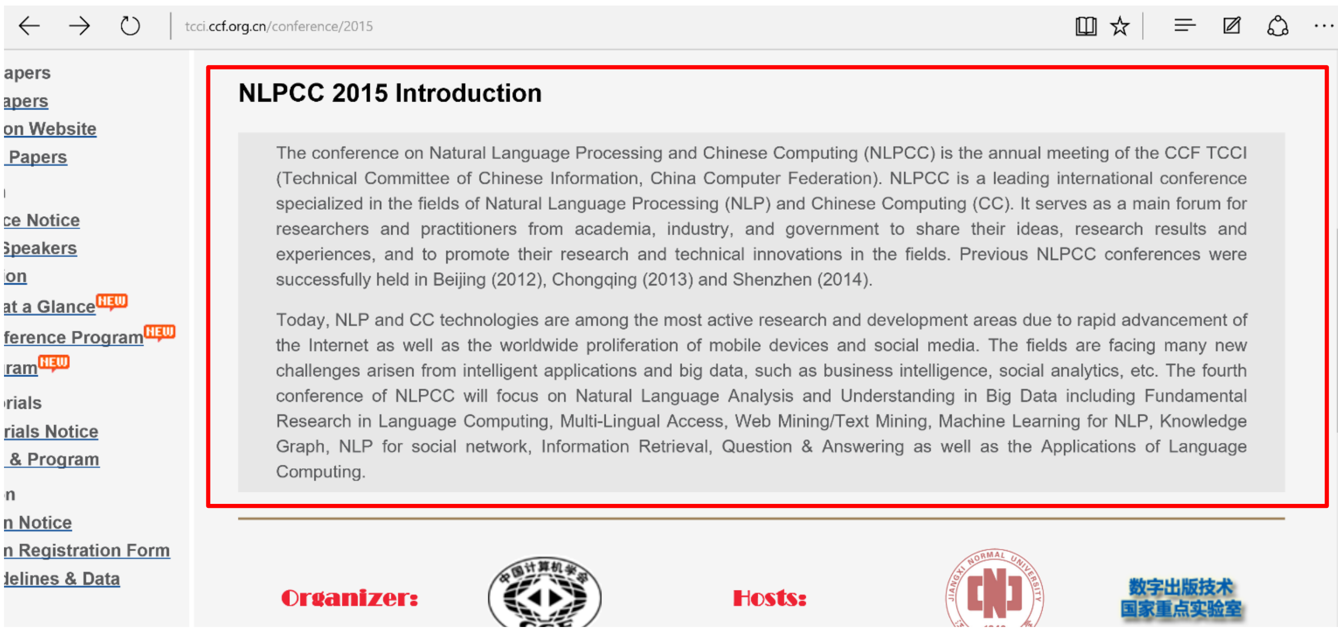
The role of NLP in IR
- How can NLP help IR?

Jian-Yun Nie

University of Montreal
nie@iro.umontreal.ca
http://www.iro.umontreal.ca/~nie

IR

- IR works mostly with texts



tcci.ccf.org.cn/conference/2015

apers
apers
on Website
Papers
ce Notice
Speakers
ion
at a Glance NEW
ference Program NEW
ram NEW
rials
rials Notice
& Program
n
n Notice
n Registration Form
lelines & Data

NLPCC 2015 Introduction

The conference on Natural Language Processing and Chinese Computing (NLPCC) is the annual meeting of the CCF TCCI (Technical Committee of Chinese Information, China Computer Federation). NLPCC is a leading international conference specialized in the fields of Natural Language Processing (NLP) and Chinese Computing (CC). It serves as a main forum for researchers and practitioners from academia, industry, and government to share their ideas, research results and experiences, and to promote their research and technical innovations in the fields. Previous NLPCC conferences were successfully held in Beijing (2012), Chongqing (2013) and Shenzhen (2014).

Today, NLP and CC technologies are among the most active research and development areas due to rapid advancement of the Internet as well as the worldwide proliferation of mobile devices and social media. The fields are facing many new challenges arisen from intelligent applications and big data, such as business intelligence, social analytics, etc. The fourth conference of NLPCC will focus on Natural Language Analysis and Understanding in Big Data including Fundamental Research in Language Computing, Multi-Lingual Access, Web Mining/Text Mining, Machine Learning for NLP, Knowledge Graph, NLP for social network, Information Retrieval, Question & Answering as well as the Applications of Language Computing.

Organizer:

Hosts:

数字出版技术
国家重点实验室

- Answer query “NLP conference”, “NLPCC 2015” …
- (Let us ignore criteria such as hyperlinks and focus on texts)

NLP

- NLP: many different aspects
  - Morphology
  - Lexical analysis
  - Phrase, collocation
  - Parsing
  - Semantic analysis
  - Discourse analysis
  - Machine translation, Question-answering, …
  - …
- They help us to understand language and texts, … thus should be useful for IR

Reality

- Only the simplest NLP techniques are used in IR
  - Stemming / lemmatization
  - E.g. *computation* → *comput*
- More sophisticated NLP techniques have not been up to their promise
  - Parsing had limited impact
  - Noun phrases were less effective than "statistical" word pairs
  - Word sense disambiguation failed to improve IR

Why?

How?

Goal of this talk

- What NLP techniques have been tried in IR?
- Why are they successful / unsuccessful?
- How can we make NLP more useful for IR?

Some exceptions

- QA: using both NLP and IR
- Cross-language IR: MT is effective for query translation

- But let us focus on the core IR task

Successes of NLP in IR

- Stemming
  - Normalization of words
  - Cut suffixes (*comput<u>ation</u>* → *comput*)
  - slight morphological transformation, E.g. Vowelization in Arabic
- Lexical processing
  - Decompounding in German
    - *windenergie* (wind-energy) → wind energie
  - Word segmentation in Chinese and Japanese
    - *杨柳青年画 = 杨柳青 年画* or *杨柳 青年 画*

- Word stemming usually leads to higher recall (better MAP)
- Decompounding and word segmentation are important

Higher level NLP?

- Noun phrases
  - Computer science (NN NN)
  - Board of directors (NN prep NN)
  - Black Monday (ADJ NN)
- Intuition: phrases are less ambiguous than single words
  - … black board … directors ≠ board of directors
  - … black board … Monday ≠ black Monday
- Solution: use phrases as additional index

$$\text{Score}(D,Q) = \lambda\, \text{Score}_{\text{Word}}(D,Q) + (1-\lambda)\, \text{Score}_{\text{Phrase}}(D,Q)$$

Early attempts [Fagan 1987] Experiments in Automatic Phrase Indexing For Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods, Cornell University.

- Syntactic phrases: follow some syntactic structures
  - information system (NN NN)
  - library of congress (NN prep NN)
- Non syntactic phrases: words frequently appear together at proximity
  - library of Smith college → college-library, Smith-college
- Non syntactic phrases > Syntactic phrases

Learning to segment a query: some attempts

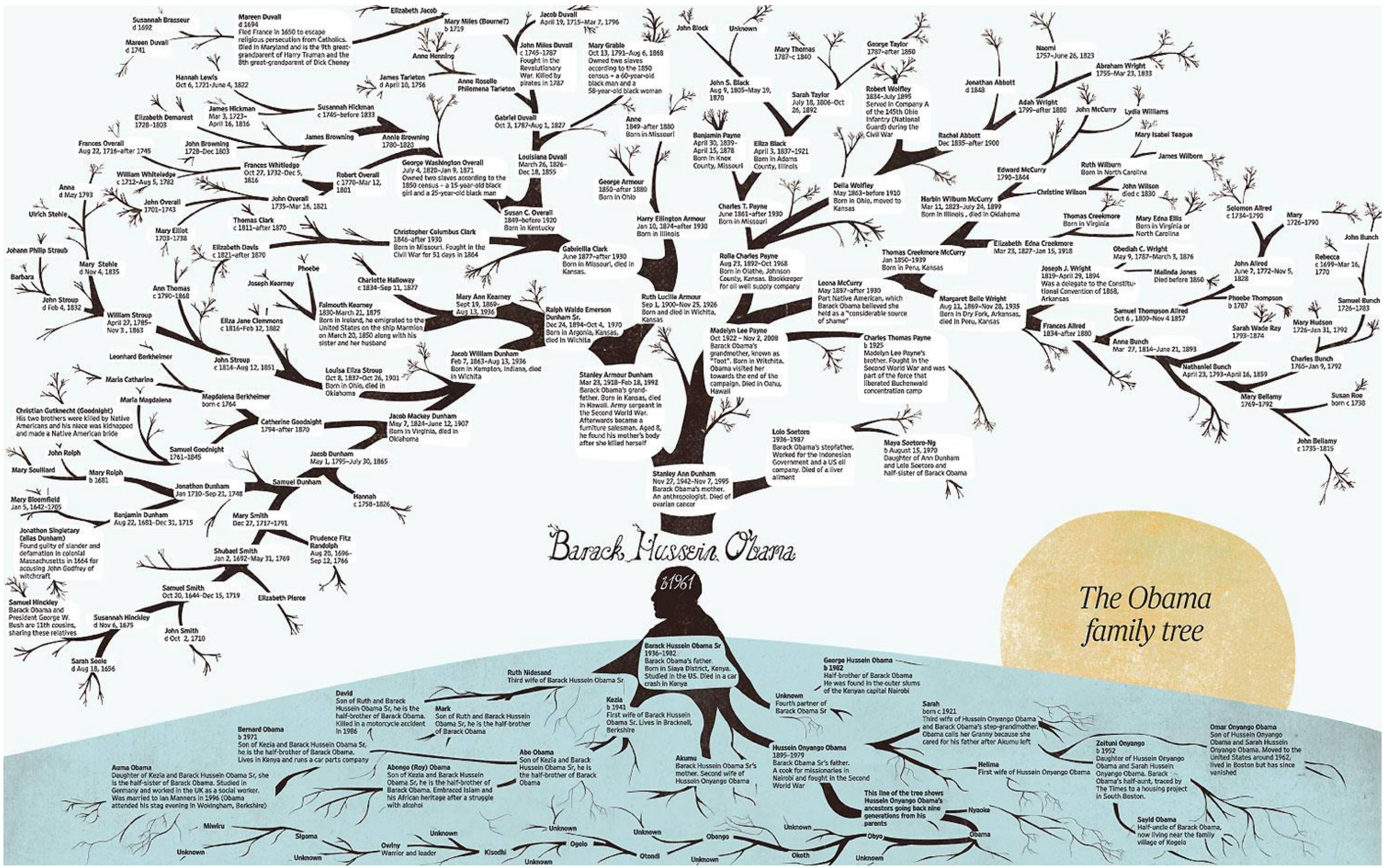
- Learn from human segmentation
  - A set of queries segmented manually
    - [book sale] in Chapters in [San Francisco]
    - Obama [family tree]
  - Train a model to segment queries
  - Accuracy in query segmentation close to 90%

    e.g. [Bergsma and Wang, 2007]


bing
obama family tree

[Obama family] and tree

Obama's [family tree]

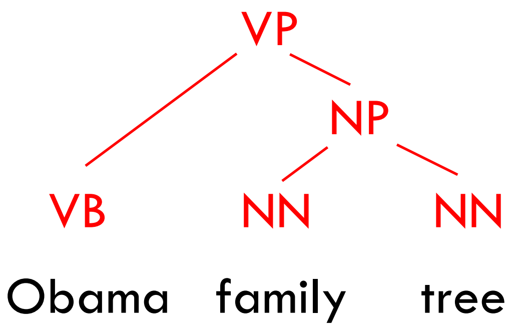


Disappointing results

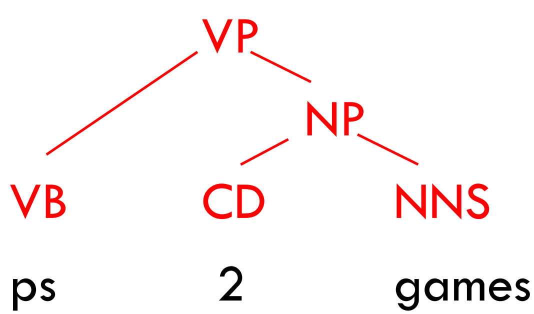
- marginal (if any) over word-based method
- Why?
  - Human segmentation ≠ utility for IR
  - [book sale] in Chapters
  - Is "book sale" better than "book" and "sale" in IR
- Key question: Should this expression in a query be founded in a relevant document?
- Should a relevant document contain "book sale"?
- Not all phrases correspond to fixed expressions
  - "Black Monday"
  - but not "NLP conference", "ps 2 games", "book sale"
- Many user queries are not grammatical

Parsing may be wrong

- Parsers make mistakes



➔ Obama [family tree]



➔ ps [2 games]

Non-linguistic phrases in IR

- Markov Random Field model [Metzler and Croft 2005]
- Any consecutive words in a query as a phrase
  - library of Smith college ➔ library-smith, smith-college
- Combining retrieval scores of
  - Single words (bag-of-words)
  - Exact phrase
  - Phrase words at proximity

$$Score(D,Q) = \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in U} \lambda_U f_U(c)$$

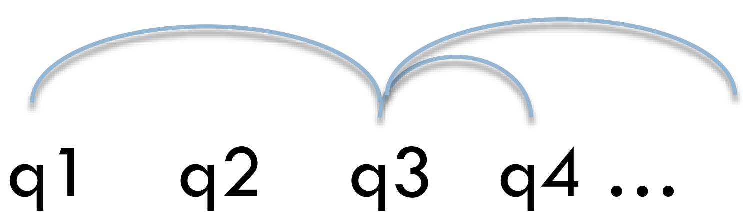
- Generally outperform bag-of-words models

IR-specific query "Parsing"

- Detect useful statistical dependences for the intended uses

⇒ Goal: Use dependent terms as a phrase to be matched exactly or at proximity

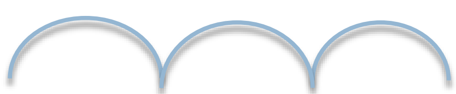
- What dependencies should be used?



Detect useful dependent pairs for IR

- Bendersky et al. 2010, WSDM

$$
\begin{aligned}
P(D|Q) \stackrel{rank}{=} & \sum_{i=1}^{k_u} w_i^t \sum_{q \in Q} g_i^u(q) f_T(q, D) + \\
& \sum_{i=1}^{k_b} w_i^b \sum_{q_j, q_{j+1} \in Q} g_i^b(q_j, q_{j+1}) f_O(q_j, q_{j+1}, D) + \\
& \sum_{i=1}^{k_b} w_i^b \sum_{q_j, q_{j+1} \in Q} g_i^b(q_j, q_{j+1}) f_U(q_j, q_{j+1}, D) \quad (3)
\end{aligned}
$$



q1 q2 q3 q4 …

  - Learning weights from judged queries based on features

Variable dependencies
(Shi and Nie, CIKM 2010, AIRS 2010)

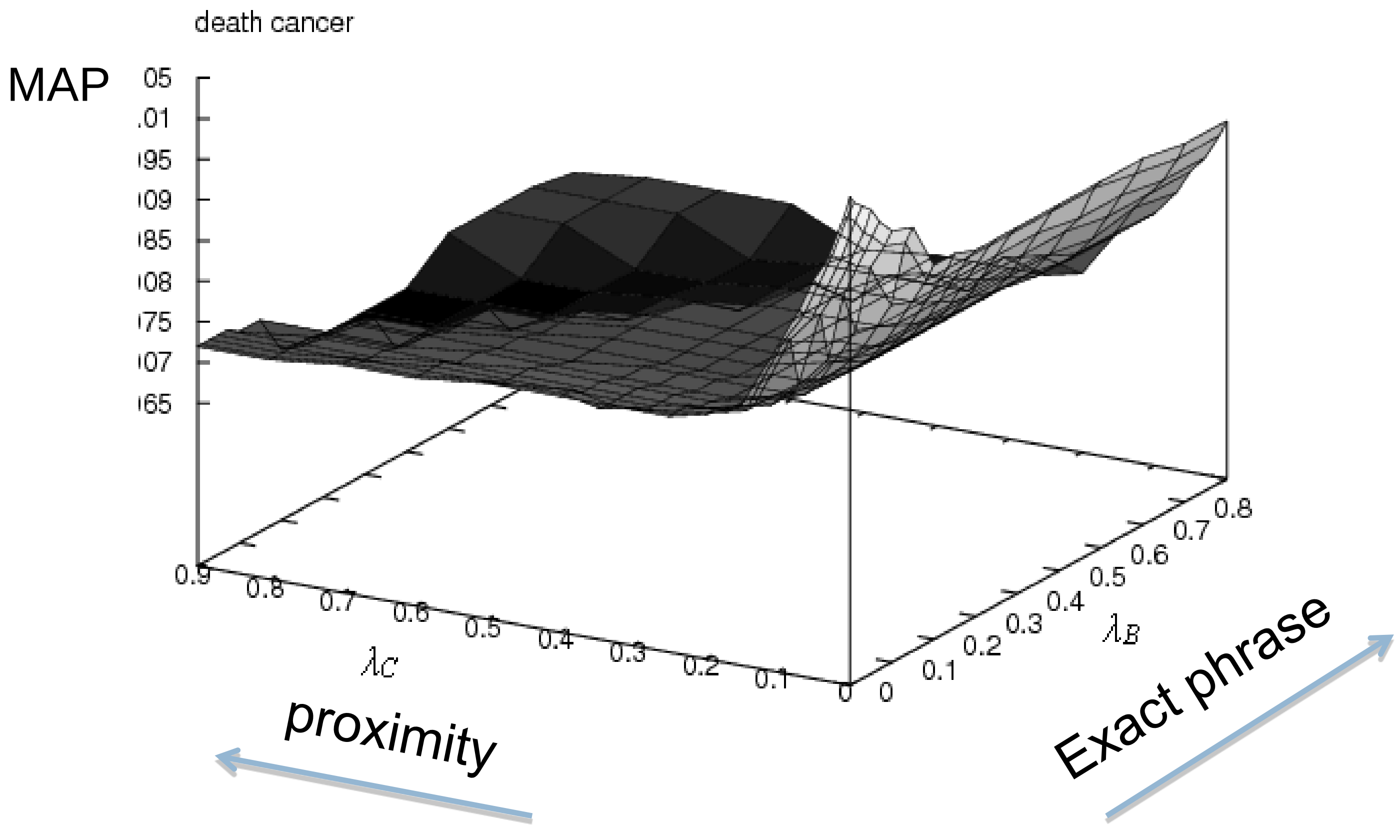
- Further extends the dependencies
  - Exact phrase (*black Monday*)
  - Proximity within window of size 2, 4, 8, 16 (book sale)

$$Score(D,Q) = \sum_{q_i \in Q} \lambda_U(q_i|Q) f_U(q_i,D) + \sum_{q_i q_{i+1} \in Q} \lambda_B(q_i,q_{i+1}|Q) f_B(q_i q_{i+1},D)$$
$$+ \sum_{w \in W} \sum_{q_i,q_j \in Q, i \neq j} \lambda_{C_w}(q_i,q_j|Q) f_{C_w}(q_i,q_j,D)$$

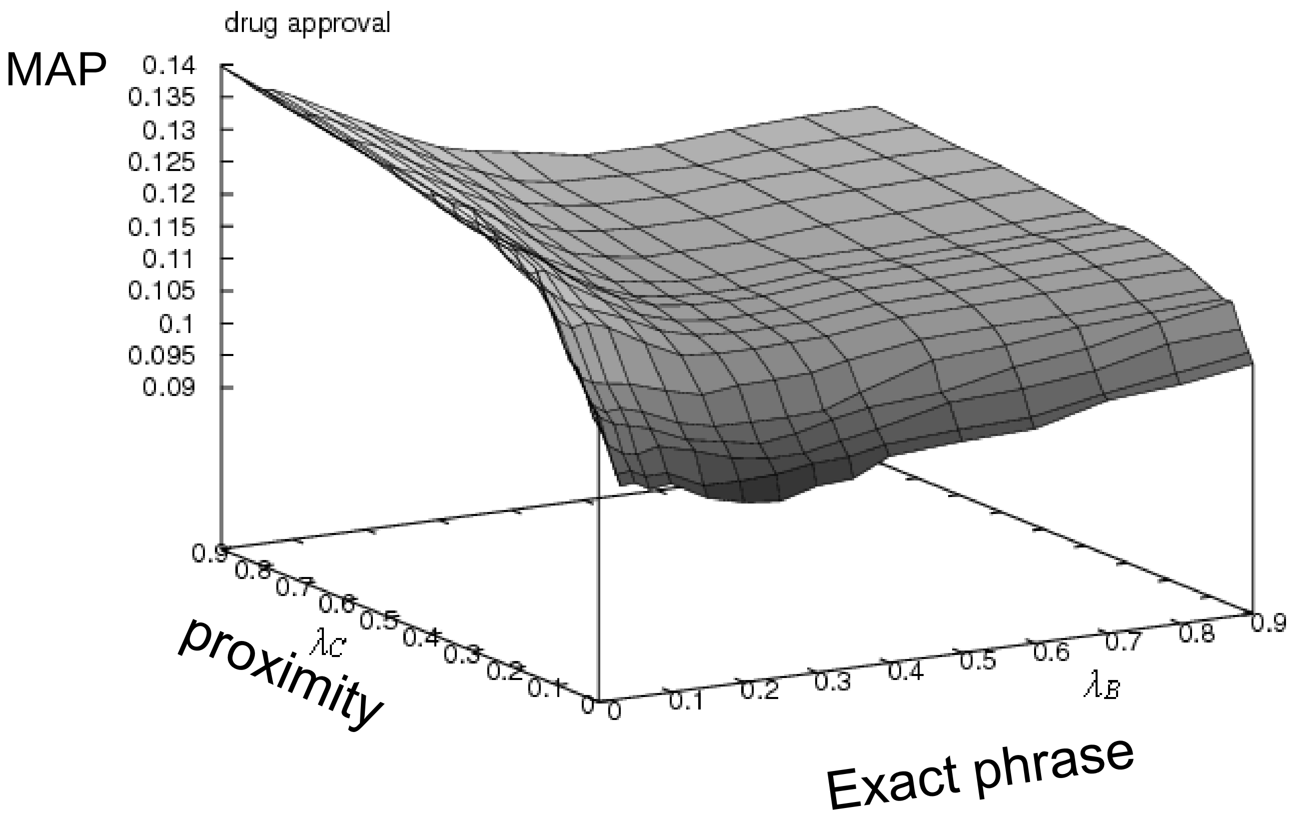
q1 q2 q3 q4 …

  - Weights of different pairs trained on judged queries

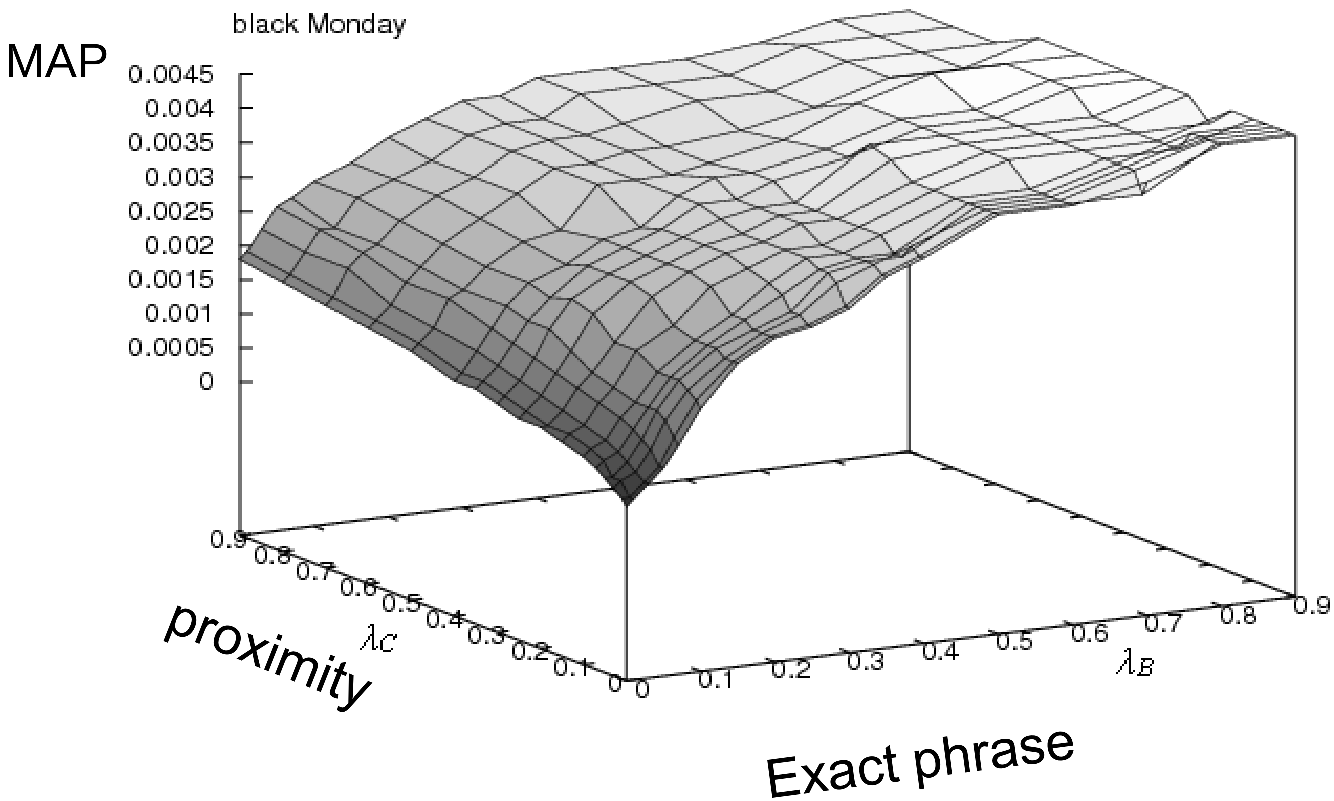
Death from cancer – No dependency



Drug approval – proximity



Black Monday – exact phrase



Summary on Query “Parsing” for IR

- Useful “phrases” for IR ≠ linguistic phrases
  - Linguistic phrases are not always useful for IR
  - Useful “phrases” may not be linguistically motivated
    - Proximity expresses flexible contextual relation
- Training a “parser” of queries on IR data
- “phrase” as useful signals of form for IR

NLP is not limited to forms

- Morphology
- Lexical analysis
- Phrase, collocation
- Parsing
- Semantic analysis

Form

- Also about Meaning
  - Word sense disambiguation
  - Semantic representations
  - …

Can NLP (about meaning) be useful for IR?

Word sense disambiguation

- WSD is still challenging
  - Accuracy 70-80% for limited vocabulary
- [Sanderson 94] (SIGIR):
  - Disambiguation at 75% works even worse than non-disambiguation
  - To be useful for IR, WSD should reach 90%
- Manual disambiguation using Wordnet does not help in IR [Voorhees 1993]

- Why IR without disambiguation may work?
  - Skewed distribution of word senses: 80% cases with most common sense
  - Context-effect of other query terms (*Java compiler*)

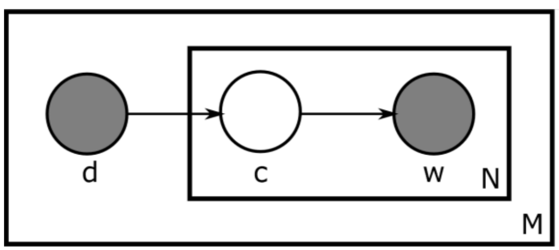
Dealing with Meaning

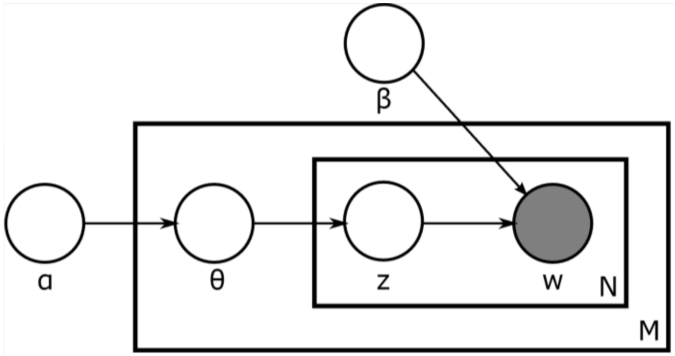
- Traditional way:
  - Define word senses
  - Classify a word occurrence
- [Lyons 1981]: *Discreteness in language is a property of form, not meaning.*
- => *Continuous representation of meaning*

Semantic Representation

- Latent semantic representation

- LSI [Deerwester et al. 1990], LDA [Wei & Croft 2006]





- Some limited improvements
- Topics seem to be too coarse for IR

LDA example

"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

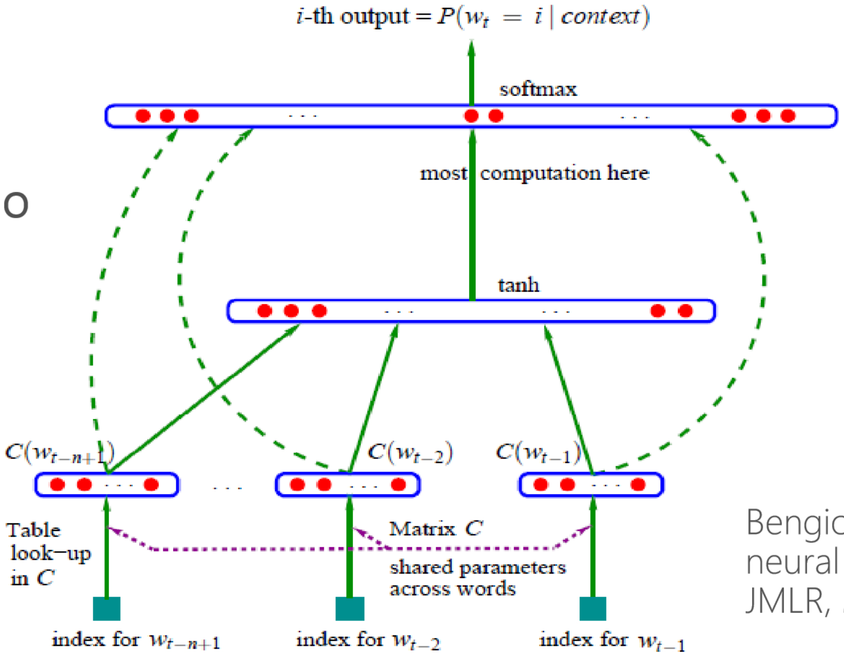
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

From [Blei et al. 2003]

Word embedding: Distributional representation

Yoshua Bengio

LM: predict the next word given the past:
e.g., $p(chases|the\ cat) = ?,\ p(says|the\ cat) = ?$

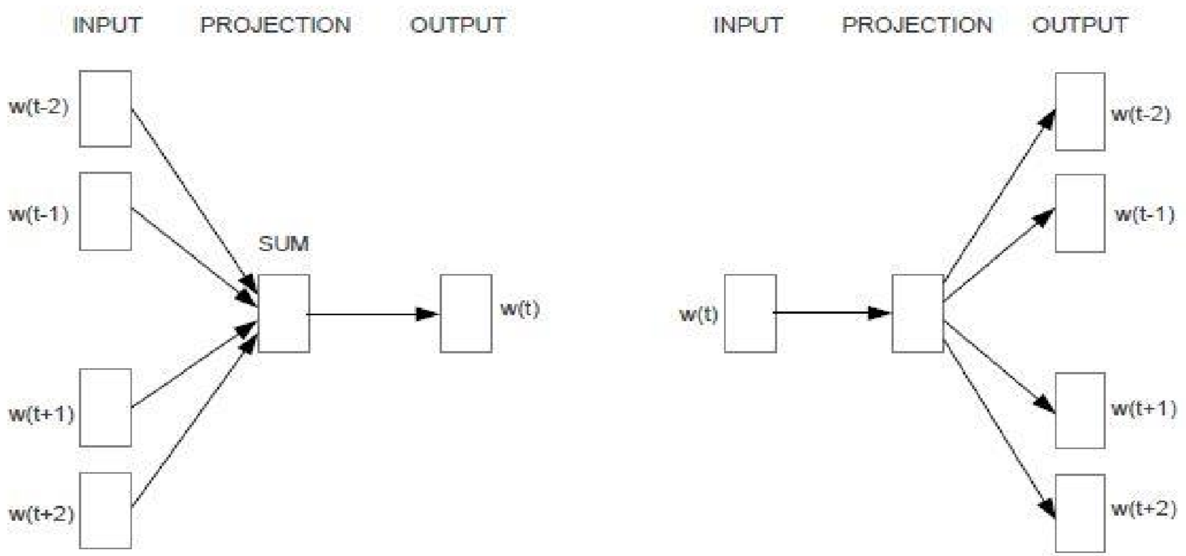


Obtain word representation that can best predict the next word

Bengio, Ducharme, Vincent, Jauvin, "A neural probabilistic language model." JMLR, 2003

Word2vec [Mikolov et al. 2013]

CBOW/Skip-gram Word Embeddings



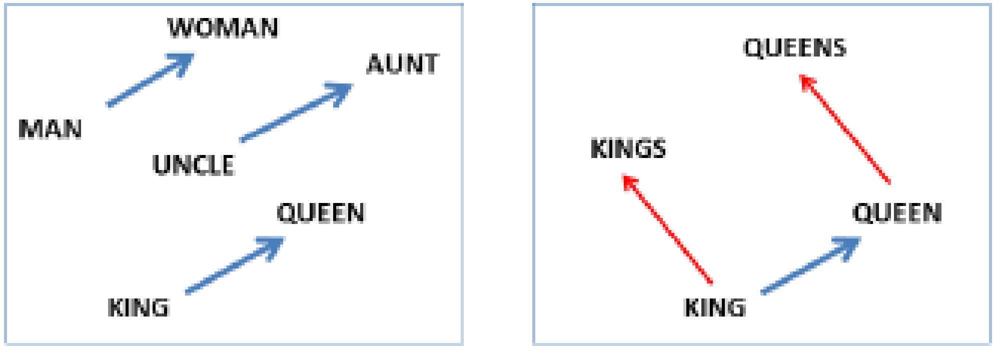
CBOW

Continuous Bag-of-Words

Skip-gram

Surprising capabilities

- Determine word similarity
  - cat ~ kitten
  - musician~singer~artist
- Analogy reasoning
  - King is to queen as man to <u>woman</u>.
  - V_{king} - v_{queen} ≈ v_{man} - v_{woman}

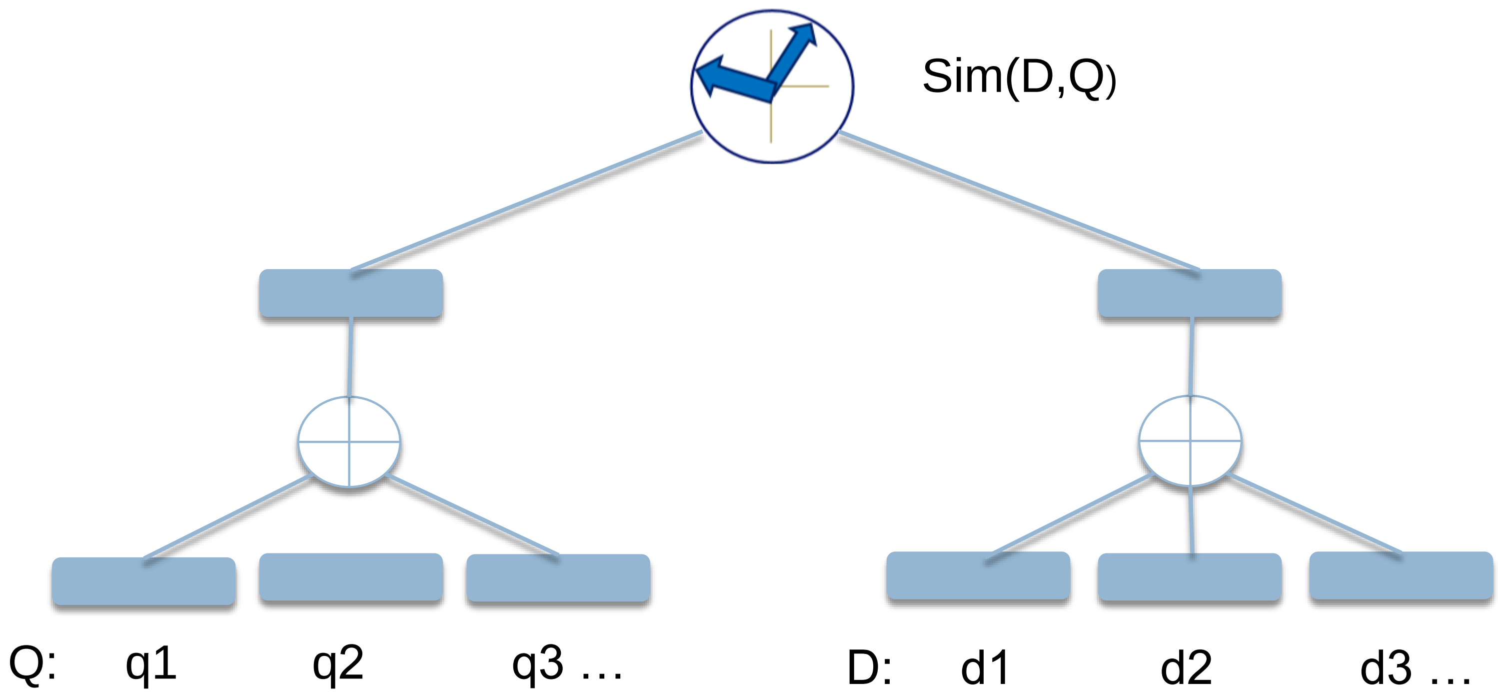


[Image credits: Mikolov et al (2013) "Efficient Estimation of Word Representation in Vector Space", *arXiv*]

- (*Firth*, J. R. 1959) You shall know a word by the company it keeps.

Generating query/doc. representation

- Generate word embedding for each word
- Sum up/average the embeddings as a global representation



Failure: Much worse than bag-of-words

pork tenderloin

Word embeddings

pork	
beef	0.735397
meat	0.717314
chicken	0.673912
sausage	0.599732
veal	0.588186
roast	0.567123
tenderloin	0.559435
sausages	0.557218
cooked	0.550576
lamb	0.539285

tenderloin	
filet	0.638106
sirloin	0.598938
loin	0.586353
roast	0.577229
steak	0.571505
pork	0.559435
venison	0.554806
grilled	0.552503
steaks	0.550434
chops	0.542049

Summing up embeddings

pork tenderloin	
pork	0.883016
tenderloin	0.883016
beef	0.678835
roast	0.647979
chicken	0.637563
veal	0.633223
meat	0.621922
sausage	0.615066
loin	0.609921
chops	0.607775

pork tenderloin

Words in rel. doc.

#pork tenderloin	
pork	790
tenderloin	495
recipes	422
recipe	214
food	176
sauce	130
minutes	112
meat	102
fat	101
add	87
low	85
roast	84
heat	83
cooking	81
easy	73
pepper	72
cook	65

Summing up embeddings

pork tenderloin	
pork	0.883016
tenderloin	0.883016
beef	0.678835
roast	0.647979
chicken	0.637563
veal	0.633223
meat	0.621922
sausage	0.615066
loin	0.609921
chops	0.607775

Obama family tree

Word embeddings

obama		family		tree	
barack	0.925472	families	0.646008	trees	0.823568
mccain	0.759077	relatives	0.615727	pine	0.528495
bush	0.757099	father	0.615119	oak	0.516302
clinton	0.70856	parents	0.613432	shrubs	0.489276
hillary	0.649792	mother	0.580256	planted	0.484893
kerry	0.614406	friends	0.578592	trunks	0.470876
rodham	0.613864	daughter	0.539657	bark	0.464572
biden	0.594085	son	0.538854	garden	0.462577
gore	0.588598	wife	0.53823	fruit	0.462216
democrats	0.56083	home	0.533894	flower	0.460653

Query embedding

obama family tree	
family	0.736234
tree	0.690256
obama	0.616687
bush	0.569031
trees	0.541988
friends	0.539409
barack	0.525719
families	0.514836
mother	0.509138
home	0.509013

Obama family tree

Words in rel. doc.

#obama family tree	
obama	14807
barack	6092
obamas	3252
family	2772
president	1797
chicago	1757
born	1584
times	1516
michelle	1479
new	1374
brother	1362
robinson	1280
dunham	1271
kenya	1269

Query embedding

obama family tree	
family	0.736234
tree	0.690256
obama	0.616687
bush	0.569031
trees	0.541988
friends	0.539409
barack	0.525719
families	0.514836
mother	0.509138
home	0.509013

Why?

- Word2vec is trained to reproduce word context
  - Context similarity ~ word sense similarity

- Semantic similarity by human vs. for IR
  - Human: Obama is similar to McCain and Bush
  - … but not for search

Success story: DSSM
– Deep Structured Semantic Model

- Using click-through data for training
  - Encode relevance relationship between query-clicked document title

- Use letter 3-grams as input rather than words
  - cat –> #-c-a c-a-t a-t-#
  - Reduce vocabulary size to about 30-50K

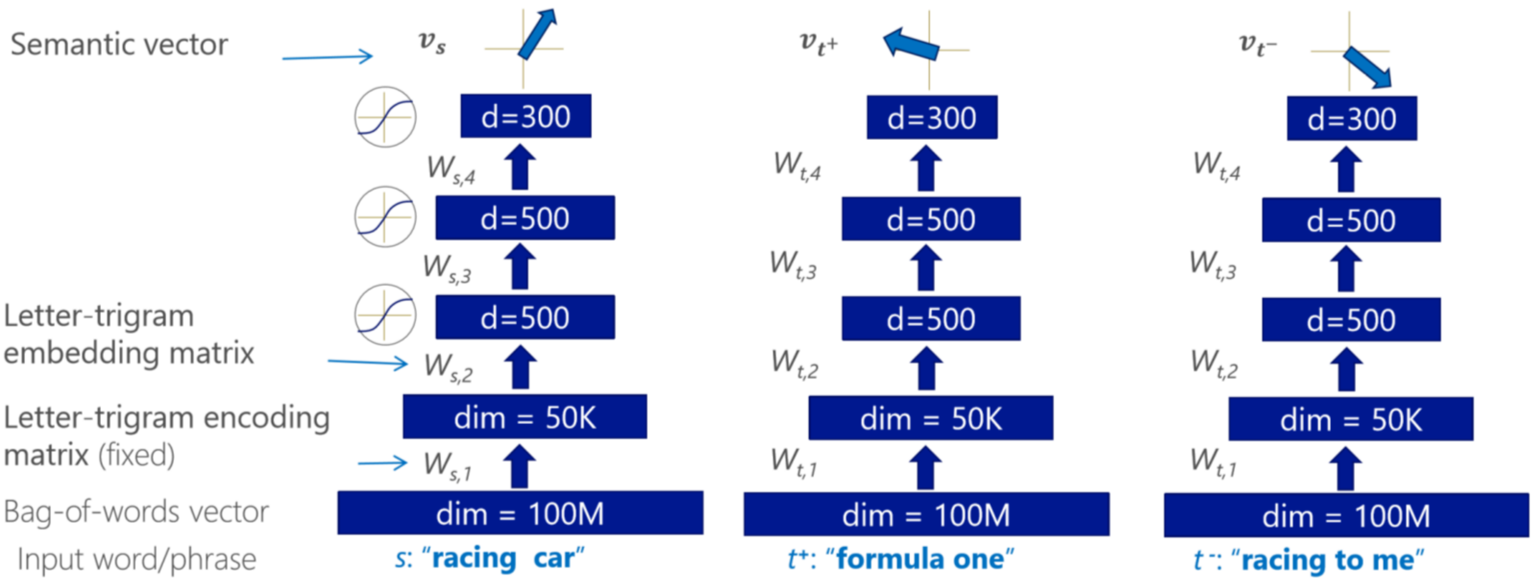
DSSM

DSSM for semantic embedding Learning

Initialization:

Neural networks are initialized with random weights

Huang, He, Gao, Deng, Acero, Heck, "Learning deep structured semantic models for web search using clickthrough data," CIKM, 2013



From [He et al. CIKM 2014 tutorial]

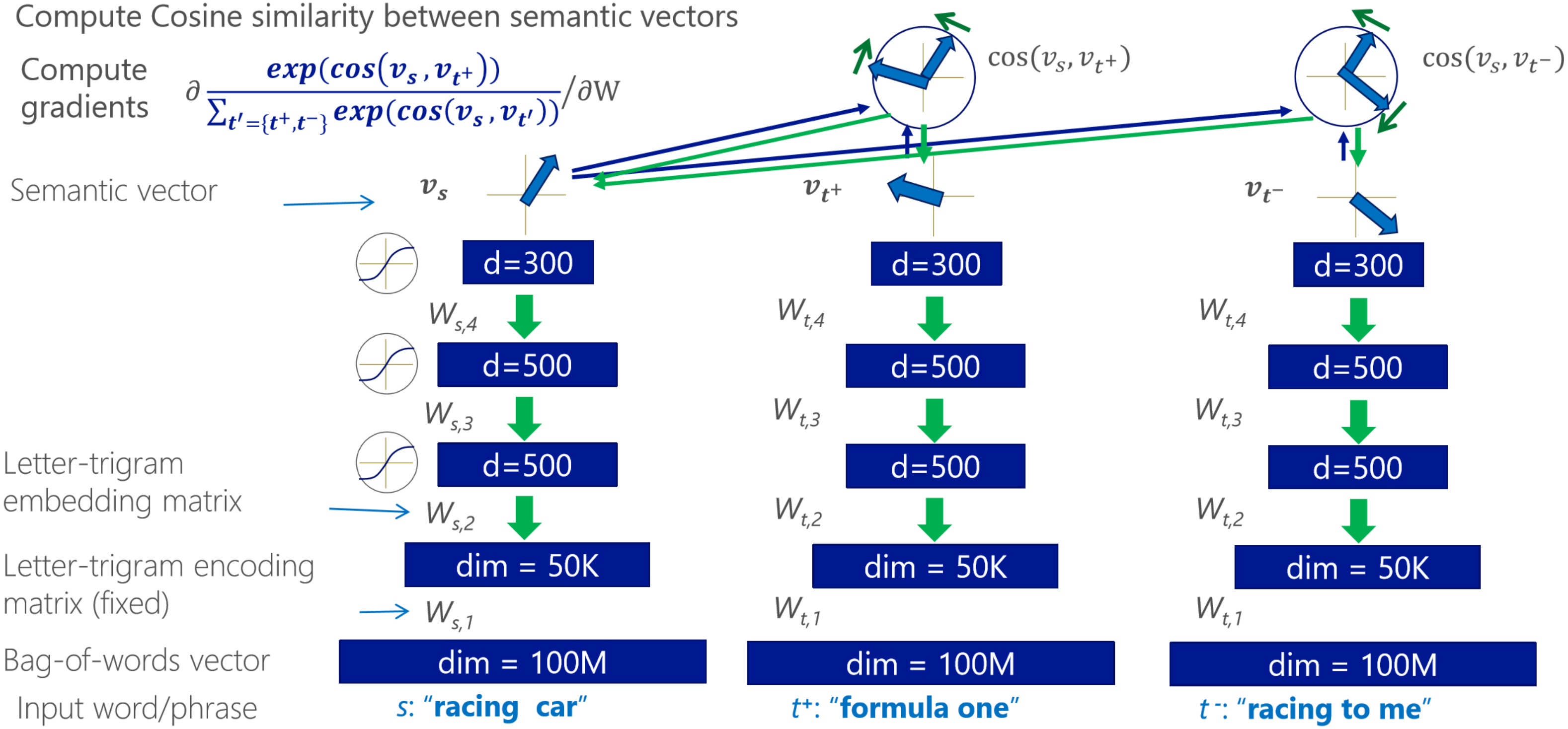
DSSM: train on click data

DSSM for semantic embedding learning

Training:

Compute Cosine similarity between semantic vectors

Compute gradients $\partial \dfrac{exp(cos(v_s, v_{t^+}))}{\sum_{t'=\{t^+,t^-\}} exp(cos(v_s, v_{t'}))} / \partial W$



From [He et al. CIKM 2014 tutorial]

Results on Web search

Model	Input dimension	NDCG@1 %
BM25 baseline	--	30.8
Probabilistic LSA (PLSA)		29.5
Auto-Encoder (Word)	40K	31.0 (+0.2)
DSSM (Word)	40K	34.2 (+3.4)
DSSM (Random projection)	30K	35.1 (+4.3)
DSSM (Letter-trigram)	30K	36.2 (+5.4)

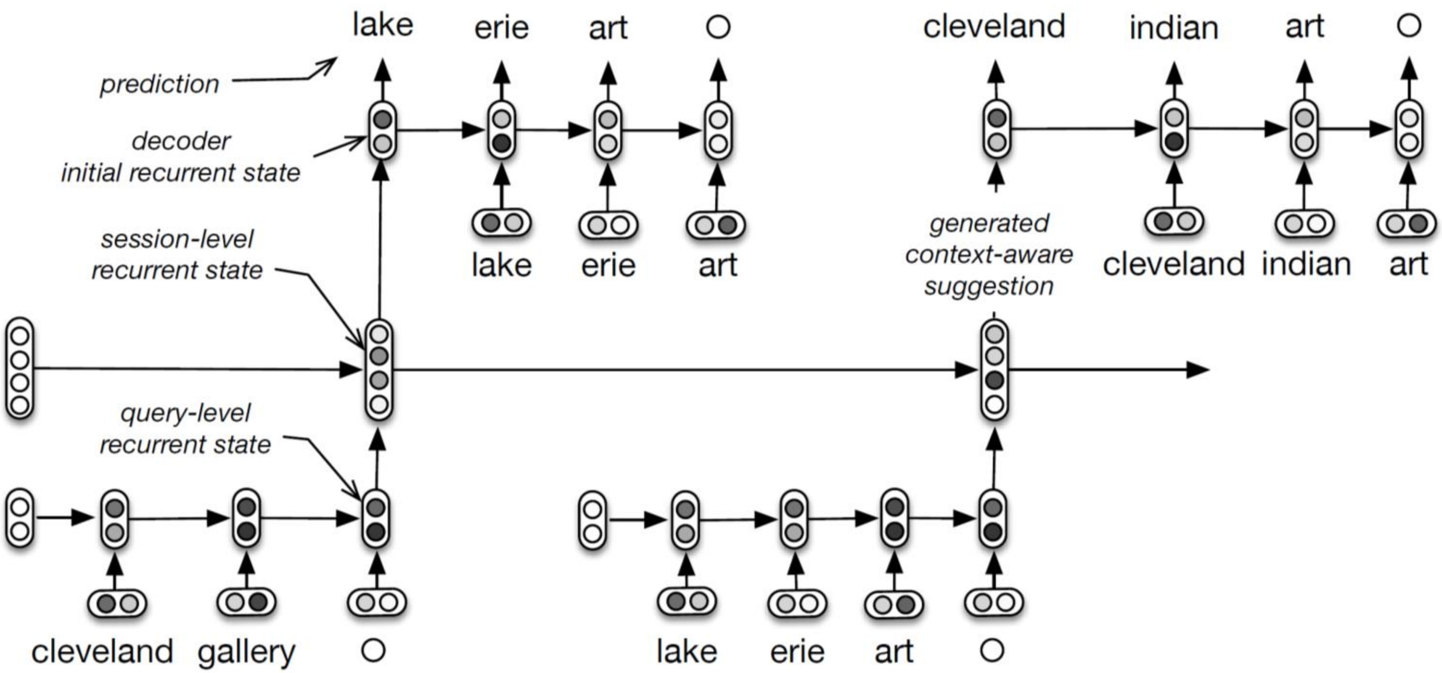
From He et al, CIKM 2014 tutorial

Hierarchical deep net for query suggestion and generation

[Sordoni et al. 2015] A hierarchical recurrent encoder-decoder for generative context-aware query suggestion, CIKM 2015)

- Learn to suggest/generate queries from query sessions
- Two layers:
  - Query embedding is generated from word embedding non-linearly (recurrent NN)
  - Query session embedding is generated from query embeddings non-linearly

HRED architecture



Use query logs to train the model:

Session: (query1, query2)

- Useful for query suggestion

Examples of generated queries

Context	Synthetic Suggestions
ace series drive	ace hardware ace hard drive hp officejet drive ace hardware series
cleveland gallery → lake erie art	cleveland indian art lake erie art gallery lake erie picture gallery sandusky ohio art gallery

Table 1: HRED suggestions given the context.

ace series drive (on Google)

cooperindustries.com/content/public/en/crouse-hinds/products/industrial_control/explosionproof_vari

ACE DG1 Series Explosionproof Variable Frequency Drives

Share

Size 1 (1 to 5 HP) Open

Click to open expanded view

The only explosionproof VFD solution utilizing NEMA 7 with active cooling.

ACE Variable Frequency Drives are highly flexible AC drives designed specifically for hazardous locations. Can be mounted next to the motor in the classified area, providing significant installation cost savings - along with the traditional VFD benefits of energy savings, speed and torque control, and system diagnostics. Now available with Eaton's PowerXL DG1 drive, up to 100 HP!

- Complete Technical Specifications
- Locate Authorized Distributor
- Contact Local Sales Representative

Evaluation on AOL

- Next query prediction
- Incorporate HRED score as feature in a L2R model (LambdaMART) with 17 other pairwise and contextual features

Method	MRR	$\Delta\%$
ADJ	0.5334	-
Baseline Ranker	0.5563	*+4.3%*
+ HRED	**0.5749**	*+7.8%/+3.3%*

Table 3: Next-query prediction results. All improvements are significant by the t-test ($p < 0.01$).

Summary on query representation

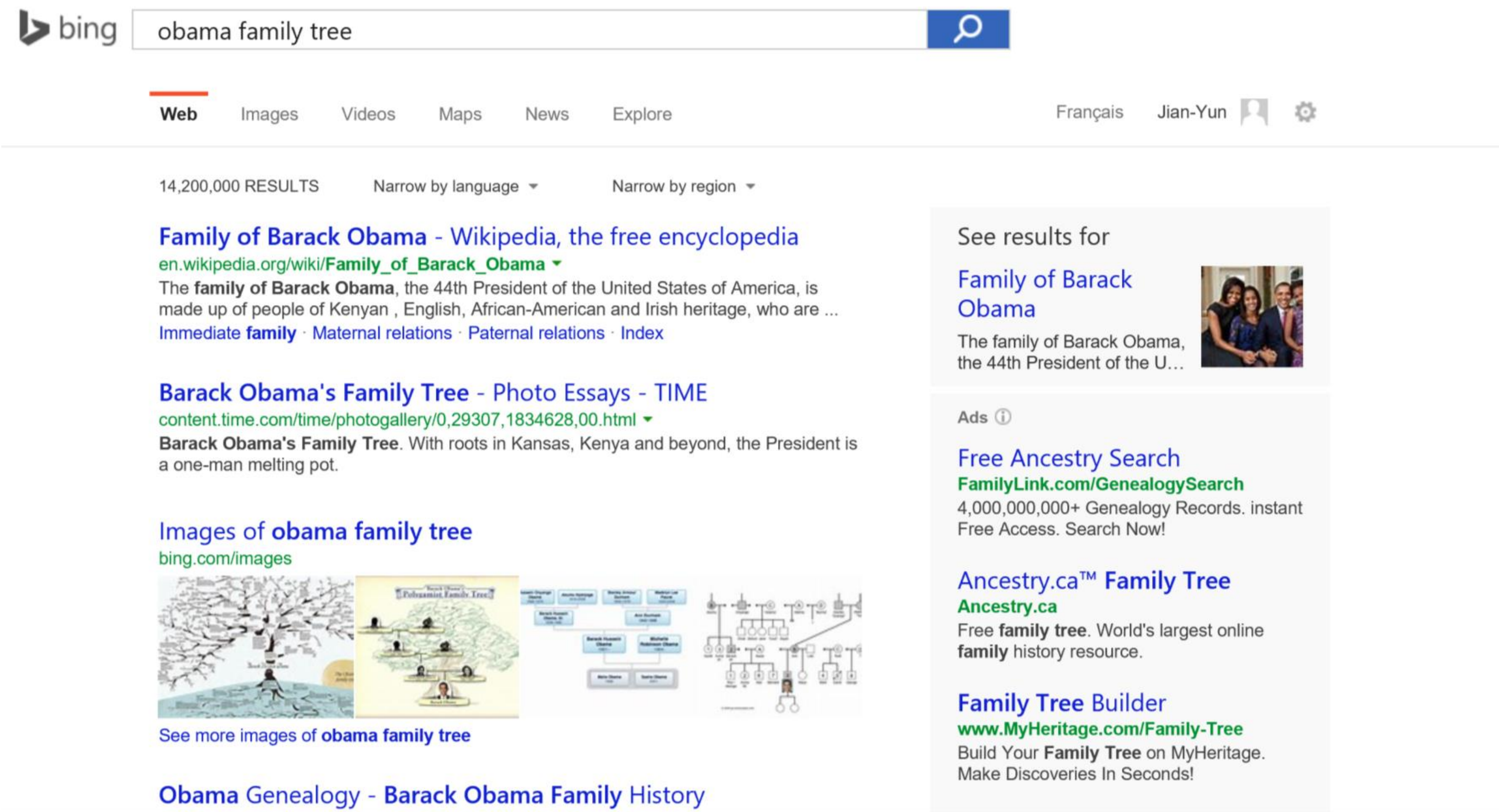
- Models trained to reproduce the text (unsupervised learning) has limited success in IR
- Models trained on IR-specific data are more successful (DSMM, HRED)

➔A representation is useful for IR if it is trained to reproduce IR data (relevance)

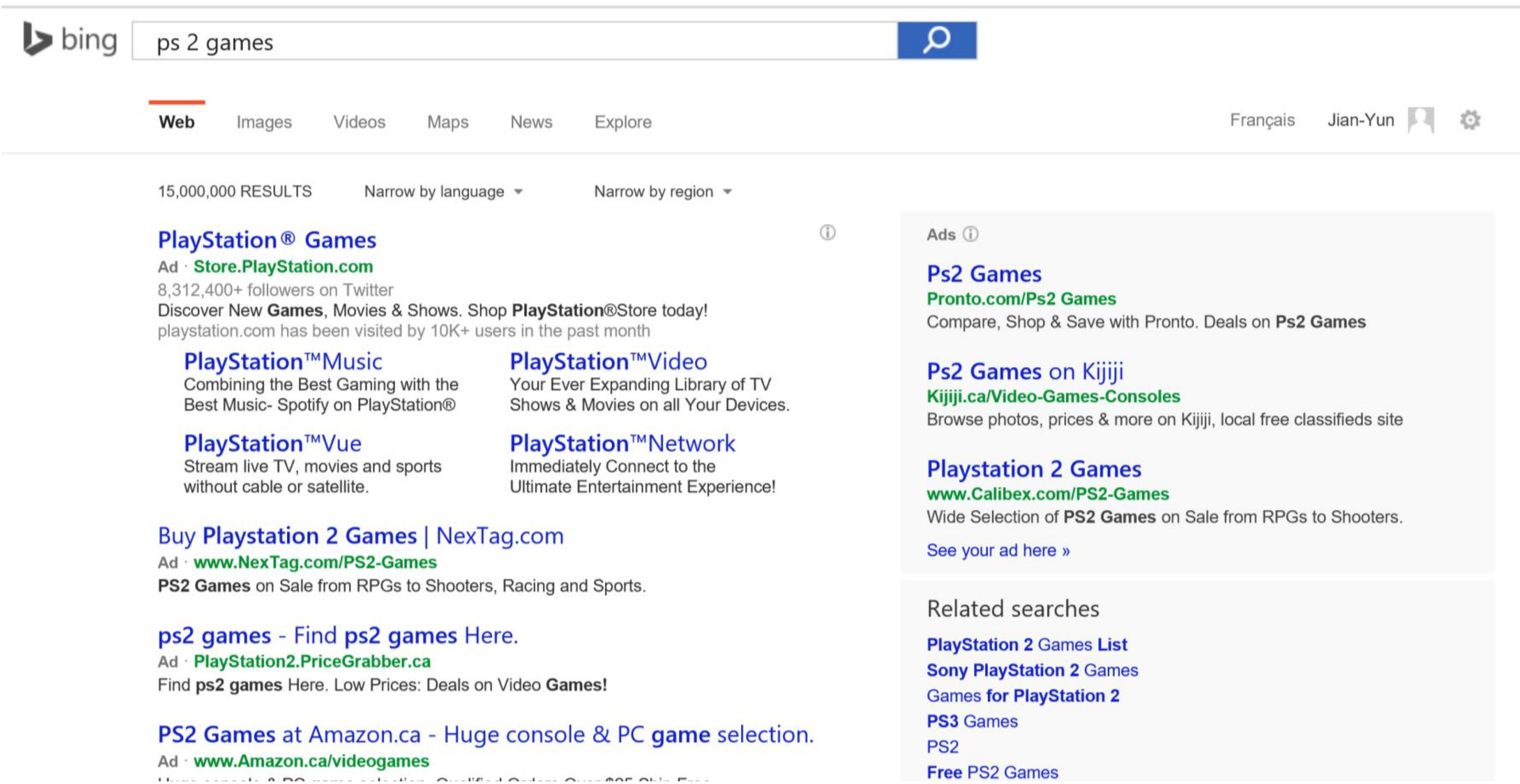
Some remarks: NLP for IR

- Off-the-shelf NLP tools are not necessarily adapted to the need of IR
- IR works with reasonably meaningful features (words) + links + query logs
  - Much more meaningful than pixels
- It is difficult to create a better representation than words

Search result is not so bad



Search result is not so bad



How can NLP help IR?

- The NLP tools should be trained for IR tasks (relevance)
  - IR-specific query “parsing”
  - Representation trained on IR data

Karl Popper: Objects can become similar or dissimilar *only* in this way – by being related to needs and interests

Thank you

nie@iro.umontreal.ca