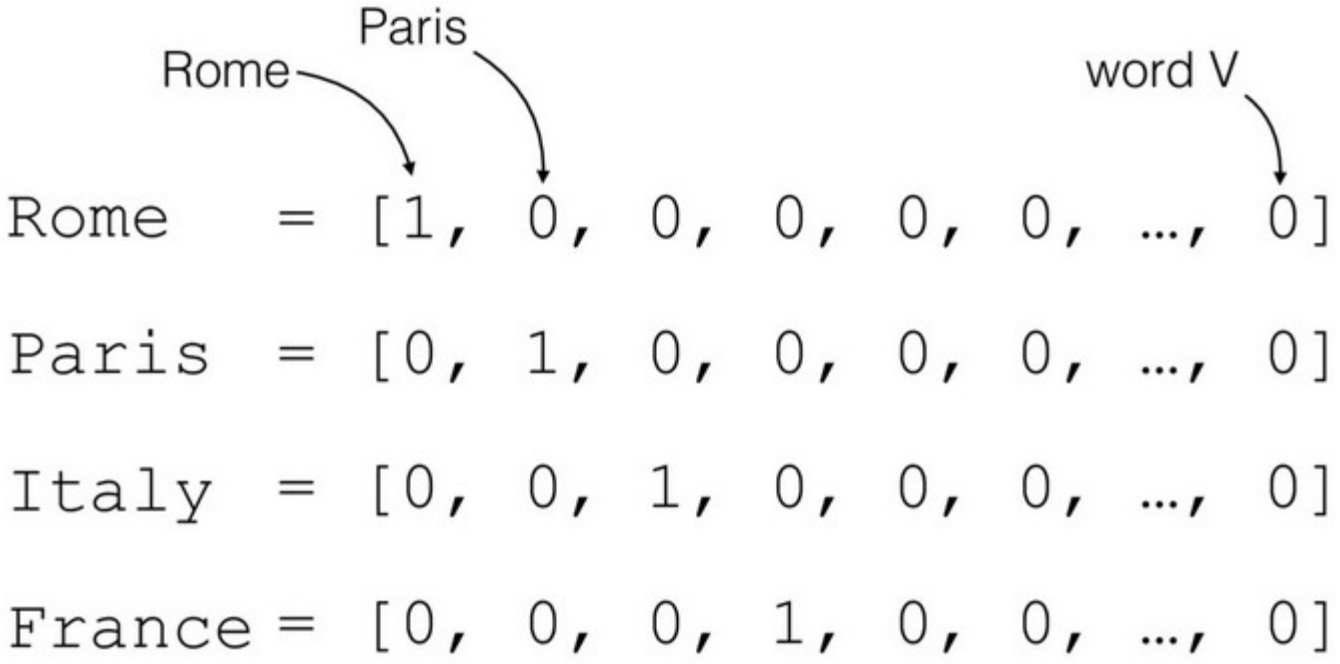
Трансформация

- Matrix decomposition, topic modeling, embeddings

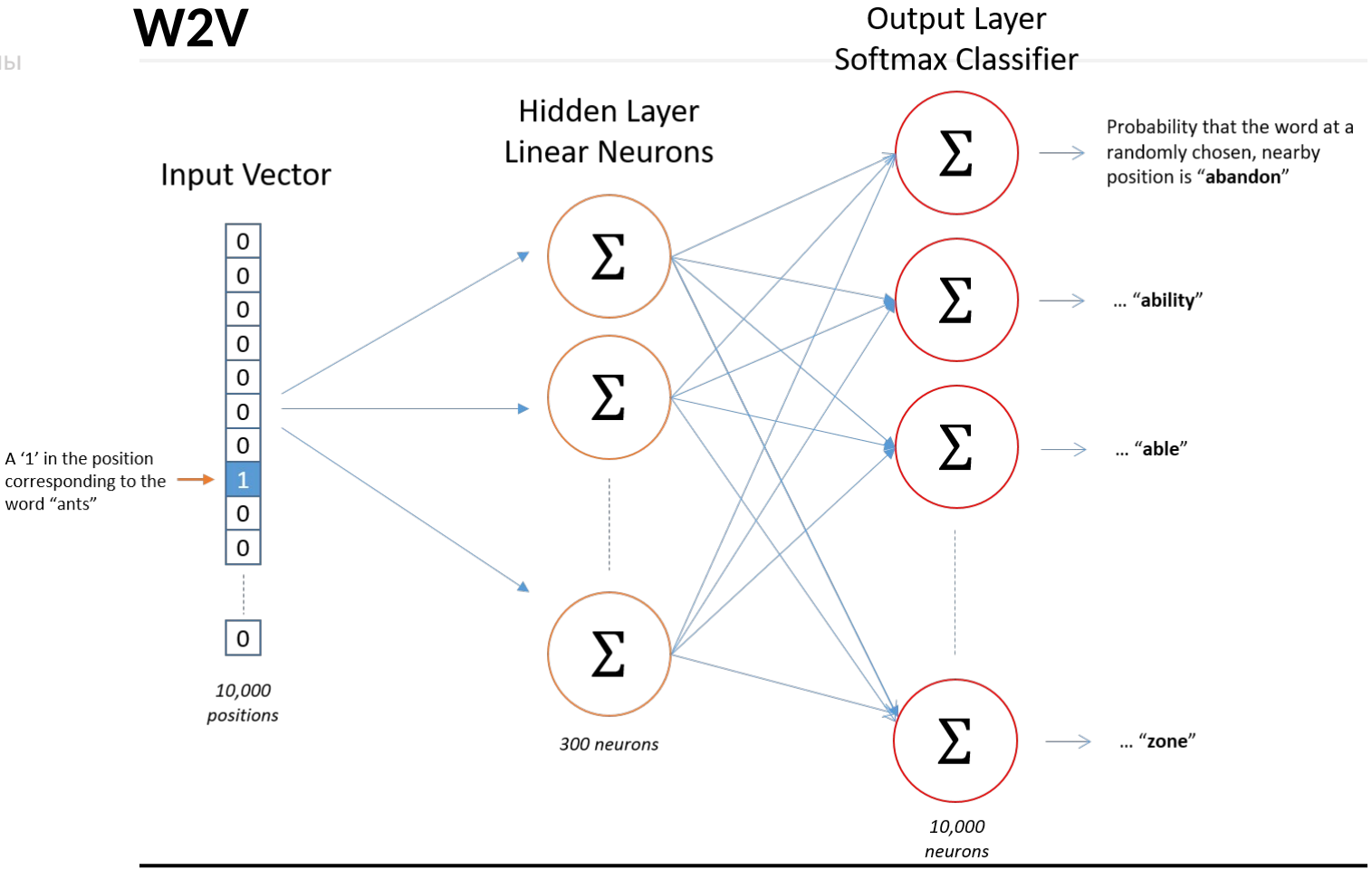
Конечный алгоритм для решения задачи

- Классификация, кластеризация, регрессия, etc

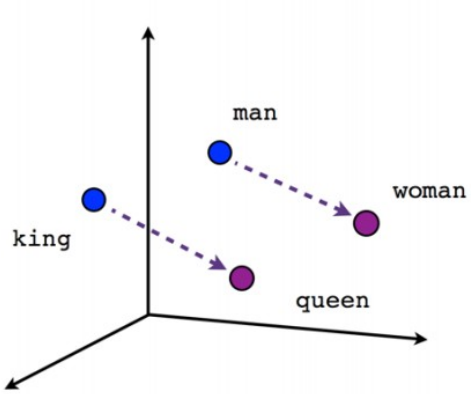
One-Hot Encoding



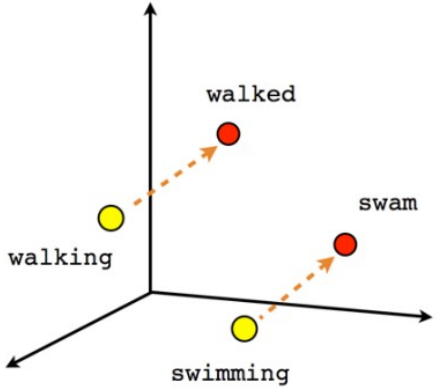
W2V


Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons

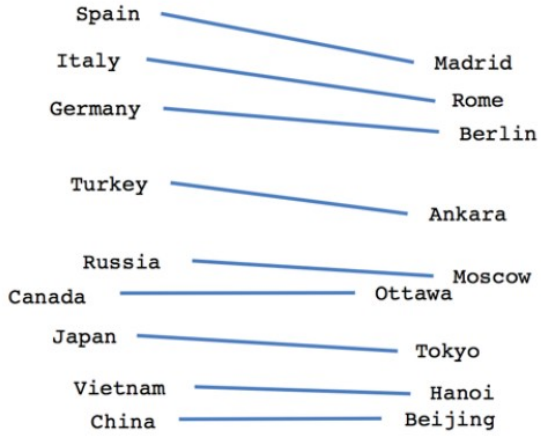
Embeddings



Male-Female



Verb tense



Country-Capital

Тематическое моделирование

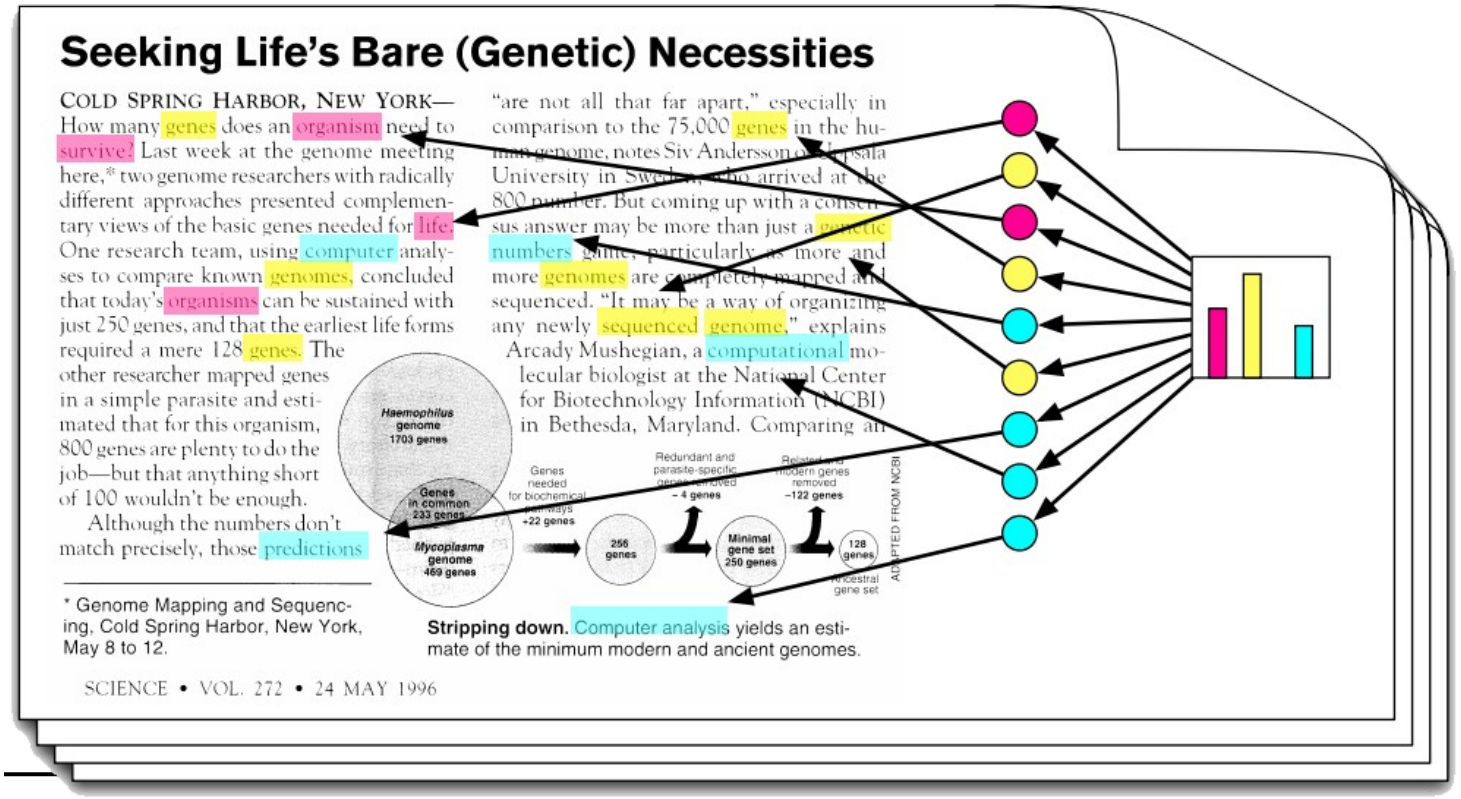
Topics *Documents* *Topic proportions & assignments*

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	



Пример тематической модели

#106: "gpu" + "performance" + "core" + "cpu" + "hardware" + "cuda" + "than" + ...

#107: "state" + "electron" + "quantum" + "algorithm" + "method" + "functional" + ...

#108: "square" + "circle" + "least" + "rectangle" + "now" + "find" + "like" + ...

#109: "correction" + "kappa" + "r2" + "energy" + "conditional" + "crf" + "exchange" + ...

#110: "f" + "h" + "g" + "p" + "\cdot" + "s" + "function" + ...

#111: "method" + "system" + "solve" + "problem" + "solver" + "equation" + "linear" + ...

#112: "state" + "action" + "reward" + "q" + "agent" + "policy" + "value" + ...

. . .

Ветки развития

Алгоритмы

Направления исследований

Специализированные и мультимодальные эмбеддинги

01.

Часто нужно уметь строить не только эмбеддинги слов и документов: изображения, метаданные, пользователи

02.

Иногда нужно захватывать структурные свойства: иерархии смыслов, иерархии документов

Направления исследований

Вероятностные методы, решающие вспомогательные задачи

01.

Вероятностные модели требуют меньше данных, легко интерпретируемы, позволяют пользователям воздействовать на работу алгоритма

02.

Иногда в терминах вероятностых моделей **легче формулировать задачи**, например DPP позволяет добиться оптимального баланса разнообразия и релевантности выдачи

Направления исследований

Разрешение синонимии и омонимии, больше лингвистики

Качественные эмбеддиги слов,
“понимающие” лингвистические особенности в конечном итоге
улучшают работу на всех задачах

State-Of-The-Art

Алгоритмы

SOTA RNN

LSTM + Attention

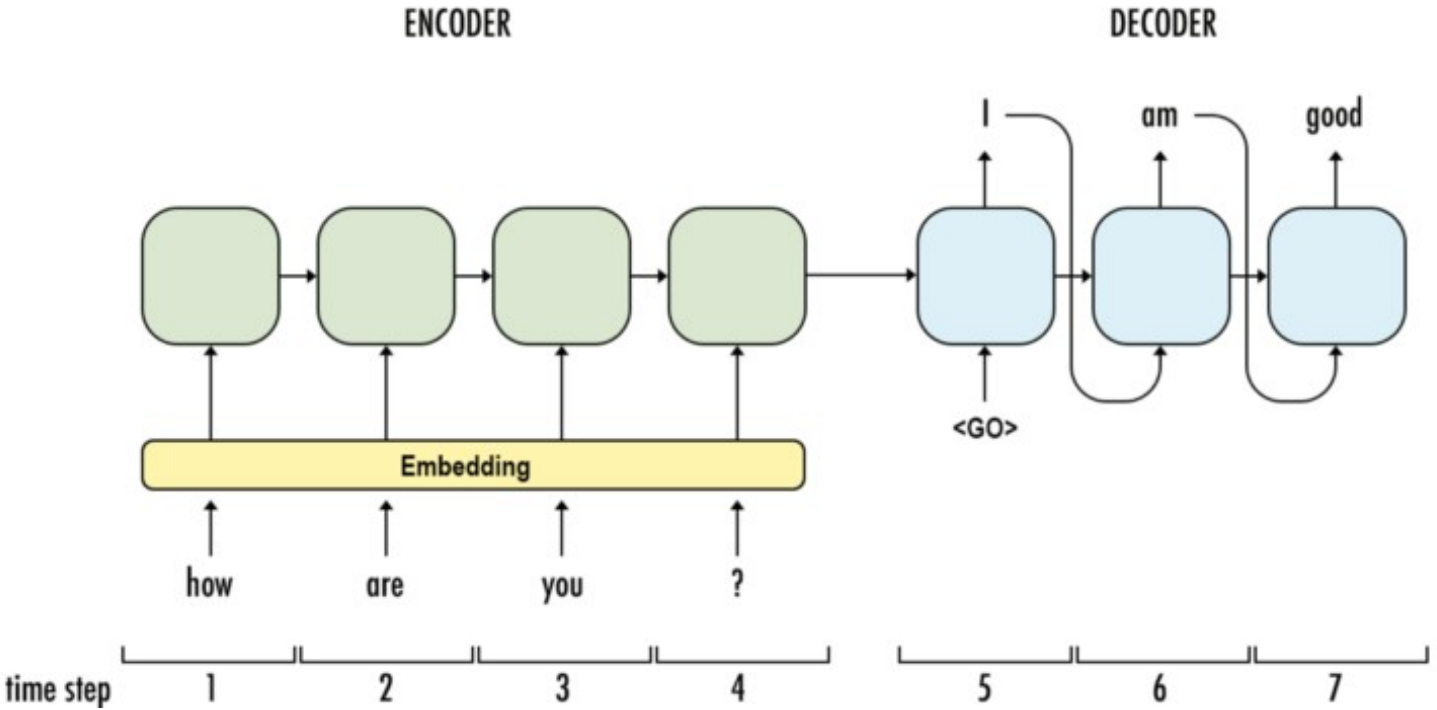
Рекуррентные сети могут “забывать” важные детали или с трудом решать задачи, которые плохо похожи друг на друга

Сети с вниманием способны “вспомнить” подходящие части последовательности, даже если они были далеко, причем именно те, которые соответствуют текущей задаче

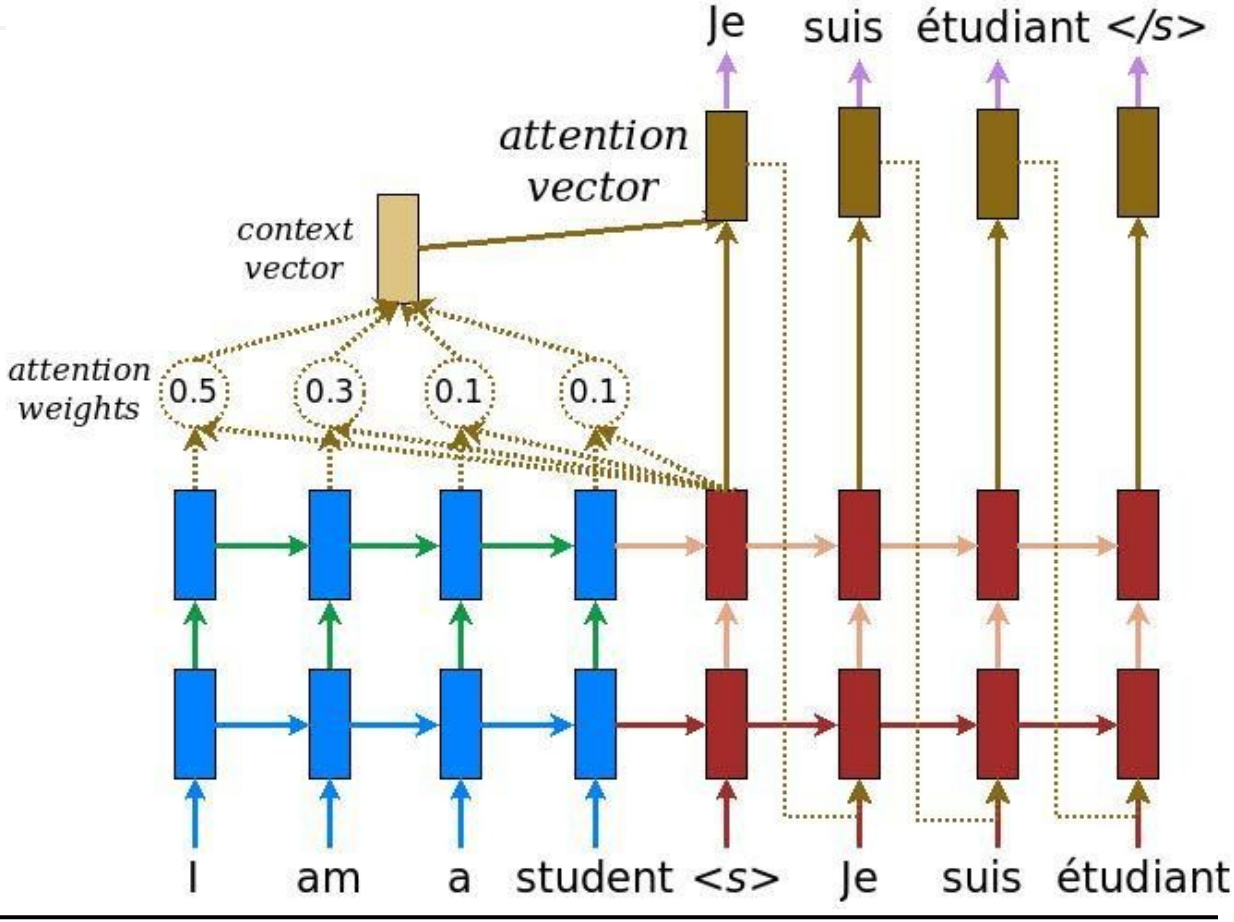
Различные модификации внимания **можно подобрать под задачу**

Некоторые модификации очень **похожи на свёрточные сети**

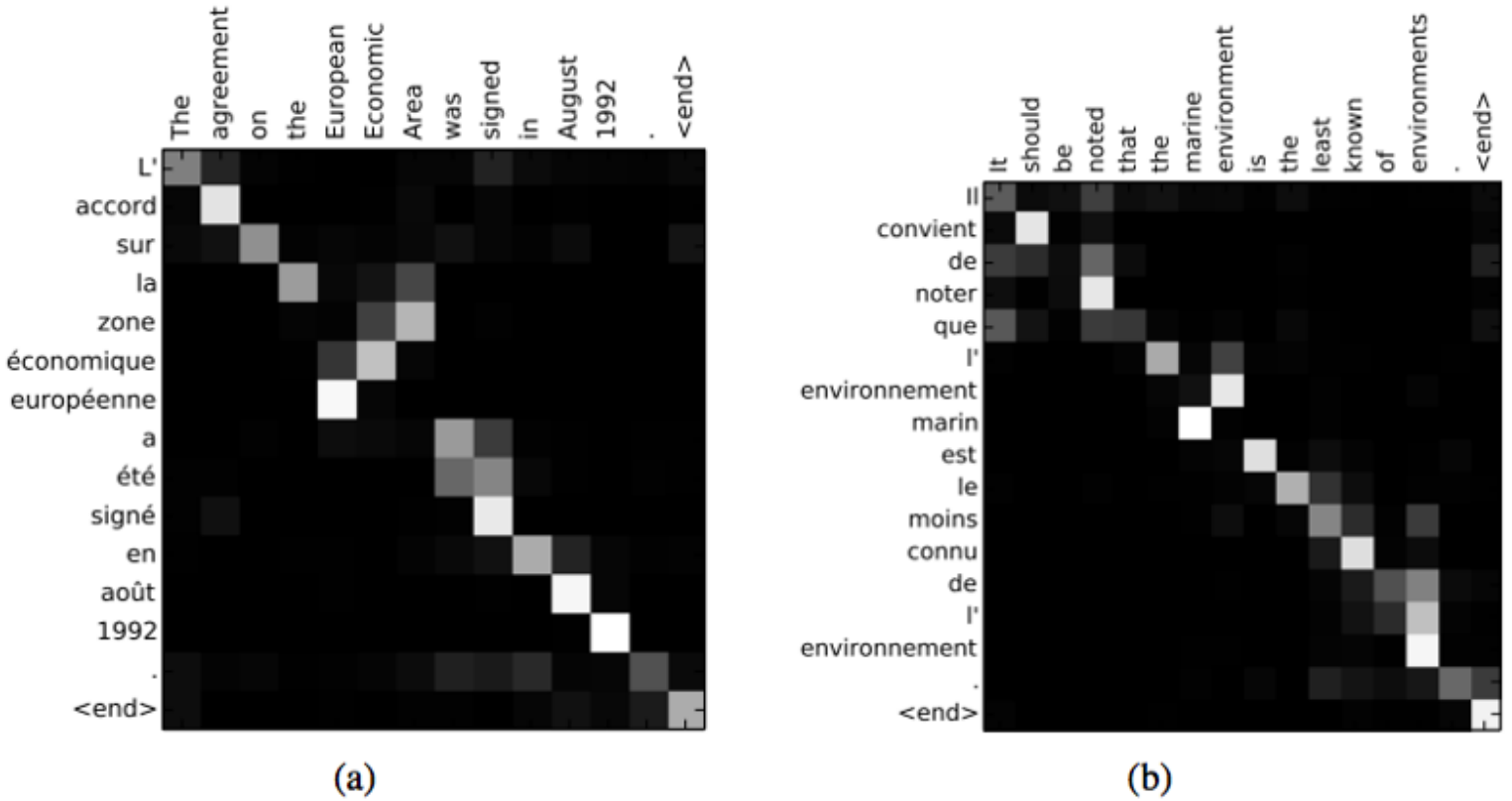
LSTM


ENCODER
DECODER
I
am
good
<GO>
Embedding
how
are
you
?
time step
1
2
3
4
5
6
7

ATTN


Je
suis
étudiant </s>
attention vector
context vector
attention weights
0.5
0.3
0.1
0.1
I
am
a
student
<s>
Je
suis
étudiant

Attention on translation task



SOTA RNN

Pointer Network

01.

Позволяют выбирать наилучший набор объектов из перечисленных в нужном порядке

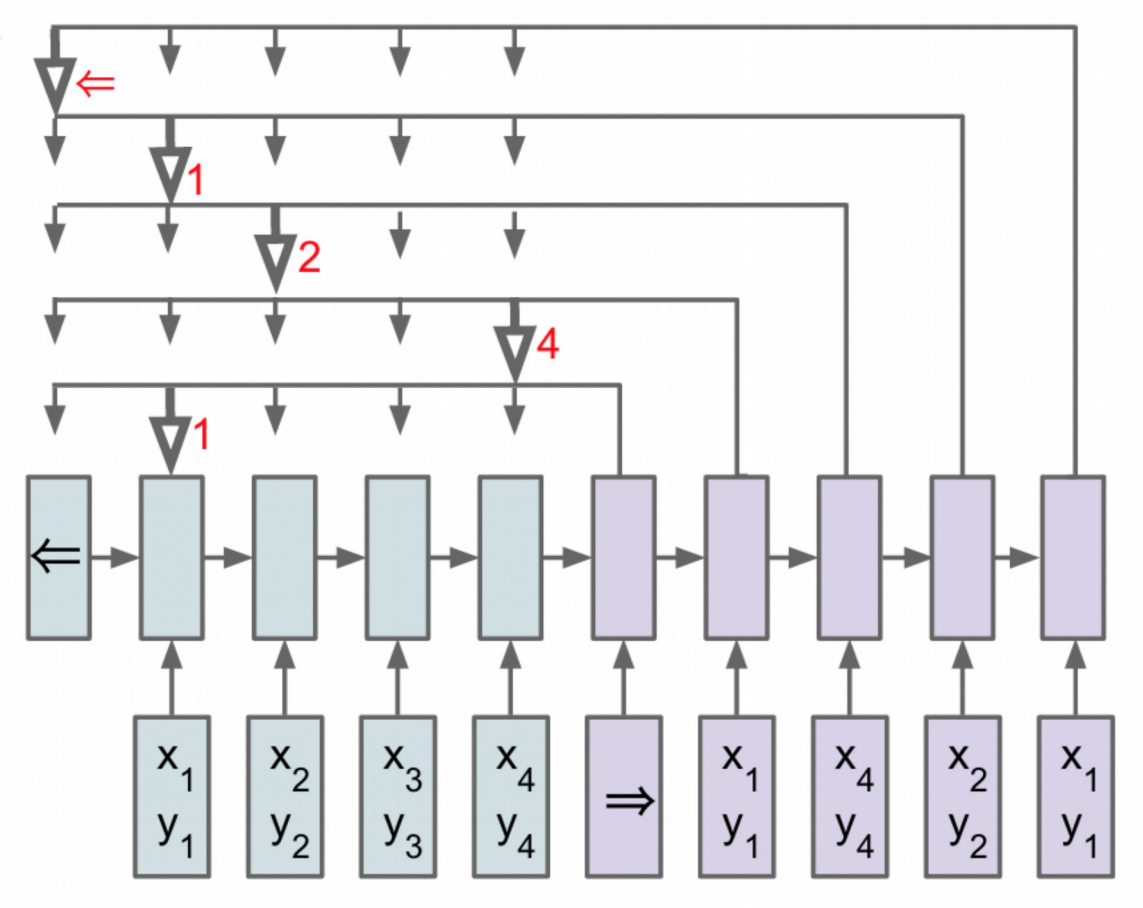
02.

Впервые были применены к задачам Коммивояжера (графы) и поиска выпуклой оболочки(геометрия)

03.

Впервые были применены к задачам Коммивояжера (графы) и поиска выпуклой оболочки(геометрия)

SOTA RNN



SOTA Embeddings

Fasttext

01.

Может работать **с минимальной предобработкой текста**

02.

Способен строить эмбеддинги даже неизвестных слов по аналогии орфорграфической структуры

SOTA Embeddings

Starspace

01.

Является скорее нетривиальным применением почти классических эмбеддингов в комлекте с эффективной реализацией

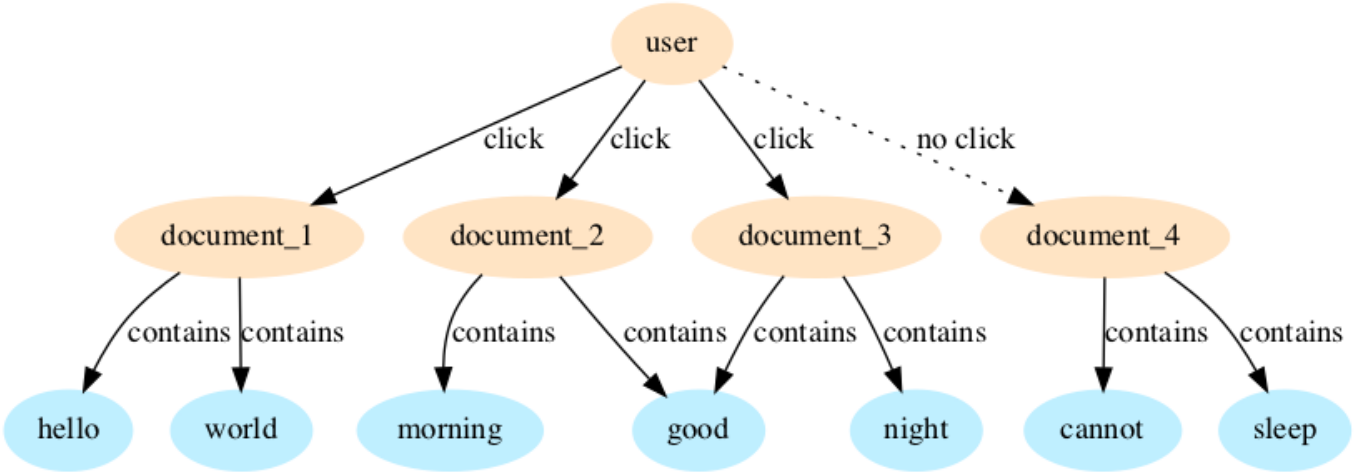
02.

Мультимодальные эмбеддинги, строит представление для всех элементов документа

03.

Можно строить карты знаний “понятие1-отношение-понятие2”

Starspace


user
click
click
click
no click
document_1
document_2
document_3
document_4
contains
contains
contains
contains
contains
contains
contains
contains
hello
world
morning
good
night
cannot
sleep

SOTA Embeddings

Poincaré

01.

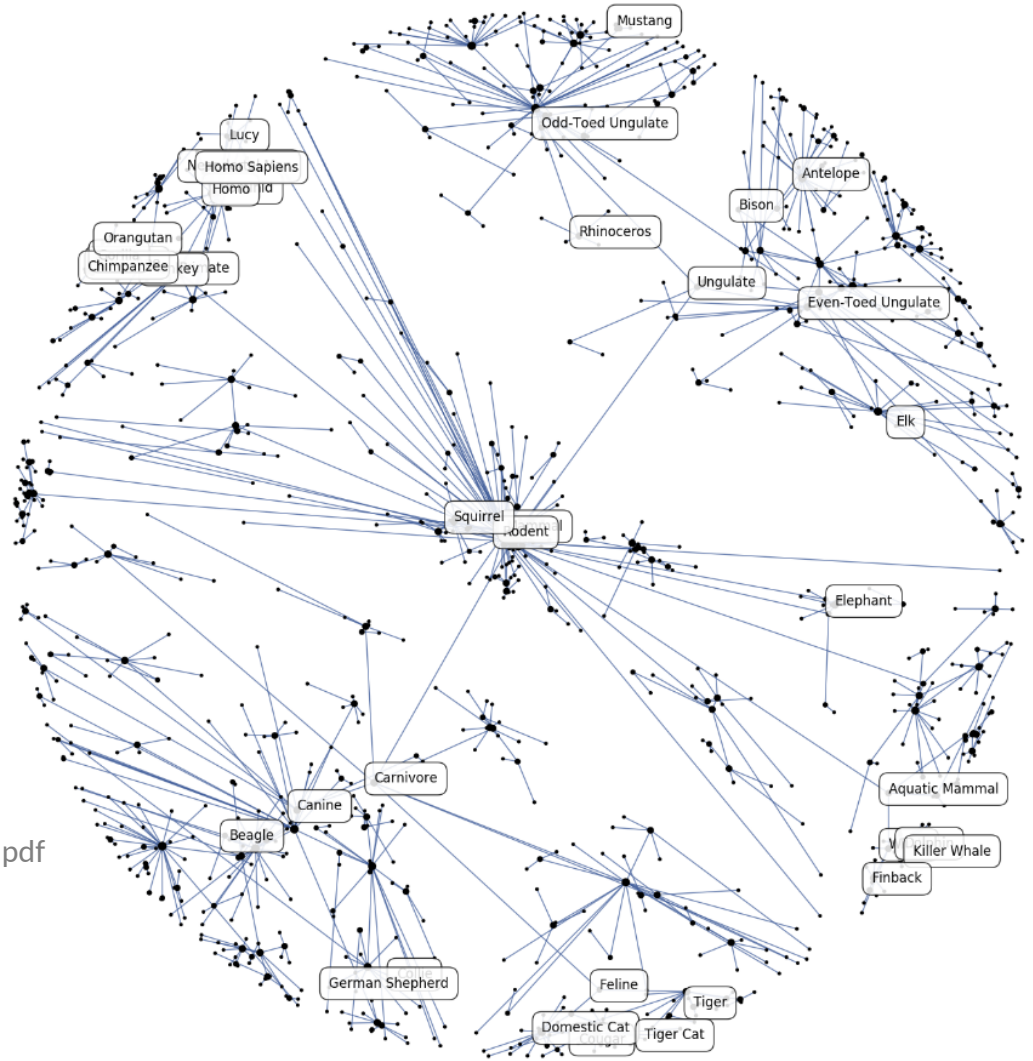
Позволяет захватить иерархическую структуру понятий “понятие1-обобщает-понятие2”

02.

Работает в нелинейных пространствах Poincaré Balls

Poincaré Embeddings for Learning Hierarchical Representation

https://arxiv.org/pdf/1705.08039.pdf



Наши направления

Алгоритмы

Сегментация документов

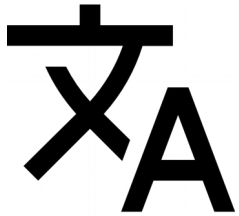
ARTM Continuous Topic Regularizer

Для поиска по большим документам **полезно нарезать их на непрерывные сегменты**, раскрывающие определенную тему или говорящие об одном(и ровно одном) факте

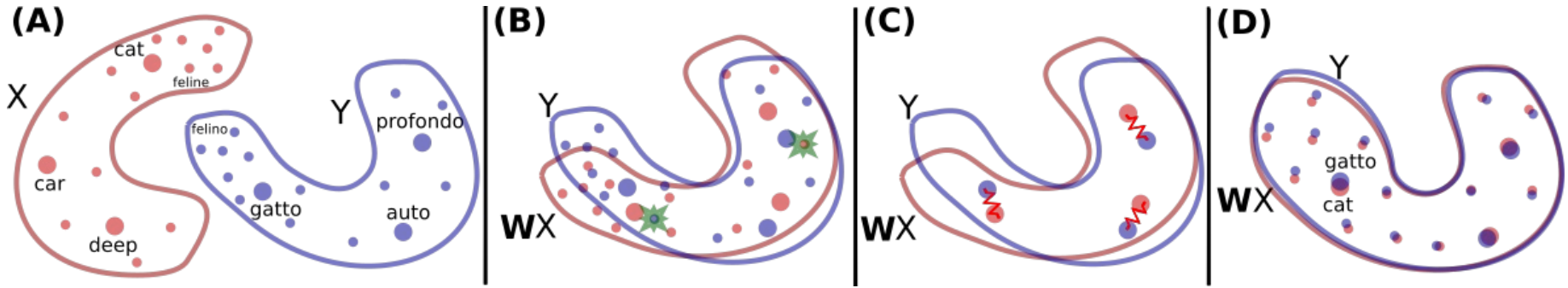
Мультиязычные эмбеддинги

*2vec

Геометрические свойства линейных эмбеддингов позволяют **без большого количества данных получить общее пространство слов** для разных языков



Мультиязычные эмбеддинги



Актуальные алгоритмы

Представление

- Tf-idf, nPMI, hashing trick

Факторизация (декомпозиция)

- PCA, LSI, LSA, pLSA

Тематическое моделирование

- pLSA, LDA, HDP, ARTM

Поиски

- BM25, HNSW, LSH

Эмбеддинги

- word2vec, glove, paragraph2vec, fasttext, starspace, poincaré

Нейросетевые подходы

- LSTM, GRU, Attention, siamese network, similarity learning

Полезный NLP-софт

Предобработка текста
(нормализация, токенизация)

- pymorphy2(ru), snowball stemmer(en), Stanford NLP(en)

Фреймворки

- sklearn, NLTK, gensim, spaCy

Узкоспециализированные фреймворки

- BigARTM, Vowpal Wabbit, Fasttext, Starspace, Muse, faiss, annoy, NMSLib, lucene, sphinx, elastic

Нейросетевые фреймворки

- Pytorch, Keras

Контакты

Штех Геннадий *
@ NAUMEN

gshtekh@naumen.ru

Gennady Shtekh

shtechgen@gmail.com
t.me/sht3ch
github.com/ShT3cH

*R&D Data Usage Department Executive