Annotation and Ontologies in the Context of IR

Eric Gaussier

Université Joseph Fourier - Laboratoire d'Informatique de Grenoble
(based on a joint work with Loïc Maisonnasse and Jean-Pierre Chevallet)

What is IR ?

- A collection of documents
- Users with information needs
- Finding (parts of) documents relevant to a given information need

Illustration

Collection of annotated medical images (CLEF 2007 - [1])

Example of an information need :

Show me chest CT images with emphysema

Specificities of IR

- The information need corresponds to one or several search topics
  - which may be imprecisely formulated
  - which need be interpreted by the IR ystem
- The overlap between topics dealt with in a document and an information need is in general partial
  $\longrightarrow$ degrees of relevance

Main components of an IR system

An IR system comprises three main modules :

1. A module for indexing information needs ($\rightarrow$ queries)
2. A module for indexing documents
3. A module for matching queries and document representations

The *bag of words* : a commonly used representation

Example *show, chest, CT, image, emphysema*

Standard IR models

For a general presentation, see ([2], [3] and [4])

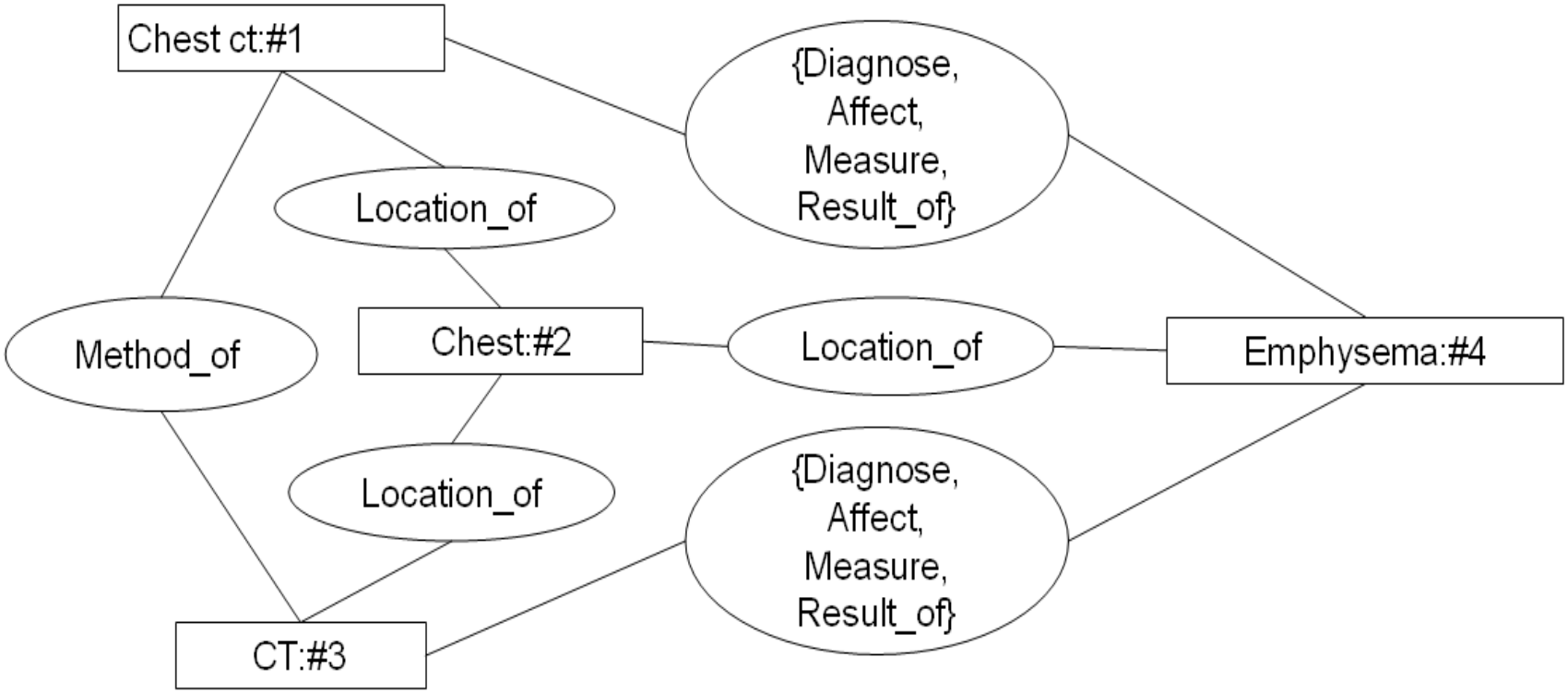
1. Boolean model
2. Vector space model
3. Probabilistic models
   - Standard probabilistic model ([5])
   - Language modeling approach ([6])
   - Divergence from randomness ([7])

Statistical approach currently widely used

What if ?

(show, CT, chest, image, emphysema)

Replaced by



What do we need ?

1. A module for buiding semantic graphs from documents and information needs

2. A module for matching semantic graphs

Building semantic graphs (1)

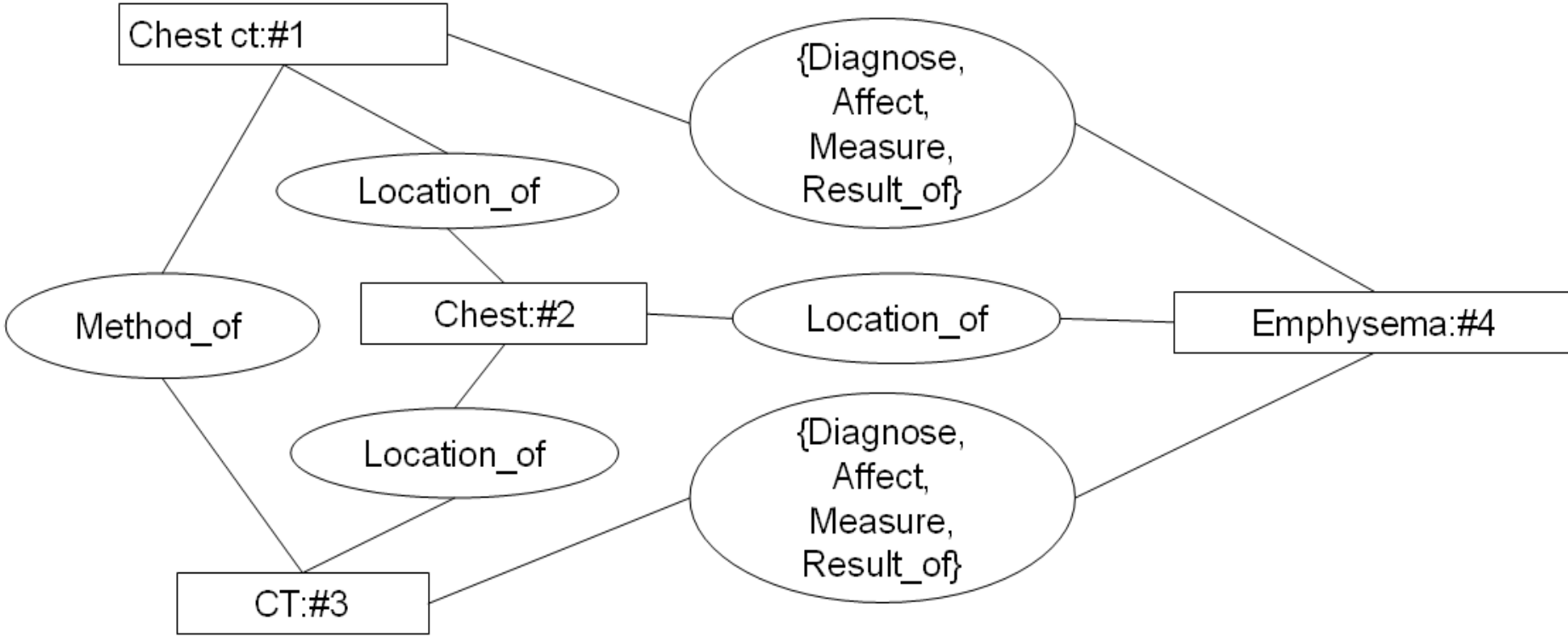
NLP processing and use of semantic resources (ontologies, thesauri)

1. Morphological analysis : POS (part-of-speech) tagging with lemmatization
2. Stopword removal : filtering empty words on the basis of their POS
3. Lexical lookup in the concept set associated with the semantic resource
4. Possibly : additional filtering on the domain
5. Relation extraction from semantic resources (on the basis of the output of a shallow parser ; using existing relations)

Building semantic graphs (2)

Available tools : MetaMap ([8]), Minipar ([9]), Treetagger ([10])

Example



Matching semantic graphs

Extending the language modeling approach to deal with semantic graphs

Queries and documents are represented as graphs :
$G_q(=< W_C^q, W_E^q >)$

Probability of the query graph to be generated by the document (graph) model M_d :

$$P(G_q|M_d) = P(W_C^q|M_d) \times P(W_E^q|W_C^q, M_d)$$

Cf. [11,12]

Recall vs precision enhancement procedures

Conceptual indexing mainly seen as a recall enhancement procedure

Relational indexing mainly seen as a precision enhancement procedure

Experimental illustration

Collection of 55,485 documents, with 85 information needs (43 used for training, 42 for testing)

	λ_u	λ_r	λ_e	Training	Test
MAP					
GLM (words)	0.4	-	-	0.222	0.218
GLM (concepts)*	0.4	-	-	0.246	0.284
GLM (relations)*	0.3	1	0.5	0.252	**0.294**
P@5					
GLM (words)	0.2	-	-	0.401	0.320
GLM (concepts)*	0.1	-	-	0.433	0.488
GLM (relations)*/**	0.1	0.3	0.5	0.457	**0.521**

Best results in bold ; */** : significant difference with *words*/*concepts* (Wilcoxon test, p=0.05)

Conclusion

1. Use of semantic resources and extended IR models improves IR system
   - Fairly straightforward use of the semantic resource
   - Complete semantic resource (meta-thesaurus and semantic network)
2. Is it a domain-specific approach ?
   - Adequacy of the resource wrt the collection under consideration
   - The case of *WordNet* and general language collections
3. Is the approach applicable to other domains ?
   - Medical and biological domains well suited (several works on semantic IR in these domains)
   - Semantic resources exist in other domains (geology, chemistry, $\cdots$) but not necessarily evaluation resources

Selected bibliography (1)

[1] H. Muller, T. Deselaers, E. Kim, J. Kalpathy-Cramer, T. M. Deserno, P. Clough and W. Hersh, *Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks*. Working Notes of the 2007 CLEF Workshop.

[2] R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, Wokingham, UK, 1999.

[3] C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. www-csli.stanford.edu/∼hinrich/information-retrieval-book.html

[4] Salton, G. (Ed), *The SMART Retrieval System. Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs (NJ), 1971.

Selected bibliography (2)

[5] S. E. Robertson, *The probability ranking principle in IR*. Journal of Documentation, Vol. **38**, 1977.
[6] J. M. Ponte and W. B. Croft, *A language modeling approach to information retrieval*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.
[7] G. Amati and C. J. van Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness*. ACM - Transactions on Information Systems, Vol. **20**, 2002.

Selected bibliography (3)

[8] A. Aronson. *Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program*. Proceedings of AMIA 2001.
[9] D. Lin, *Dependency-based Evaluation of MiniPar*. Proceedings of the Workshop on the Evaluation of Parsing Systems, 1998.
[10] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[11] L. Maisonnasse, E. Gaussier and J.-P. Chevallet, *Revisiting the dependence language model for information retrieval*. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (poster session).

Selected bibliography (4)

[12] L. Maisonnasse, E. Gaussier and J.-P. Chevallet, *Multiplying concept sources for graph modeling*. In C. Peters, V. Jijkoun, T. Mandl, H. Muller, D.W. Oard, A. Peñas, V. Petras, D. Santos, (Eds.) : Advances in Multilingual and Multimodal Information Retrieval. LNCS #5152. Springer-Verlag, Berlin (2008).

[13] Y. Huang, H. J. Lowe and W. Hersh, *A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports*. Proceedings of the Conference of the American Medical Informatics Association, 2003.

[14] S. Vintar, P. Buitelaar and M. Volk, *Relations in Concept-Based Cross-Language Medical Information Retrieval*. Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM), 2003.