Обзор статей для комплинг исследования научно-популярных текстов

Анастасия Кузнецова, Анна Лапидус,
Юлия Коломенская, Ксения Самойленко

НИУ ВШЭ

24 января 2017

План

1. Topic Modeling
2. Term extraction
3. Named Entity Recognition
4. Readability

План доклада

1. Topic Modeling
2. Term extraction
3. Named Entity Recognition
4. Readability

Topic Modeling

Задача: определить тематику документов в коллекции

- Тема - вероятностное распределение слов
- Документ - вероятностное распределение тем

Latent Dirichlet allocation. Blei D. M., Ng A. Y., Jordan M. I. Journal of Machine Learning Research. 2003

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA

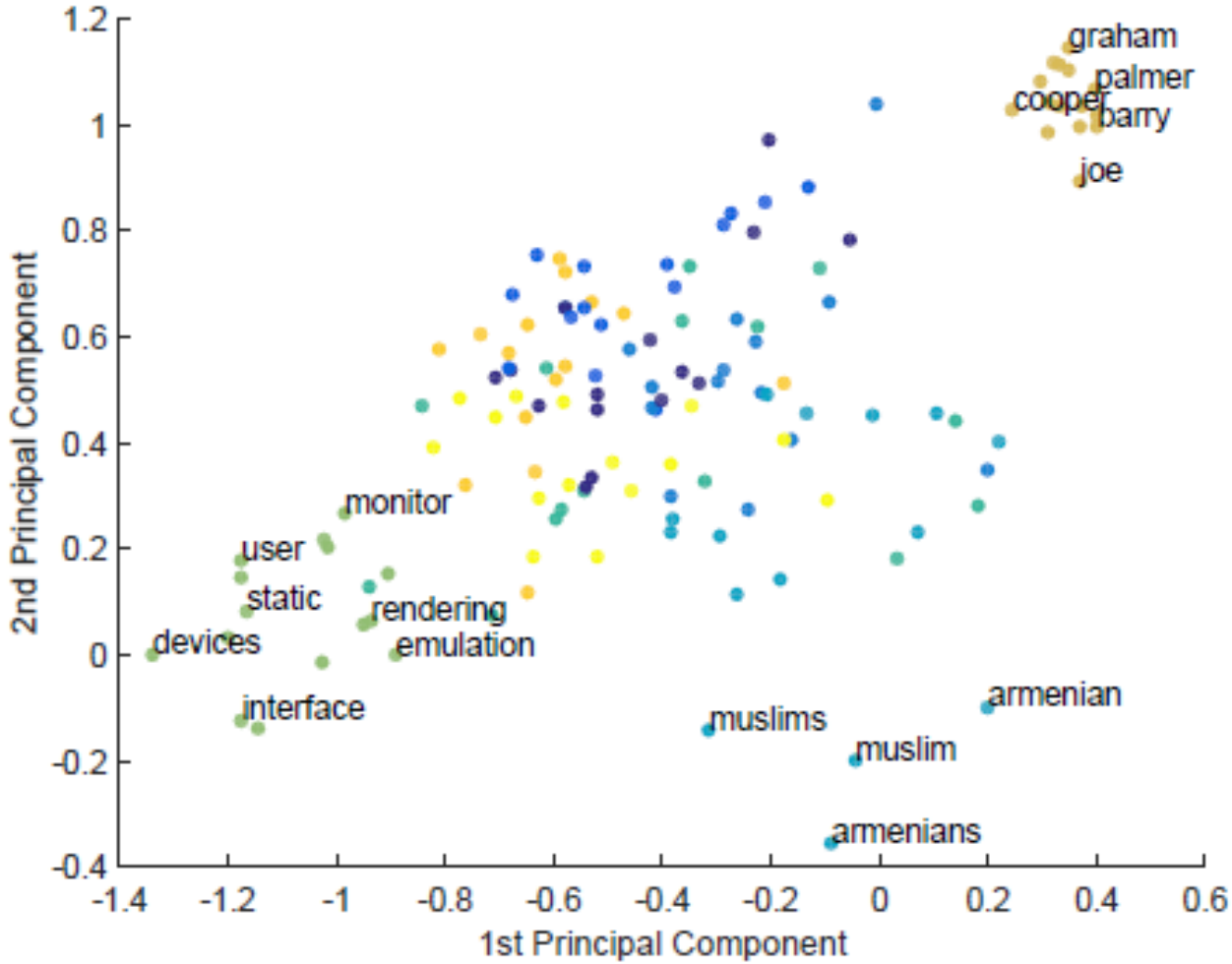
- Выделяет скрытые темы
- Определяет частотные тематические слова
- Bag of Words
- Темы не когерентны

Word Embeddings

Gaussian LDA for Topic Models with Word Embeddings
Rajarshi Das, Manzil Zaheer, Chris Dyer

- Вместо дискретного распределения на множестве слов - распределение на множестве word embeddings
- Семантическая связь между тематическими словами
- Позволяет выделять новые тематические слова

Gaussian LDA topics								
hostile	play	government	people	university	hardware	scott	market	gun
murder	round	state	god	program	interface	stevens	buying	rocket
violence	win	group	jews	public	mode	graham	sector	military
victim	players	initiative	israel	law	devices	walker	purchases	force
testifying	games	board	christians	institute	rendering	tom	payments	machine
provoking	goal	legal	christian	high	renderer	russell	purchase	attack
legal	challenge	bill	great	research	user	baker	company	operation
citizens	final	general	jesus	college	computers	barry	owners	enemy
conflict	playing	policy	muslims	center	monitor	adams	paying	fire
victims	hitting	favor	religion	study	static	jones	corporate	flying
rape	match	office	armenian	reading	encryption	joe	limited	defense
laws	ball	political	armenians	technology	emulation	palmer	loans	warning
violent	advance	commission	church	programs	reverse	cooper	credit	soldiers
trial	participants	private	muslim	level	device	robinson	financing	guns
intervention	scores	federal	bible	press	target	smith	fees	operations
0.8302	0.9302	0.4943	2.0306	0.5216	2.3615	2.7660	1.4999	1.1847
Multinomial LDA topics								
turkish	year	people	god	university	window	space	ken	gun
armenian	writes	president	jesus	information	image	nasa	stuff	people
people	game	mr	people	national	color	gov	serve	law
armenians	good	don	bible	research	file	earth	line	guns
armenia	team	money	christian	center	windows	launch	attempt	don
turks	article	government	church	april	program	writes	den	state
turkey	baseball	stephanopoulos	christ	san	display	orbit	due	crime
don	don	time	christians	number	jpeg	moon	peaceful	weapons
greek	games	make	life	year	problem	satellite	article	firearms
soviet	season	clinton	time	conference	screen	article	served	police
time	runs	work	don	washington	bit	shuttle	warrant	control
genocide	players	tax	faith	california	files	lunar	lotsa	writes
government	hit	years	good	page	graphics	henry	occurred	rights
told	time	ll	man	state	gif	data	writes	article
killed	apr	ve	law	states	writes	flight	process	laws
0.3394	0.2036	0.1578	0.7561	0.0039	1.3767	1.5747	-0.0721	0.2443

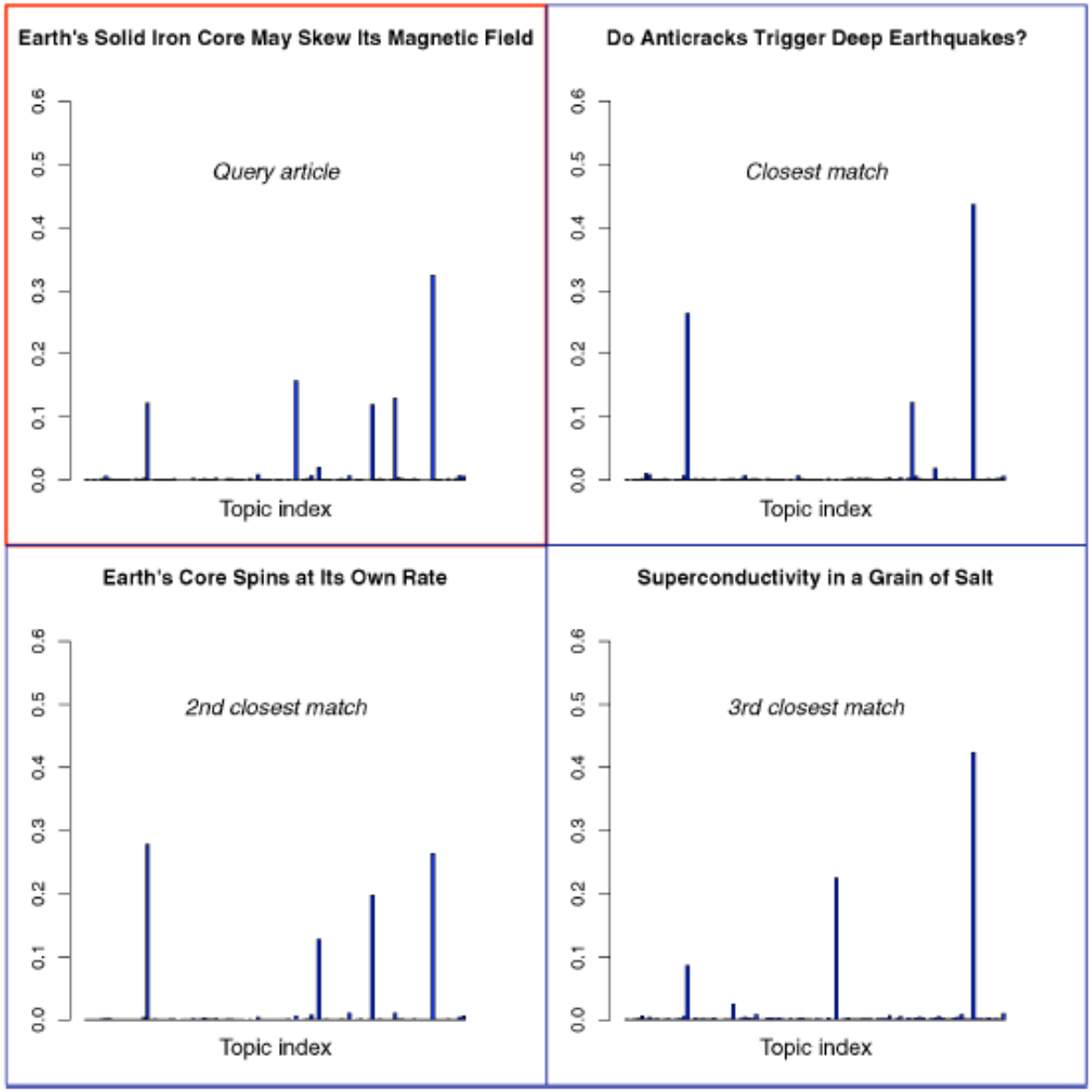

graham
palmer
cooper
barry
joe
monitor
user
static
rendering
devices
emulation
interface
armenian
muslims
muslim
armenians
2nd Principal Component
1st Principal Component
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-1.4
-1.2
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6

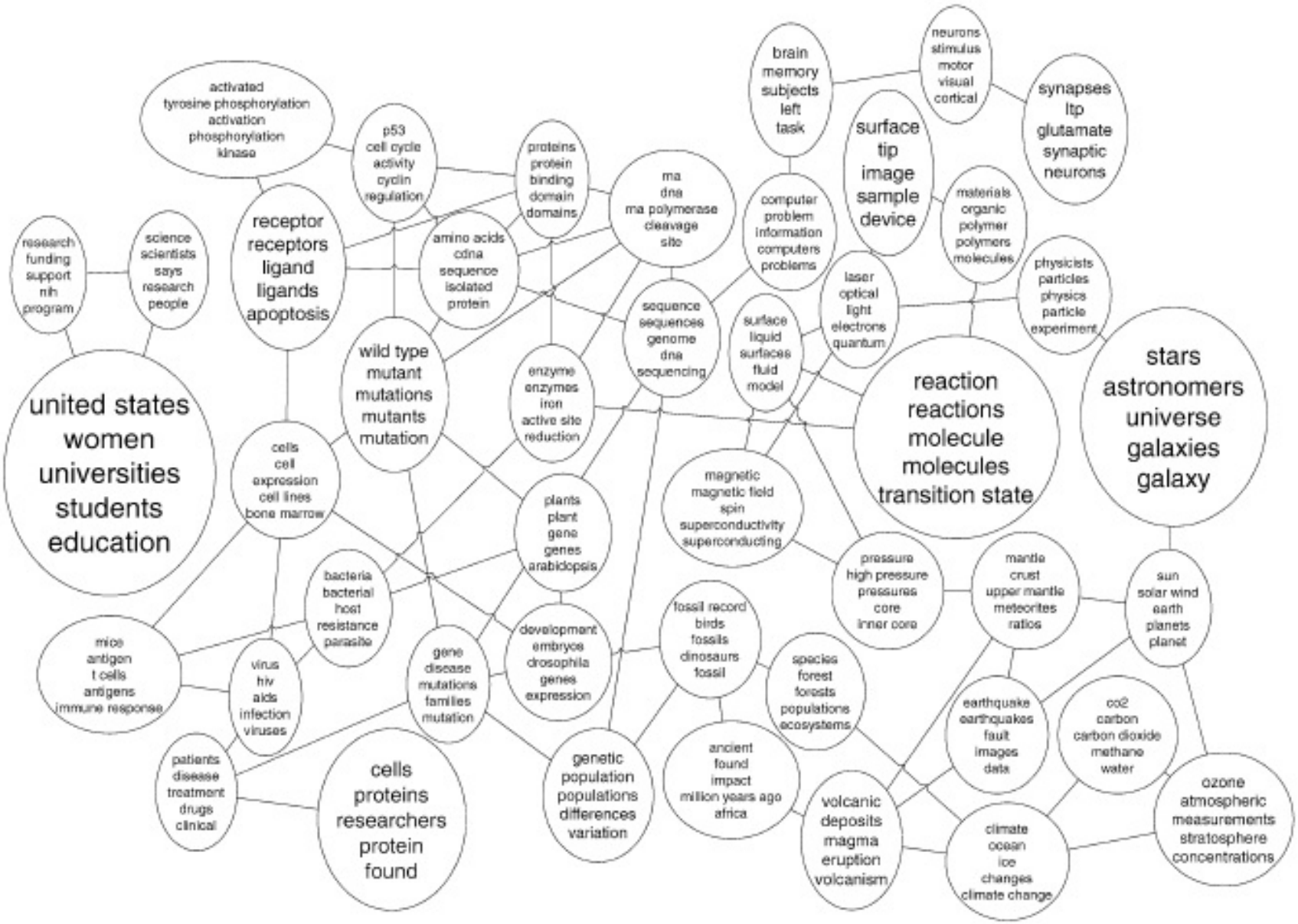
Другие модели

- Мультимодальные тематическое моделирование
  Tag-weighted topic model for mining semi-structured documents.
  Li S., Li J., Pan R.
- Модель коррелированных тем
  A correlated topic model of Science. Blei D., Lafferty J.

A correlated topic model of Science. Blei D., Lafferty J.

- Логнормальное многомерное распределение позволяет учитывать взаимосвязь между темами
- Построен граф тематик статей журнала Science
- На основе верятностых мер (расстояние Хеллингера) определяются наиболее близкие статьи


Earth's Solid Iron Core May Skew Its Magnetic Field
Query article
Topic index
Do Anticracks Trigger Deep Earthquakes?
Closest match
Topic index
Earth's Core Spins at Its Own Rate
2nd closest match
Topic index
Superconductivity in a Grain of Salt
3rd closest match
Topic index
0.0
0.1
0.2
0.3
0.4
0.5
0.6


activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
ma dna ma polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
computer problem information computers problems
materials organic polymer polymers molecules
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
sequence sequences genome dna sequencing
laser optical light electrons quantum
physicists particles physics particle experiment
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
united states women universities students education
cells cell expression cell lines bone marrow
plants plant gene genes arabidopsis
magnetic magnetic field spin superconductivity superconducting
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
development embryo drosophila genes expression
fossil record birds fossils dinosaurs fossil
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
species forest forests populations ecosystems
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

План доклада

1. Topic Modeling
2. Term extraction
3. Named Entity Recognition
4. Readability

Term Extraction

Term extraction (recognition, identification, acquisition) - преследует различные цели:

- создание онтологии;
- создание глоссария;
- составление указателей (indexing)
- научные исследования и пр.

Принцип работы большинства экстракторов:

- Предобработка текстов
- Отбор кандидатов
- Ранжирование или сортировка кандидатов
- Валидация (сравнение выбранных показателей с определенным пороговым значением)

Лингвистические характеристики термина:

- Синтаксическая структура (в основном существительные или именные группы) (для n-грамм) *(нужен POStagger, отбирается слишком много кандидатов -> stop lists)* (1) ((Adj|Noun)+|((Adj|Noun)*(NounP rep(Adj|Noun*))Noun (2) Noun1(Adj|(P rep(Det)?)?Noun2|V Inf)
- Морфологическая структура (сложносоставные термины, латинские аффиксы)
- Типичная структура контекстного окружения (определения (vector spaces), пояснения (regexp))

Статистические характеристики термина:

- Частотность
- Co-occurrence measure (the Dice coefficient, Pointwise Mutual Information (PMI), Log-Likelihood ratio)

nested terms: floating point arithmetic -> floating point, BUT point arithmetic

Дистрибутивные характеристики термина:

- Между словами в n-граммах (unithood)
- Внутри документа или в коллекции документов (termhood, -> tf-idf)
- Weirdness (specific corpus domain vs general corpus)

Оценка работы экстрактора:

- сложно разметить золотой стандарт
- используются уже существующие словари и онтологии
- проверка результатов вручную

"It seems to be a general truth that results vary a lot with the corpora and evaluation methods used. For a different Wikipedia corpus, Hjelm (2009) found precision values as low as 12-13 pct. while in Zhang et al. (2008) they are around and above 90 pct".

План доклада

1. Topic Modeling
2. Term extraction
3. Named Entity Recognition
4. Readability

NER в крупных и малоресурсных языках

Малоресурсные языки – это языки, для которых не существует размеченных корпусов NE, либо такие корпуса очень малы, что делает невозможным обучение на них.

Типы исследований:

- Разметка параллельных корпусов;
- Обучение модели на текстах без разметки.
- Расширение существующих корпусов;

Методы разметки:

- Conditional Random Fields (CRF)
- Long Short Term Memory (LSTM)
- Word embeddings

State of the art

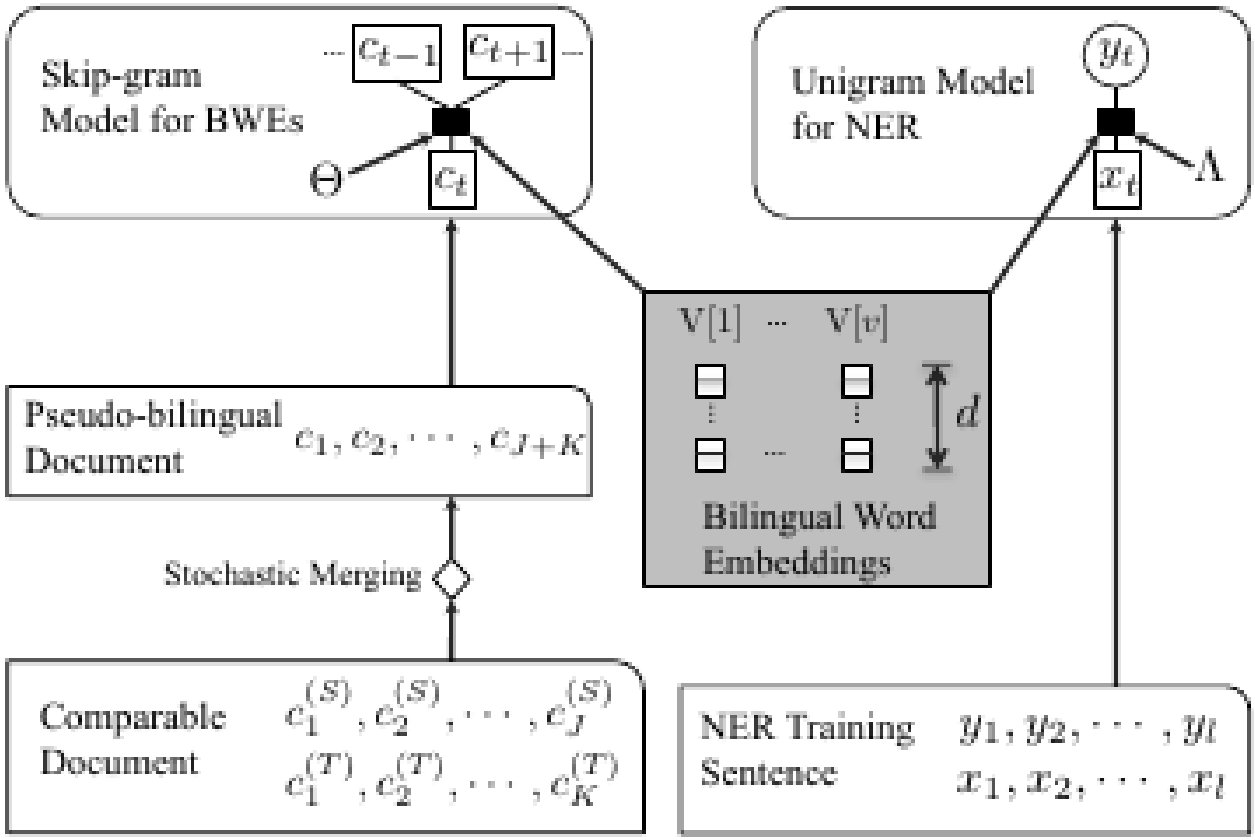
Stanford NER tagger (Lample et al.): LSTM-CRF модель, которая работает свекторно-символьным представлением слов, полученных при обучении на размеченных корпусах текстов.
Работает только накрупных языках.

Разметка при помощи параллельных корпусов

Dingquan Wang, Nanyun Peng, Kevin Duh. A Multi-task Learning Approach to Adapting Bilingual Word Embeddings for Cross-lingual Named Entity Recognition, 2017.

- Проекция NE из размеченного языка на менее разработанный с использованием параллельного корпуса текстов на основе билингвальных эмбедингов
- Проекция NE в сравнительном корпусе с английского на китайский язык

Multitask Model



Query	Type	Top 8 results in English
NBA	ORG	**NBA**, rebounds, **Knicks**, **Lakers**, Lewiston-Porter, **76ers**, guard-forward, **Celtics**
NBA		**NBA**, **Lakers**, Gervin, rebounds, **Celtics**, **Cavaliers**, **Knicks**, **All-Defensive**
西班牙	LOC	**Spain**, Spanish, **Nogueruelas**, Rosanes, **Mazarete**, **Ólvega**, Marquesado, **Montija**
Spain		**Spain**, Rosanes, **Cenicientos**, **Madrid**, Sorita, **Alcahozo**, **Nogueruelas**, **Villaralto**
希腊	LOC	**Greece**, Greek, **Achaia**, annalistic, heroized, Gigantomachy, Hecabe, river-god
Greece		Hachadoor, **Greece**, Demoorjian, **Safranbolu**, **Scicli**, **Holasovice**, **Sighisoara**, **Litomysl**
卡卡	PER	**Kakashi**, **Moure**, Uzumaki, cosplayed, **Uchiha**, humanizing, **Yens**, hilarious
Kaka		**Kakashi**, **Kaka**, **Moure**, **Nedved**, **Suazo**, **Batistuta**, Uzumaki, **Quagliarella**

Word embeddings. Обучение без разметки

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, Steven Skiena. POLYGLOT -NER: Massive Multilingual Named Entity Recognition, 2015.

- Разметка NE в 40 языках с использованием Word embeddings (совместная встречаемость слов в тексте), которые используются в качестве признаков при обучении.
- Отсутствующие теги частей речи размечались методом точного соответствия.
- Оценка потерь. Оптимизация при помощи стохастического градиентного спуска.

PERSON

Language	Sentence	Translation
English	**Simien** was traded from the Heat along with **Antoine Walker** and **Michael Doleac** to the **Minnesota Timberwolves** on October 24, 2007, for **Ricky Davis** and **Mark Blount**.	-
Hungarian	**Dimitri** beszélt egy utat **Rómába**.	**Dimitri** talked about a trip to **Rome**.
Spanish	**Pešek** nació en **Praga** y estudió dirección de orquesta piano en la **Academia de Artes** allí, con **Václav Smetacek**	**Pešek** born in **Prague** and studied orchestra direction piano at the **Academy of Arts** there, **Vaclav Smetacek**.
Russian	Уроженец **Рио-де-Жанейро**, **Хосе Родригес Тринидади** использовал сокращенную форму как своим сценическим псевдонимом.	A native of **Rio de Janeiro**, **Jose Rodriguez Trindade** used as a shortened form of his stage name.
Korean	**도널드 스미스** 는 1976 년에 **자이드 압둘 아지즈** 에 그의 이름 을 바꿨다 .	**Donald Smith** in 1976 changed his name to **Zaid Abdul Aziz**.

Ошибки. PER/ORG, LOC/ORG

Russian	Русский Федерация под руководством Владимира Путина в приложении Крым.	Russian Federation under the leadership of Vladimir Putin annexed Crimea.
Korean	검찰은 유 병출 ..유엔, 서울 에서 운영 ..제주.. 해양 (주)의 소유자의 홈을 급습했다.	Prosecutors raided the home of Yoo Byung- un, the owner of Jeju Marine Co. Ltd, which operates in Seoul.
French	En 1970 , Burgess a été le candidat républicain succès pour le lieutenant - gouverneur et a servi deux mandats , de 1971 à 1975 .	In 1970, Burgess was the successful Republican candidate for Lieutenant - Governor and served two terms from 1971 to 1975.
Turkish	1979 yılında , o folha sol ve kısa ömürlü Jornal da República ' da Mino Carta ile çalışmaya başladı .	In 1979, he left folha and short-lived Jornal da República began working with Mino Carta.
Arabic	وكان البابا يتحدث عن القدس وإلى جانبه الرئيس الفلسطيني محمود عباس بعد وصوله مباشرة إلى مدينة بيت لحم .	The Pope speaks of Jerusalem and to his part, Palestinian President Mahmoud Abbas after his arrival directly to the city of Bethlehem.
Indonesian	Ia lahir di Totowa, New Jersey dan meninggal di Brooklyn, New York.	He was born in Totowa, New Jersey and died in Brooklyn, New York.
Chinese	周 先生 是 ..四川省 党委 书记 成为 中国 公安部 负责人 在2003 年 以前 。	Mr. Zhou was the party secretary in Sichuan province before becoming head of China's Public Security Ministry in 2003.
Greek	Ήταν ο μόνος Αμερικανός για να χρησιμεύσει ως Πρέσβης στη Γαλλία, της Δημοκρατίας της Γερμανίας και το Ηνωμένο Βασίλειο.	He was the only American to serve as Ambassador to France, the Republic of Germany and the United Kingdom.

Оценка качества

- Тексты CONLL для крупных языков.
- Distant Evaluation при помощи машинного перевода.

Результаты

DEV	English	Spanish	Dutch
POLYGLOT-NER$_{S_7}$	62.9	56.7	53.2
POLYGLOT-NER$_{S_6+S_7}$	**73.3**	59.3	59.7
Nothman et al. [22]	67.9	**60.7**	**62.2**
TEST			
POLYGLOT-NER$_{S_7}$	58.5	58.5	51.5
POLYGLOT-NER$_{S_6+S_7}$	**71.3**	**63.0**	59.6
Nothman et al. [22]	61.3	61.0	**64.0**

Table 6: Cross-domain performance measured by Exact F_1 on TEST and DEV sections of CONLL corpora.

План доклада

1. Topic Modeling
2. Term extraction
3. Named Entity Recognition
4. Readability

Readability

Readability — сумму всех элементов текстового материала, котоыре влияют на понимание текста, скорость прочтения и уровень интереса к материалу

Основные методики

- Подсчет соотношения длины предложений и количества слов (FRE, FKR);
- Подсчет соотношения длины предложений и количества сложных слов (NDC)
- Применение машинного обучения

Horacio Saggion, 2017. **Виды признаков для оценки ридабилити**

- Лексико-семантические (словарные) признаки: относительная частота слов, оценка вида и кол-ва токенов, вероятностные лингвистические модели
- Психолингвистические признаки: возраст понимания, конкретность, полисемия
- Синтаксические признаки — длина предложения, уровни в деревьях
- Дискурсивные признаки — кореерентные связи, именнованные сущности, плотность текста
- Семантические и прагматические признаки: использование идиом, культурных отсылок и образов, тип текста (мнение, сатира и т.д.)

A Multi-task Approach to Predict Likability of Books

Suraj Maharjan, Manuel Montes-y-Gomez, Thamar Solorio, John Arevalo and Fabio A. Gonzalez **Как оценить ридабилити художественного текста?**

- Создали нейронную сеть и научили ее предсказывать потенциальную успешность книги
- Использовали классические метрики и машинное обучение
- Обучили 25 моделей
- Наилучший результат — 71
- И его показала не нейросеть...

Какие еще бывают исследования?

- Изменение сложности научных текстов за последние 200 лет
- Измерение ридабилити узкоспецаиилизрованных медицинских текстов
- Измерение ридабилити текстов о финансах и юриспруденции