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
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

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Weirdness</th>
<th>Relevance</th>
<th>Domain Pertinence</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>0.1</td>
<td>0.6</td>
<td>0.8</td>
<td>...</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>...</td>
</tr>
</tbody>
</table>

Candidate selection by voting
| 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 Conference                      |
| 53  | Search engine                                            |
| 54  | Natural language system                                   |
| 55  | Spoken language                                          |
| 56  | Dependency structure                                      |
| 57  | Latent semantic analysis                                  |
| 58  | Natural language understanding system                     |
| 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

{  
  "honorific": null,
  "givenName": "John",
  "middleInitial": "P",
  "familyName": "McCrae"
}

John P. McCrae

McCrae, J.P.
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
Step 7 - Connect Authors

Topic 1: TF-IAF: 0.3
Topic 2: TF-IAF: 0.7
Topic 3: TF-IAF: 0.5

TF-IAF(T; r) = \sum_{Doc \text{ if } T \in Doc, r \in \text{Authors(Doc)}} \text{TF-IDF}(T)
Step 8 - Author Similarity

Step 9 - Topic Similarity

Cosine

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 +

Expertise Mining

Qun Liu

Topics

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

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)

1. Improving Statistical Machine Translation using Lexicalized Rule Selection
   2008 - Zhongjun He, Qun Liu, Shouxun Lin
   2008 - Wei-Bin Liang, Haitao Mi, Qun Liu
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

Heterogeneous graph

- Metadata links
- Term Extraction
- Topic Similarity
- Author Similarity
Industry Applications

Content Analysis for Book Recommendation

Semantic Search on Digital News Archives
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
  o Term Extraction
    • Based on Astrakhantsev 2017
  o Taxonomy Learning
    • Use TExEval datasets (WordNet, E
    • New datasets that are taxonomic, ACM Computing Classification Sys
• Then: New **algorithms** :)
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