Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search

Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, Narayan Bhamidipati

Yahoo Labs - US & London

Search Solutions 2015 - London, UK

Query Rewriting

Try our state-of-the-art system for query rewriting.

bank of america loans

Result	Relevance
bank_of_america_loans_personal	73.3375%
bank_of_america_personal_loans	70.0765%
bank_of_america_car_loans	68.5054%
bank_of_america_home_loans	68.2654%
bank_of_america_auto_loans	65.7220%
bank_of_america_mortgage	65.6517%
bank_of_america_personal_loans_online	64.9183%
bank_of_america_loan	64.1622%
bank_of_america_personal_loan	62.3031%
chase_personal_loans	62.23%
bank_of_america_student_loans	61.3762%

Input Type:

Search Query

Output Type:

Search Query

Search

Query Rewriting

Try our state-of-the-art system for query rewriting.

bank of america loans

Result	Relevance
bank_of_america_loans_personal	73.3375%
bank_of_america_personal_loans	70.0765%
bank_of_america_car_loans	68.5054%
bank_of_america_home_loans	68.2654%
bank_of_america_auto_loans	65.7220%
bank_of_america_mortgage	65.6517%
bank_of_america_personal_loans_online	64.9183%
bank_of_america_loan	64.1622%
bank_of_america_personal_loan	62.3031%
chase_personal_loans	62.23%
bank_of_america_student_loans	61.3762%

Input Type:

Search Query

Output Type:

Search Query

Search

Query Rewriting for Advertising

bidterms

Books at Amazon - Free 2-Day Shipping w/ Amazon Prime - Amazon.com
www.amazon.com/**Books**
Low Prices on Millions of **Books**
★★★★½ rating for amazon.com

book

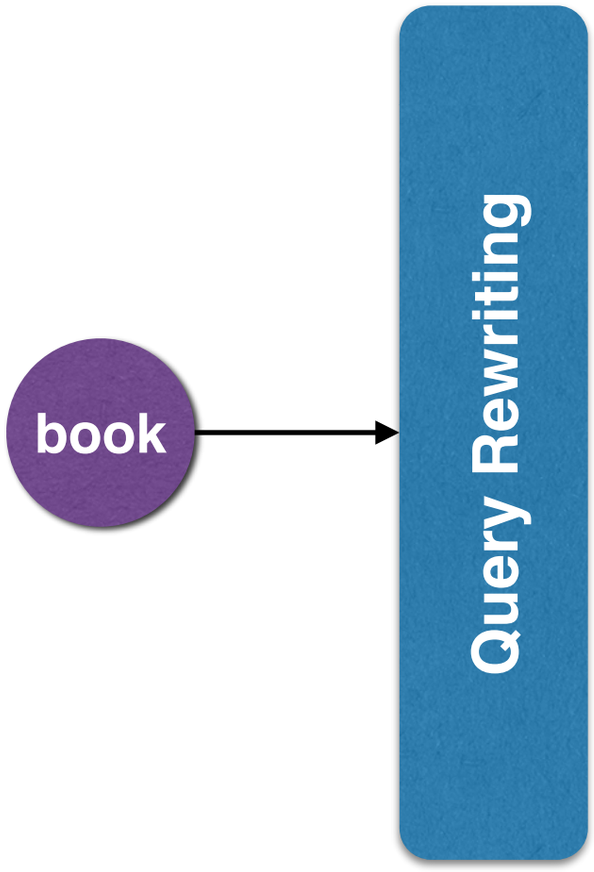
bidterms

Cheap Audio **Books**
TheWorks.co.uk/audio-books Ad
4.5 ★★★★½ user rating for theworks.co.uk
Audio **Books** : Save at The Works Discount Store

bidterms

Kobo™ Free eBooks
KoboBooks.com Ad
Millions Of Bestsellers And Free eBooks. Read On Almost Any Device!

Query Rewriting for Advertising



bidterms

Books at Amazon - Free 2-Day Shipping w/ Amazon Prime - Amazon.com
www.amazon.com/**Books**
Low Prices on Millions of **Books**
★★★★½ rating for amazon.com

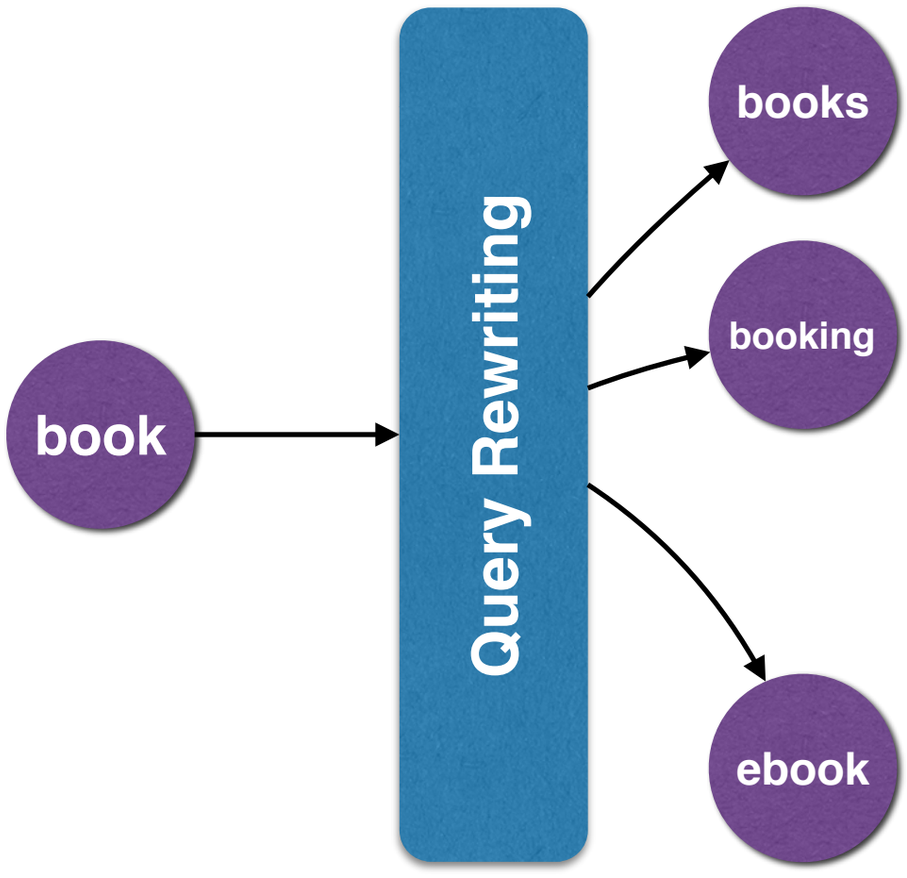
bidterms

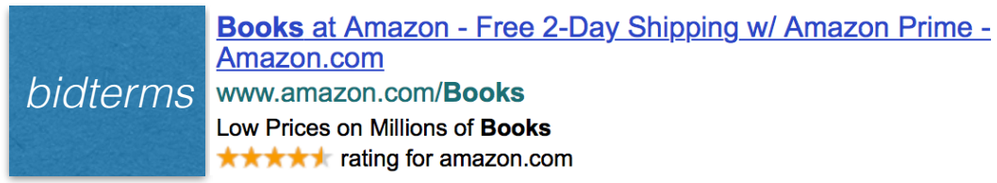
Cheap Audio **Books**
TheWorks.co.uk/audio-books Ad
4.5 ★★★★½ user rating for theworks.co.uk
Audio **Books** : Save at The Works Discount Store

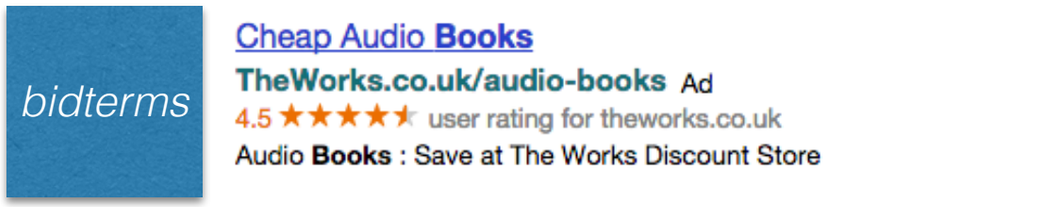
bidterms

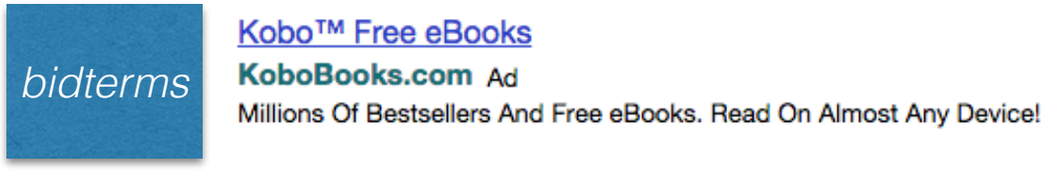
Kobo™ Free eBooks
KoboBooks.com Ad
Millions Of Bestsellers And Free eBooks. Read On Almost Any Device!

Query Rewriting for Advertising



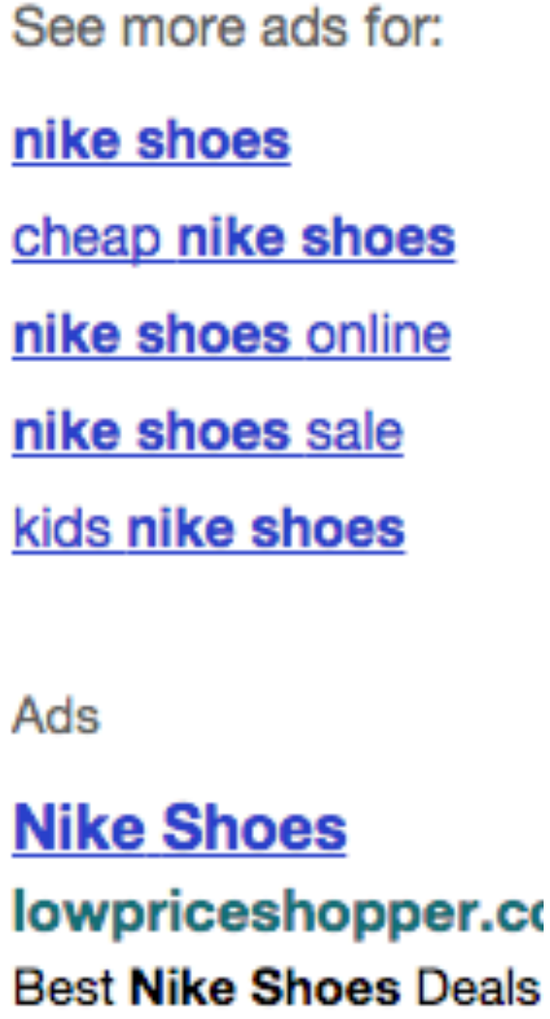






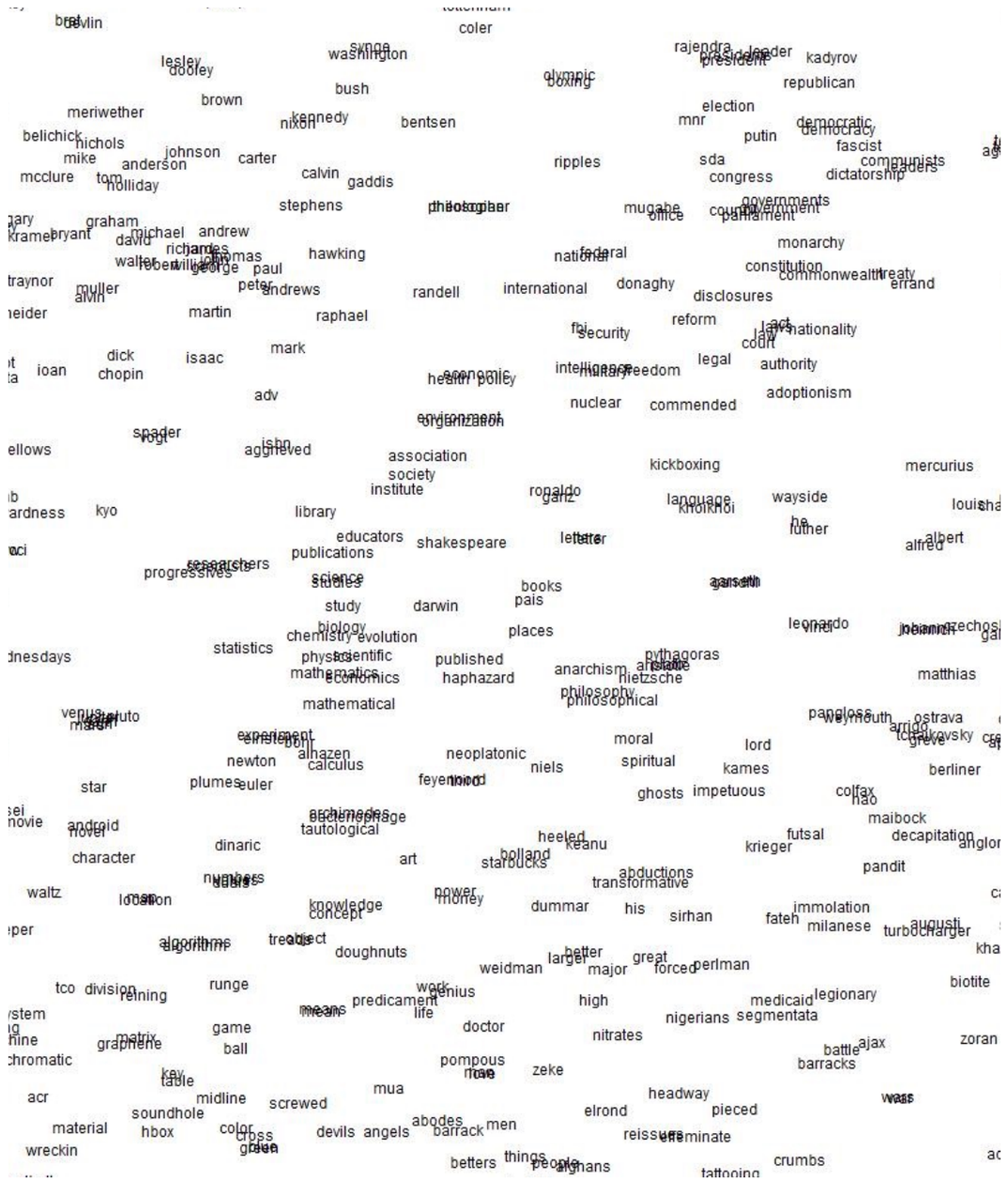
(Lots of) Previous Work

- Graph-based Methods
  - Click Graph
  - Query-Flow Graph
  - Term-Query Graph
  - …
- Clustering queries from history
- Supervised Learning-based methods
- …



Using Word Embeddings

- Sessions represent well the interactions of users with a search engines:
  - *queries*
  - *clicks*: algo results and ads
- We generate a distributed representation of actions in a common vector space:
  - inspired by *word2vec*

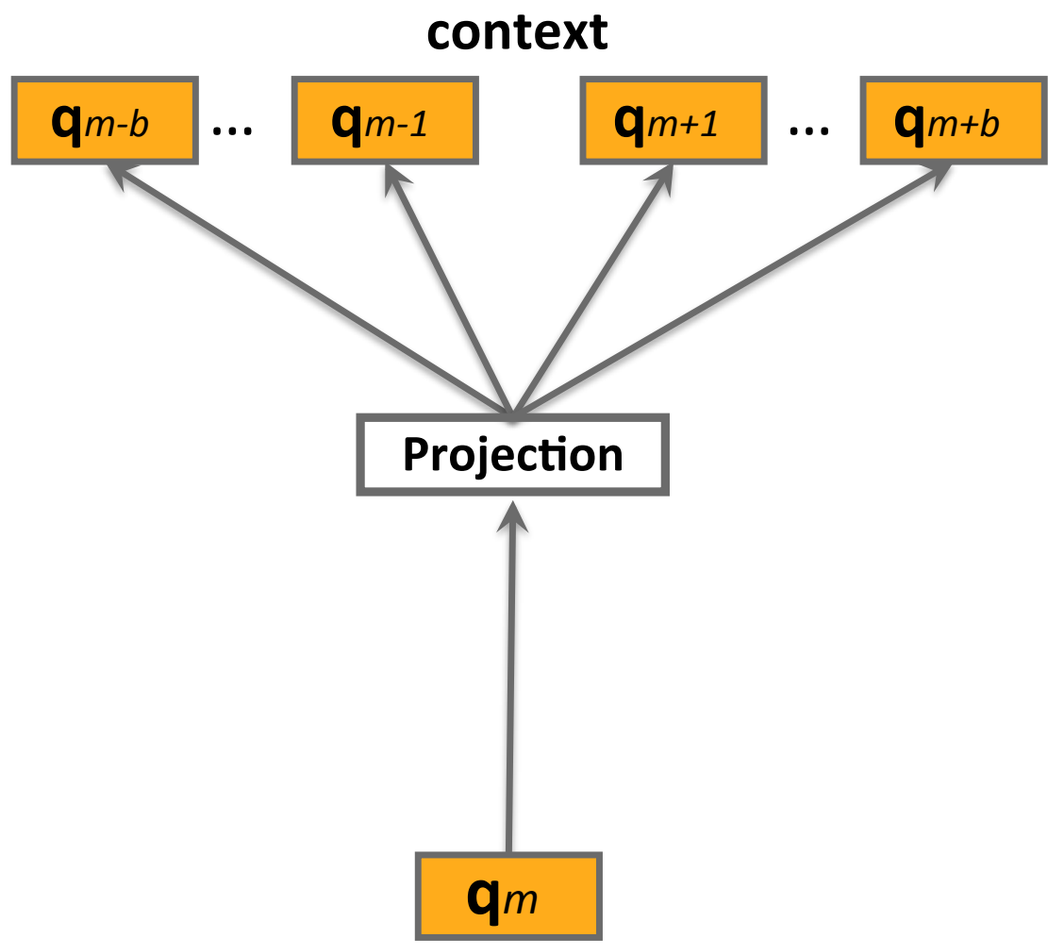


Models for Embedding Interactions

- Context2Vec
- Content2Vec
- Context-Content2Vec

Model 1: Context2vec

- Session as a sentence:
  - actions are words (e.g., ‘nike shoes’)
  - vectors are learned on each action/word



$$\mathcal{L} = \sum_{s \in \mathcal{S}} \sum_{q_m \in s} \sum_{-b \leq i \leq b, i \neq 0} \log \mathbb{P}(q_{m+i} | q_m).$$

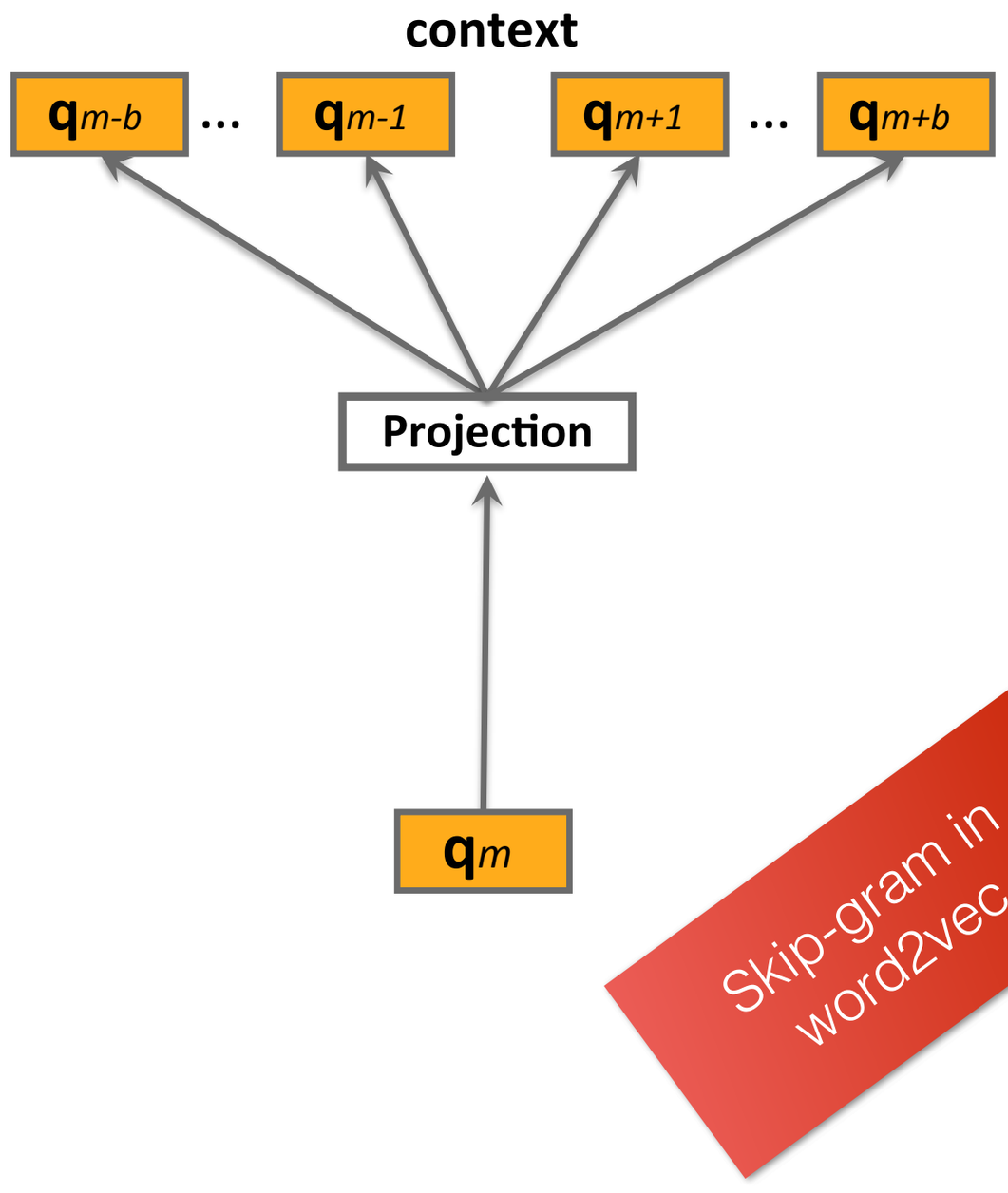
$$\mathbb{P}(q_{m+i} | q_m) = \frac{\exp(\mathbf{v}_{q_m}^\top \mathbf{v}'_{q_{m+i}})}{\sum_{q=1}^{V} \exp(\mathbf{v}_{q_m}^\top \mathbf{v}'_q)}$$

Model 1: Context2vec

- Session as a sentence:
  - actions are words (e.g., ‘nike shoes’)
  - vectors are learned on each action/word

$$\mathcal{L} = \sum_{s \in \mathcal{S}} \sum_{q_m \in s} \sum_{-b \leq i \leq b, i \neq 0} \log \mathbb{P}(q_{m+i} | q_m).$$

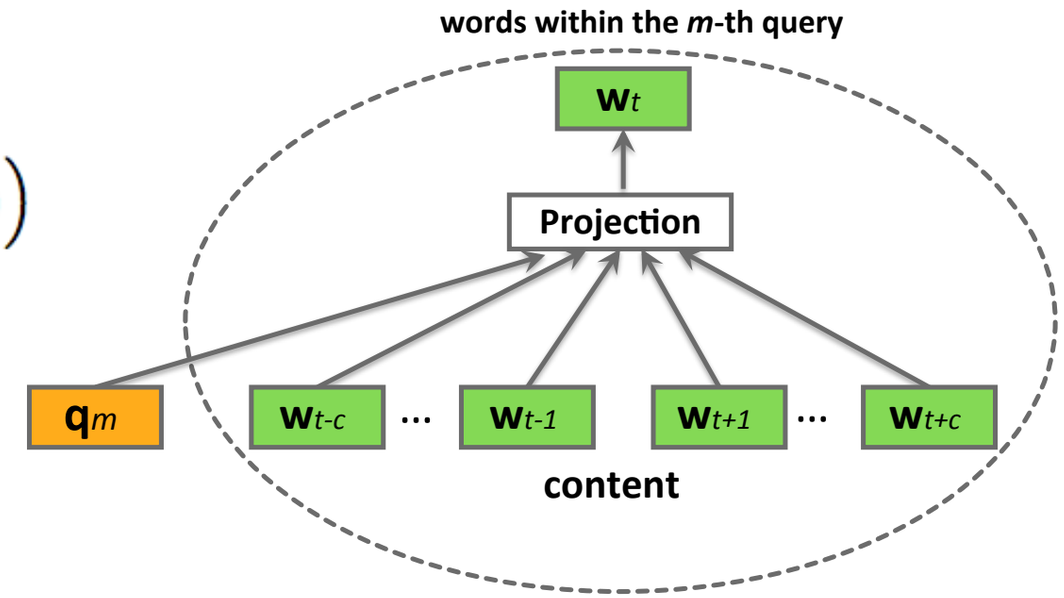
$$\mathbb{P}(q_{m+i} | q_m) = \frac{\exp(\mathbf{v}_{q_m}^\top \mathbf{v}'_{q_{m+i}})}{\sum_{q=1}^{V} \exp(\mathbf{v}_{q_m}^\top \mathbf{v}'_q)}$$



Model 2: Content2Vec

- Consider the “content” of the action

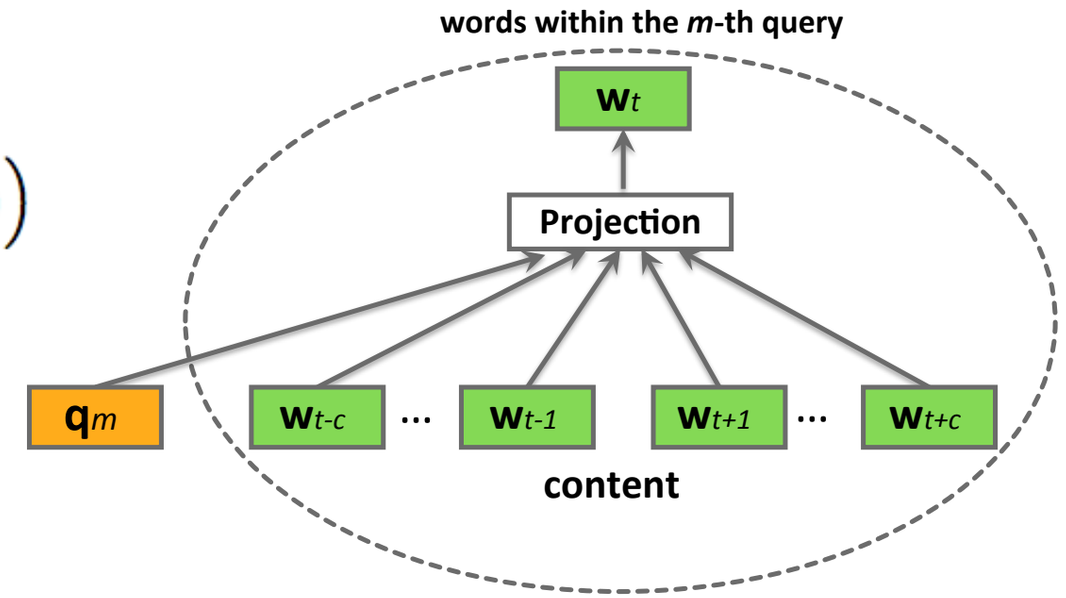
$$\mathcal{L} = \sum_{s \in \mathcal{S}} \Big(\sum_{q_m \in s} \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Big)$$



Model 2: Content2Vec

- Consider the “content” of the action

$$\mathcal{L} = \sum_{s \in \mathcal{S}} \Big(\sum_{q_m \in s} \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Big)$$



$$\mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) = \frac{\exp(\bar{\mathbf{v}}^\top \mathbf{v}'_{w_{mt}})}{\sum_{w=1}^{V} \exp(\bar{\mathbf{v}}^\top \mathbf{v}'_w)}$$

$$\bar{\mathbf{v}} = \frac{1}{2c+1} (\mathbf{v}_{q_m} + \sum_{-c \le j \le c, j \neq 0} \mathbf{v}_{w_{m,t+j}})$$

$$\mathbb{P}(q_m | w_{m1} : w_{mT_m}) = \frac{\exp(\bar{\mathbf{v}}_m^\top \mathbf{v}'_{q_m})}{\sum_{w=1}^{V} \exp(\bar{\mathbf{v}}_m^\top \mathbf{v}'_w)}$$

$$\bar{\mathbf{v}}_m = \frac{1}{T_m} \sum_{t=1}^{T_m} \mathbf{v}_{w_{mt}}$$

Model 2: Content2Vec

- Consider the "content" of the action

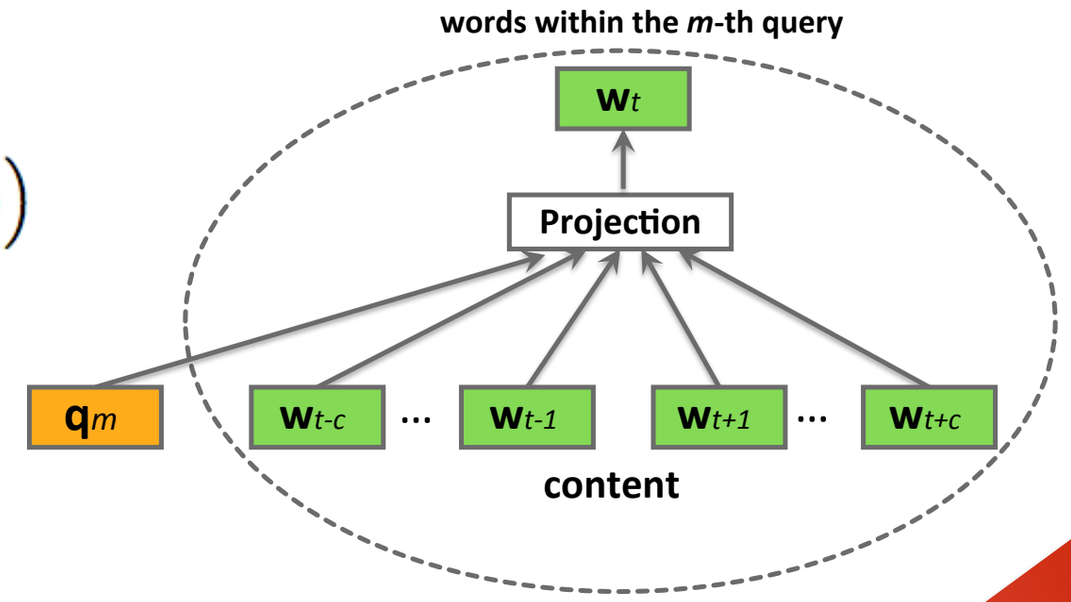
$$\mathcal{L} = \sum_{s \in \mathcal{S}} \Big(\sum_{q_m \in s} \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Big)$$

$$\mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) = \frac{\exp(\bar{\mathbf{v}}^\top \mathbf{v}'_{w_{mt}})}{\sum_{w=1}^{V} \exp(\bar{\mathbf{v}}^\top \mathbf{v}'_w)}$$

$$\bar{\mathbf{v}} = \frac{1}{2c+1} (\mathbf{v}_{q_m} + \sum_{-c \le j \le c, j \neq 0} \mathbf{v}_{w_{m,t+j}})$$

$$\mathbb{P}(q_m | w_{m1} : w_{mT_m}) = \frac{\exp(\bar{\mathbf{v}}_m^\top \mathbf{v}'_{q_m})}{\sum_{w=1}^{V} \exp(\bar{\mathbf{v}}_m^\top \mathbf{v}'_w)}$$

$$\bar{\mathbf{v}}_m = \frac{1}{T_m} \sum_{t=1}^{T_m} \mathbf{v}_{w_{mt}}$$

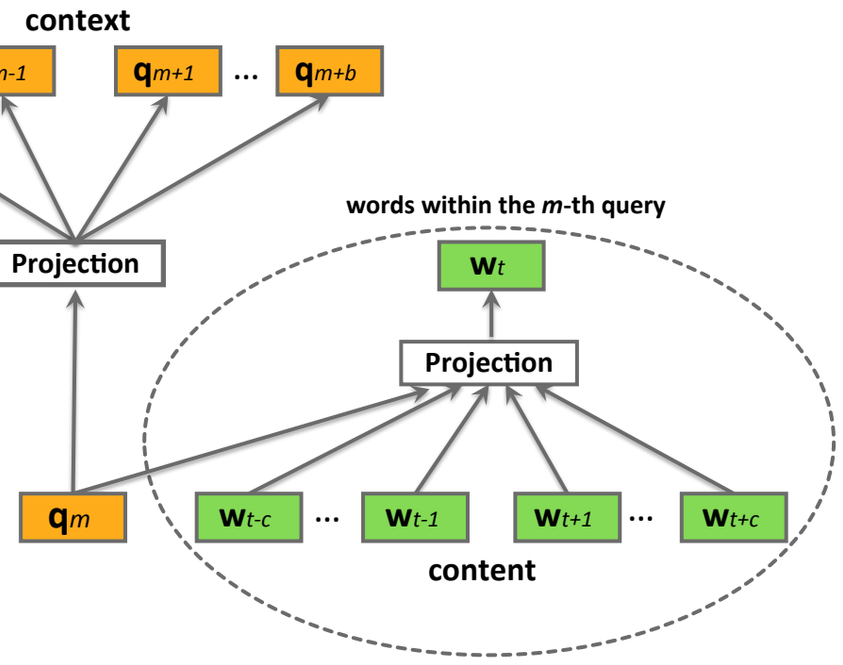


Similar to Paragraph2Vec

Model 3: Context-Content2Vec

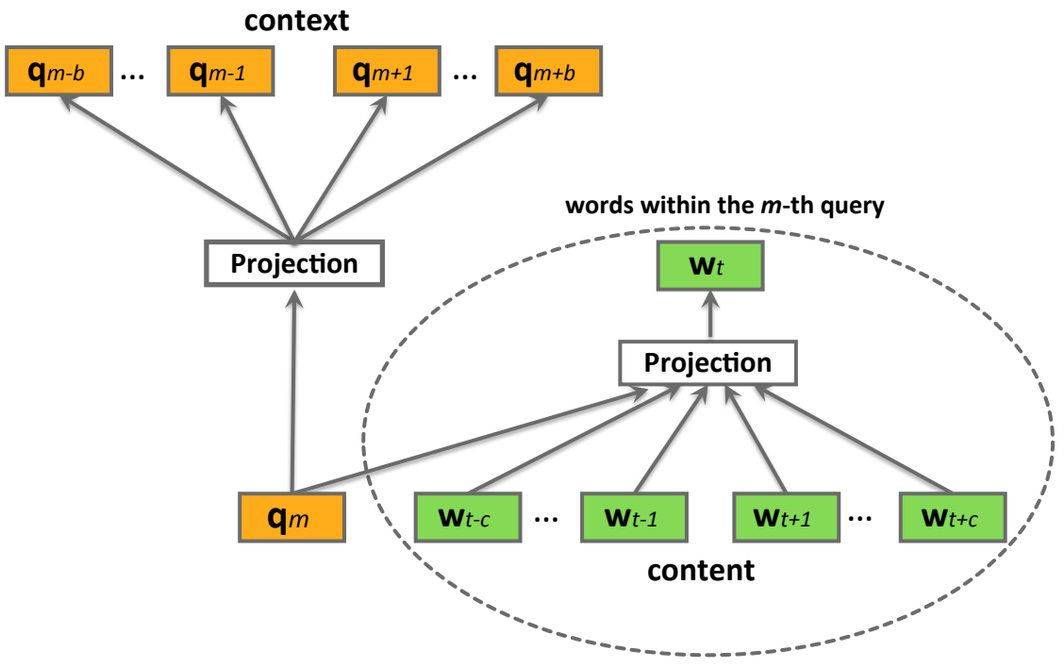
- A combination of the previous two models
- When an action is not popular we give more importance to its content

$$\mathcal{L} = \sum_{s \in \mathcal{S}} \sum_{q_m \in s} \Bigg(\sum_{-b \leq i \leq b, i \neq 0} \log \mathbb{P}(q_{m+i} | q_m) + \alpha_m \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Bigg)$$



Model 3: Context-Content2Vec

- A combination of the previous two models
- When an action is not popular we give more importance to its content



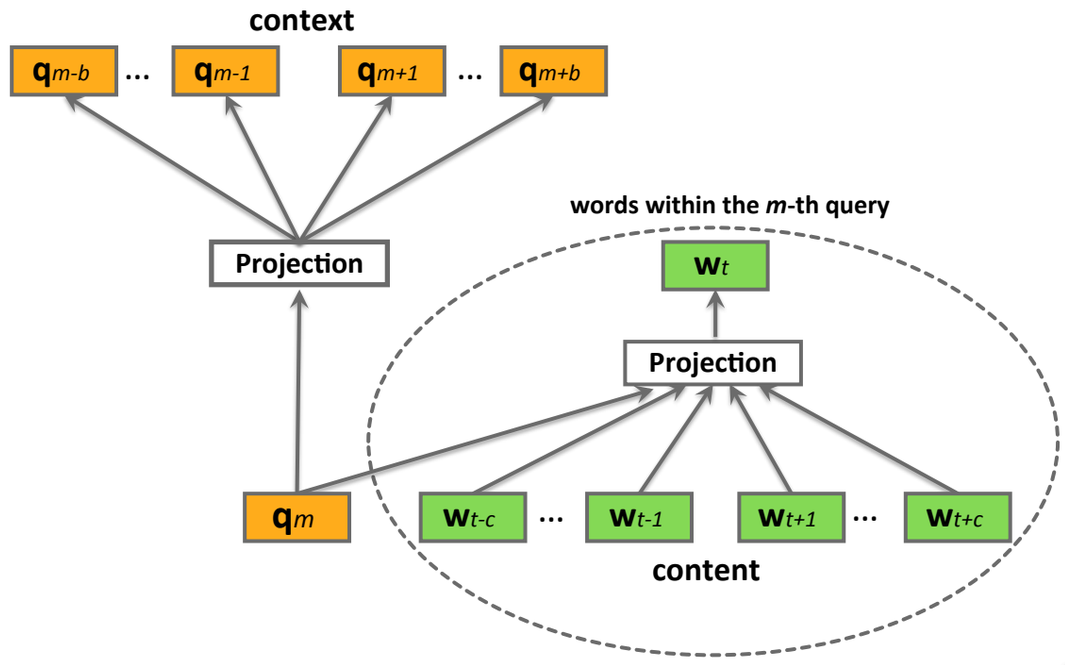
$$\mathcal{L} = \sum_{s \in \mathcal{S}} \sum_{q_m \in s} \Bigg(\sum_{-b \le i \le b, i \neq 0} \log \mathbb{P}(q_{m+i} | q_m) + \alpha_m \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Bigg)$$

$$\alpha_m = \frac{1}{\log(1+K_m)}$$

K_m is the frequency of the K-th actions

Model 3: Context-Content2Vec

- A combination of the previous two models
- When an action is not popular we give more importance to its content



$$\mathcal{L} = \sum_{s \in \mathcal{S}} \sum_{q_m \in s} \Bigg(\sum_{-b \le i \le b, i \neq 0} \log \mathbb{P}(q_{m+i} | q_m) + \alpha_m \log \mathbb{P}(q_m | w_{m1} : w_{mT_m}) + \sum_{w_{mt} \in q_m} \log \mathbb{P}(w_{mt} | w_{m,t-c} : w_{m,t+c}, q_m) \Bigg)$$

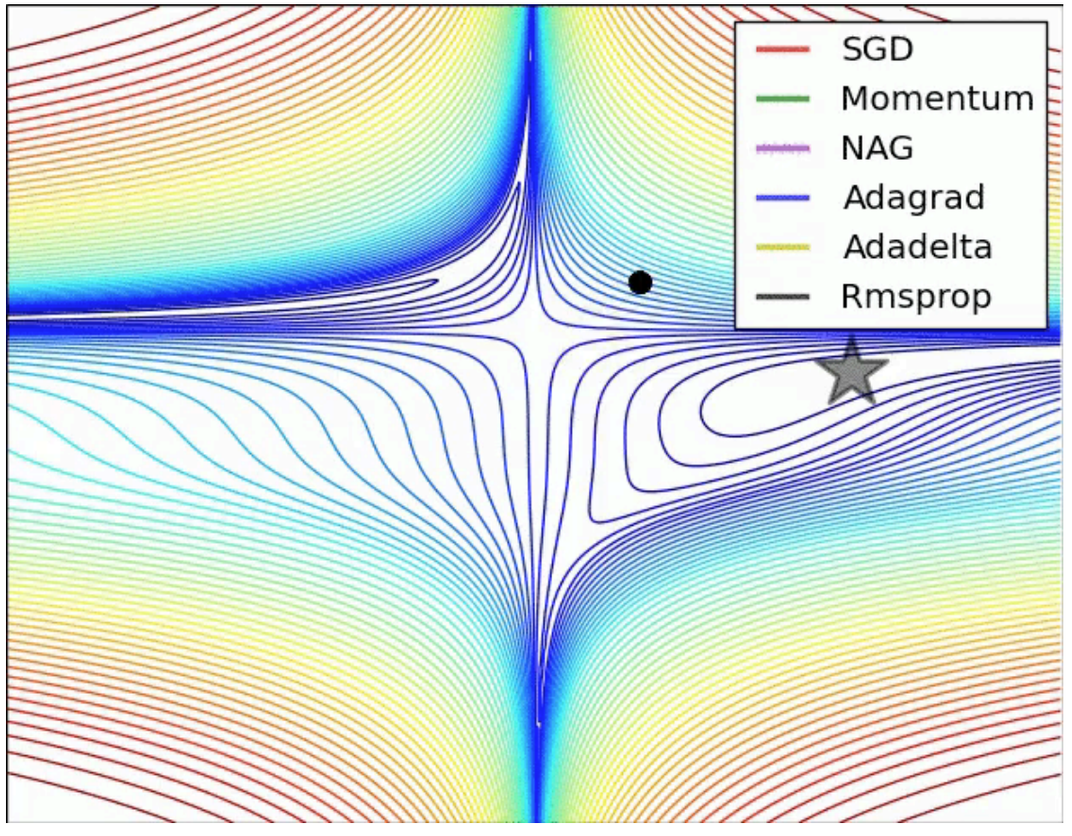
$$\alpha_m = \frac{1}{\log(1+K_m)}$$

K_m is the frequency of the K-th actions

Original Contribution

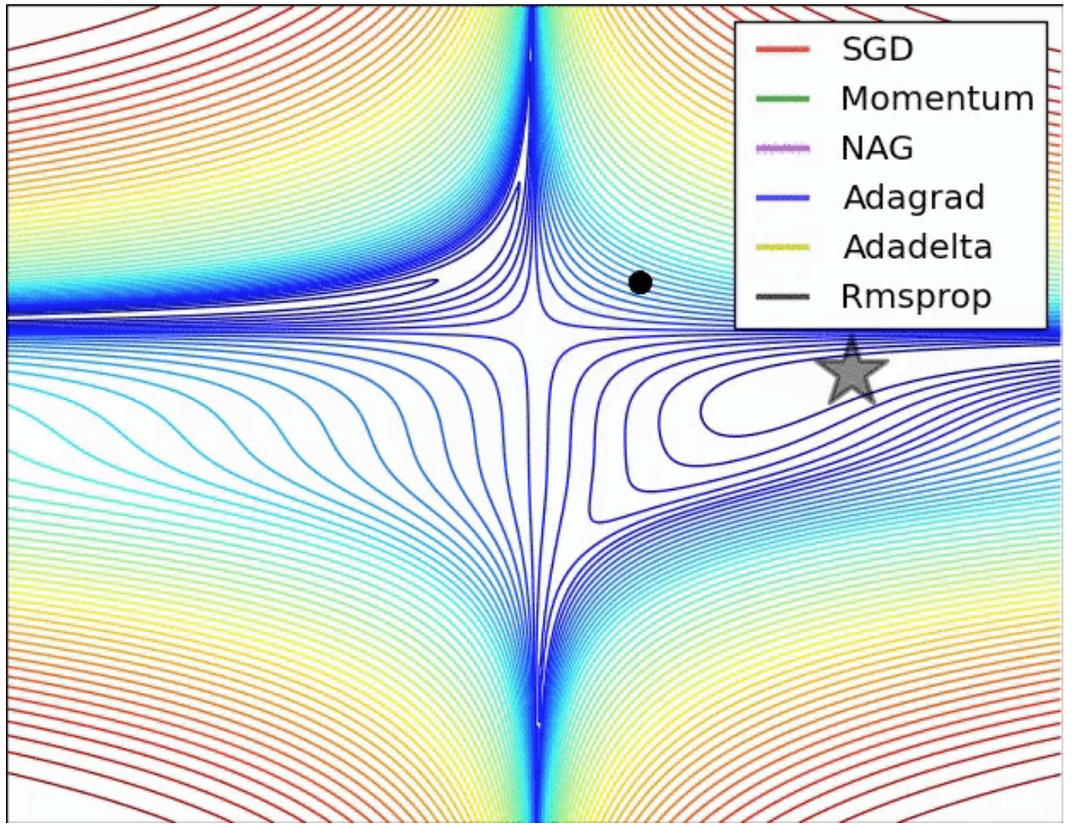
Optimization

- Stochastic Gradient Ascent
- Negative Sampling:
  - 10 Negative samples
  - Negatives chosen in order to deal with clicks and skips
- Dimension D = 300
- Context neighborhood = 5
- Content neighborhood = 7



Optimization

- Stochastic Gradient Ascent
- Negative Sampling:
  - 10 Negative samples
  - Negatives chosen in order to deal with clicks and skips
- Dimension D = 300
- Context neighborhood = 5
- Content neighborhood = 7



Experiments

- Session Data
  - *12 billion* sessions collected on the US website of Yahoo Search
- Query Content Data
  - from the *45 million* most frequent queries
- Click Data
  - For all the sessions we collected the clicks on algo and sponsored results
- Query Flow Graph (QFG) as baseline

Anecdotes

Query	cx2vec	cx-cn2vec
minnesota insurance exam crossword puzzles	satellite tv otego satellite tv menominee satellite tv west end satellite tv townsend satellite tv lake sara	minnesota insurance minnesota insurance license practice exams online insurance exam crossword puzzles colorado insurance exam crossword puzzles online minnesota insurance exam crossword puzzles
microwave food safety	staphylococcal enteritis definition salmonella enteritis definition listeria monocytogenes prevention e coli cdc preventing cross contamination	microwave oven food safety microwave baby food safety microwave food safety studies microwave food safety issues foodsafety.com
what to cook in cast iron skillet	steak sauce substitute montreal seasoning ingredients ground turkey breakfast sausage how to cook steak on cast iron skillet reseason cast iron	cast iron skillet recipes how to cook with cast iron skillet how to cook in a cast iron skillet how to cook with a cast iron skillet chicken in cast iron skillet
iphone 6 repair services	mp3attic music donar ovulos en elche credit a la consommation rapide smart phone repair service social security disability bronx ny	at&t iphone repair service iphone 5c repair service iphone repair iphone service repair iphone repair services

The Effect of Ad Clicks

makeup (no ads)	**makeup (with ads)**
makeup tips	lipstick
fashion makeup	**mac makeup**
make up	**makeup sets**
makeup pictures	eye shadow
makeup images	**makeup covergirl**
makeup tutorial	makeup items

The Effect of Ad Clicks

makeup (no ads)	makeup (with ads)
makeup tips	lipstick
fashion makeup	**mac makeup**
make up	**makeup sets**
makeup pictures	eye shadow
make	
make	

snowboarding (no ads)	snowboarding (with ads)
snowbaording	snowboards
snow boarding	**snowboarding gear**
snowboarding information	**burton snowboarding**
snowboarding jumps	**snowboard deals**
snowboard pics	**snowboards on sale**
shaun white snowboarding	snowboarding mountains

The Effect of Ad Clicks

makeup (no ads)	makeup (with ads)
makeup tips	lipstick
fashion makeup	**mac makeup**
make up	**makeup sets**
makeup pictures	eye shadow
make	
make	

snowboarding (no ads)	snowboarding (with ads)
snowbaording	snowboards
snow boarding	**snowboarding gear**
snowboarding information	**burton snowboarding**
snowboarding jump	
snowboard pics	
shaun white snowb	

seafood (no ads)	seafood (with ads)
sea food	**seafood restaurant**
crab legs	**seafood restaurants**
best seafood	crab shack
oysters	**seafood market**
lobster recipes	sea food
seafood market	**sea food menu**

A Broader Comparison

Original	$\mathbf{QFG_{ad+link}}$	cn2vec	cx-cn2vec	$\mathbf{cx\text{-}cn2vec_{ad}}$
wedding budget calculator	wedding budget	monthly budget calculator	**wedding planning checklist**	**wedding budget worksheet**
	wedding cost calculator	online budget calculator	wedding budget template	wedding vendors
	wedding calculator	**budget wedding**	**wedding budget worksheet**	**the knot**
	wedding cost breakdown	budget calculator free	wedding checklist printable	wedding planning checklist
	wedding budget worksheet	average wedding budget	wedding costs average	**wedding wire**
gmat prep classes	gmat prep	gmat prep online	**gmat preparation courses**	**gmat study books**
	gmat classes	gmat prep courses	sample gmat tests	**gmat test prep classes**
	gmat	**gmat online prep**	**gmat prep class**	**gmat prep class**
	kaplan gmat course	online gmat prep	which is easier gre or gmat	**kaplan gre courses**
	kaplan gmat	best gmat prep courses	how much is the gmat	free gmat sample tests
how to build a fence	building a fence	how to build fence	**how to build a fence on a hill**	how to build a fence minecraft
	build your own fence	build fence	how to fence a yard	fancy fences and gates
	how to build a wood fence	build a fence	how to build a fence gate	**how to build a metal fence**
	do it yourself fence	how to build a cheap fence	how to build a brick wall fence	**home depot com fencing**
	how to build a privacy fence	build your own fence	**how to build a fence video**	**back yard fences**
solar panels	**solar panels for homes**	solar panels for your home	**solar power**	**solar power**
	solar electric panels	solar panels for residential homes	**solar energy**	**solar panels for homes**
	solar power	**solar panels for**	**solar panels for homes**	**solar panels on sale**
	ebay solar panels	solar panels on ebay	**solar power systems**	**solar panel kits**
	how to make solar panels	davis solar	solar panel	**solar panels for sale**

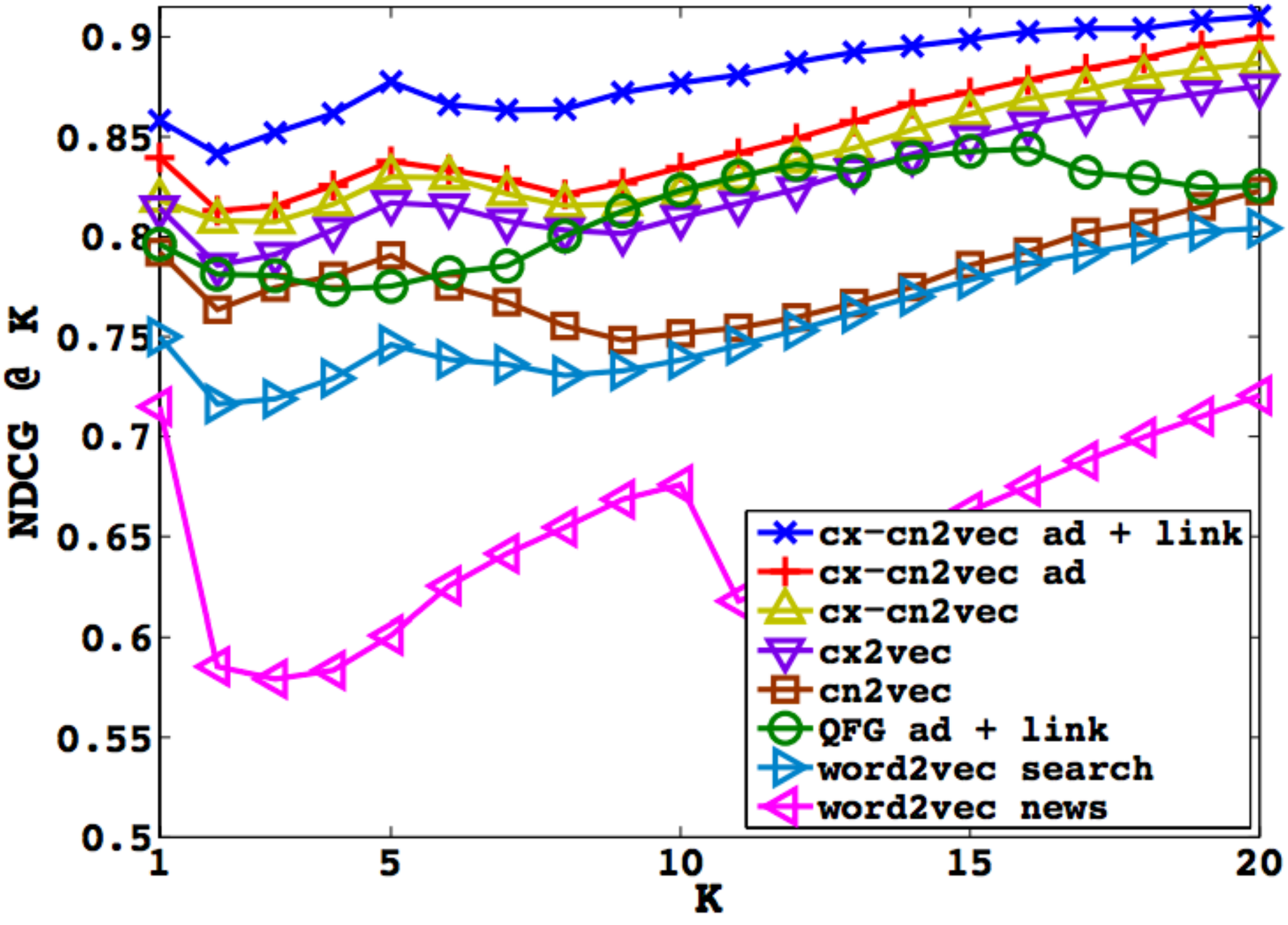
Editorial Evaluation

In-dictionary

grade	**pairs**	cn2vec	cx2vec	cx-cn2vec	cx-cn2vec$_{ad}$	**cx-cn2vec$_{ad+link}$**	word2vec$_{news}$	word2vec$_{search}$	QFG$_{ad+link}$
Excellent	1,518	0.630 (0.136)	0.658 (0.107)	0.669 (0.107)	0.668 (0.114)	**0.733 (0.094)**	0.818 (0.146)	0.648 (0.151)	0.329 (0.667)
Good	5,531	0.599 (0.136)	0.621 (0.125)	0.637 (0.100)	0.632 (0.104)	**0.683 (0.097)**	0.770 (0.152)	0.614 (0.155)	0.205 (0.574)
Fair	4,021	0.550 (0.167)	0.565 (0.129)	0.577 (0.124)	0.566 (0.130)	**0.605 (0.132)**	0.749 (0.190)	0.567 (0.173)	0.114 (0.366)
Bad	4,229	0.398 (0.196)	0.363 (0.170)	0.349 (0.184)	0.336 (0.187)	**0.425 (0.179)**	0.517 (0.280)	0.395 (0.201)	0.166 (0.584)
avg. p-value	-	1.39e-15	5.44e-26	8.27e-28	2.99e-30	**1e-100**	4.121e-07	1.014e-14	0.013
Excellent	2,119	0.791 (0.166)	0.623 (0.141)	0.628 (0.134)	0.623 (0.143)	**0.668 (0.147)**	0.824 (0.145)	0.790 (0.157)	-
Good	11,305	0.752 (0.155)	0.587 (0.135)	0.592 (0.130)	0.584 (0.137)	**0.612 (0.141)**	0.796 (0.145)	0.756 (0.156)	-
Fair	11,146	0.715 (0.136)	0.561 (0.145)	0.565 (0.139)	0.558 (0.146)	**0.584 (0.147)**	0.769 (0.175)	0.707 (0.141)	-
Bad	7,849	0.635 (0.199)	0.383 (0.211)	0.387 (0.209)	0.382 (0.212)	**0.410 (0.208)**	0.509 (0.311)	0.602 (0.208)	-
avg. p-value	-	4.99e-26	1.137e-27	4.0423e-32	1.196e-30	**2.926e-40**	2.038e-16	9.388e-22	-

Out-of-dictionary

NDCG@K Values



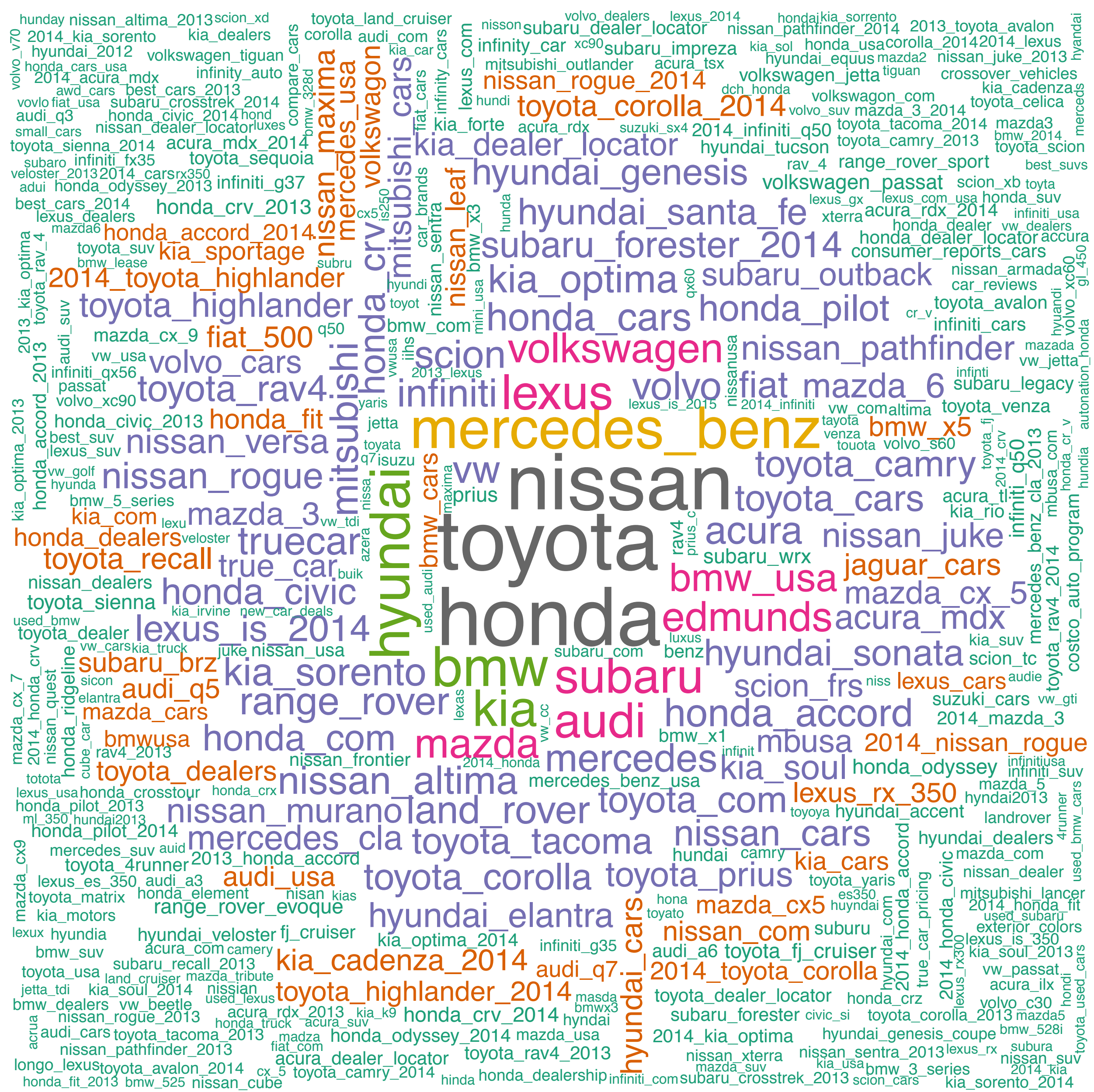
Experiments on TREC Data

Method	Editorial grade	Levenshtein dist.
$\text{QFG}_{\text{ad+link}}$	1.0441	11.70
$\text{word2vec}_{\text{news}}$	0.9189	10.91
$\text{word2vec}_{\text{search}}$	0.9492	11.32
cn2vec	0.9571	11.37
cx2vec	1.1273	13.79
cx-cn2vec	1.1343	13.13
$\text{cx-cn2vec}_{\text{ad}}$	1.2281	13.62
$\text{cx-cn2vec}_{\text{ad+link}}$	**1.2457**	13.25

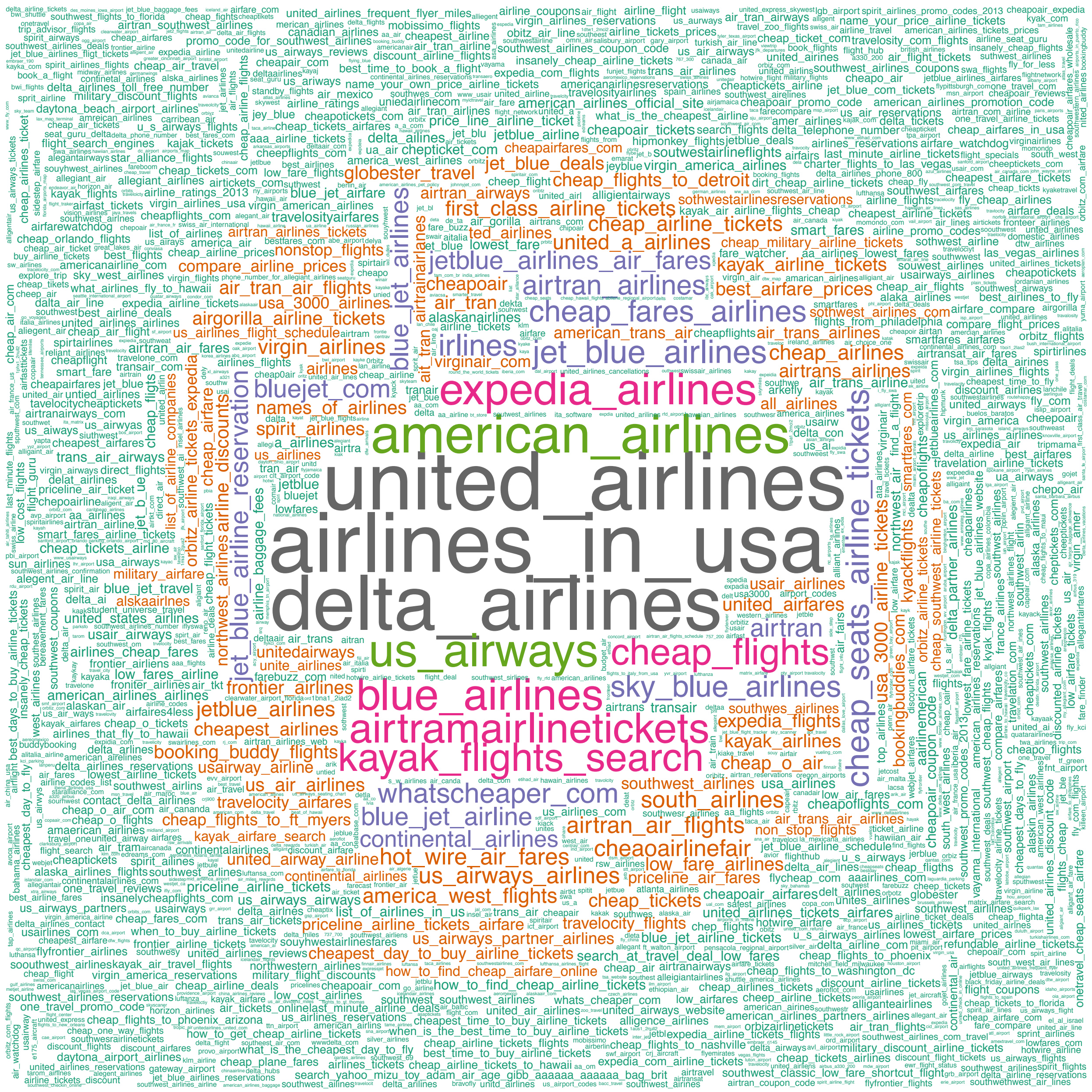
Coverage and Estimated Improvements

	In-house data			TREC data		
Method	**Coverage**	**Revenue potential**	**eCPM**	**Coverage**	**Revenue potential**	**eCPM**
$QFG_{ad+link}$	1.00	1.00	1.00	1.00	1.00	1.00
$word2vec_{news}$	0.76	0.39	0.46	0.32	0.66	0.73
$word2vec_{search}$	0.87	0.58	0.77	0.57	0.84	0.75
cn2vec	0.89	0.62	0.84	0.59	0.84	0.74
cx2vec	1.16	1.80	1.41	1.41	1.16	1.20
cx-cn2vec	1.18	1.86	1.38	1.44	1.21	1.19
$cx\text{-}cn2vec_{ad}$	**1.20**	**1.89**	**1.60**	**1.52**	**1.35**	**1.31**
$cx\text{-}cn2vec_{ad+link}$	1.18	1.88	1.45	1.50	1.28	1.22

Queries similar to ads (ads category = ‘cars’)



Queries similar to ads (ads category = 'air travel')



Conclusions and Future Work

- We presented the first query rewriting mechanism using word embeddings
  - We evaluated the proposed methods using both in-house and publicly available TREC data sets
  - When compared to the current state-of-the-art approaches, we showed that context-content2vec generates the most relevant query rewrites, while at the same time maintains high level of ad coverage
- Future:
  - Better sessionization algorithms
  - Navigational query detection to bias our method less towards generating those queries


Gracias