

# Non-Local Evidence for Expert Finding

**Krisztian Balog** and Maarten de Rijke



ISLA, University of Amsterdam  
<http://ilps.science.uva.nl>

# Non-Local Evidence for Expert Finding

- Task
  - Find the right person with the appropriate skills and knowledge
  - Given a topic, rank expert *candidates*

# Non-Local Evidence for Expert Finding

- Existing approaches to expert finding
- Compute associations between candidates and topics, based on their co-occurrence in
  - documents
  - text-snippets

# Non-Local Evidence for Expert Finding

- Our aim
  - Identify and integrate non-local sources of evidence into existing expert finding models
  - Evidence that is not available from an individual page or text snippet

# Outline

- Retrieval model
- Experimental setting
- Identifying and integrating non-local evidence
- Results
- Conclusions

# Retrieval Model

- The problem of experts finding is stated as:
  - What is the probability of a candidate  $ca$  being an expert given the query topic  $q$ ?

$$p(ca|q) = \frac{p(q|ca) \cdot p(ca)}{p(q)}$$

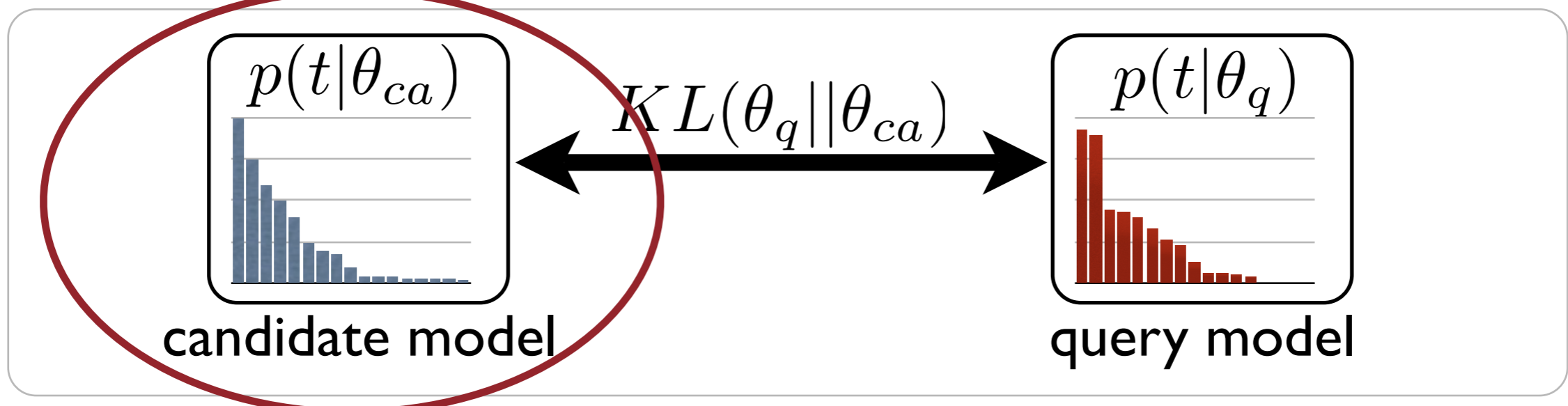
$$p(ca|q) \propto \underbrace{p(q|ca)} \cdot \underbrace{p(ca)}$$

How likely the candidate would produce the query

The *a priori* belief that the candidate is an expert

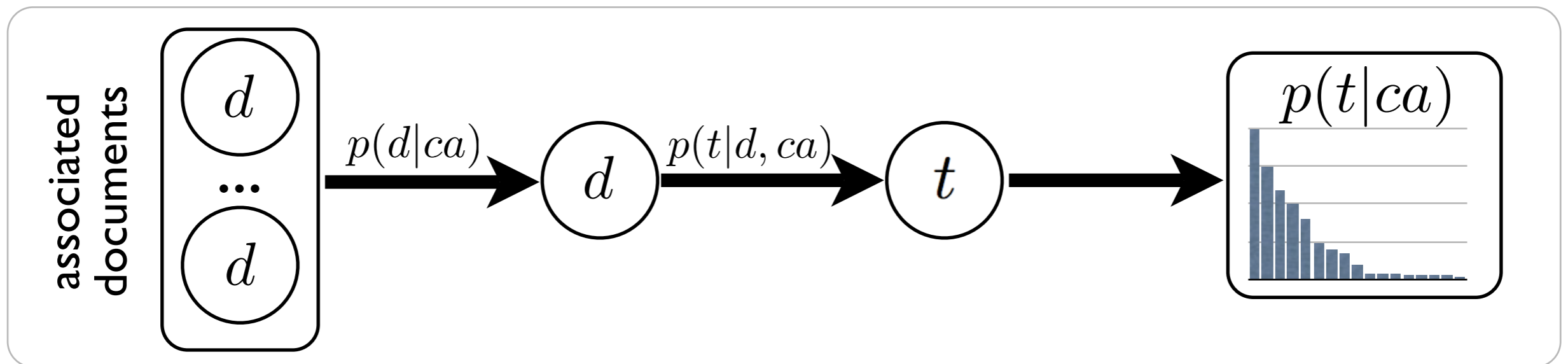
# Retrieval Model (2)

- How likely the candidate would produce the query?  $p(q|ca)$
- Generative language modeling approach
  - Both the candidate and the query are represented as a multinomial probability distribution over terms



# Candidate model

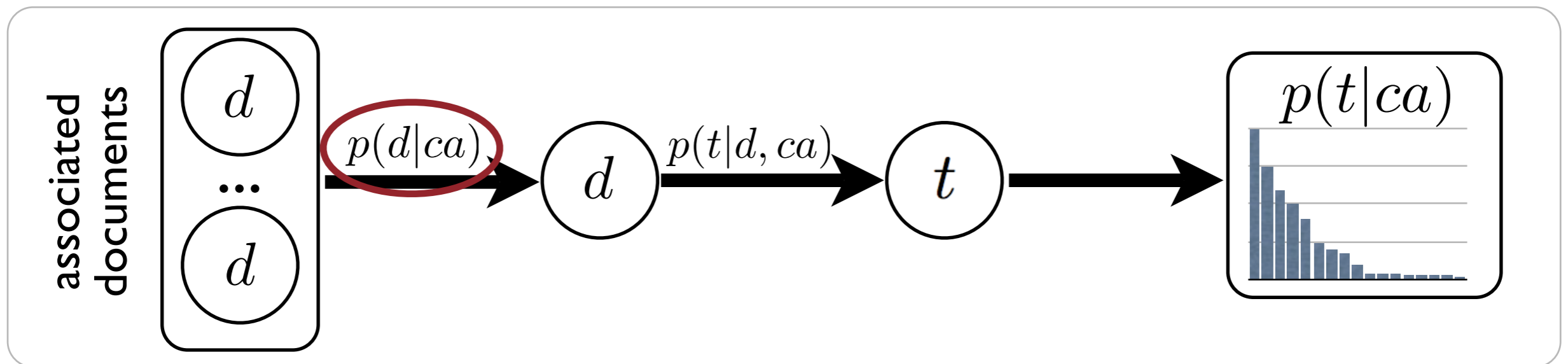
$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t)$$





# Candidate model

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t)$$



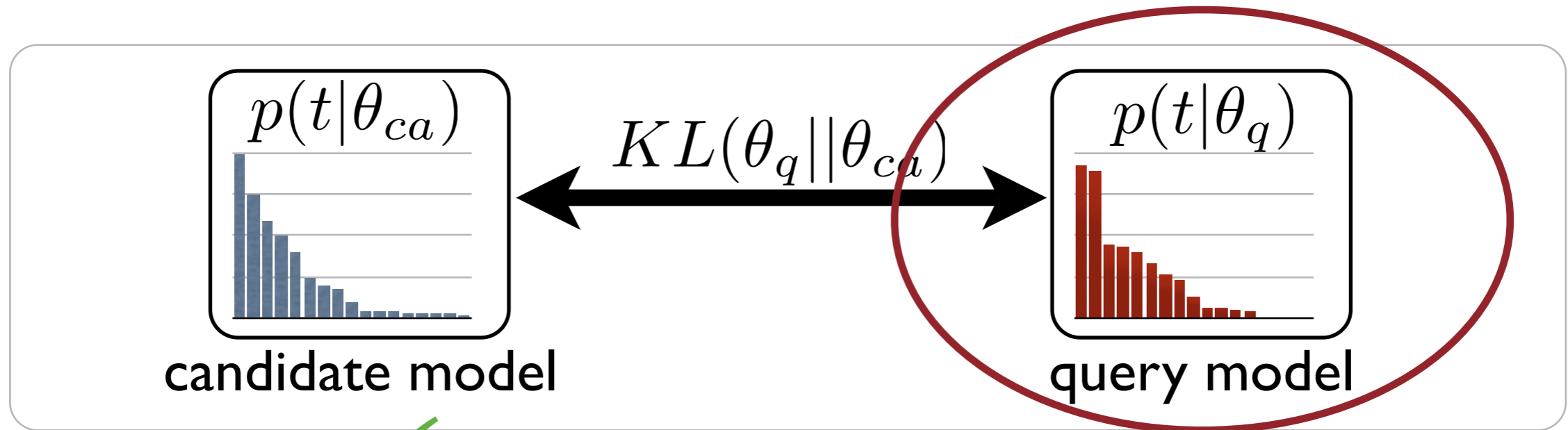
Document-candidate associations:  
*Boolean model*

$$p(d|ca) = \begin{cases} 1, & \underline{n(ca, d)} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The number of times  $ca$  is recognized in document  $d$



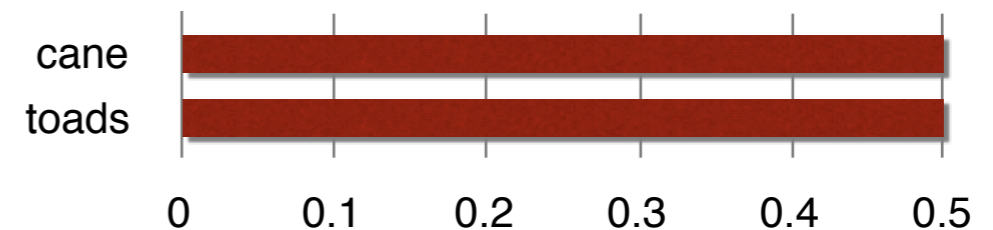
# Query Model



*Baseline query model (BL)*

Probability mass assigned uniformly across query terms

Example query: *cane toads*



# Outline

- Retrieval model
- Experimental setting
- Identifying and integrating non-local evidence
- Results
- Conclusions

# Experimental Setting

- TREC 2007 Enterprise Track
  - Document collection: web crawl of CSIRO (~370.000 docs, 4.2 GB)
  - 50 topics
  - Candidate identification
    - No canonical list is given in advance
    - E-mail addresses follow Firstname.Lastname@csiro.au format
    - Occurrences are replaced with a unique id

# Setting the Baseline

- Boolean document-candidate associations
  - All candidates mentioned in the document are equally important, and vice versa
- Baseline query
  - All query terms are equally important
- Uniform priors
  - All candidates are equally likely to be experts

# Non-Local Evidence

- Document-candidate associations
- Query model
- Candidate priors

# Document-candidate Associations

- Importance of a candidate given a document  $p(d|ca)$ 
  - So far: all candidates are equally important
  - Estimate the strength of the association based on
    - How many times the candidate is mentioned in the document
    - How many other documents the candidate is related to



# Document-candidate Associations (2)

- Lean document representation
  - Document contains only candidate mentions
- Use a term weighting scheme that combines the candidate's (local) frequency in the document and its global frequency

$$p(d|ca) \propto TF.IDF(d, ca)$$

# Document-candidate Associations (3)

	$ca_1$	...	$ca_i$	...	$ca_n$
$d_1$					
...					
$d_j$					
...					
$d_m$					

- Weight of the candidate in the document is computed in two ways

1) Number of occurrences

$$n(ca, d)$$

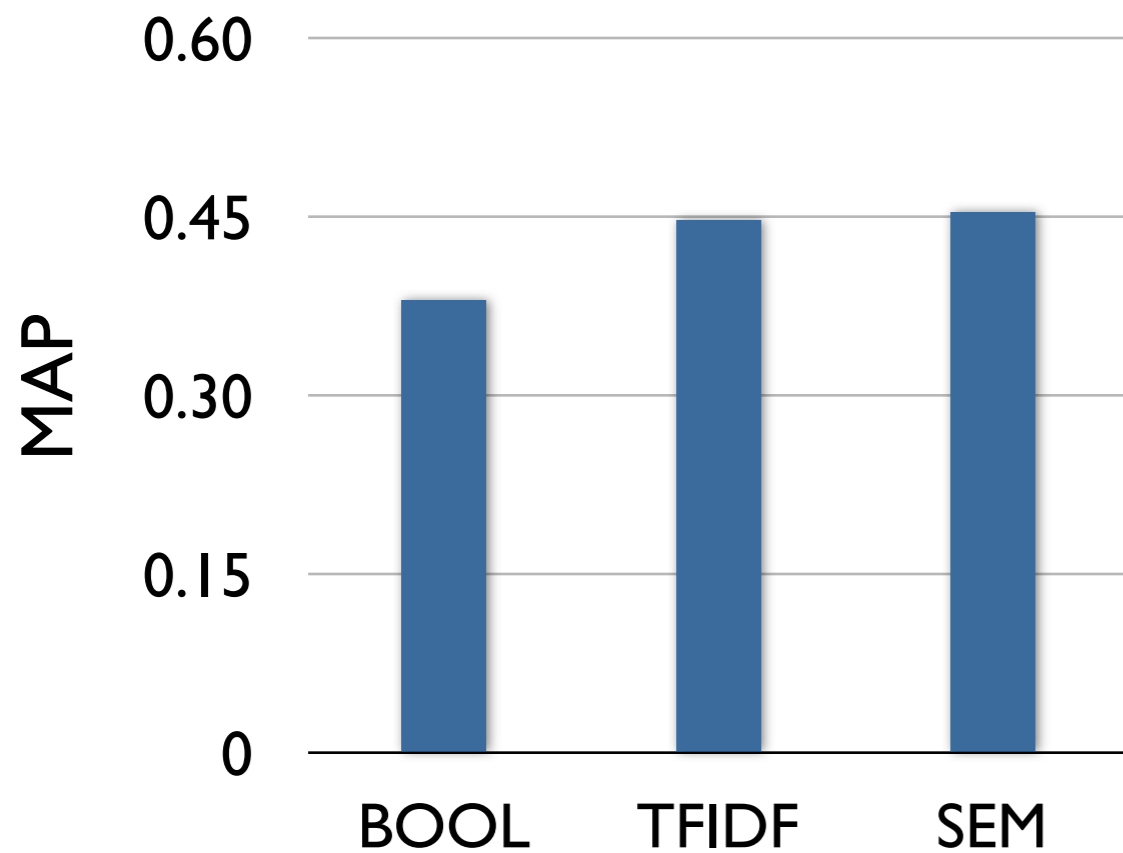
TFIDF

2) Semantic relatedness

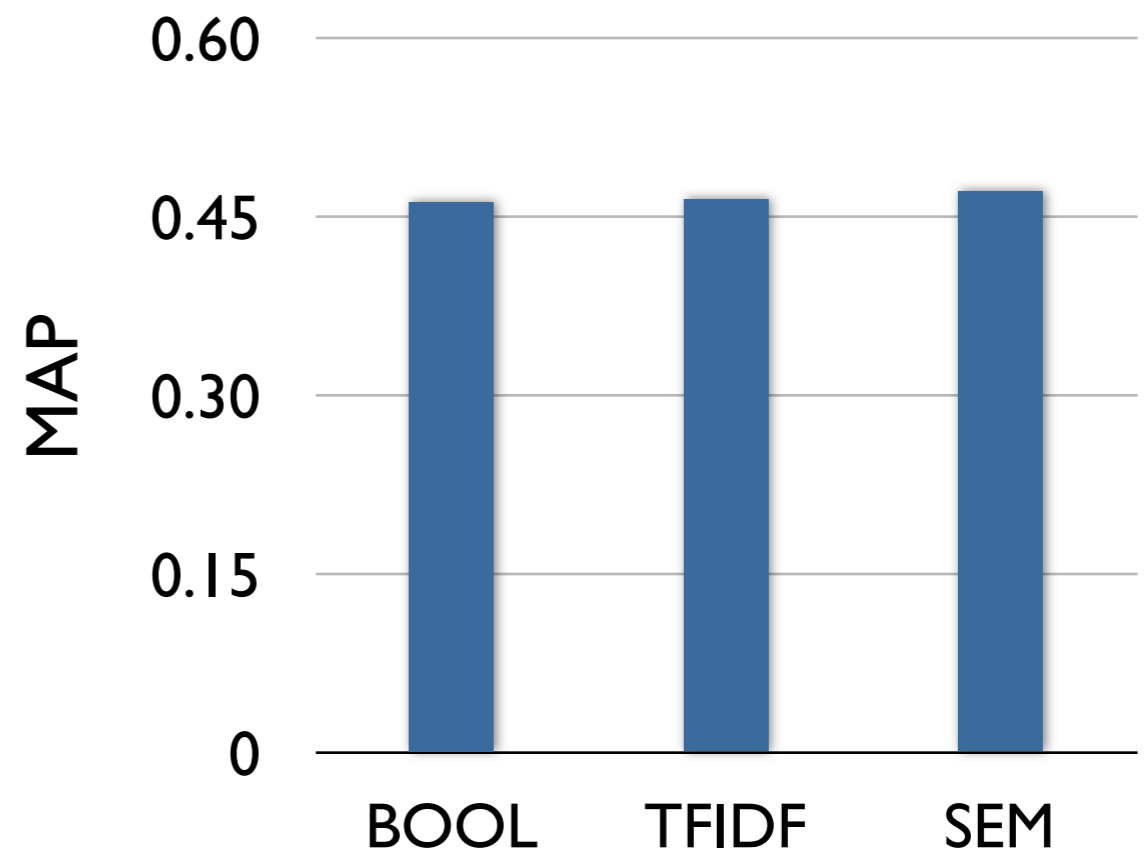
SEM

$$n'(ca, d) = \begin{cases} \text{KL}(\theta_{ca} || \theta_d), & n(ca, d) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

# Results



Document-based model



Proximity-based model

# Query Models

- TREC 2007 Enterprise track simulates a type of click-based system
- A few examples of key documents are provided with the topic description
- [Balog et. al., 2008] propose an effective method for constructing a query model by sampling terms from example documents

# Example Topic

<top>

<num>CE-039</num>

<query>**cane toads**</query>

<narr>

Cane toads were introduced into Australia in a failed bid to control Australian native beetles. [...] Resources describing cane toads, invasive species, pest management, biological control would all be relevant to the topic.

</narr>

<page>CSIRO141-14983789</page>

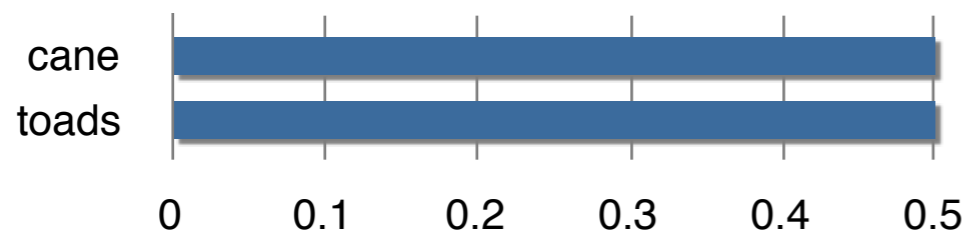
<page>CSIRO139-09015831</page>

<page>CSIRO134-11651748</page>

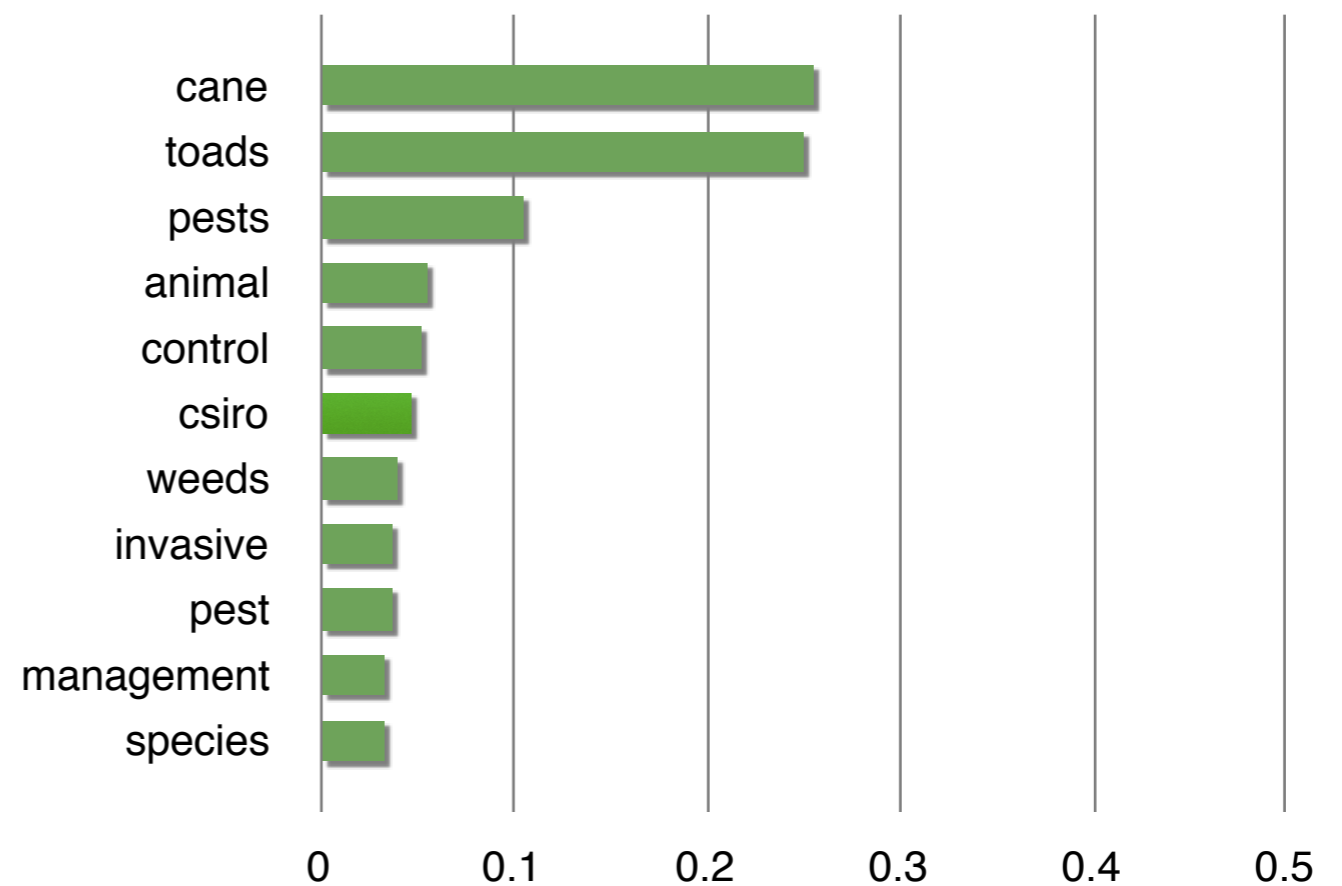
</top>

# Example Query Model

Baseline (BL)



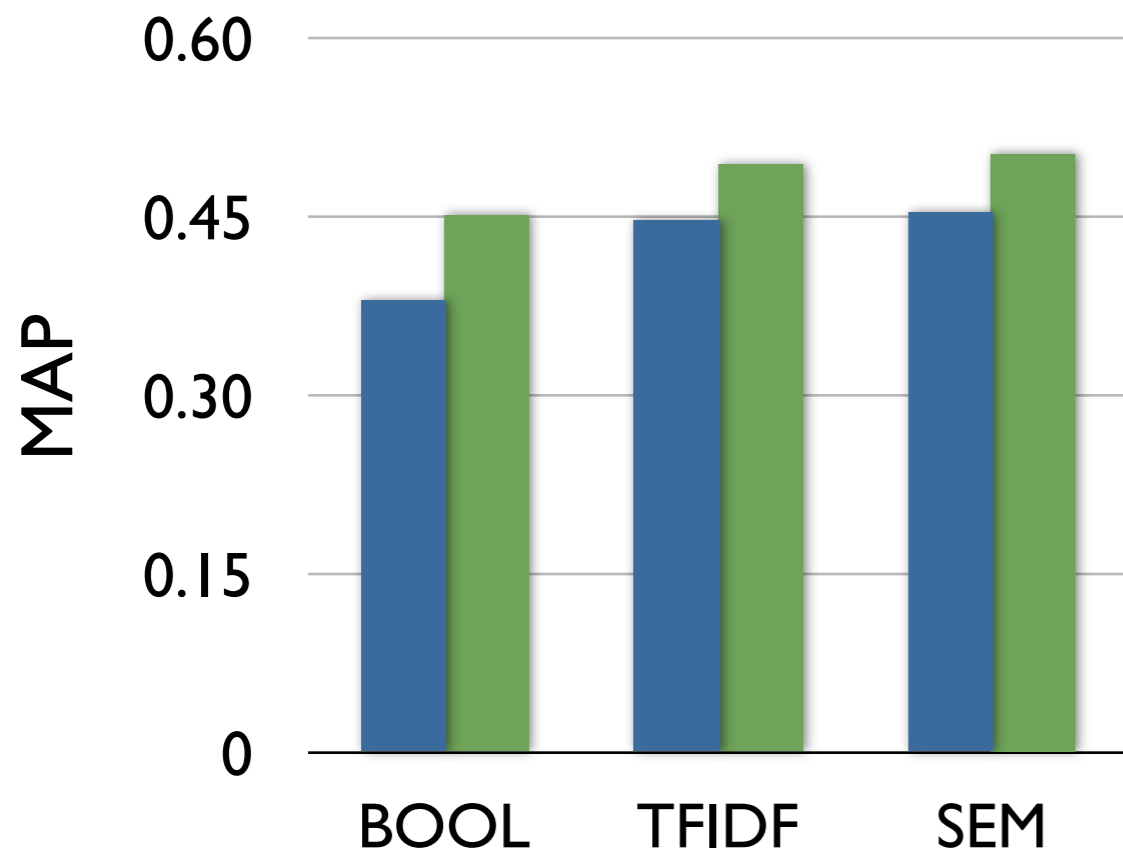
Example docs (EX)



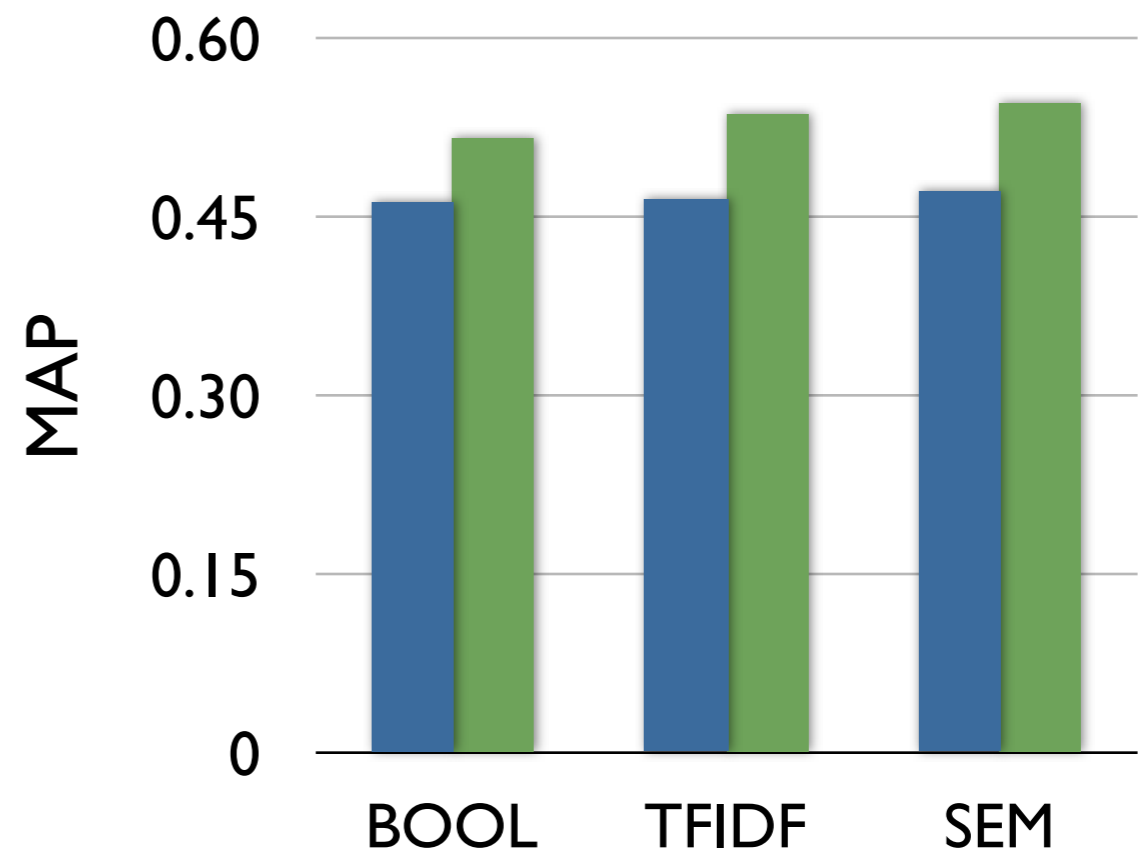
# Results

■ BL query model

■ EX query model



Document-based model



Proximity-based model

# Candidate Priors

**CSIRO**

About CSIRO | Doing Business | Media | Events | Explore & Educate | Publications | Careers | CSIRO Shop | Contact Us

**FOOD**

Home - Farming & Food - Food

## Overview

### Food Science Australia

Food Science Australia researchers work across all stages of the food production chain; from providing quality ingredients and processing innovations to providing objective consumer insights to organisations that market food products.

- ~ [Consumer and market access research](#)
- ~ [Nutrition and healthy foods](#)
- ~ [Product delivery](#)
- ~ [Production and assembly](#)
- ~ [Facilities](#)

Food Science Australia is a joint venture of CSIRO and the Victorian Government.

It assists in every stage of the food processing business system, from concept through formulation and processing, pilot production, packaging, shelf-life assessment and scale-up to full production.

#### CONSUMER AND MARKET ACCESS RESEARCH

Consumer and market access research includes:

- conducting sensory perception and consumer preference assessments
- assessing risk and product compliance of product claim, and chemical and microbiological contamination
- assessing nutritional properties of food ingredients and products
- providing advice on market entry and regulatory matters.

#### NUTRITION AND HEALTHY FOODS

Our research into nutrition and healthy foods includes:

- food science to promote health and vitality
- substantiation of product health claims

#### FAST FACTS

- ▲ Food Science Australia is a joint venture of CSIRO and the Victorian Government
- ▲ Food Science Australia assists the food industry
- ▲ It has expertise from concept through to formulation and processing, pilot production, packaging, shelf life assessment and scale-up to full production

#### PRIMARY CONTACT

**Mr John Smith**  
Communication Manager  
Food Science Australia  
Phone: 61 8 8303 8857  
Fax: 61 8 8303 8837  
Email: [John.M.Smith@csiro.au](mailto:John.M.Smith@csiro.au)

#### LOCATIONS

**Food Science Australia - North Ryde**  
Riverside Corporate Park  
11 Julius Avenue  
North Ryde NSW 2113  
Australia  
PO Box 52  
North Ryde NSW 1670  
Australia

**Food Science Australia - Werribee**



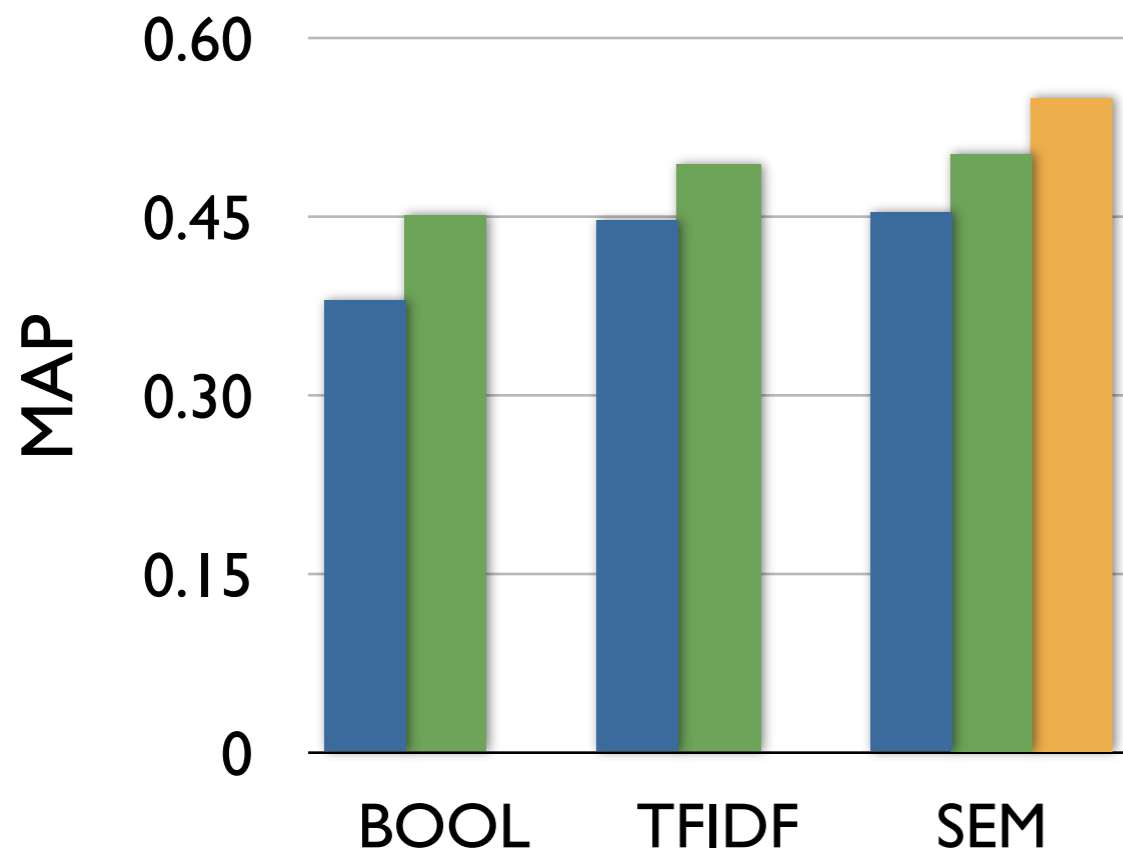
# Candidate priors

- Encodes organizational knowledge
- Extracted names and positions from contact boxes
- Filtering out science communicators (SC) based on position information
  - communication officer/manager/advisor
  - manager public communications

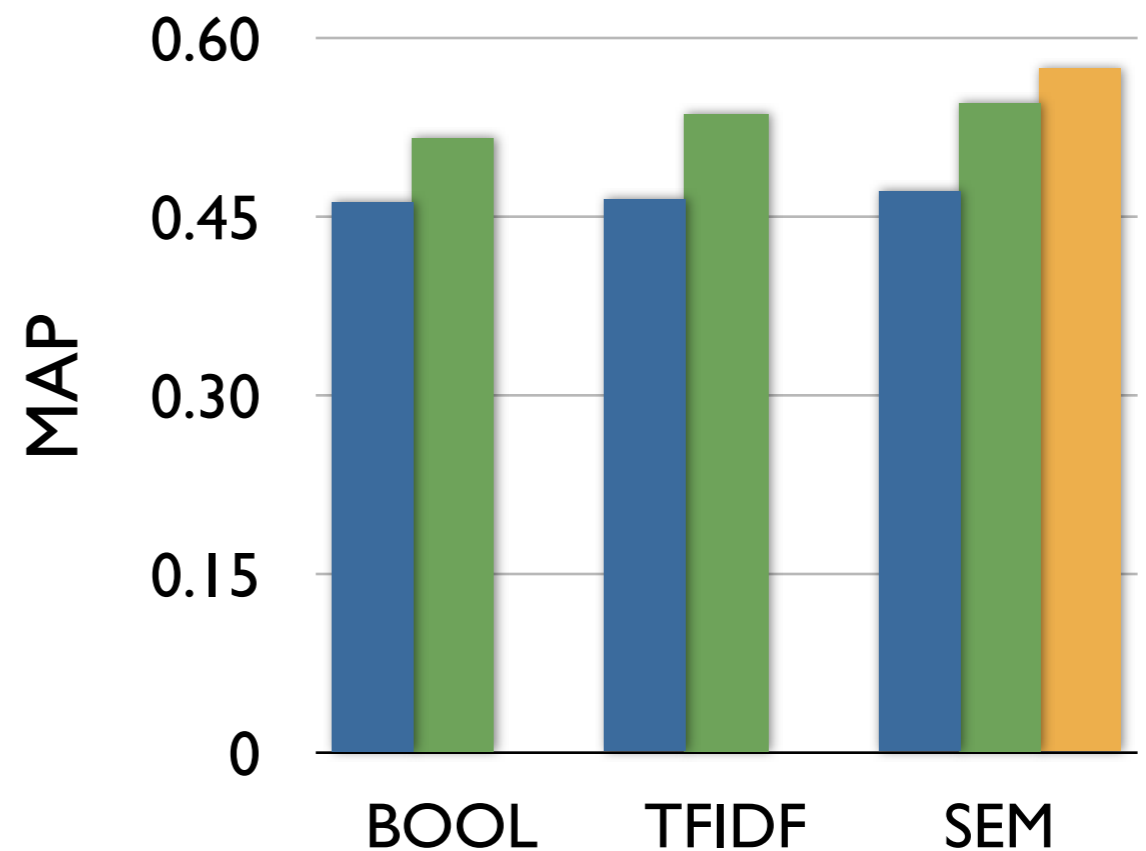
$$p(ca) = \begin{cases} 1, & ca \notin SC, \\ 0, & ca \in SC. \end{cases}$$

# Results

■ BL query model    ■ EX query model    ■ SC prior



Document-based model



Proximity-based model

# How good is it?

Method	Run type	MAP
TREC 2007 best	automatic	0.4632
TREC 2007 best	feedback	0.3660
TREC 2007 best	manual	0.4787
Voting model [1]	automatic	0.3519
Relevance prop. [2]	automatic	0.4319
Baselines in this paper		
Document-based model	automatic	0.3801
Proximity-based model	automatic	0.4633

[1] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *ECIR 2008*.

[2] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling relevance propagation for the expert search task. In *TREC 2007*.

# How good is it? (2)

Method	Run type	MAP
<b>Document-based model</b>		
Baseline	automatic	0.3801
Document-cand. assoc.	automatic	0.4541
Query model	feedback	0.5044
Candidate priors	feedback	0.5506
<b>Proximity-based model</b>		
Baseline	automatic	0.4633
Document-cand. assoc.	automatic	0.4735
Query model	feedback	0.5465
Candidate priors	feedback	0.5747


# Conclusions

- Identified a number of non-local sources of evidence for expert finding
- Complemented existing document and proximity-based approaches to incorporate non-local evidence
- Showed significant improvements over a very competitive baseline
- Outperformed existing state-of-the-art

# Further Work

- Non-local evidence within documents
  - Recognize and exploit internal document structure

# Future Work (2)



About CSIRO | Doing Business | Media | Events | Explore & Educate | Publications | Careers | CSIRO Shop | Contact Us

WATCH & LISTEN

Home - Showcases - Watch & Listen

Search

Search all CSIRO  
 Search this site

**Showcases** ▾  
**Flagships** ▾  
**Divisions** ▾

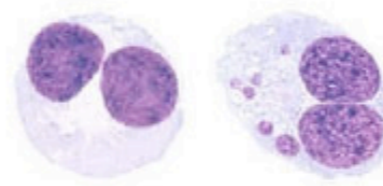
Astronomy & Space ▾  
Energy ▾  
Environment ▾  
Farming & Food ▾  
Health & Wellbeing ▾  
Information & Communication Technology ▾  
Manufacturing ▾  
Materials ▾  
Mining & Minerals ▾  
Transport & Infrastructure ▾

[Subscription Information](#)

## Video

### DNA Doctor: Catalyst, ABC interview

In this video CSIRO's Dr Michael Fenech says that damage to the genome is a fundamental disease that can be diagnosed and treated. (8:00)



The cell on the left is normal but the one on the right shows signs of genetic damage. The damaged DNA appears as six micronuclei in the cell.

[Low bandwidth Streaming video](#)  
[Broadband Streaming video](#)

[Download Windows Media Player](#)

[View Transcript](#) 22 KB

Dr Michael Fenech says we should consider damage to the genome as a fundamental disease that can be diagnosed and treated.

In this video ABC Reporter for the television program *Catalyst*, Mr Paul Willis, acts a guinea pig to test Dr Fenech's theories.

The video also features an interview with Professor Bruce Armstrong at the University of Sydney, Sydney, NSW, Australia, about the likelihood of being able to repair our genomes.

CSIRO has completed negotiations with a private company to make the genome health analysis test described in this *Catalyst* interview available to the general public on a commercial basis together with advice on dietary patterns and/or supplements that may assist in prevention of DNA damage.

The launch of the Reach 100 clinic in early July 2007, highlighted the role of preventative health and dietary methods of reducing cancer risk factors.

Learn more about CSIRO's work [Preventing diseases and detecting them sooner](#).

#### PRIMARY CONTACT

**Dr Michael Fenech**  
Theme Director - Food and Nutrition  
Food Science Australia  
Phone: 61 8 8303 8880  
Alt Phone: 61 8 8303 8800  
Fax: 61 8 8303 8899  
Email: [Michael.Fenech@csiro.au](mailto:Michael.Fenech@csiro.au)

#### EDITOR'S CHOICE

- ▴ [Dr Michael Fenech: keeping our genes healthy](#)
- ▴ [The genome health and nutrigenomics project](#)
- ▴ [Welcome to the world of personalised nutrition \(Media release 9 Dec 05\)](#)

# Non-Local Evidence for Expert Finding

[K.Balog@uva.nl](mailto:K.Balog@uva.nl)

<http://www.science.uva.nl/~kbalog>