FONDAZIONE
BRUNO KESSLER

Keyword Search Through Semantic Artifacts: an Introduction

Mauro Dragoni

Fondazione Bruno Kessler (FBK), Shape and Evolve Living Knowledge Unit (SHELL)

https://shell.fbk.eu/index.php/Mauro_Dragoni - dragoni@fbk.eu

KEYSTONE Meeting, Marseille

February, 22nd 2016

FONDAZIONE
BRUNO KESSLER

Outline

1. On your marks and get set…
2. A general approach: pros and cons of concept-based structured representations
3. Ontology-based IR platforms

FONDAZIONE
BRUNO KESSLER

Before to start…

- What is an ontology?
- What is a machine-readable dictionary?
- What about ambiguity?
- Terms vs. concepts, is everything clear?

FONDAZIONE BRUNO KESSLER

What is an ontology?

- "the branch of philosophy which deals with the nature and the organization of reality"
- "an ontology is an explicit specification of a conceptualization" [Gruber1993]
  - **conceptualization**: abstract model of the world
  - **explicit specification**: model described by using unambiguous language
- domain ontology
- upper ontology
  - example: DOLCE [Guarino2002]

FONDAZIONE
BRUNO KESSLER

Ontology Components

- Classes: entities describing objects common characteristics (for example: “Agricultural Method”).
- Individuals: entities that are instances of classes (for example “Multi Crops Farming” is an instance of “Agricultural Method”).
- Properties: binary relations between entities (for example “IsAffectedBy”).
- Attributes (or DataType Properties): characteristics that qualify individuals (for example “Has Name”).

FONDAZIONE
BRUNO KESSLER

Hierarchies

- Concepts can be organized in subsumptions hierarchies
- Meaning: every sub-concepts is also a super-concept
- Examples:
  - "Intensive Farming" is-a "Agricultural Method"
  - "Agricultural Method" is-a "Method"
- Concept hierarchies are generally represented by using tree structures

FONDAZIONE BRUNO KESSLER

Attributes and Properties

- Properties: binary relations between classes
  - Domain and co-domain: classes to which individuals need to belong to be in relation
  - Example: “Agriculture” <isAffectedBy> “Agriculture Pollution”
- Attributes: binary relations between an individual and values (not other entities)
  - Domain: class to which the attribute is applied
  - Co-domain: the type of the value (for example “String”)
- Properties and Attributes can be organized in hierarchies.

FONDAZIONE
BRUNO KESSLER

Steps for building an ontology

- To identify the classes of the domain.
- To organize them in a hierarchy.
- To define properties and attributes.
- To define individuals, if there are.

FONDAZIONE BRUNO KESSLER

Why ontologies are useful?

- Ontologies provide:
  - common dictionary of terms;
  - a shared and formal interpretation of the domain.
- Ontologies permit to:
  - solve ambiguities;
  - share knowledge (not only between humans, but also between machines);
  - use automatic reasoning techniques.

FONDAZIONE
BRUNO KESSLER

Use of ontologies in IR

- Exploit metadata
- Entity linking
  - "which president …" → "Barack Obama is-a President"
- Extraction of triples from text
  - applying NLP parsers for extracting dependencies

FONDAZIONE BRUNO KESSLER

What is an thesaurus?

- A "coarse" version of ontologies
- Generally, 3 kinds of relations are represented:
  - hierarchical (generalization/specialization)
  - equivalence (synonymity)
  - associative (other kind of relationships)
- Extensive tool used for query expansion approaches [Bhogal2007, Grootjen2006,Qiu1993,Mandala2000]

FONDAZIONE BRUNO KESSLER

Machine-readable dictionaries

- A dictionary in an electronic form.
- The power of MRD is characterized by word senses. [Kilgariff1997, Lakoff1987, Ruhl1989]

- **Identity of meaning**: synonyms [Gove1973]
- **Inclusion of meaning**: hyponymy or hyperonymy; troponymy [Cruse1986, Green2002, Fellbaum1998]
  - transitive relationship
- **Part-whole meaning**: meronymy (has part), holonymy (part of) [Green2002, Cruse1986, Evens1986]
- **Opposite meaning**: antonymy

FONDAZIONE
BRUNO KESSLER

and now…

… let's see how we can exploit this within an information retrieval system…

FONDAZIONE BRUNO KESSLER

Motivations and Challenges

- Considering how information is usually represented and classified.
  - Documents and Queries are represented using terms.
  - Indexing:
    - terms are extracted from each document;
    - terms frequency of each document is computed (TF);
    - terms frequency over the entire index is computed (IDF).
  - Searching:
    - the vector space model is used to computed the similarity between documents and queries;
    - queries are generally expanded to increase the recall of the system.

FONDAZIONE BRUNO KESSLER

Drawbacks of the Term-Based representation – 1/2

- The "semantic connections" between terms in documents and queries are not considered.

- Different vector positions may be allocated to the synonyms of the same term:
  - the importance of a determinate *concept* is distributed among different vector components;
    - information loss.

FONDAZIONE BRUNO KESSLER

Drawbacks of the Term-Based representation – 2/2

- The query expansion has to be used carefully.
  - It is more easy to increase the recall of a system with respect to its precision. Which is better? [Abdelali2007]

- In the worst case, the size of a document vector could be close to the number of terms used in the repository:
  - in general, the number of concepts is less than the number of words;
  - the time needed to compare documents is higher;

FONDAZIONE
BRUNO KESSLER

Intuition Behind

- Using concepts to represent the terms contained in documents and queries. [Dragoni2012b]

1. Documents and Queries may be represented in the same way.
2. The issue related to how many and which terms have to be used for query expansion is not considered.
3. The size of a concept vector is generally smaller than the size of a term vector.

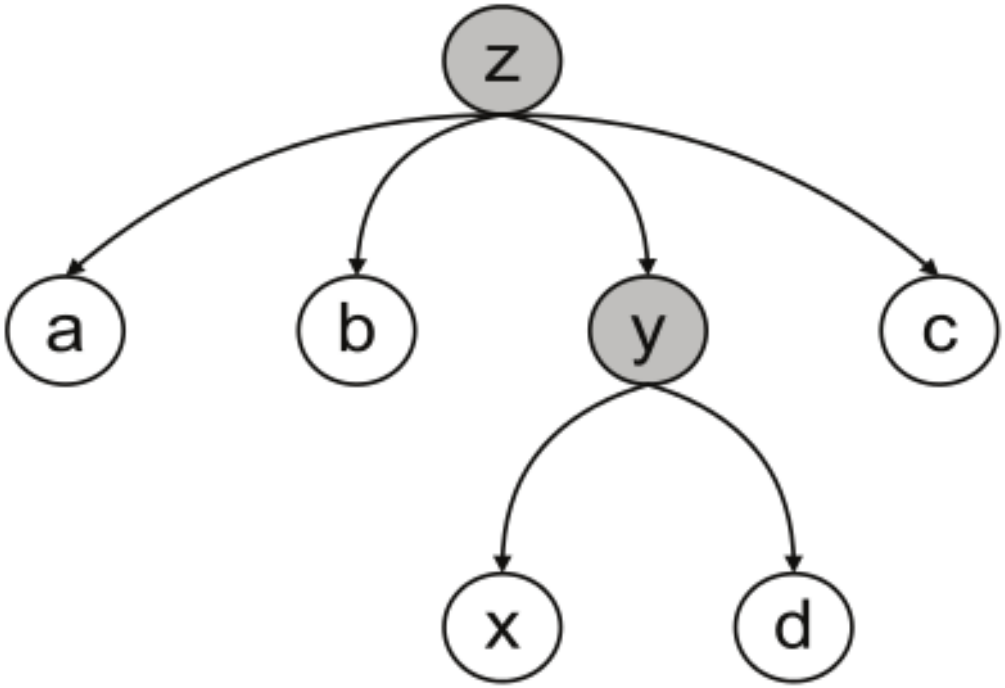
- IMPORTANT: This is not a query expansion technique !!!

FONDAZIONE
BRUNO KESSLER

a first simple example …

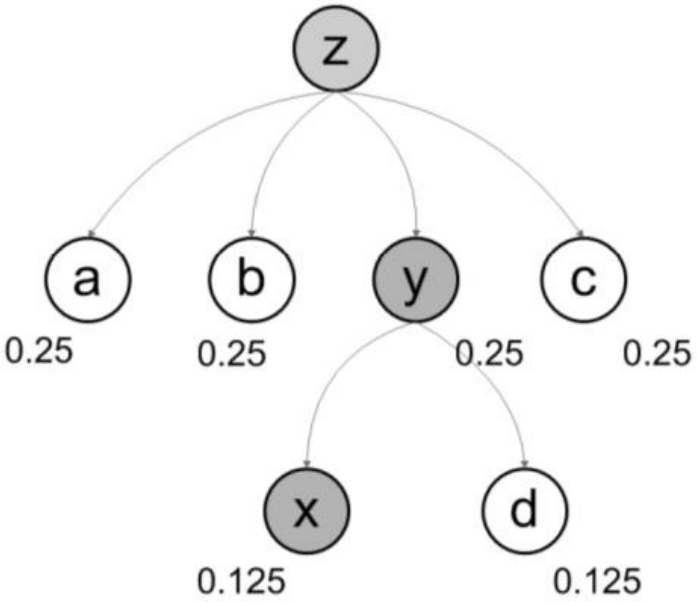
- a close vocabulary:

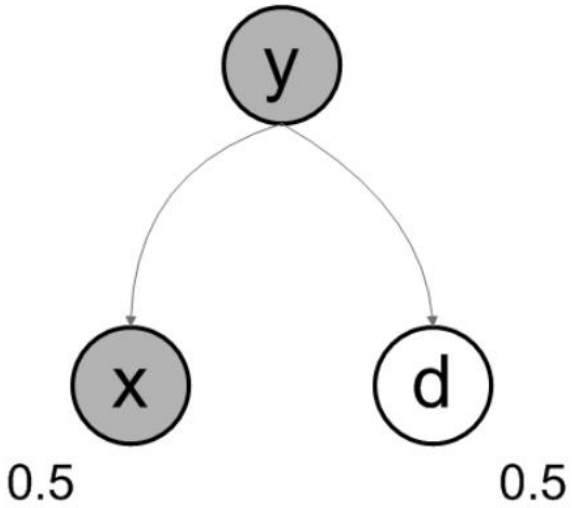
$$I = \{a, b, c, d, x\}$$



a first simple example …

- a close vocabulary:





FONDAZIONE
BRUNO KESSLER

a first simple example …

- how to compute concept weights?

$$N(c) = \mathrm{occ}(c) + \sum_{c \in \mathrm{Path}(c,\ldots,\top)} \sum_{i=2}^{\mathrm{depth}(c)} \frac{\mathrm{occ}(c_i)}{\prod_{j=2}^{i} \|\mathrm{children}(c_j)\|}$$

$$\mathrm{info}(b_i) = \frac{N_{\mathrm{doc}}(b_i)}{N_{\mathrm{rep}}(b_i)}$$

FONDAZIONE BRUNO KESSLER

a first simple example …

- how is weighted each concept of the vocabulary?

$$z = (0.25, 0.25, 0.25, 0.125, 0.125)$$
$$a = (1.0, 0.0, 0.0, 0.0, 0.0)$$
$$b = (0.0, 1.0, 0.0, 0.0, 0.0)$$
$$c = (0.0, 0.0, 1.0, 0.0, 0.0)$$
$$y = (0.0, 0.0, 0.0, 0.5, 0.5)$$
$$d = (0.0, 0.0, 0.0, 1.0, 0.0)$$
$$x = (0.0, 0.0, 0.0, 0.0, 1.0)$$

- suppose to have the document "xxyyyz"

$$D_1 = (2 * \bar{x}) + (3 * \bar{y}) + \bar{z} = (0.25, 0.25, 0.25, 1.625, 3.625)$$

FONDAZIONE BRUNO KESSLER

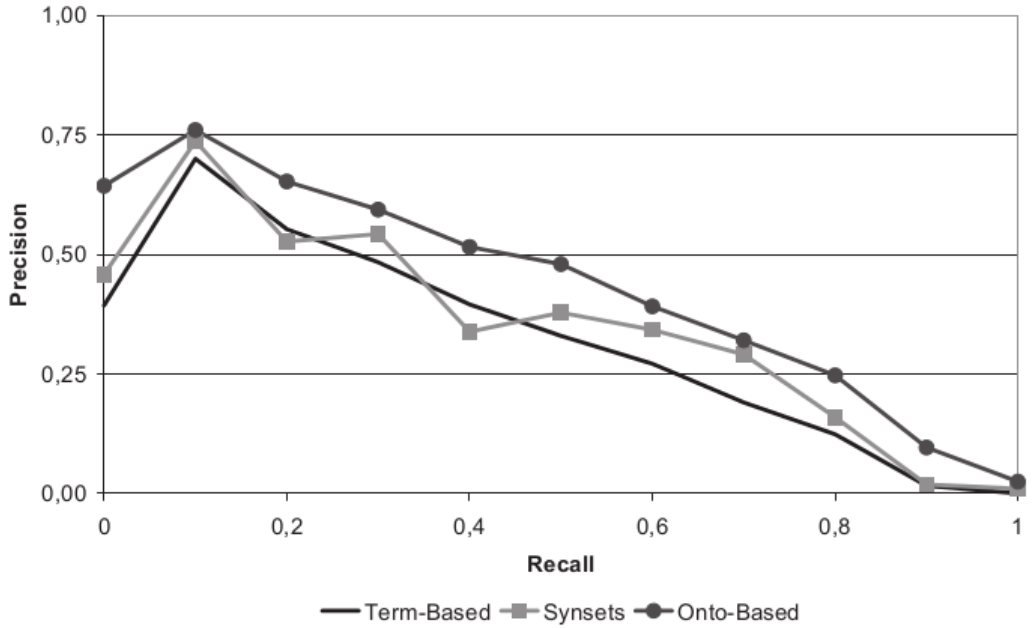
… that we evaluated

- Experiments on the MuchMore Collection (http://muchmore.dfki.de)
  - The collection contains numerous medical terms.
  - The term-based representations is advantaged over the semantic representation.

- Experiments on the TREC Ad-Hoc Collection:
  - Results have been compared with the IRS presented at TREC-7 and TREC-8 conference
  - Only the systems that implements a semantic representation of queries have been considered.
  - Over dozens of runs, the three systems that performs better at recall 0.0 have been chosen. [Spink2006]

MuchMore Collection

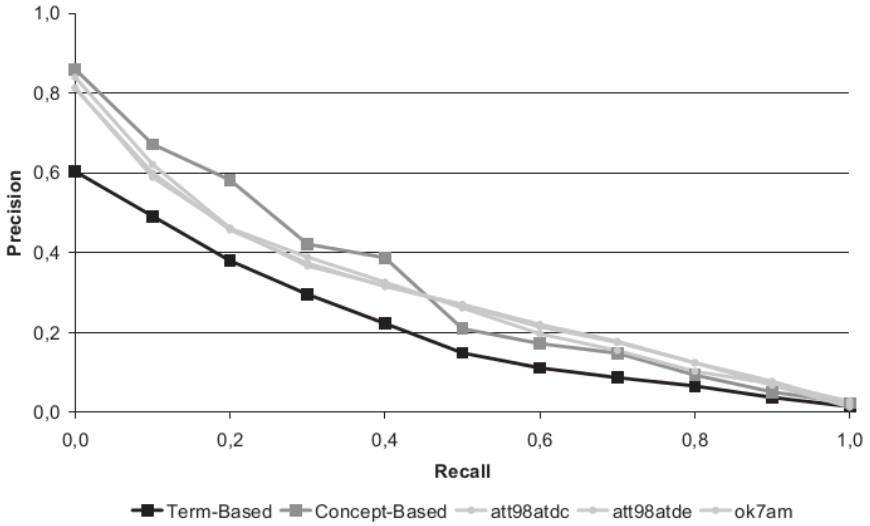
System	P@5	P@10	P@15	P@30	MAP
Term-Based	0.544	0.480	0.405	0.273	0.449
Synset-Based	0.648	0.484	0.403	0.309	0.459
Conceptual Indexing	0.770	0.735	0.690	0.523	0.449
Ontology Indexing	**0.784**	**0.765**	**0.728**	**0.594**	**0.477**



FONDAZIONE BRUNO KESSLER

TREC-7

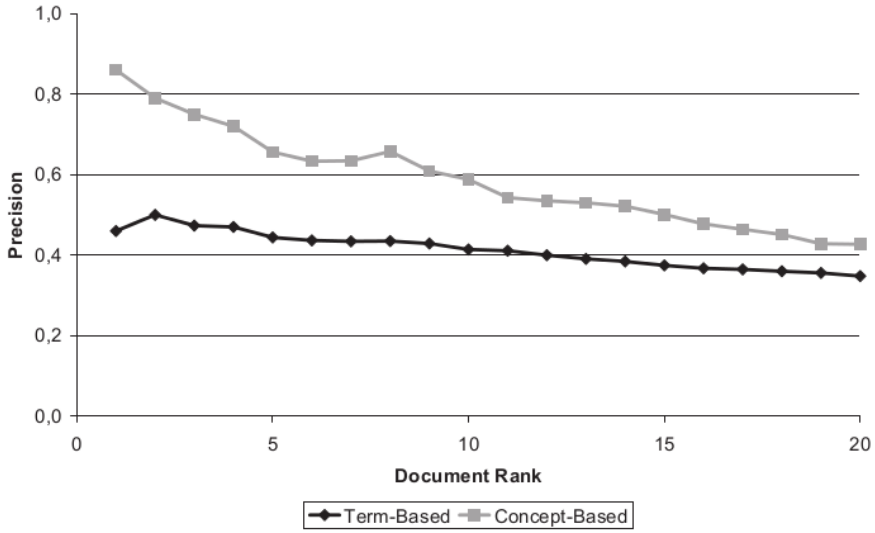
System	P@5	P@10	P@15	P@30	MAP
Term-Based	0.444	0.414	0.375	0.348	0.199
AT&T Labs 1	0.644	0.558	0.499	**0.419**	0.296
AT&T Labs 2	0.644	0.558	0.497	0.413	0.294
City University, Sheffield, Microsoft	0.572	0.542	**0.507**	0.412	0.288
Ontology Indexing	**0.656**	**0.588**	0.501	0.397	**0.309**



FONDAZIONE BRUNO KESSLER

TREC-7

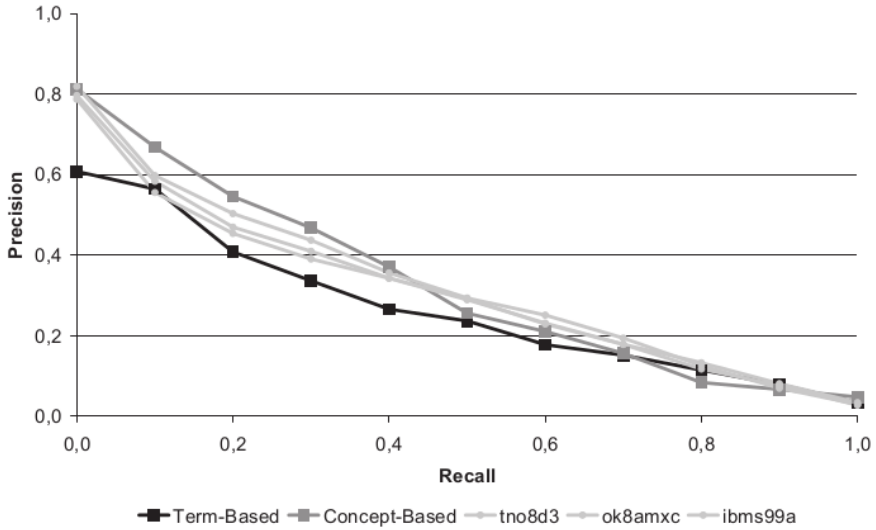
System	P@5	P@10	P@15	P@30	MAP
Term-Based	0.444	0.414	0.375	0.348	0.199
AT&T Labs 1	0.644	0.558	0.499	**0.419**	0.296
AT&T Labs 2	0.644	0.558	0.497	0.413	0.294
City University, Sheffield, Microsoft	0.572	0.542	**0.507**	0.412	0.288
Ontology Indexing	**0.656**	**0.588**	0.501	0.397	**0.309**



FONDAZIONE BRUNO KESSLER

TREC-8

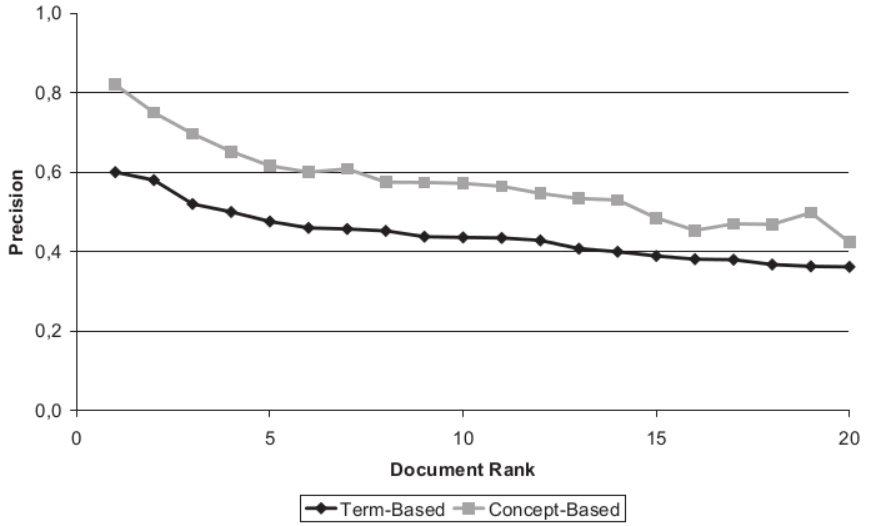
System	P@5	P@10	P@15	P@30	MAP
Term-Based	0.476	0.436	0.389	0.362	0.243
IBM Watson	0.588	0.504	0.472	0.410	0.301
Microsoft Research	0.580	0.550	**0.499**	**0.425**	**0.317**
TwentyOne	0.500	0.454	0.433	0.368	0.292
Ontology Indexing	**0.616**	**0.572**	0.485	0.415	0.315



FONDAZIONE BRUNO KESSLER

TREC-8

System	P@5	P@10	P@15	P@30	MAP
Term-Based	0.476	0.436	0.389	0.362	0.243
IBM Watson	0.588	0.504	0.472	0.410	0.301
Microsoft Research	0.580	0.550	**0.499**	**0.425**	**0.317**
TwentyOne	0.500	0.454	0.433	0.368	0.292
Ontology Indexing	**0.616**	**0.572**	0.485	0.415	0.315



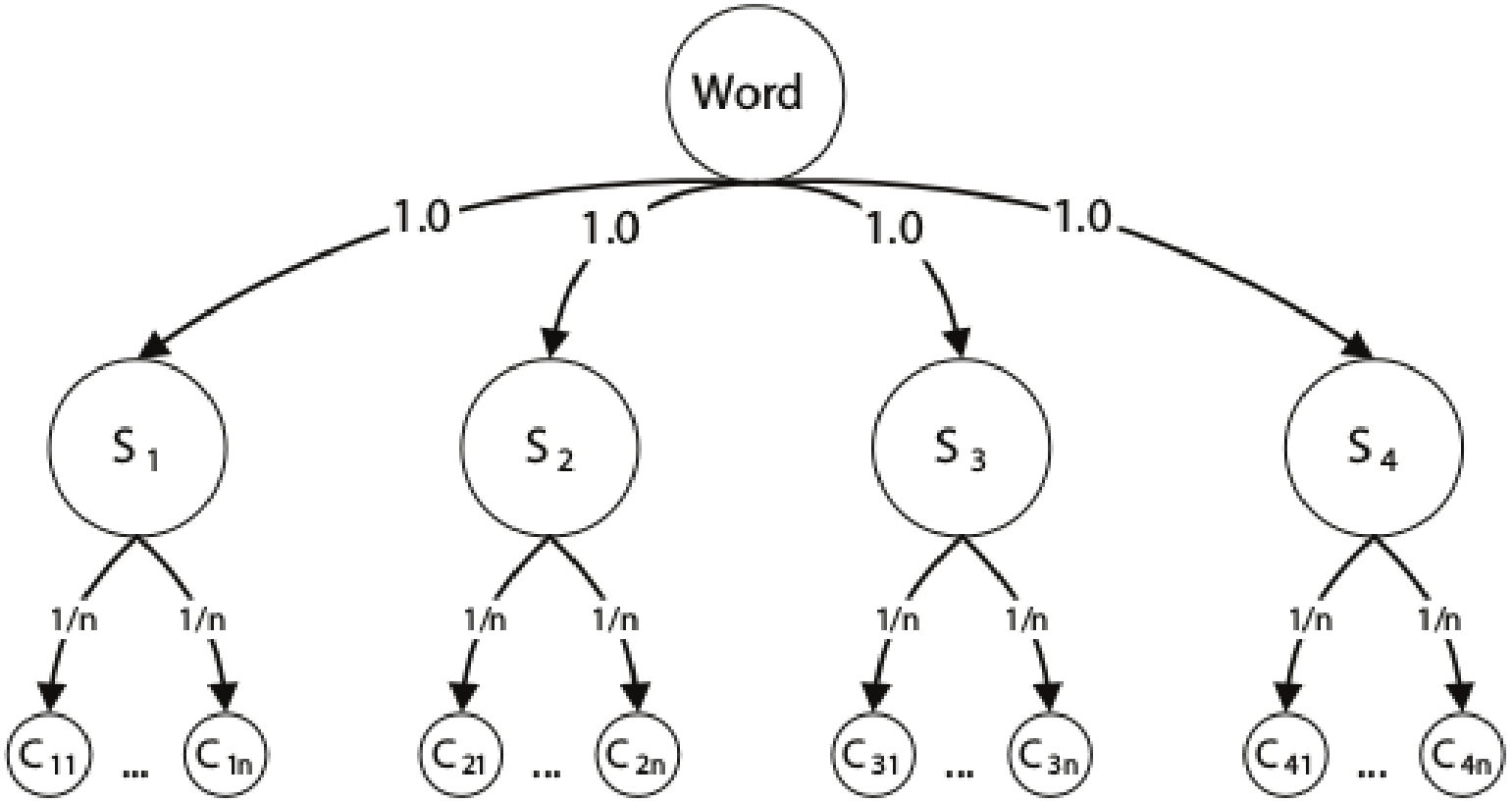
FONDAZIONE BRUNO KESSLER

Some considerations

- Two drawbacks have been identified:
  - The absence of some terms in the ontology, (in particular terms related to specific domains like biomedical, mechanical, business, etc.), may affects the final retrieval result.
    - a more complete knowledge base is needed.
  - Term ambiguity. By using a Word Sense Disambiguation approach, concepts associated with incorrect senses would be discarded or weighted less.
    - a Word Sense Disambiguation algorithm is required: but it has to be used carefully.

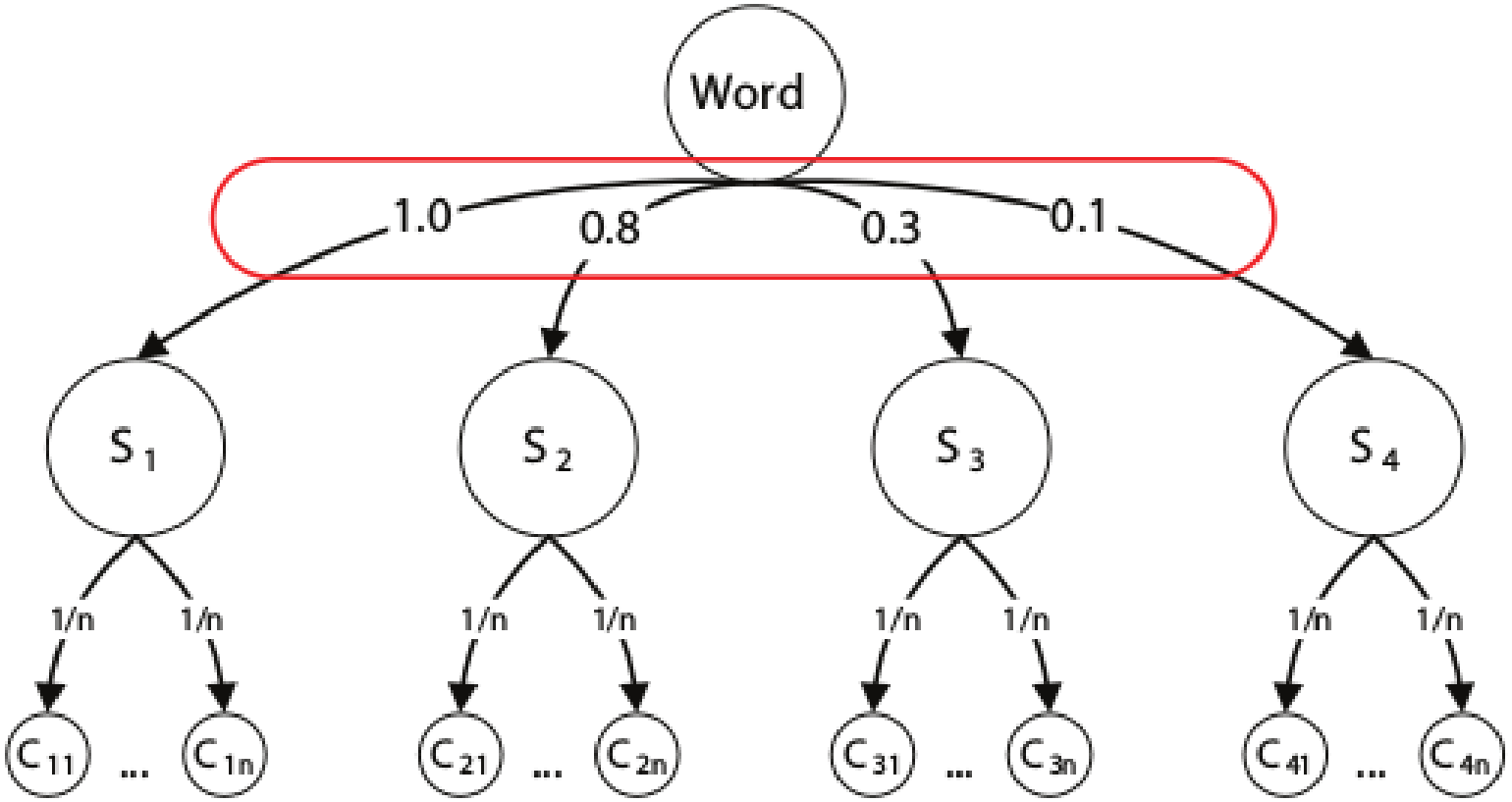
FONDAZIONE
BRUNO KESSLER

Few words on disambiguation



FONDAZIONE
BRUNO KESSLER

Few words on disambiguation



FONDAZIONE BRUNO KESSLER

Ontology enhanced IR

- Enrichment of documents (and queries) with information coming from semantic resources
  - information expansion: adding synonyms, antonyms, … not new but still helpful
  - annotations: relation or association between a semantic entity and a document
- Most of the information expansion systems are based on WordNet and the Roget's Thesaurus
- Systems using annotations are interfaced with the Linked Open Data cloud, and mainly with Freebase and Wikipedia

FONDAZIONE BRUNO KESSLER

Classification of Semantic IR approaches

Criterion	Approaches
Semantic knowledge representation	• Statistical [Deerwester1990] • Linguistic conceptualization [Gonzalo1998, Mandala1998,Giunchiglia2009] • Ontology-based [Guha2003,Popov2004]
Scope	• Web search [Finin2005,Fernandez2008] • Limited domain repositories [Popov2004] • Desktop search [Chirita2005]
Query	• Keyword query [Guha2003] • Natural language query [Lopez2009] • Controlled natural language query [Bernstein2006, Cohen2003] • Structured query based on ontology query language [notes]
Content retrieved	• Data retrieval • Information retrieval
Content ranking	• No ranking • Keyword-based ranking [Guha2003] • Semantic-based ranking [Stojanovic2003]

FONDAZIONE BRUNO KESSLER

Limitation of Semantic IR approaches – 1/2

Criterion	Limitation	IR	Semantic
Semantic knowledge representation	• No exploitation of the full potential of an ontological language, beyond those that could be reduced to conventional classification schemes.	x	(Partially)
Scope	• No scalability to large and heterogeneous repositories of documents.		x
Goal	• Boolean retrieval models where the Information Retrieval problem is reduced to a data retrieval task.		x
Query	• Limited usability		x

FONDAZIONE BRUNO KESSLER

Limitation of Semantic IR approaches – 2/2

Criterion	Limitation	IR	Semantic
Content retrieved	• Focus on textual content: no management of different formats (multimedia)	(Partially)	(Partially)
Content ranking	• Lack of semantic ranking criterion. The ranking (if provided relies on keyword-based approaches.	x	x
Coverage	• Knowledge incompleteness. [Croft1986]	(Partially)	x
Evaluation	• Lack of standard evaluation frameworks. [Giunchiglia2009]		x

FONDAZIONE
BRUNO KESSLER

So… at the end…

- Ontologies in IR is still a controversial topic
- Personal Opinion: to combine structured and unstructured representation seems to be the most suitable solution
- Pay attention to the kind of queries performed by users
- Aggregation of results

FONDAZIONE
BRUNO KESSLER

Mauro Dragoni

https://shell.fbk.eu/index.php/Mauro_Dragoni
dragoni@fbk.eu