



TECHNISCHE  
UNIVERSITÄT  
DRESDEN

BIOTEchnologisches Zentrum



# Word-sense disambiguation in biomedical ontologies

Dimitra Alexopoulou, Michael Schroeder

Ontologies are widely used

for data retrieval & analysis from disparate sources

### **Gene Ontology (GO)**

for gene product description (process, component, function)

### **Medical Subject Headings (MeSH)**

for biomedical information & document indexing

### **Open Biomedical Ontologies (OBO foundry)**

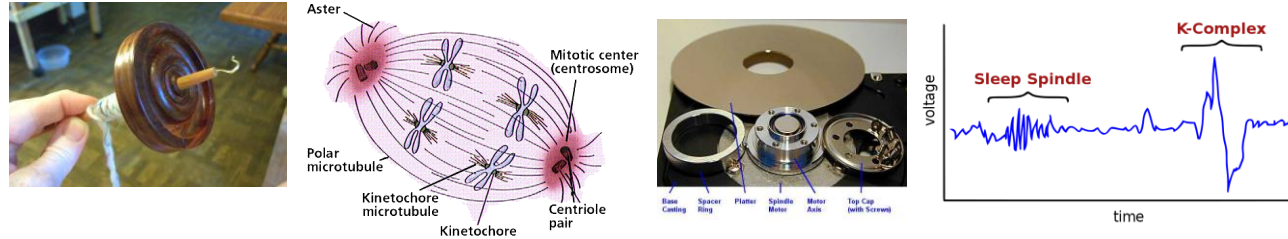
> 60 ontologies

(anatomy, pathology, experimental methods...)

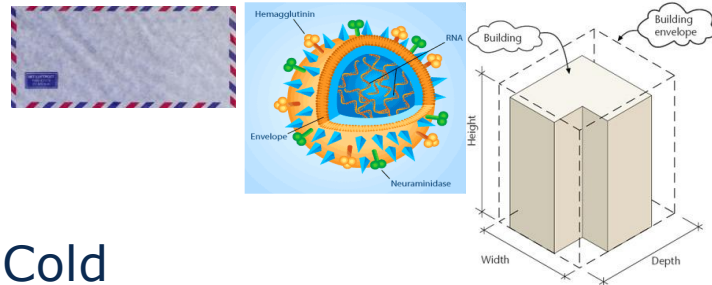
in literature search! (GoPubMed.org)

Ambiguity is also widespread  
 100% match of the term in text, but different meaning

## Spindle



## Envelope



## Cold



## Bank



## Example: development

in developmental biology:

*"...expression of multiple genes involved in **development**..."*

in biochemistry:

*"...the **development** of the Na<sup>+</sup> gradient..."*

in medicine:

*"...the **development** of the patient care plan..."*

in human resources:

*"...staff **development** and continuing education instructors..."*

in software development:

*"...Bioconductor, an open **development** software project for..."*

Problem:

Which sense to choose?

Granularity

- **developmental biology vs. all others?**
- development in **biomedicine vs. all others?**
- development in **biochemistry vs. medicine?**
- correct identification of **all possible senses?**

Decide + /- wrt Ontology sense (GO, MeSH, UMLS)

# Word Sense Disambiguation

## Approaches in biomedicine

for resolving biomedical abbreviations (up to 98% accuracy)  
& gene name normalization (up to 86% for human genes)

### **Knowledge-based**

use dictionaries, thesauri, ontologies  
e.g., MeSH, UMLS, GOA, UniProt, EntrezGene  
cos similarity, metadata

*Schijvenaars et al. 2005*  
*Humphrey et al. 2006*  
*Hakenberg et al. 2008*  
*Farkas 2008*

### **Supervised**

learn a classifier from labeled training sets  
naïve Bayes, decision trees, SVM,  
word/term co-occurrences

*Hatzivassiloglou et al. 2001*  
*Ginter et al. 2004*  
*Liu et al. 2002, 2004*  
*Gaudan et al. 2005*  
*Pahikkala et al. 2005*

**Unsupervised** (based on unlabeled corpora)  
noun co-occurrences, Markov clustering  
WordNet

*Dorow & Widdows 2003*

# Word Sense Disambiguation In biomedicine

High accuracy (up to 98%)

**BUT** easier tasks:

- resolving biomedical abbreviations (long form usually in text)
- gene name normalization (metadata usually available)

# Word Sense Disambiguation In biomedicine

There are no approaches which exploit  
inference over ontologies,  
term similarity, and metadata

## 3 disambiguation approaches

### Term Cooc

identify sense by direct & inferred co-occurrences

`cell death' is part of `development'

`cell death' occurs often with `cell proliferation'

→ `cell proliferation' is indicative of developmental biology

### Closest Sense

find shortest path of cooc terms to one of the UMLS senses

e.g. `patient care plan' is far away from `development'

### MetaData

train a classifier on metadata (journal, title, data)

e.g. Michael Brand's papers are all about developmental biology

**"Biomedical word sense disambiguation with ontologies and metadata:  
automation meets accuracy"**

*Alexopoulou et al., 2009, BMC Bioinformatics*

# Term Cooc approach

Typical abstract:

"...during early eye **development** in *Drosophila*...

**Organogenesis** involves an initial surge of **cell proliferation**, leading to **differentiation**. This is followed by **cell death**..."

Principle: Terms have best friends

What is a **best friend**, e.g. for development?

1. Number of co-occurrences

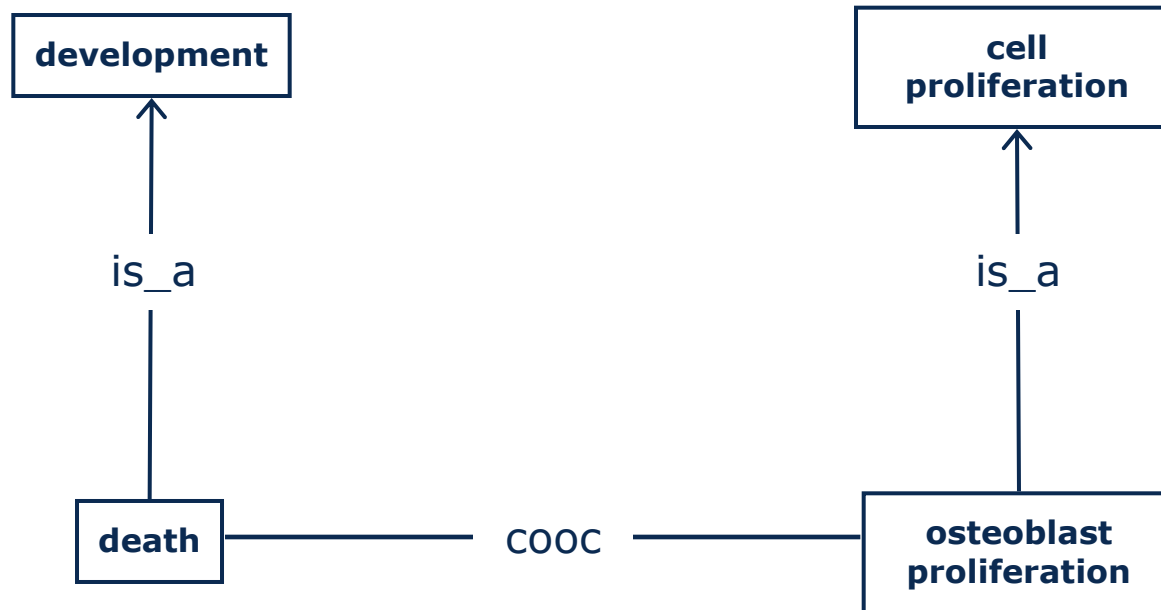
**Top cooc terms:** cell, nucleus, protein binding

2. Log-odds ratio:  $\log \frac{A-B_{cooc}}{A_{oc} B_{oc}}$

**Best friends:** cell proliferation, differentiation, transcription initiation factor activity

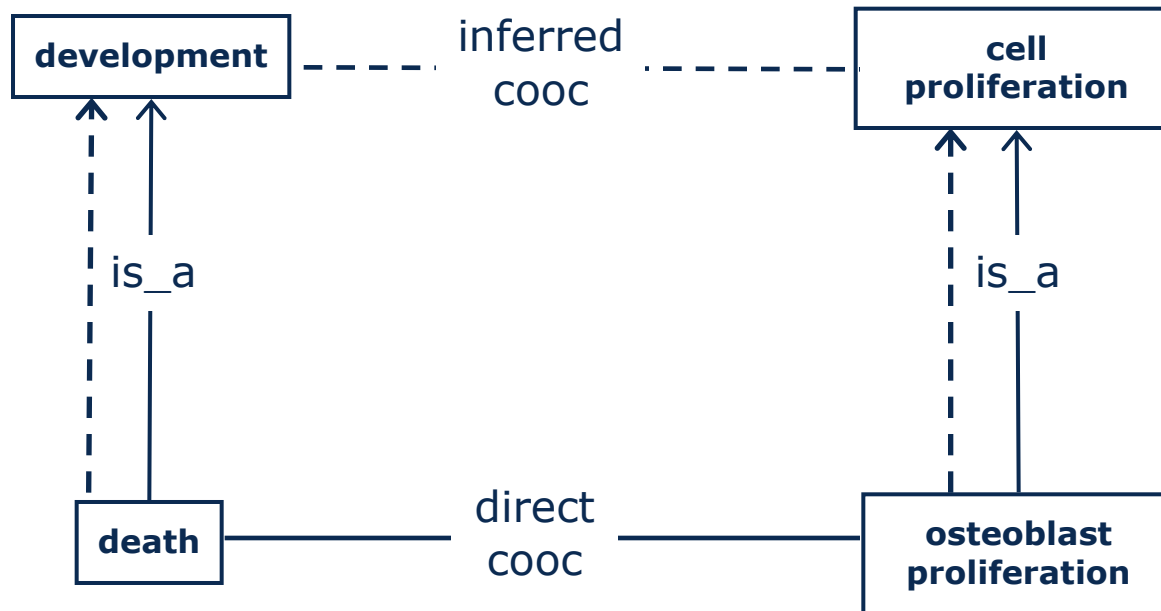
# Inferred Cooc

## Boosting co-occurrences using the ontology



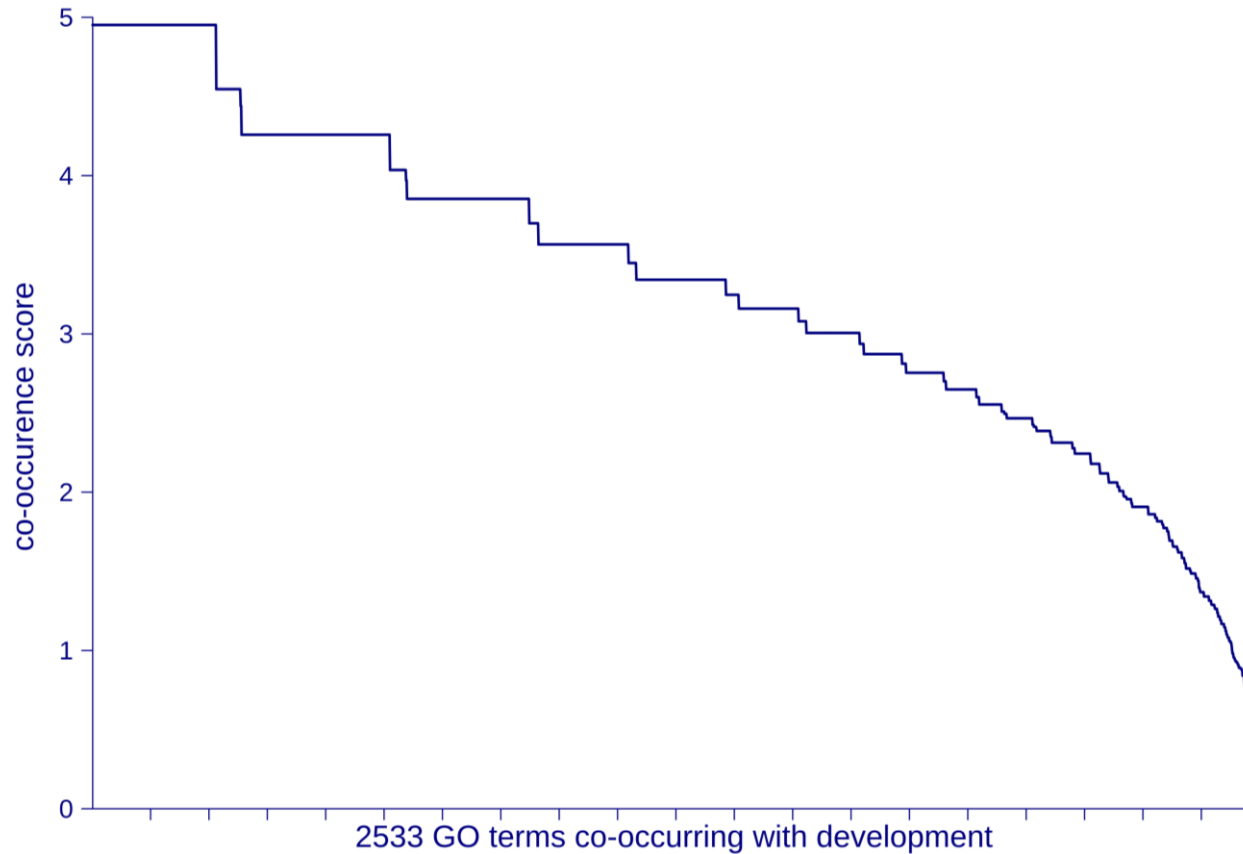
# Inferred Cooc

## Boosting co-occurrences using the ontology



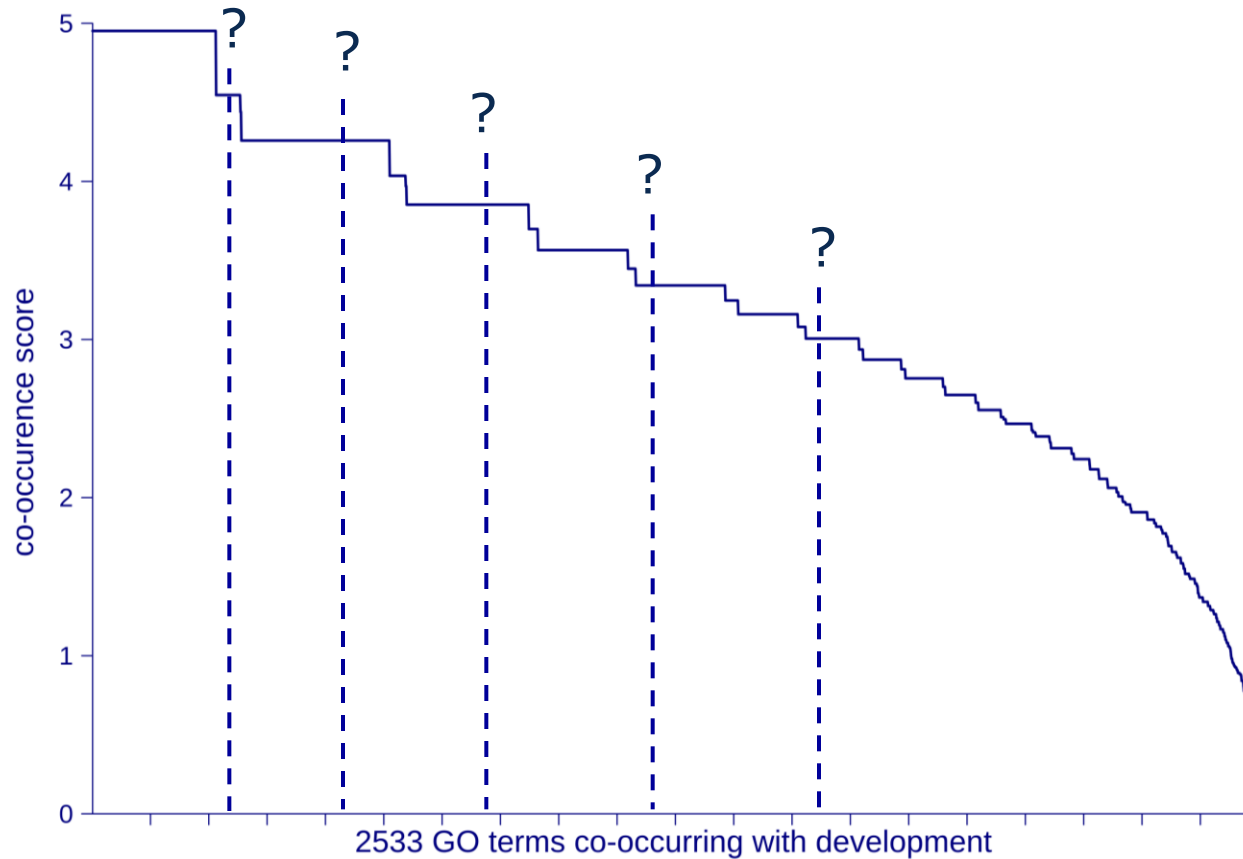
# Setting the threshold

Problem: where to cut?



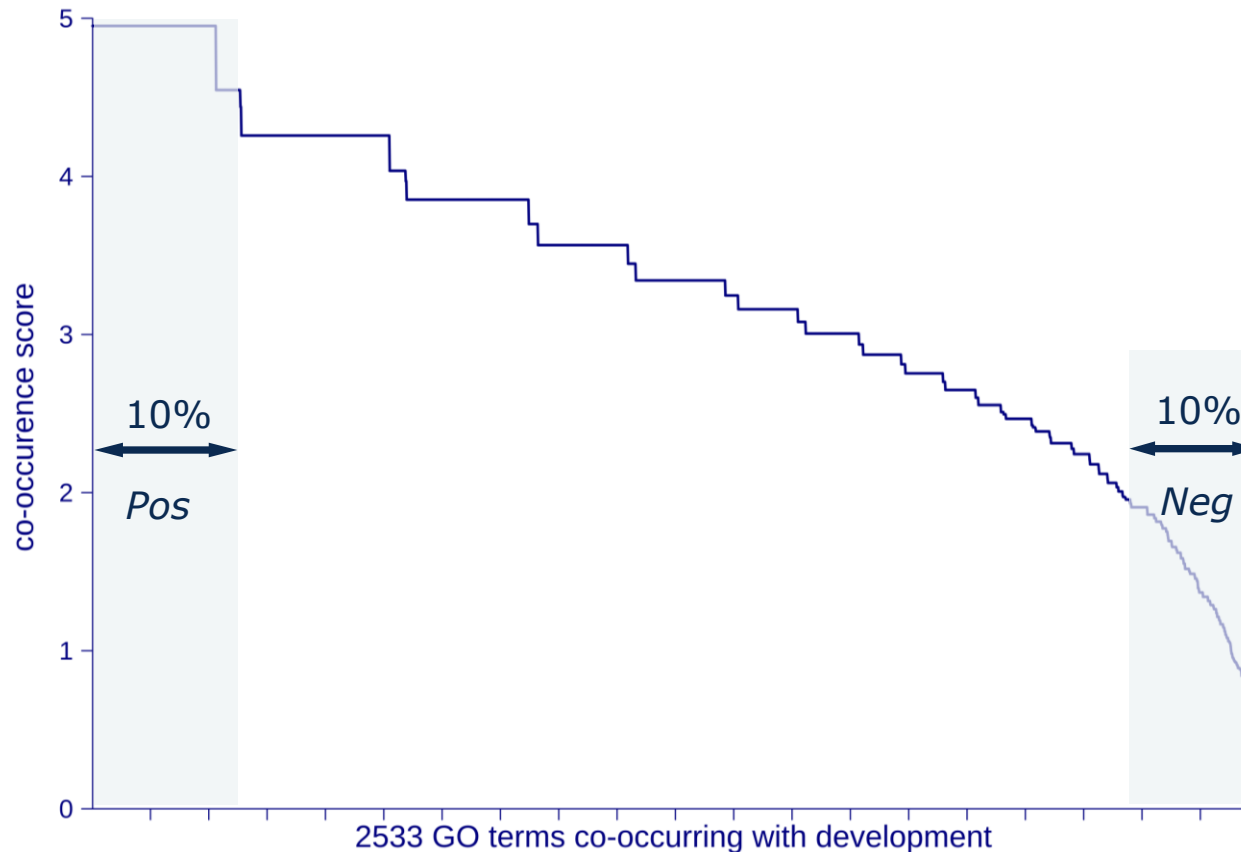
# Setting the threshold

Problem: where to cut?



# Setting the threshold

Solution: label left 10% *Pos*, right 10% *Neg*



Use *Pos/Neg* as training examples & iteratively train SVM

## Closest Sense approach

Computes the shortest path of cooc terms in the document to one of the senses of the ambiguous term

Relies on:

- A formal model (i.e. ontology)
- Semantic distances using:
  - hierarchy of concepts
  - hierarchy of properties
- Similarities (using semantic distances) between the ambiguous term & its neighbours in the text

## Closest Sense example

"I also tracked **lipid profiles**, **HBA1C**, **blood pressure**, **body mass index**, **hostility**, and nicotine use"

3 senses for '**blood pressure**':

Organism Function   Diagnostic Procedure   Lab or Test Result

**lipid profile** is closest to Diagnostic Procedure

**HBA1C** is distant to all senses

**body mass index** is closest to Diagnostic Procedure

**hostility** is closest to Organism Function

Overall: Diagnostic Procedure is closest to all terms in context

## MetaData approach

Metadata make indirect use of human intelligence

- Editors choose articles for journals  
e.g. journal 'Development' is about 'biological development'
- Authors continuously work on a subject  
e.g. all articles of 'Michael Brand' are on 'biological development'
- Maximum Entropy modeling:  
provided training examples, extract set of relationships
- Training example: sentence (manual collection)
- Features: n-tuples of word stems + metadata
- Metadata: journal, paper title, publication period

## MetaData examples

Thrush the mouth disease [MeSH]

(+) *Journal of Oral Hygiene*

Inhibition in Psychology [MeSH]

(+) journal '*Physiol Behav*', 'conditioned stimulus', 'emotion regulation', 'temperamental', journal '*J Abnorm Psychol*'

(-) diabetes, pH, tumor, antibody, enzyme, protein, membrane

Development [GO]

(+) signal transduction, kinase, embryo, neuron, stage

# 7 ambiguous terms

	Term	# senses in WordNet	# senses in Wiktionary	some senses in PubMed
<b>GO</b>	Development	9	7	<b>maturation</b> , staff -, algorithm -, method -
	Spindle	5	6 (4 N, 2 V)	<b>mitotic</b> , sleep spindles, spindle-shaped cells, muscle spindles
	Nucleus	6	6	<b>cell nucleus</b> , caudate nucleus
	Transport	11 (6 N, 5 V)	10 (7 N, 3 V)	<b>protein transport</b> , patient-, trans. of virus cultures
<b>MeSH</b>	Thrush	3	3	<b>mouth disease</b> , songbird
	Lead	31 (17 N, 14 V)	10 (8 N, 2 V)	<b>heavy metal</b> , measurement, "to result in"
	Inhibition	4	1	<b>psychological inhibition</b> , metabolic -

## 3 benchmark datasets

2600 *manually* curated documents for 7 terms

**[GO]** development, spindle, nucleus, transport

**[MeSH]** thrush, lead, inhibition

### 1. Expert manual dataset: 100 true & 100 false docs per term

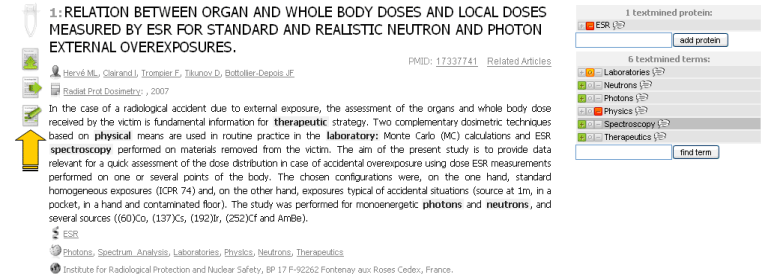
(give 50% chance to each appearance of the term to be true or false)

### 2. Non-expert manual dataset:

**1500** docs, automatic annotations

**manually** confirmed by non-experts

(unbalanced)



1: RELATION BETWEEN ORGAN AND WHOLE BODY DOSES AND LOCAL DOSES MEASURED BY ESR FOR STANDARD AND REALISTIC NEUTRON AND PHOTON EXTERNAL OVEREXPOSURES.

PMID: 17337741 Related Articles

thervé M., Clairand J., Tromper F., Thunoz D., Bittler-Depois J.F.

Radial Prot Dosimetry. 2007

In the case of a radiological accident due to external exposure, the assessment of the organs and whole body dose received by the victim is fundamental information for **therapeutic** strategy. Two complementary dosimetric techniques based on **physical** means are used in routine practice in the **laboratory**: Monte Carlo (MC) calculations and ESR **spectroscopy** performed on materials removed from the victim. The aim of the present study is to provide data relevant for a quick assessment of the dose distribution in case of accidental overexposure using dose ESR measurements performed on one or several points of the body. The chosen configurations were, on the one hand, standard homogeneous exposures (ICRP 74) and, on the other hand, exposures typical of accidental situations (source at 3m, in a pocket, in a hand and contaminated floor). The study was performed for monoenergetic **photons** and **neutrons**, and several sources ( $^{60}\text{Co}$ ,  $^{137}\text{Cs}$ ,  $^{192}\text{Ir}$ ,  $^{252}\text{Cf}$  and AmBe).

ESR

Photons; Spectrum Analysis; Laboratories; Physics; Neutrons; Therapeutics

Institute for Radiological Protection and Nuclear Safety, BP 17 F-92262 Fontenay aux Roses Cedex, France.

### 3. Semi-automatic dataset: 16,600 docs,

- cluster docs, add/delete clusters with significant phrases
- **manual** assignment of sense to each cluster

# 3 benchmark datasets

2600 *manually* curated documents for 7 terms



**1: RELATION BETWEEN ORGAN AND WHOLE BODY DOSES AND LOCAL DOSES MEASURED BY ESR FOR STANDARD AND REALISTIC NEUTRON AND PHOTON EXTERNAL OVEREXPOSURES.**



Hervé ML, Clairand J, Trompier F, Tikunov D, Bottollier-Depois JF

PMID: [17337741](#) [Related Articles](#)



Radiat Prot Dosimetry, 2007



In the case of a radiological accident due to external exposure, the assessment of the organs and whole body dose received by the victim is fundamental information for **therapeutic** strategy. Two complementary dosimetric techniques based on **physical** means are used in routine practice in the **laboratory**: Monte Carlo (MC) calculations and ESR **spectroscopy** performed on materials removed from the victim. The aim of the present study is to provide data relevant for a quick assessment of the dose distribution in case of accidental overexposure using dose ESR measurements performed on one or several points of the body. The chosen configurations were, on the one hand, standard homogeneous exposures (ICPR 74) and, on the other hand, exposures typical of accidental situations (source at 1m, in a pocket, in a hand and contaminated floor). The study was performed for monoenergetic **photons** and **neutrons**, and several sources ((60)Co, (137)Cs, (192)Ir, (252)Cf and AmBe).



ESR

[Photons](#), [Spectrum Analysis](#), [Laboratories](#), [Physics](#), [Neutrons](#), [Therapeutics](#)

Institute for Radiological Protection and Nuclear Safety, BP 17 F-92262 Fontenay aux Roses Cedex, France.

## 2. Non-expert manual dataset:

**1500** docs, automatic annotations

**manually** confirmed by non-experts

(unbalanced)

1 textmined protein:

+ ESR

6 textmined terms:

+ Laboratories

+ Neutrons

+ Photons

+ Physics

+ Spectroscopy

+ Therapeutics

## 3 benchmark datasets

2600 *manually* curated documents for 7 terms

**[GO]** development, spindle, nucleus, transport

**[MeSH]** thrush, lead, inhibition

### 1. Expert manual dataset: 100 true & 100 false docs per term

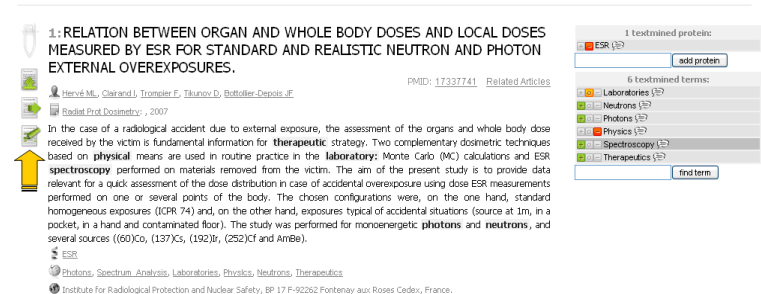
(give 50% chance to each appearance of the term to be true or false)

### 2. Non-expert manual dataset:

**1500** docs, automatic annotations

**manually** confirmed by non-experts

(unbalanced)



1: RELATION BETWEEN ORGAN AND WHOLE BODY DOSES AND LOCAL DOSES MEASURED BY ESR FOR STANDARD AND REALISTIC NEUTRON AND PHOTON EXTERNAL OVEREXPOSURES.

Herz M, Cibirka J, Tromper F, Tisano D, Buttler-Depois JF. PMID: 17337741 Related Articles

Radiat Prot Dosimetry. 2007

In the case of a radiological accident due to external exposure, the assessment of the organs and whole body dose received by the victim is fundamental information for **therapeutic** strategy. Two complementary dosimetric techniques based on **physical** means are used in routine practice in the **laboratory**: Monte Carlo (MC) calculations and ESR **spectroscopy** performed on materials removed from the victim. The aim of the present study is to provide data relevant for a quick assessment of the dose distribution in case of accidental overexposure using dose ESR measurements performed on one or several parts of the body. The chosen configurations were, on the one hand, standard homogeneous exposures (ICRP 74) and, on the other hand, exposures typical of accidental situations (source at 1m, in a pocket, in a hand and contaminated floor). The study was performed for monoenergetic **photons** and **neutrons**, and several sources ( $^{60}\text{Co}$ ,  $^{137}\text{Cs}$ ,  $^{192}\text{Ir}$ ,  $^{252}\text{Cf}$  and  $\text{AmBe}$ ).

ESR

Photons; Spectrum Analysis; Laboratories; Physics; Neutrons; Therapeutics

Institute for Radiological Protection and Nuclear Safety, BP 17 F-92262 Fontenay aux Roses Cedex, France.

1 text mined protein:  
 ESR (1/2)

6 text mined terms:  
 Laboratories (1/2)  
 Neutrons (1/2)  
 Photons (1/2)  
 Physics (1/2)  
 Spectroscopy (1/2)  
 Therapeutics (1/2)

### 3. Semi-automatic dataset: 16,600 docs,

- cluster docs, add/delete clusters with significant phrases
- **manual** assignment of sense to each cluster

# Results (% *f*-measure)

Term	Method									Avg
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79	90	80	95	<b>98</b>	100	77	78	<b>85</b>
Nucleus	89	94	81	78	75	<b>95</b>	99	91	77	<b>87</b>
Transport	83	71	90	89	88	<b>94</b>	98	91	88	<b>88</b>
Thrush	88	94	87	82	<b>78</b>	81	94	94	58	84
Lead	36	53	89	49	<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

# Results (% *f*-measure)

Term	Method									
	Closest Sense	Term Cooc				MetaData				
		CS	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	Avg	
Development	87	71	57	<b>79</b>	96	80	80	<b>79</b>		
Spindle	75	79	90	80	95	<b>98</b>	100	77	78	<b>85</b>
Nucleus	89	94	81	78	75	<b>95</b>	99	91	77	<b>87</b>
Transport	83	71	90	89	88	<b>94</b>	98	91	88	<b>88</b>
Thrush	88	94	87	82	<b>78</b>	81	94	94	58	84
Lead	36	53	89	49	<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**GO terms**

# Results (% *f*-measure)

Term	Method									
	Closest Sense		Term Cooc				MetaData			Avg
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79	90	80	95	<b>98</b>	100	77	78	<b>85</b>
Nucleus	89	88	88	88	75	<b>95</b>	99	91	77	<b>87</b>
Transport	85	85	85	89	88	<b>94</b>	98	91	88	<b>88</b>
Thrush	87	94	87	82	<b>78</b>	81	94	94	58	84
Lead	36	53	89	49	<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**MeSH terms**

# Results (% *f*-measure)

Term	Method									Avg
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	80	95	80	95	<b>98</b>	100	77	78	<b>85</b>
Nucleus	87	87	91	89	88	<b>94</b>	99	91	77	<b>87</b>
Transport	83	71	90	89	88	<b>94</b>	98	91	88	<b>88</b>
Thrush	88	94	87	82	<b>78</b>	81	94	94	58	84
Lead	36	53	89	49	<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**CS1: only hierarchy of senses**  
**CS2: senses + properties + optimal weights**

# Results (% *f*-measure)

Term	Method									Avg
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79	91	89	89	89	99	77	78	<b>85</b>
Nucleus	89	94	88	88	88	88	99	91	77	<b>87</b>
Transport	83	71	91	89	89	89	98	91	88	<b>88</b>
Thrush	88	94	87	82	<b>78</b>	81	94	94	58	84
Lead	36	53	89	49	<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**Delevopment and Lead were harder to disambiguate**

# 7 ambiguous terms

	Term	# senses in WordNet	# senses in Wiktionary	some senses in PubMed
<b>GO</b>	Development	9	7	<b>maturation</b> , staff -, algorithm -, method -
	Spindle	5	6 (4 N, 2 V)	<b>mitotic</b> , sleep spindles, spindle-shaped cells, muscle spindles
	Nucleus	6	6	<b>cell nucleus</b> , caudate nucleus
	Transport	11 (6 N, 5 V)	10 (7 N, 3 V)	<b>protein transport</b> , patient-, trans. of virus cultures
<b>MeSH</b>	Thrush	3	3	<b>mouth disease</b> , songbird
	Lead	31 (17 N, 14 V)	10 (8 N, 2 V)	<b>heavy metal</b> , measurement, "to result in"
	Inhibition	4	1	<b>psychological inhibition</b> , metabolic -

# Results (% *f*-measure)

Term	Method									Avg
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79				<b>98</b>	100	77	78	<b>85</b>
Nucleus	89	94				<b>95</b>	99	91	77	<b>87</b>
Transport	83	71				<b>94</b>	98	91	88	<b>88</b>
Thrush	88	94				81	94	94	58	84
Lead	36	53	89	49	92	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	80	84	73	82	84	<b>96</b>	81	70	80

**MetaData performs best**  
(with high quality training data)

# Results (% *f*-measure)

Term	Method									
	Closest Sense		Term Cooc				MetaData			Avg
	CS1	CS2	cooc	inf cooc	cooc+ SVM	inf cooc+ SVM	High	Med	Low	
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79	90	80	95	<b>98</b>	100	77	78	<b>85</b>
Nucleus	89					<b>95</b>	99	91	77	<b>87</b>
Transport	83					<b>94</b>	98	91	88	<b>88</b>
Thrush	88					<b>81</b>	94	94	58	84
Lead	36	53	89	49		81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	80	84	73	82	84	<b>96</b>	<b>81</b>	<b>70</b>	80

**MetaData performance strongly depends on training data quality**

# Results (% *f*-measure)

Term	Method									
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	Avg
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79							78	<b>85</b>
Nucleus	89	94							77	<b>87</b>
Transport	83	71							88	<b>88</b>
Thrush	88	94							58	84
Lead	36	53	89		<b>93</b>	81	85	36	14	<b>60</b>
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	82	84	96	81	70	80

**Closest Sense  
better with  
hierarchy of senses  
+ hierarchy of properties  
+ optimal weights**

# Results (% f-measure)

Term	Method									
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	Avg
Development	87	86	74	71	57	<b>79</b>	96	80	80	<b>79</b>
Spindle	70	79	90	80	95	<b>98</b>		77	78	<b>85</b>
Nucleus	89	94	81	78	75	<b>95</b>	9	81	77	<b>87</b>
Transport	83	71	90	89	88	<b>94</b>				
Thrush	88	94	87	82	<b>78</b>	81				
Le		83	89	49	<b>93</b>	81				
Ir		84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**GO terms**

**Term Cooc performs best with inferred cooc & SVM**

# Results (% *f*-measure)

Term	Method									
	Closest Sense		Term Cooc				MetaData			
	CS1	CS2	cooc	inf cooc	cooc+SVM	inf cooc+SVM	High	Med	Low	Avg
Development	87	86	74	71	57	79	96	80	80	79
Spindle	77	77	80	80	95	98	96	77	78	85
Nucleus	88	88	78	78	75	95	96	81	77	87
Transport	85	71	90	89	88	94	96	81	77	87
Thrush	88	94	87	82	<b>78</b>	81	96	81	77	87
Lead	36	53	89	49	<b>93</b>	81	96	81	77	87
Inhibition	66	84	77	62	<b>85</b>	58	100	95	97	80
<b>Avg</b>	74	<b>80</b>	84	73	<b>82</b>	84	<b>96</b>	81	70	80

**MeSH terms**

**Term Cooc performs worse with inferred cooc & SVM**

# Results (% *f*-measure)

Term	Method			Avg		
	Closest	...	...			
Development	<p><b><u>GO is "tall &amp; thin"</u></b></p> <ul style="list-style-type: none"> <li>• few children per node</li> <li>• many levels</li> <li>• max # of levels = 19</li> </ul> <p><b><u>MeSH is "short &amp; fat"</u></b></p> <ul style="list-style-type: none"> <li>• many children per node</li> <li>• few levels</li> <li>• max # of levels = 9 (<i>MeSH 2007</i>)</li> </ul>			79		
Spindle				85		
Nucleus				87		
Transport				88		
Thrush				84		
Lead				60		
Inhibition				80		
<b>Avg</b>				74	70	80

# Conclusion

**MetaData:** 80-100% success rate, ~10% due to metadata

+ no ontology needed

- high quality training data needed

- prone to overfitting (when few training examples)

**Term Cooc:** ~ 80% success rate, up to 5-15% due to inference

+ (almost) no training data needed

- well modelled ontology needed

**Closest Sense:** ~ 80% success rate

+ no training data

- large, well modelled ontology needed

- sense needs to be in the ontology

# Take home message

## Cautious recommendation

Ontology + inference **can** improve WSD

**BUT** depend on the type of the ontology (size, type of relations)

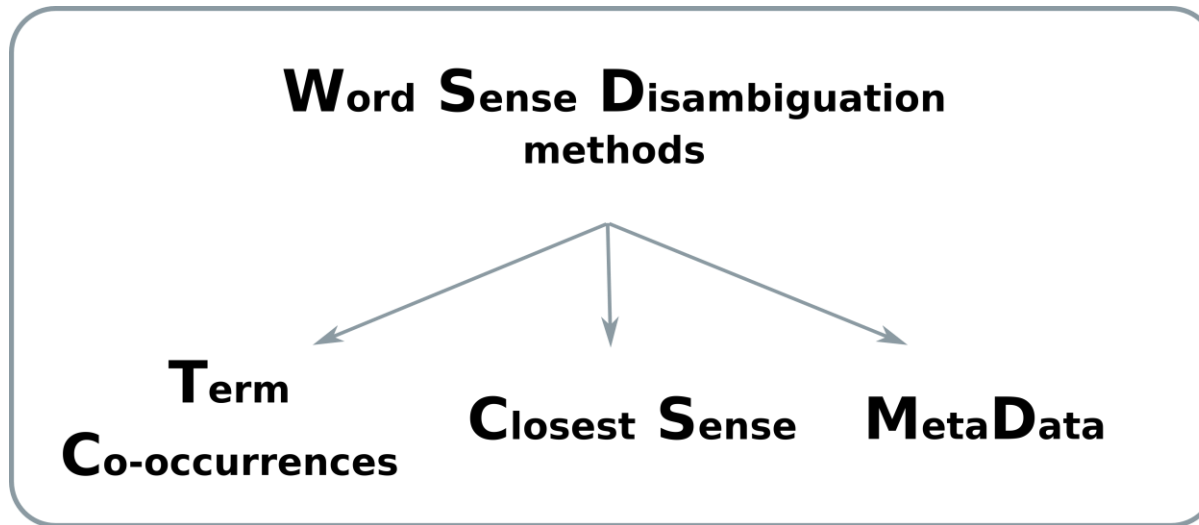
Metadata works **well** as indirect human input

**BUT** we need to obtain it in a scalable way

## Limits – open questions

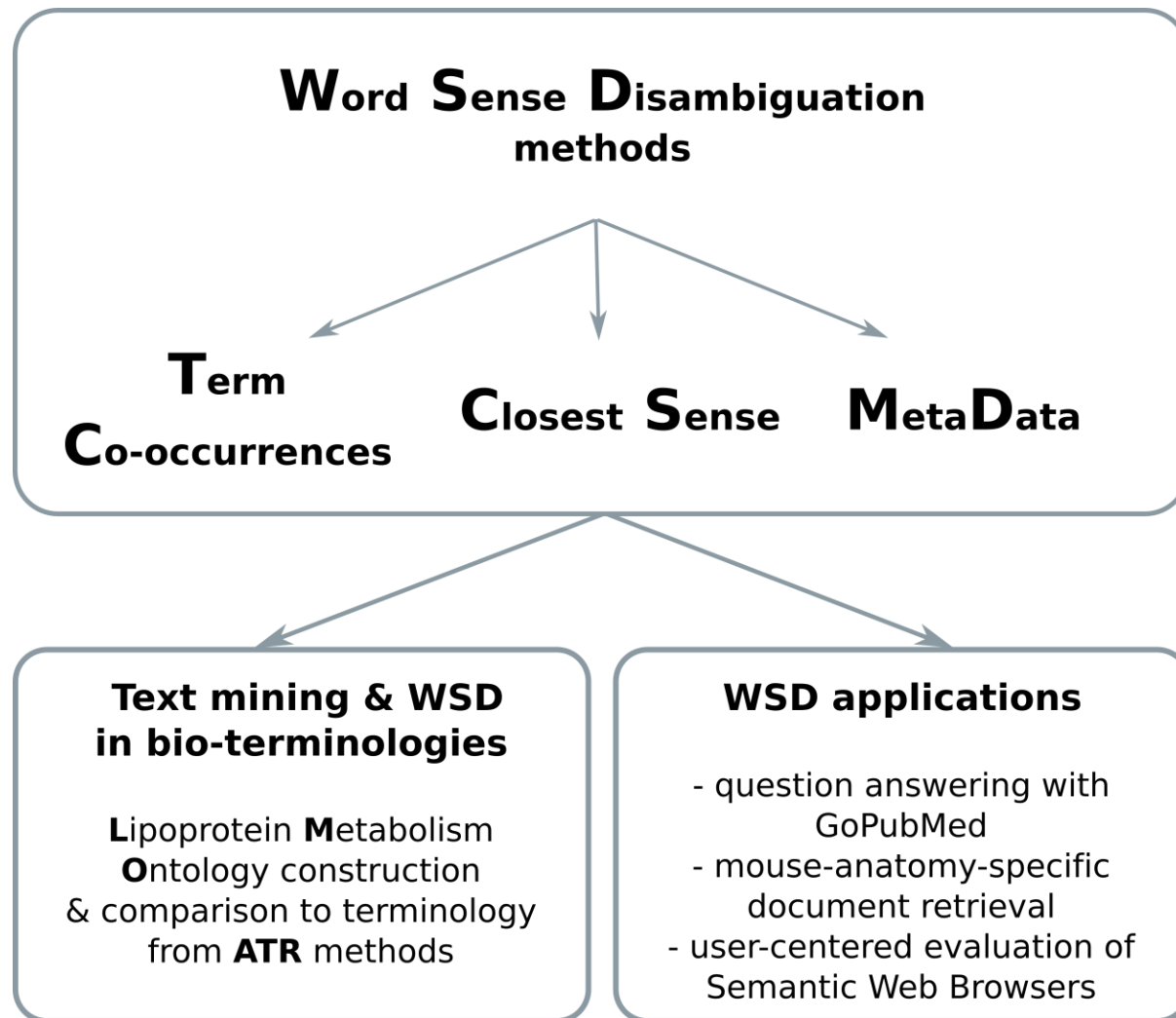
- one abstract = one sense?  
e.g. "...Bioconductor, an open source and **development** software...on a heart **development** dataset..."
- only 7 terms, 2600 docs;  
how can we collect more documents?
- how to identify ambiguous terms automatically?

# My Contributions



- Term cooc – Inferred cooc
- Manual document collection
- Semi-automated document collection

# Overview of PhD work



# Acknowledgments

B. Andreopoulos, J. Hakenberg, A. Doms (BIOTEC, TU Dresden)

K. Khelif, F. Gandon (INRIA Sophia-Antipolis)

T. Wächter, H. Dietze

G. Tsatsaronis (BIOTEC, TU Dresden)

Prof. M. Schroeder

[www.transinsight.com](http://www.transinsight.com)



# Publications

**“Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy”**, Alexopoulou et al., *BMC Bioinformatics* 2009, 10:28

**“Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering”**, Andreopoulos et al., *International Journal of Data Mining and Bioinformatics* 2008, 2(3): 193-215

**“Terminologies for text-mining; an experiment in the lipoprotein metabolism domain”**, Alexopoulou et al., *BMC Bioinformatics* 2008, 9:S4

**“A User-Centred Evaluation Framework for the SeaLife Semantic Web Browsers”**, Oliver et al., *BMC Bioinformatics* 2009, (SWAT4LS workshop special issue)