Centre for
Data Analytics

Insight

Saffron
Knowledge Extraction Framework

Topic extraction, expert finding and trend analysis from scientific literature

John P. McCrae

Insight Centre for Data Analytics

National University of Ireland Galway

Knowledge Extraction from Text

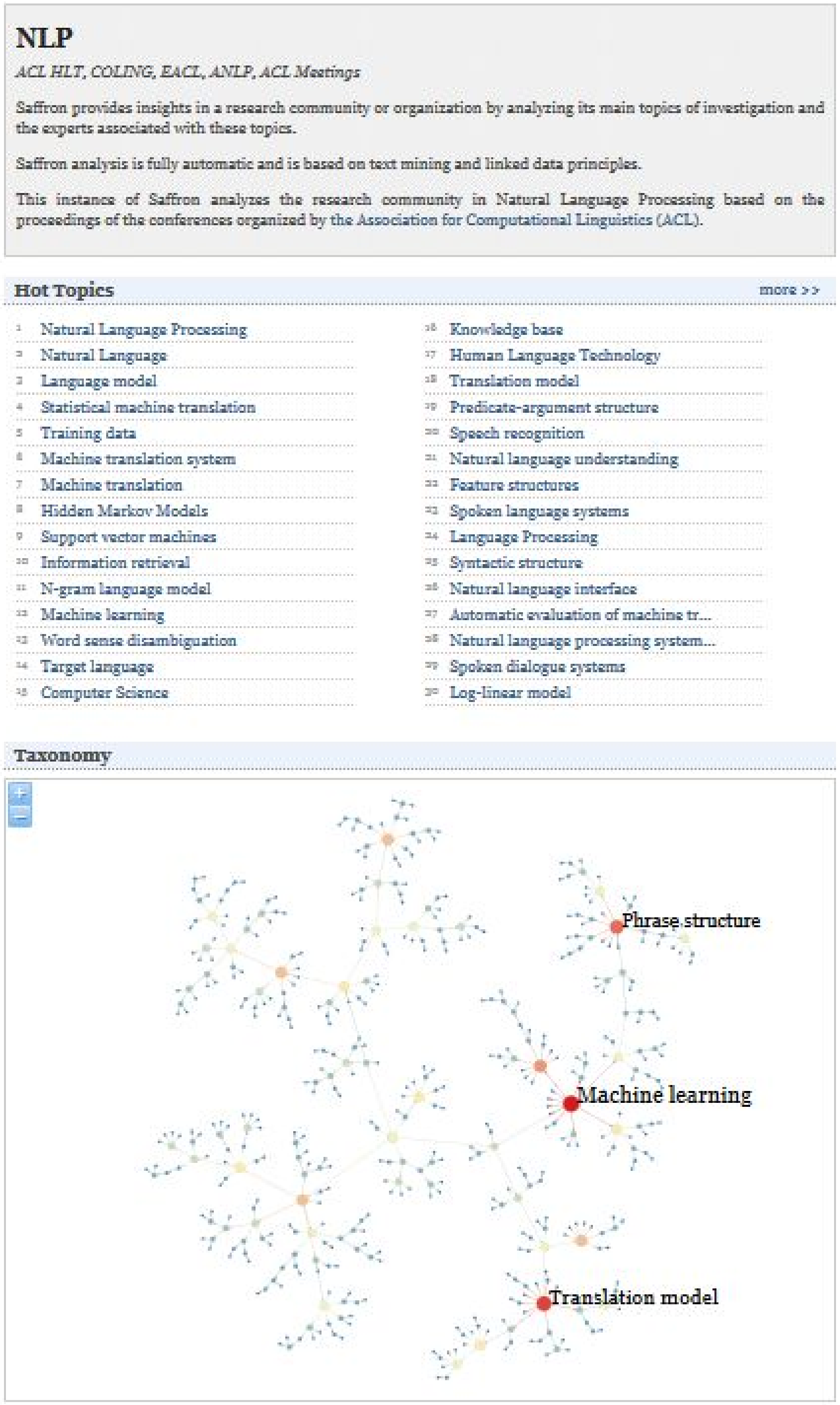
- with Saffron -

… act or process of **retrieving awareness** or understanding **of someone** or **something**, such as facts, **information**, **descriptions**, or **skills** out of text for further data processing

... usually followed by data transformation and possibly the **addition of metadata** prior to export to another stage in the data workflow …

Original Use Case: Expert Finding

ACL Anthology
A Digital Archive of Research Papers in Computational Linguistics



NLP

ACL HLT, COLING, EACL, ANLP, ACL Meetings

Saffron provides insights in a research community or organization by analyzing its main topics of investigation and the experts associated with these topics.

Saffron analysis is fully automatic and is based on text mining and linked data principles.

This instance of Saffron analyzes the research community in Natural Language Processing based on the proceedings of the conferences organized by the Association for Computational Linguistics (ACL).

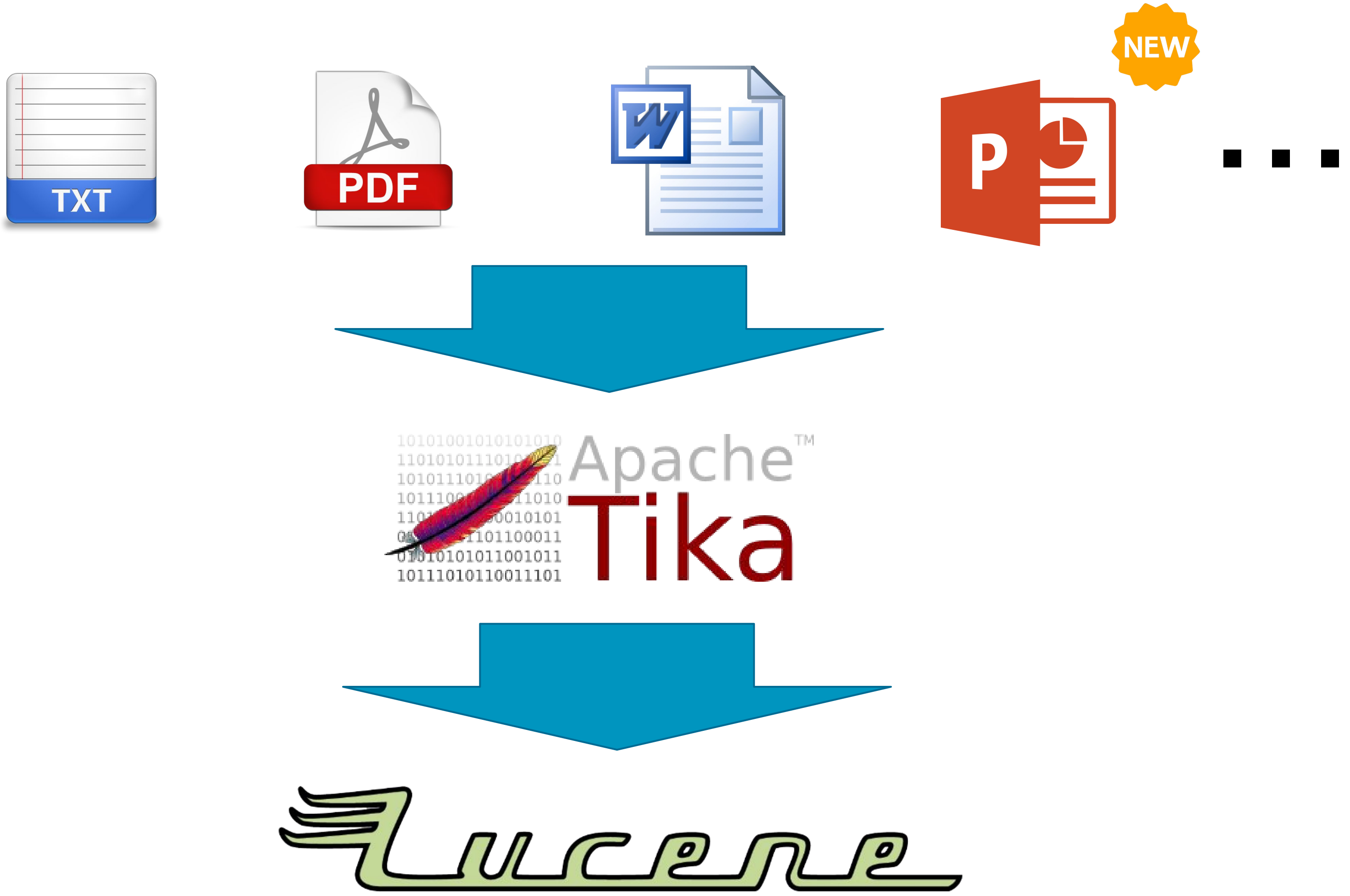
Hot Topics more >>

1. Natural Language Processing
2. Natural Language
3. Language model
4. Statistical machine translation
5. Training data
6. Machine translation system
7. Machine translation
8. Hidden Markov Models
9. Support vector machines
10. Information retrieval
11. N-gram language model
12. Machine learning
13. Word sense disambiguation
14. Target language
15. Computer Science
16. Knowledge base
17. Human Language Technology
18. Translation model
19. Predicate-argument structure
20. Speech recognition
21. Natural language understanding
22. Feature structures
23. Spoken language systems
24. Language Processing
25. Syntactic structure
26. Natural language interface
27. Automatic evaluation of machine tr...
28. Natural language processing system...
29. Spoken dialogue systems
30. Log-linear model

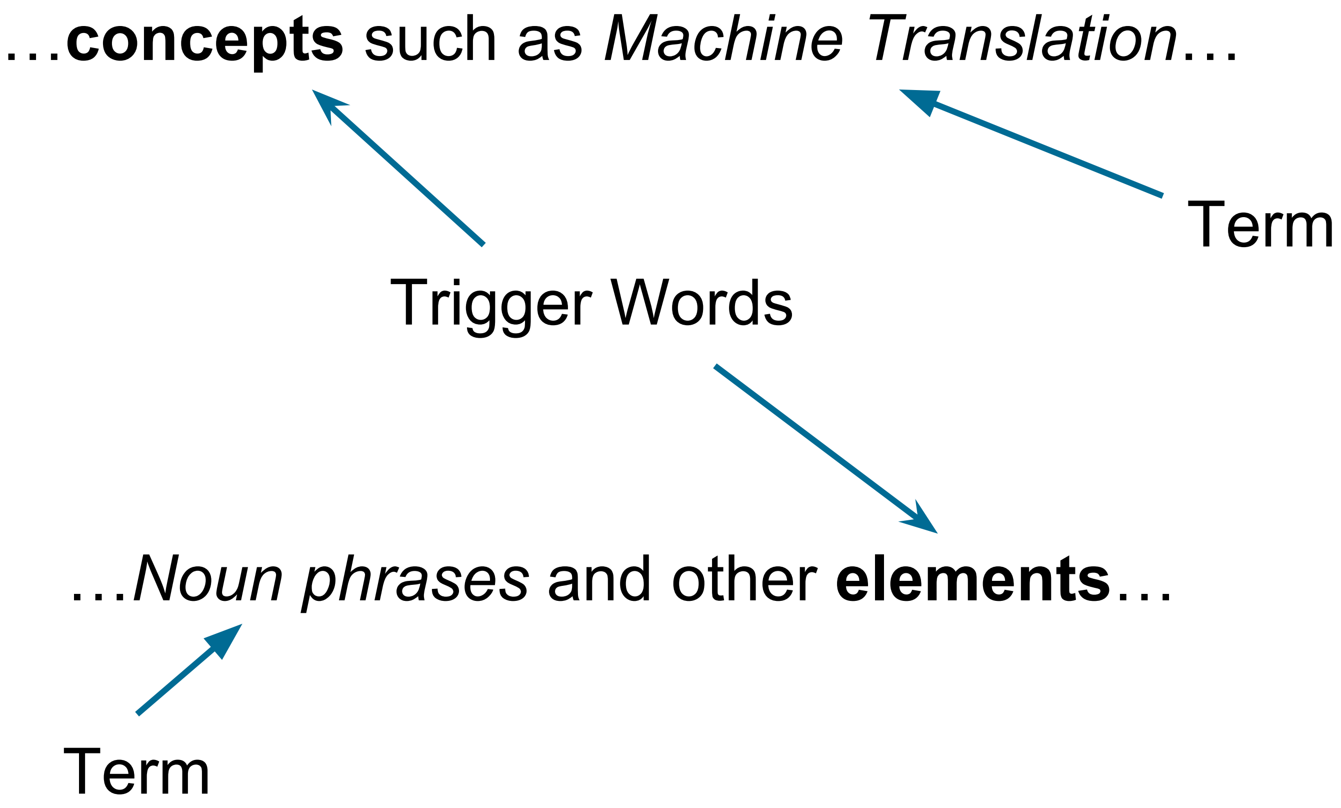
Taxonomy

Architecture

Step 1 - Corpus Indexing



Step 2 - Domain Modelling



…**concepts** such as *Machine Translation*…

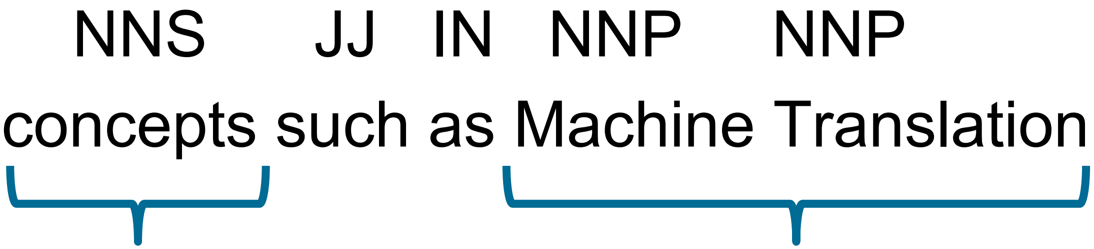
Term

Trigger Words

…*Noun phrases* and other **elements**…

Term

Step 3 - Topic (term) Extraction



NNS JJ IN NNP NNP

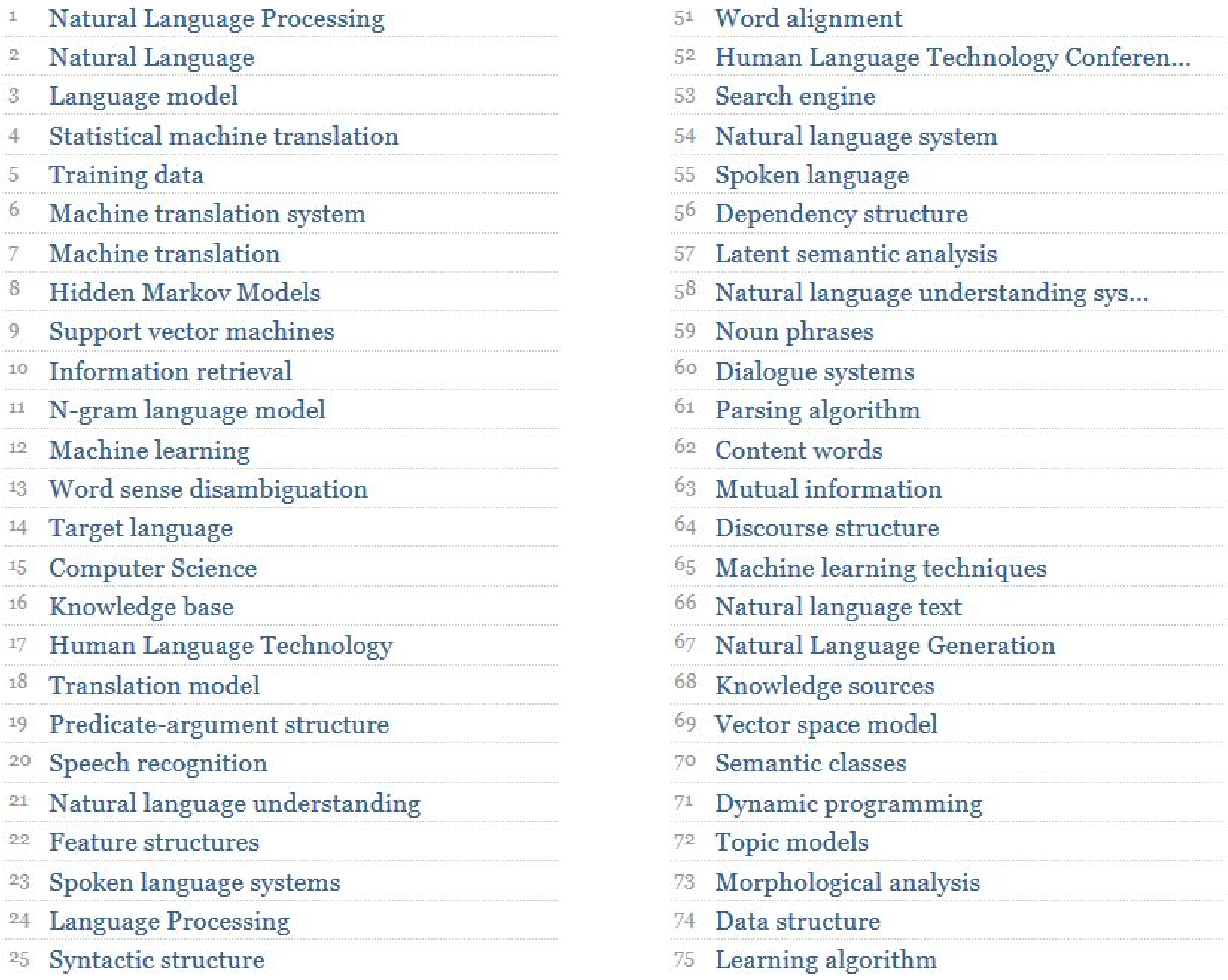
concepts such as Machine Translation

Candidate	Weirdness	Relevance	Domain Pertinence	...
Concepts	0.1	0.6	0.8	...
Machine Translation	0.8	0.7	0.7	...

NEW

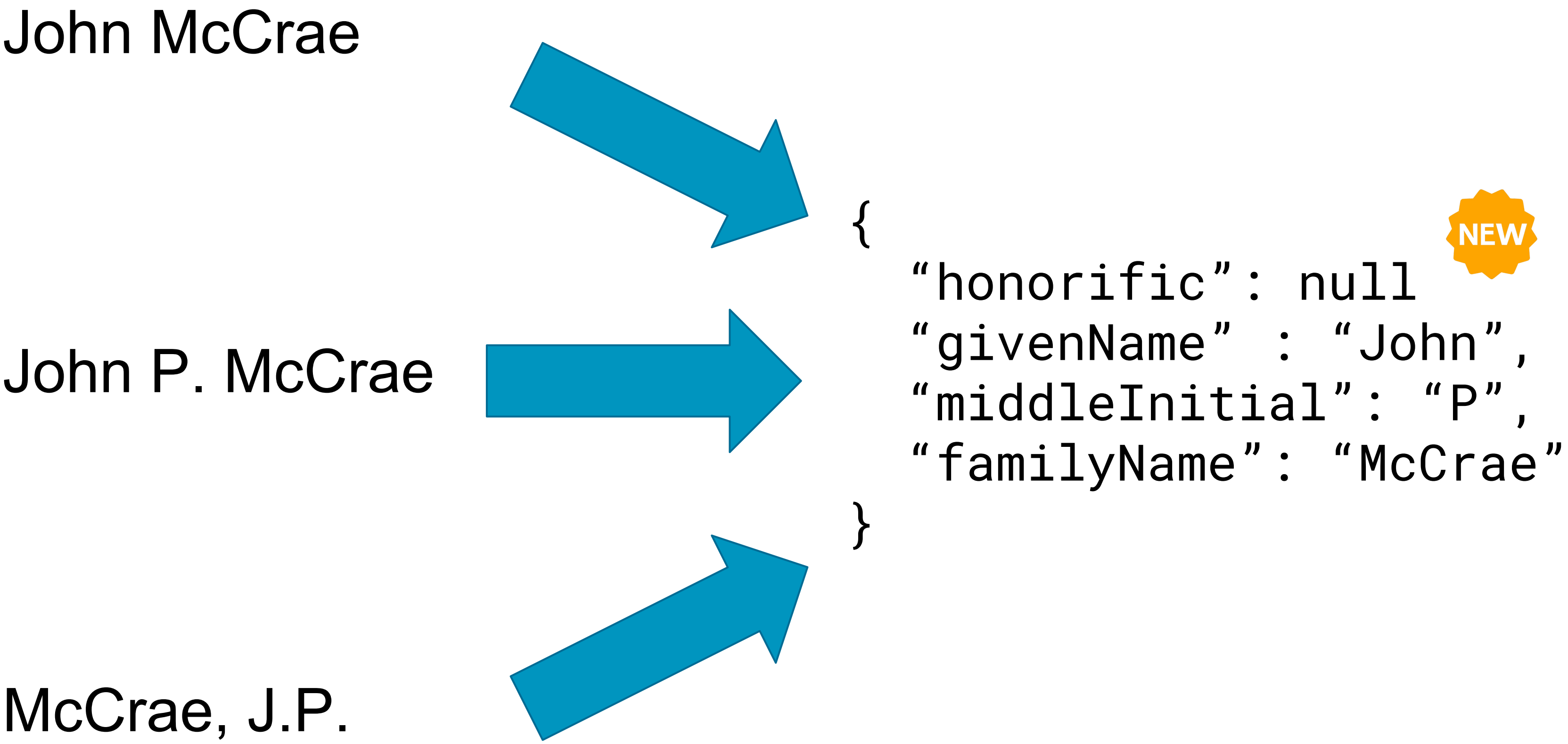
Candidate selection
by voting

Term Extraction – ACL Anthology



1 Natural Language Processing
2 Natural Language
3 Language model
4 Statistical machine translation
5 Training data
6 Machine translation system
7 Machine translation
8 Hidden Markov Models
9 Support vector machines
10 Information retrieval
11 N-gram language model
12 Machine learning
13 Word sense disambiguation
14 Target language
15 Computer Science
16 Knowledge base
17 Human Language Technology
18 Translation model
19 Predicate-argument structure
20 Speech recognition
21 Natural language understanding
22 Feature structures
23 Spoken language systems
24 Language Processing
25 Syntactic structure
51 Word alignment
52 Human Language Technology Conferen...
53 Search engine
54 Natural language system
55 Spoken language
56 Dependency structure
57 Latent semantic analysis
58 Natural language understanding sys...
59 Noun phrases
60 Dialogue systems
61 Parsing algorithm
62 Content words
63 Mutual information
64 Discourse structure
65 Machine learning techniques
66 Natural language text
67 Natural Language Generation
68 Knowledge sources
69 Vector space model
70 Semantic classes
71 Dynamic programming
72 Topic models
73 Morphological analysis
74 Data structure
75 Learning algorithm

Step 4 - Author Consolidation



John McCrae

John P. McCrae

McCrae, J.P.

NEW

```
{
   “honorific”: null
   “givenName” : “John”,
   “middleInitial”: “P”,
   “familyName”: “McCrae”
}
```

Step 5 - DBpedia Lookup

“Machine Translation”

http://dbpedia.org/resource/Machine_translation

Step 6 - Topic Statistics

Topic Generality

$$g(t) = \sum_{d \in \text{corpus}} \frac{PMI(t;d)}{p(t,d)}$$

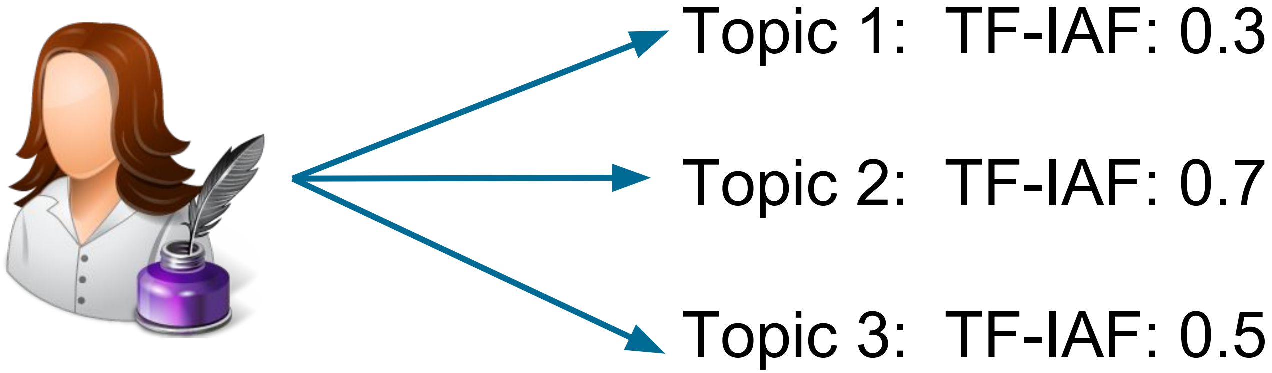
Weaknesses:

- Favours common terms
- Denormalized PMI?

⇒ Multi-factor metric NEW

Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Step 7 - Connect Authors



Topic 1: TF-IAF: 0.3

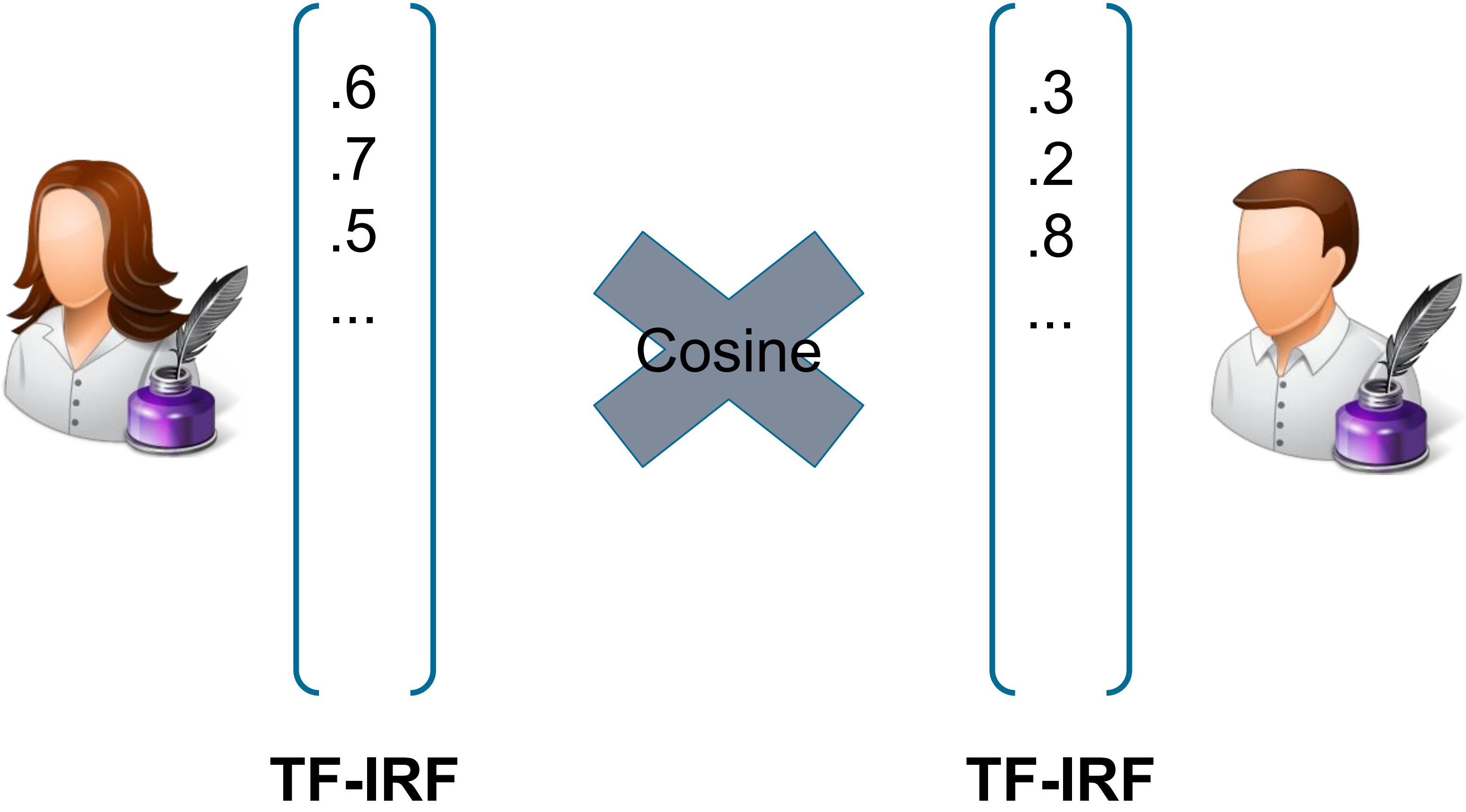
Topic 2: TF-IAF: 0.7

Topic 3: TF-IAF: 0.5

$$\text{TF-IAF}(T;r) = \sum_{Doc \text{ if } T\in\text{Doc},r\in\text{Authors}(\text{Doc})} \text{TF-IDF}(T)$$

Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Step 8 - Author Similarity



$$\begin{bmatrix} .6 \\ .7 \\ .5 \\ \dots \end{bmatrix}$$

Cosine

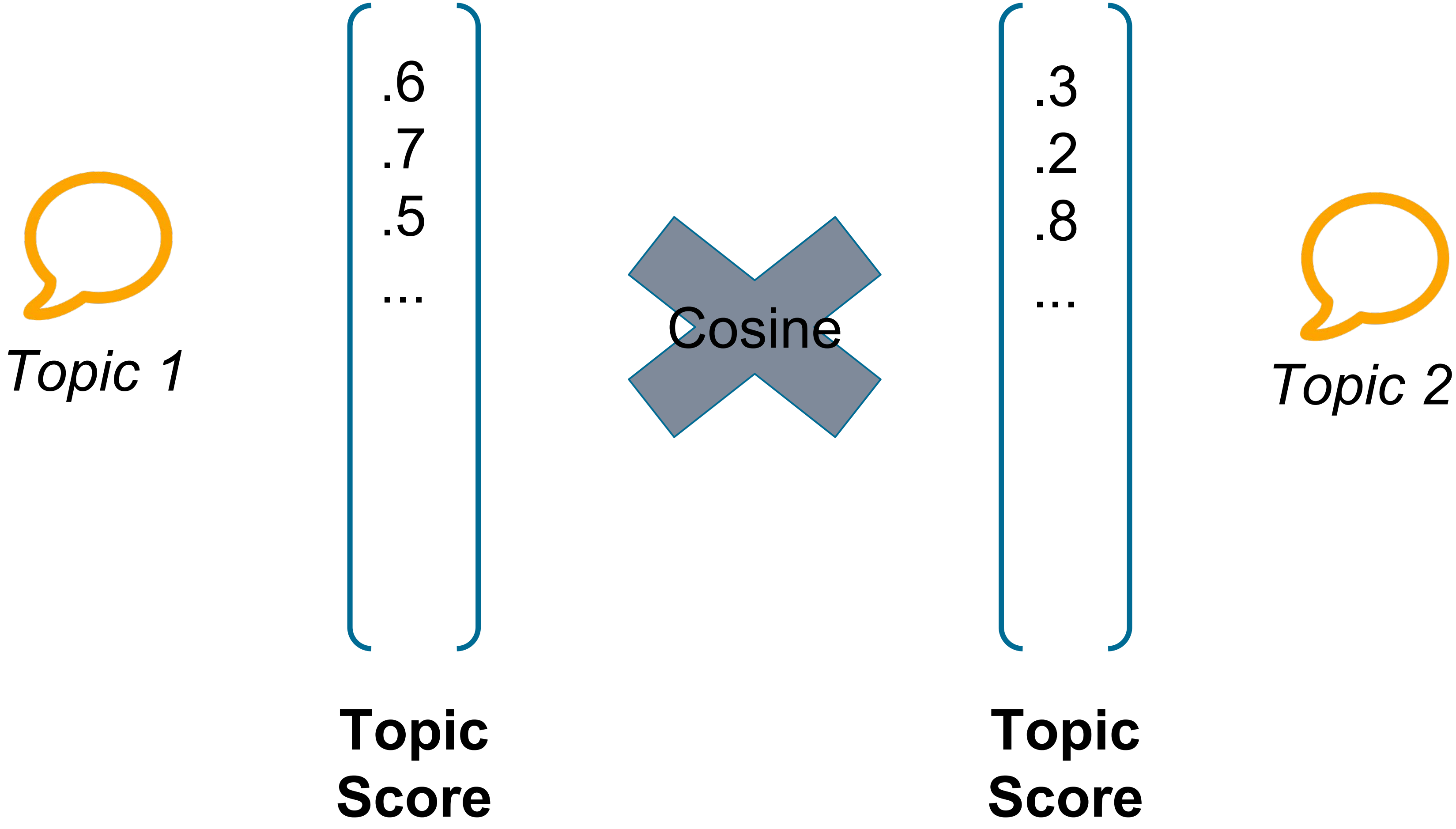
$$\begin{bmatrix} .3 \\ .2 \\ .8 \\ \dots \end{bmatrix}$$

TF-IRF

TF-IRF

Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Step 9 - Topic Similarity



Topic 1

.6
.7
.5
...

Topic Score

Cosine

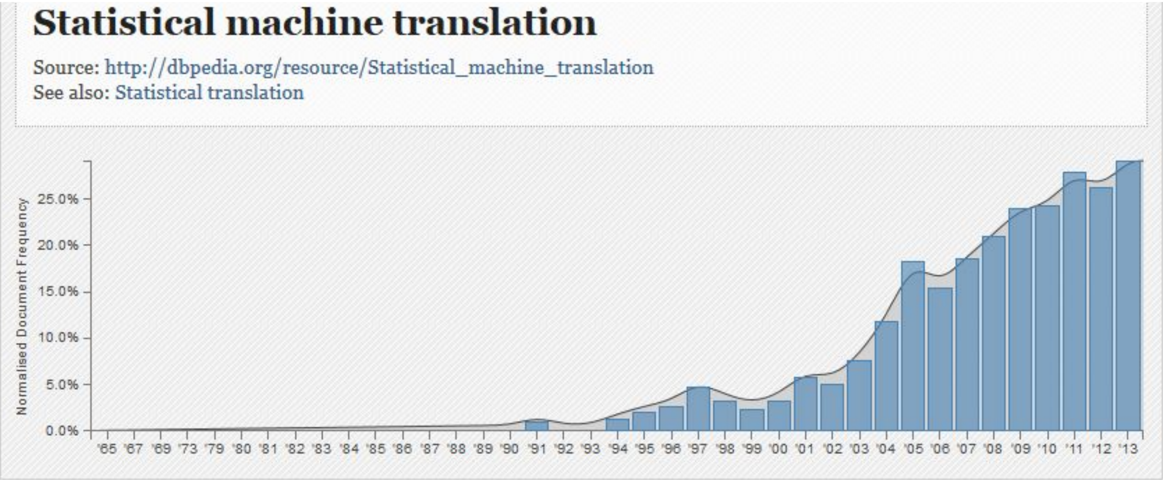
.3
.2
.8
...

Topic Score

Topic 2

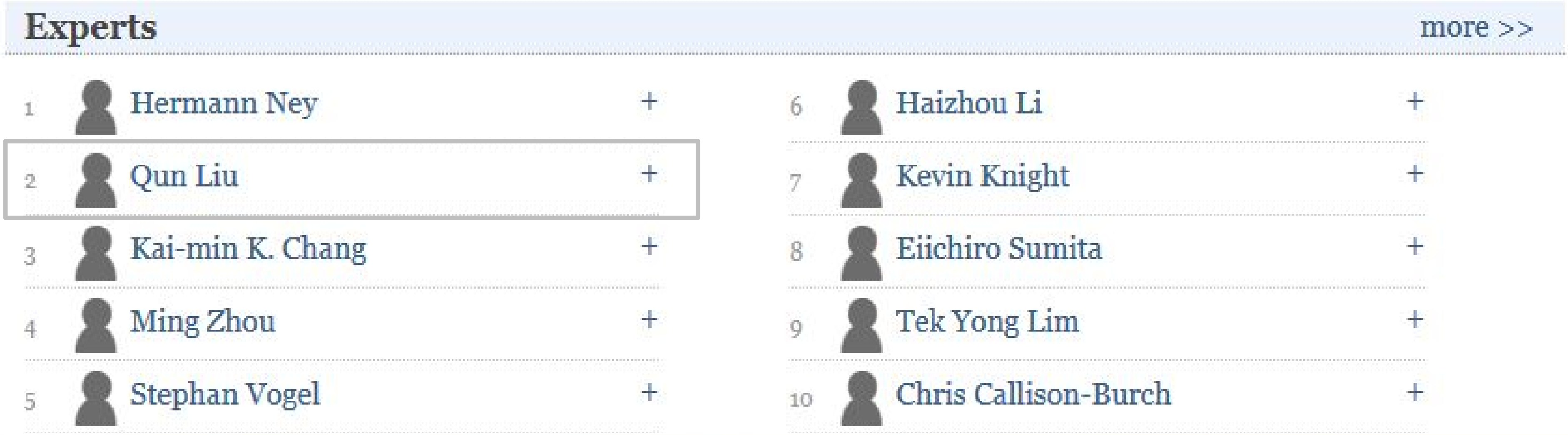
Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Expertise Mining



Statistical machine translation

Source: http://dbpedia.org/resource/Statistical_machine_translation
See also: Statistical translation

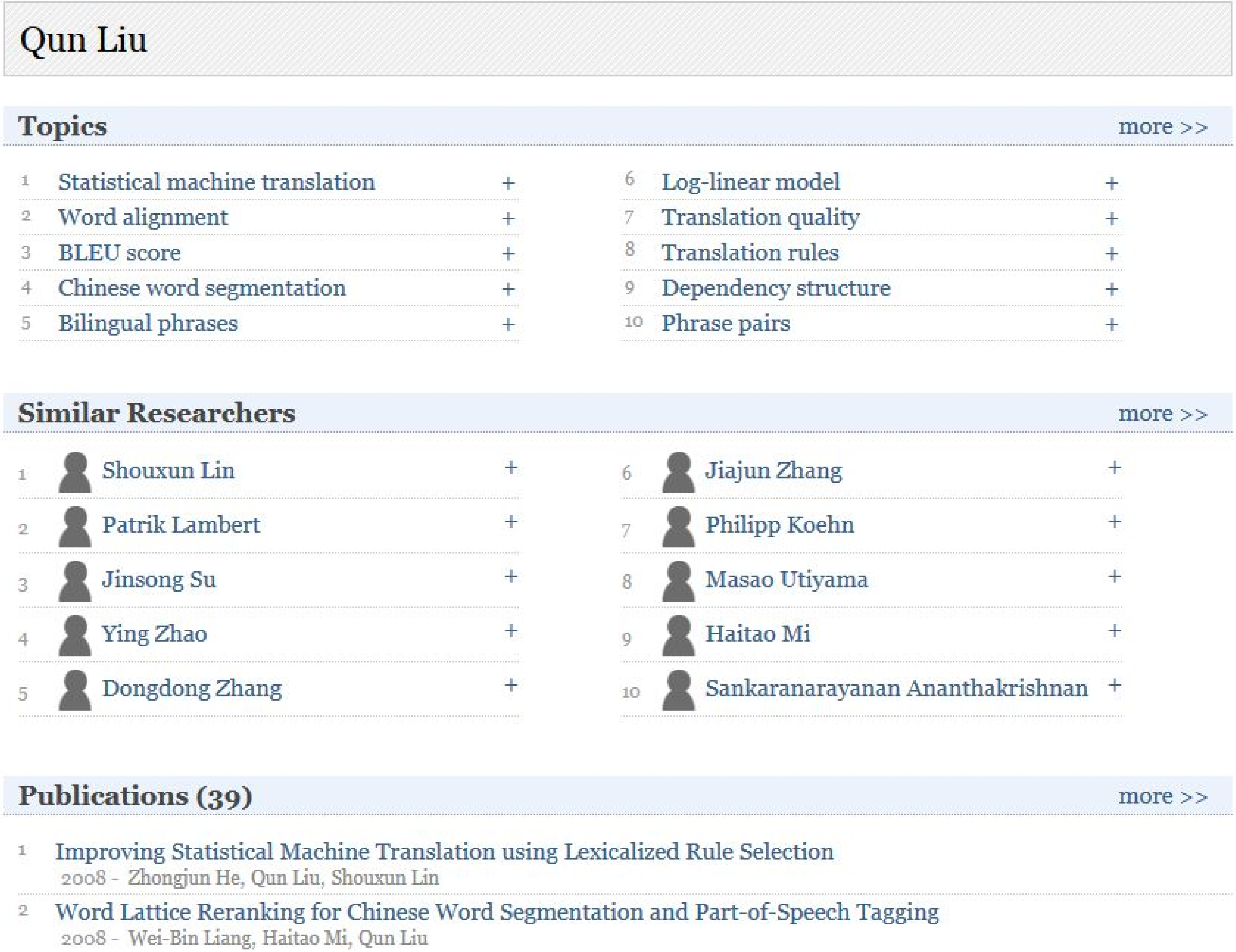


Experts

1 Hermann Ney +
2 Qun Liu +
3 Kai-min K. Chang +
4 Ming Zhou +
5 Stephan Vogel +
6 Haizhou Li +
7 Kevin Knight +
8 Eiichiro Sumita +
9 Tek Yong Lim +
10 Chris Callison-Burch +

Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Expertise Mining



Qun Liu

Topics

more >>

1 Statistical machine translation +
2 Word alignment +
3 BLEU score +
4 Chinese word segmentation +
5 Bilingual phrases +
6 Log-linear model +
7 Translation quality +
8 Translation rules +
9 Dependency structure +
10 Phrase pairs +

Similar Researchers

more >>

1 Shouxun Lin +
2 Patrik Lambert +
3 Jinsong Su +
4 Ying Zhao +
5 Dongdong Zhang +
6 Jiajun Zhang +
7 Philipp Koehn +
8 Masao Utiyama +
9 Haitao Mi +
10 Sankaranarayanan Ananthakrishnan +

Publications (39)

more >>

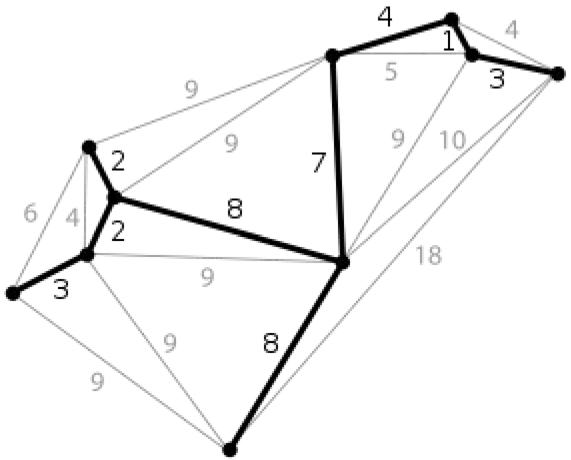
1 Improving Statistical Machine Translation using Lexicalized Rule Selection
2008 - Zhongjun He, Qun Liu, Shouxun Lin

2 Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging
2008 - Wei-Bin Liang, Haitao Mi, Qun Liu

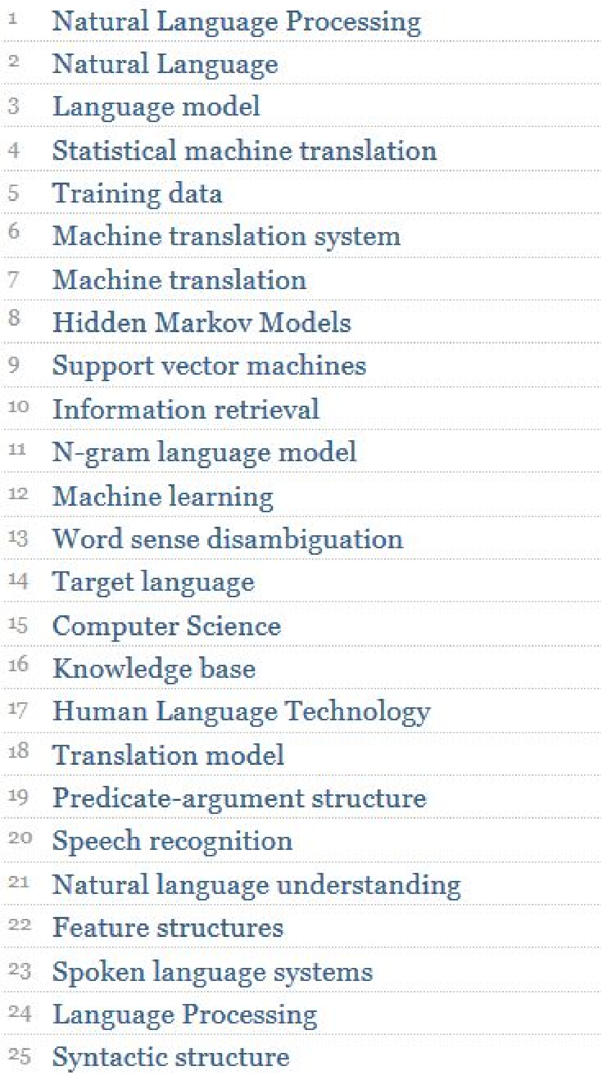
Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Step 10 - Taxonomy Construction

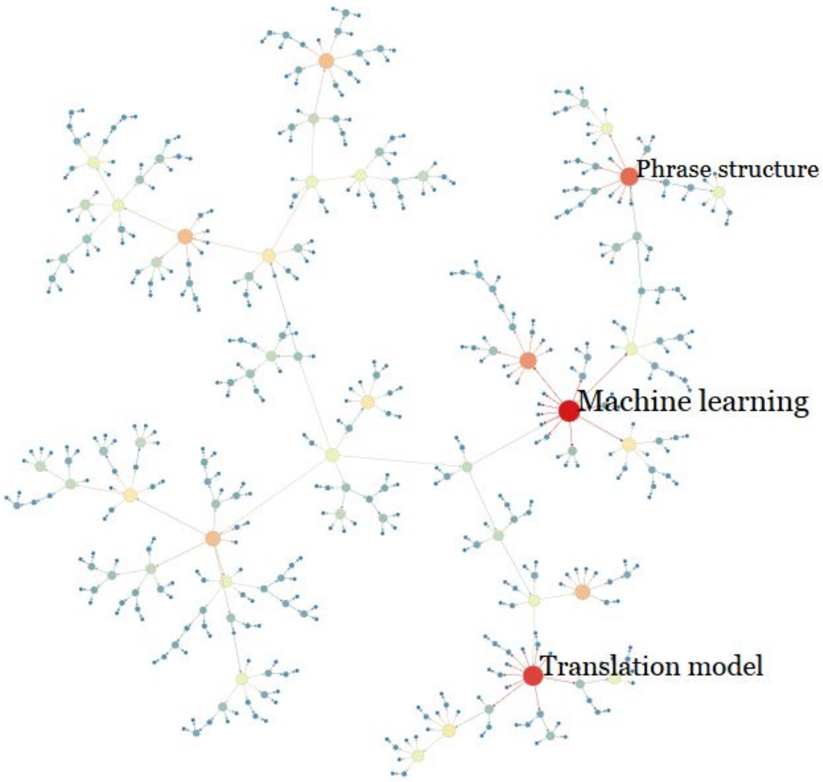
- Reduce topic-topic graph to directed acyclic graph
    - Simpler hierarchical structure for corpus
- Minimum spanning tree
- Directed to ensure most general nodes are at the top



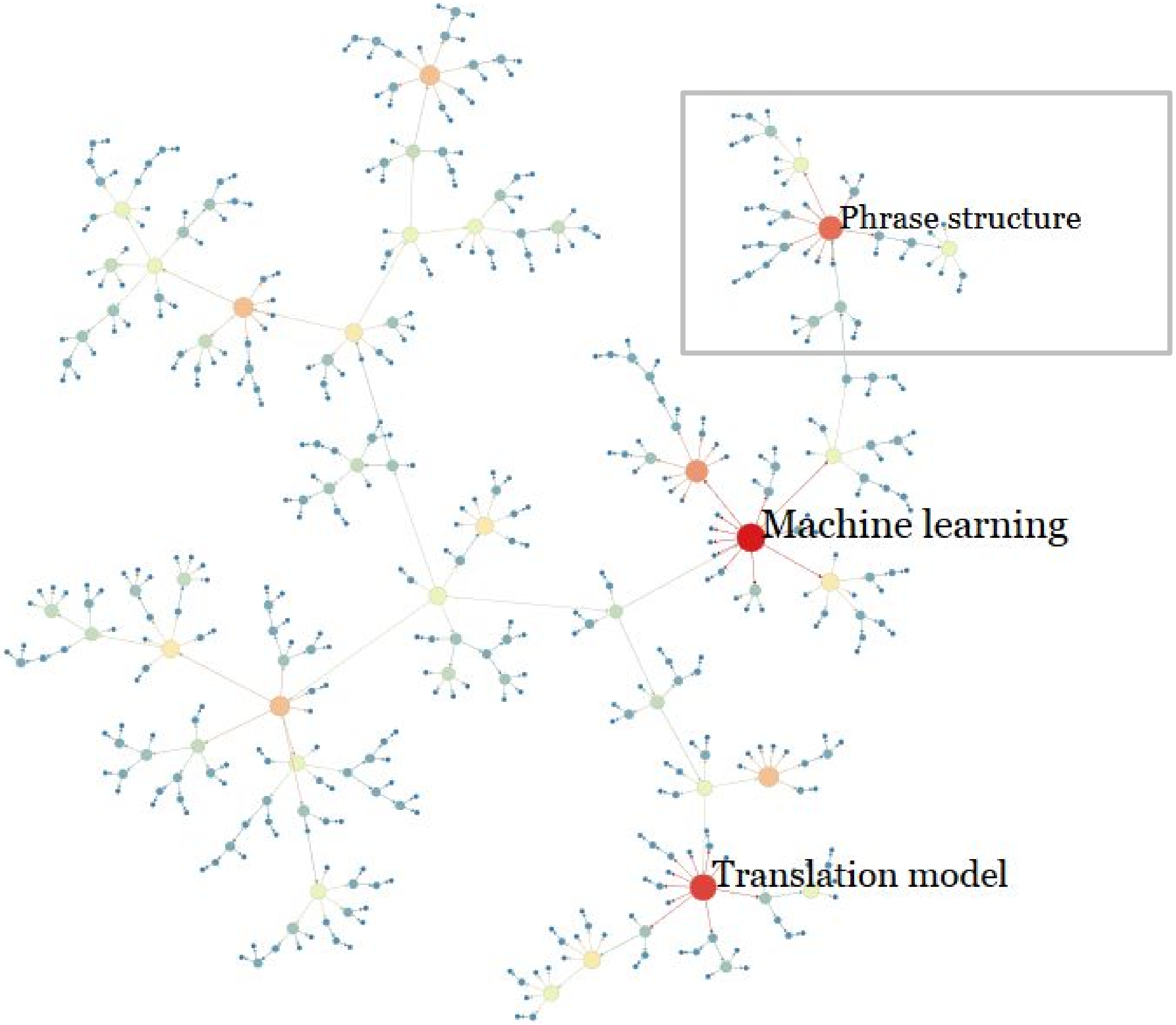
Terms to Taxonomy - ACL Anthology



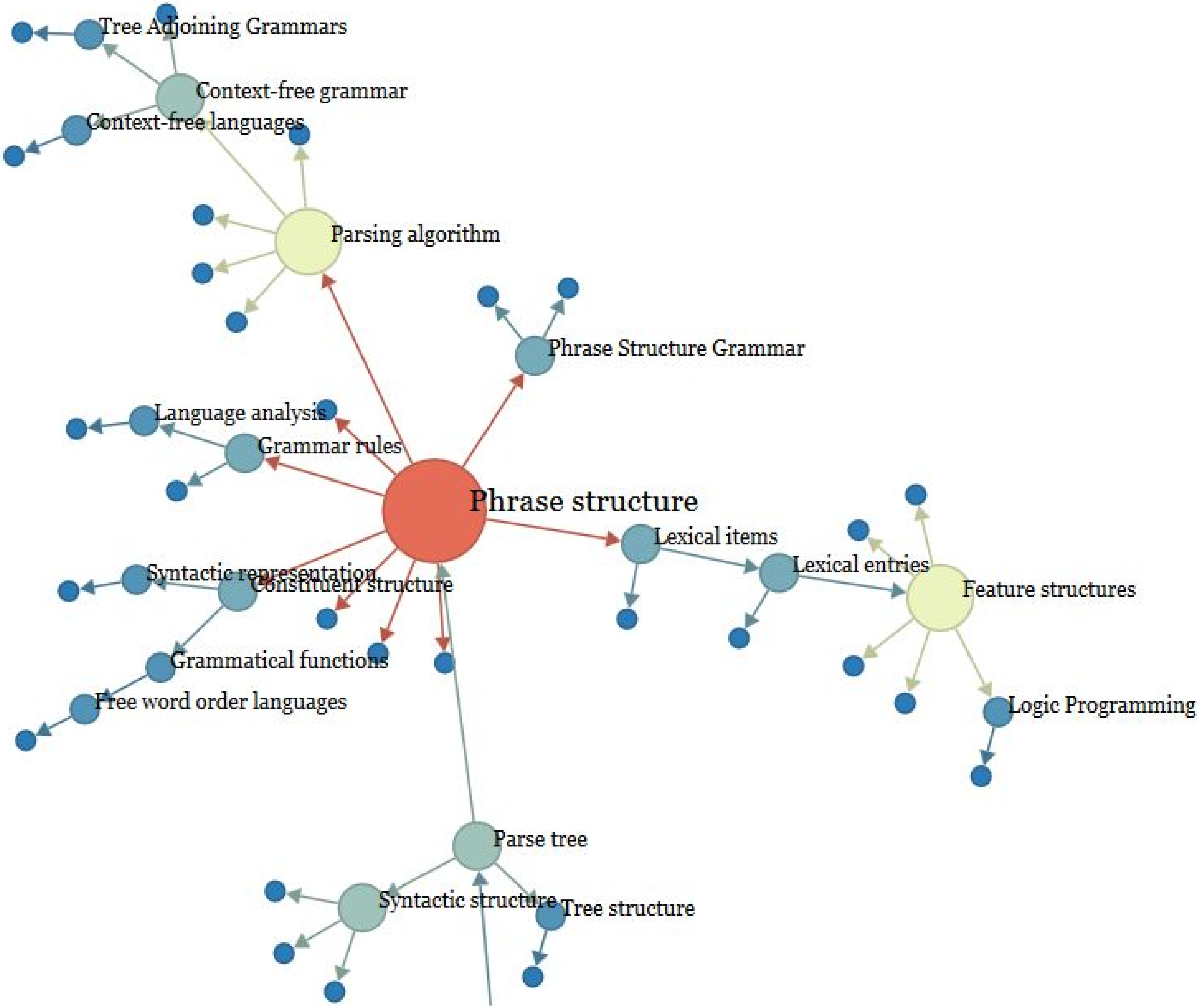
1 Natural Language Processing
2 Natural Language
3 Language model
4 Statistical machine translation
5 Training data
6 Machine translation system
7 Machine translation
8 Hidden Markov Models
9 Support vector machines
10 Information retrieval
11 N-gram language model
12 Machine learning
13 Word sense disambiguation
14 Target language
15 Computer Science
16 Knowledge base
17 Human Language Technology
18 Translation model
19 Predicate-argument structure
20 Speech recognition
21 Natural language understanding
22 Feature structures
23 Spoken language systems
24 Language Processing
25 Syntactic structure



Taxonomy Extraction – ACL Anthology

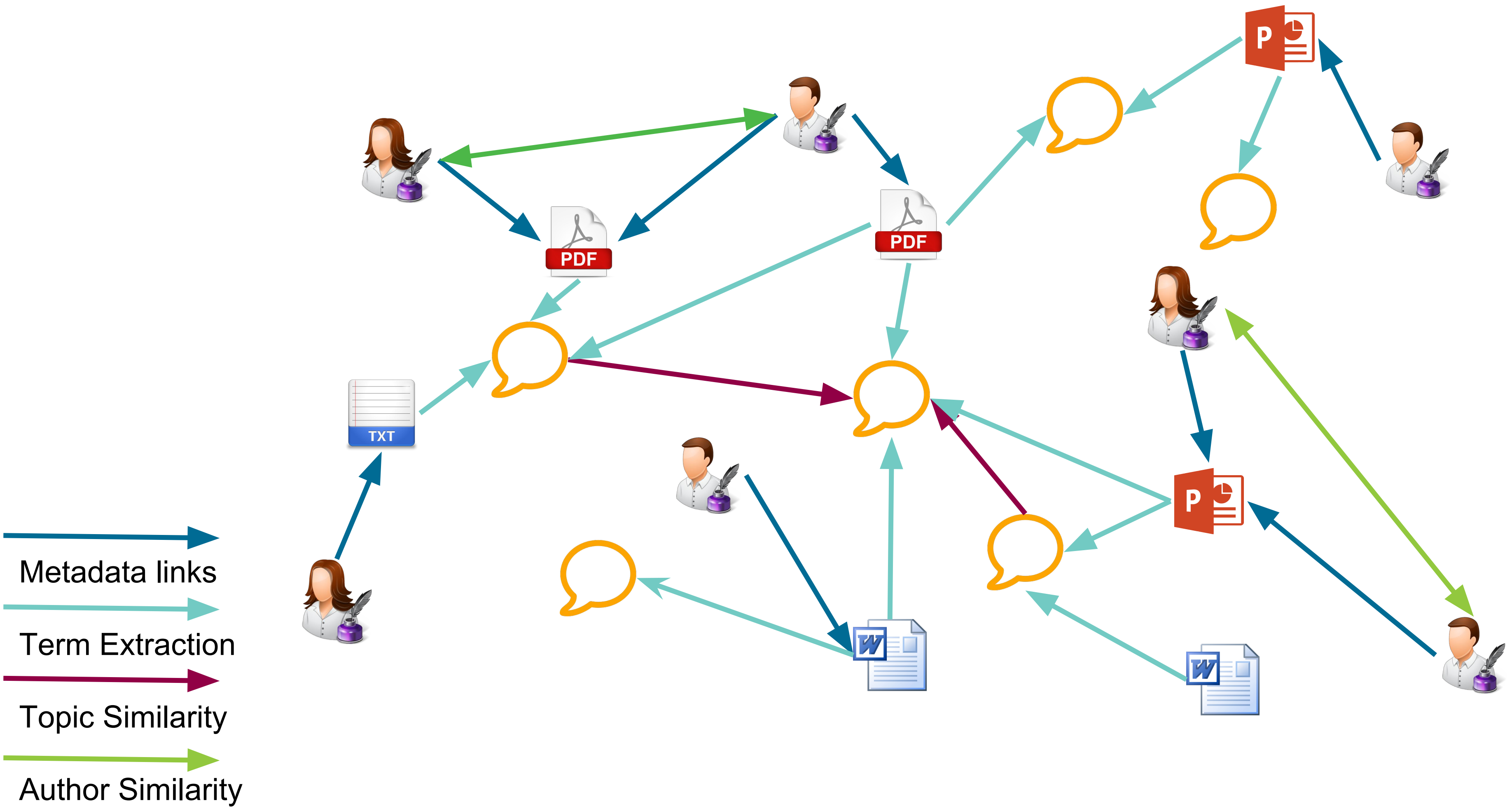


Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway



Georgeta Bordea (2013) Domain adaptive extraction of topical hierarchies for Expertise Mining. PhD Thesis, National University of Ireland, Galway

Heterogeneous graph



Metadata links

Term Extraction

Topic Similarity

Author Similarity

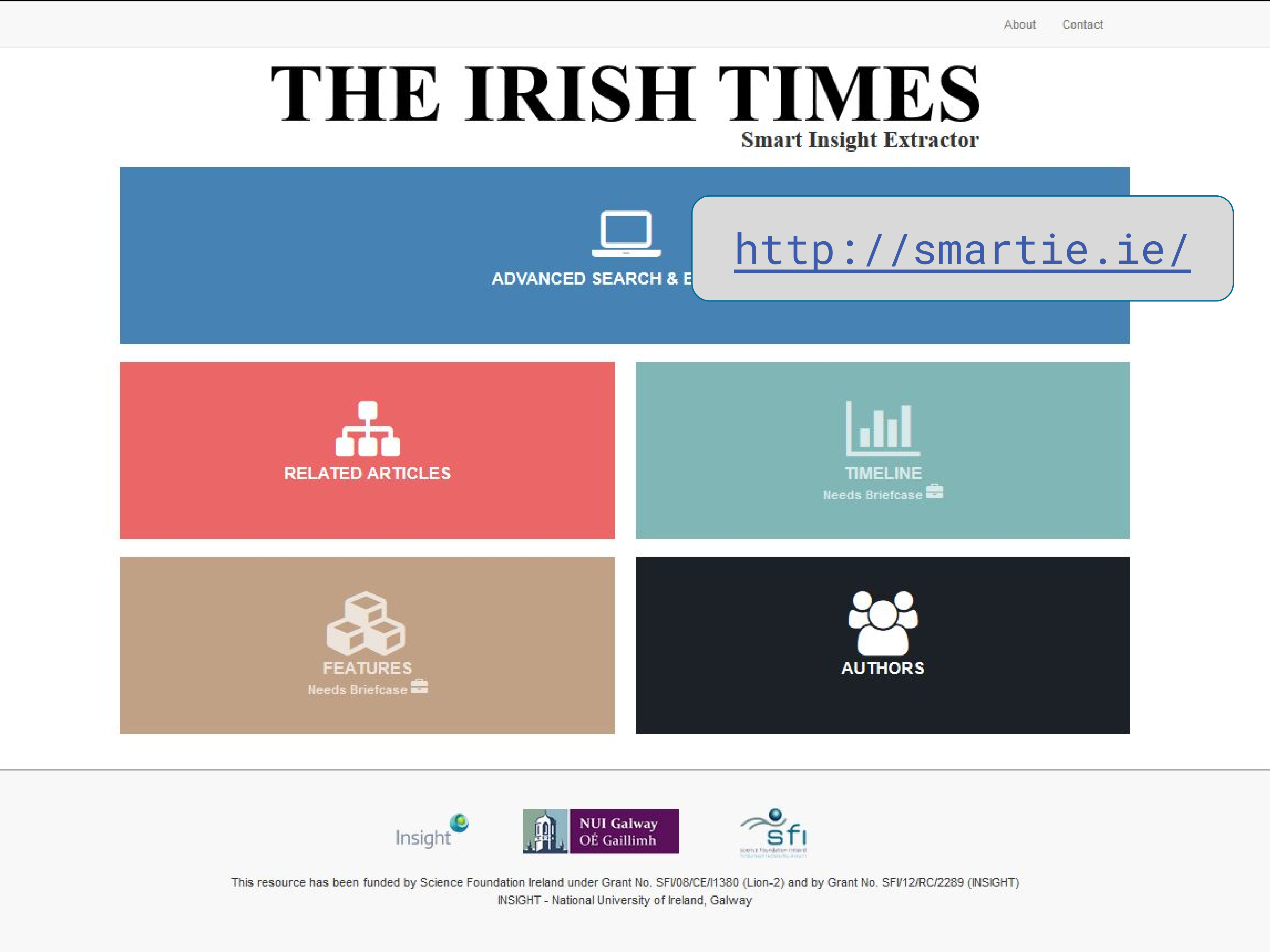
Industry Applications

Content Analysis for Book Recommendation

Kennys.ie
5 million books
...free delivery worldwide!

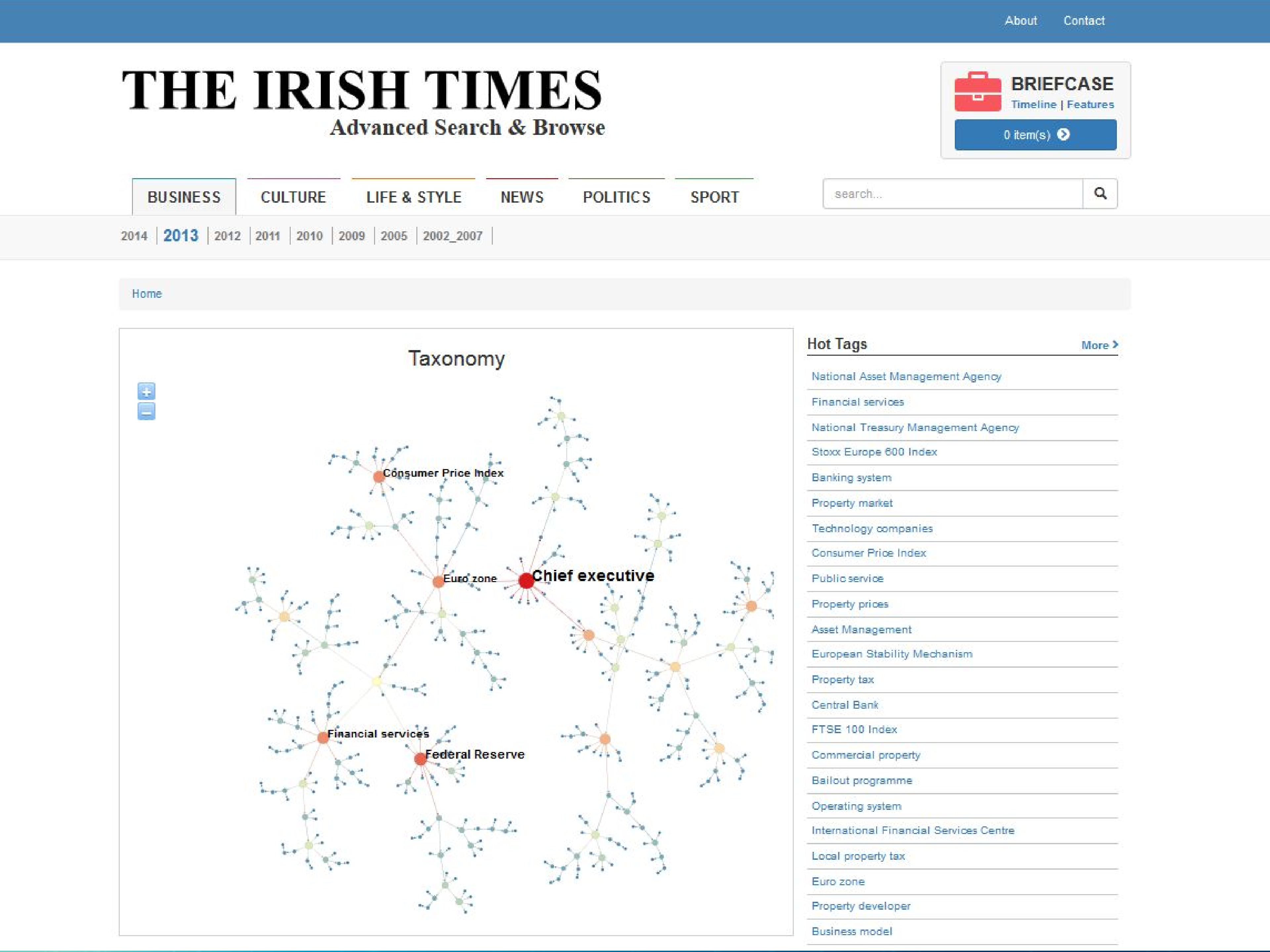
IT THE IRISH TIMES
irishtimes.com

Semantic Search on Digital News Archives



http://smartie.ie/

This resource has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2) and by Grant No. SFI/12/RC/2289 (INSIGHT)
INSIGHT - National University of Ireland, Galway



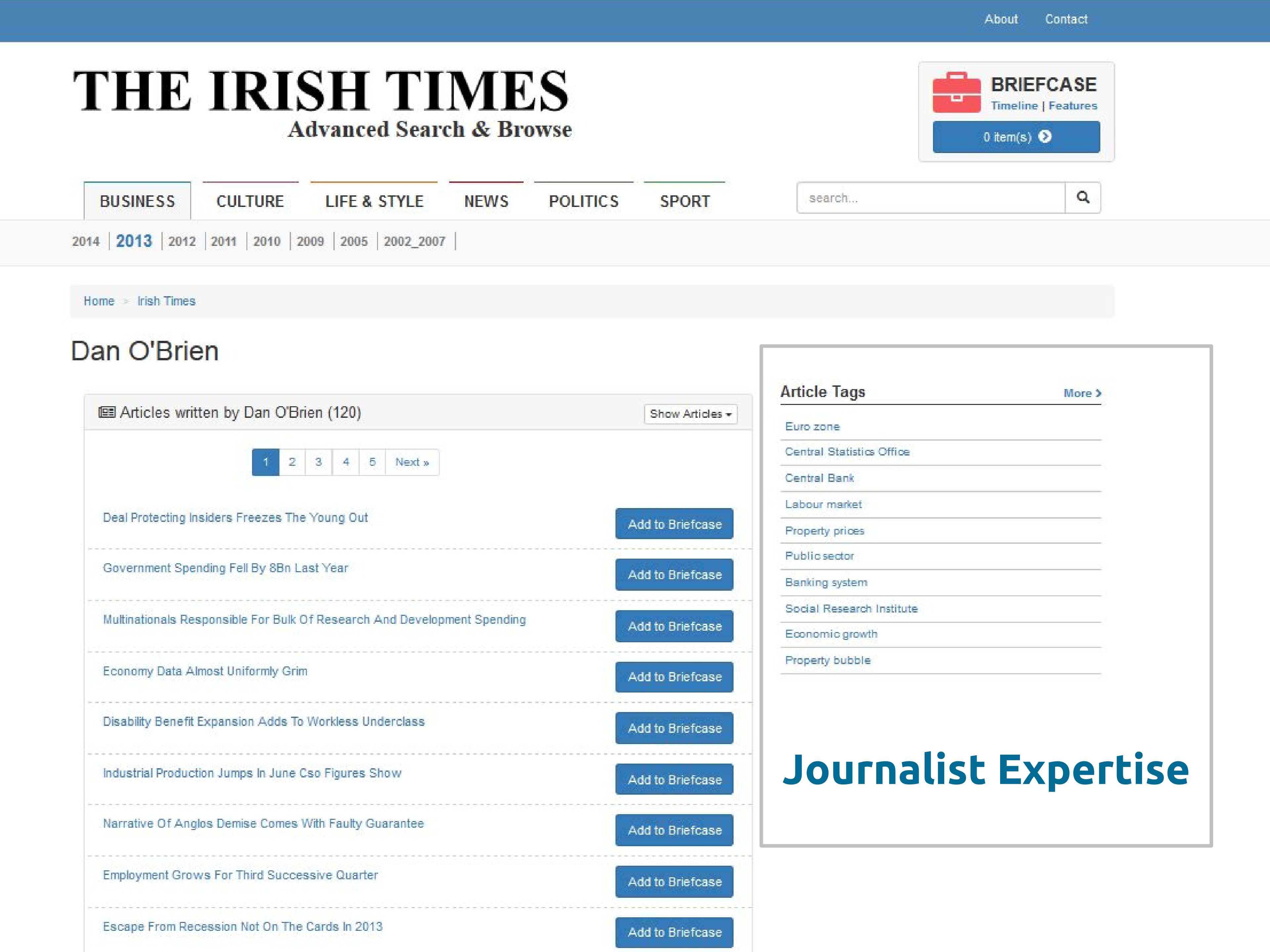
THE IRISH TIMES

Advanced Search & Browse

Taxonomy

Hot Tags

- National Asset Management Agency
- Financial services
- National Treasury Management Agency
- Stoxx Europe 600 Index
- Banking system
- Property market
- Technology companies
- Consumer Price Index
- Public service
- Property prices
- Asset Management
- European Stability Mechanism
- Property tax
- Central Bank
- FTSE 100 Index
- Commercial property
- Bailout programme
- Operating system
- International Financial Services Centre
- Local property tax
- Euro zone
- Property developer
- Business model



Dan O'Brien

Articles written by Dan O'Brien (120)

- Deal Protecting Insiders Freezes The Young Out — Add to Briefcase
- Government Spending Fell By 8Bn Last Year — Add to Briefcase
- Multinationals Responsible For Bulk Of Research And Development Spending — Add to Briefcase
- Economy Data Almost Uniformly Grim — Add to Briefcase
- Disability Benefit Expansion Adds To Workless Underclass — Add to Briefcase
- Industrial Production Jumps In June Cso Figures Show — Add to Briefcase
- Narrative Of Anglos Demise Comes With Faulty Guarantee — Add to Briefcase
- Employment Grows For Third Successive Quarter — Add to Briefcase
- Escape From Recession Not On The Cards In 2013 — Add to Briefcase

Article Tags

- Euro zone
- Central Statistics Office
- Central Bank
- Labour market
- Property prices
- Public sector
- Banking system
- Social Research Institute
- Economic growth
- Property bubble

Journalist Expertise

Towards Saffron 3

- **Saffron** was developed primarily by Georgeta Bordea, Barry Coughlan (and many others)
- Technical improvements
  - One language (Java), one database (Lucene), one build system (Maven) etc.
  - Refactor code with existing libraries
    - V2.0: 14,500 Java LoC, 35,919 Python LoC
    - V3.0: 7,000 Java LoC

Towards Saffron 3

- **Saffron** has attracted a lot of research and commercial attention
- But, **Saffron** is more importantly a research project.
- Next Step: Establish new baseline for
    - Term Extraction
        - Based on Astrakhanstev 2017
    - Taxonomy Learning
        - Use TExEval datasets (WordNet, E
        - New datasets that are taxonomic, ACM Computing Classification Sys
- Then: New **algorithms** :)

N. Astrakhantsev. ATR4S: Toolkit with State-of-the-art Automatic Terms Recognition Methods in Scala. https://arxiv.org/abs/1611.07804
TExEval @ SemEval 2016: http://alt.qcri.org/semeval2016/task13/

Conclusion

- Big document collections are **hard to understand**
  - In Academia
  - In Industry
- **Taxonomies** are the natural way to explore datasets
  - Evaluating the quality of a taxonomy is very hard
- Author metadata for documents lets us understand and **find experts**
- **Heterogeneous** graphs give new options for exploring document collections