43rd European Conference on Information Retrieval

Doctoral Consortium

Deep Learning System for Biomedical Relation Extraction Combining External Sources of Knowledge

Diana Sousa

LASIGE, Faculty of Sciences, University of Lisbon, Portugal

March 28 to April 1, 2021

Presentation Outline

- Motivation
- Background
- State-of-the-art
- Problem Statement
- Research Questions
- Research Work
  - Baseline system + ontologies
  - Phenotype-Gene Relations (PGR) improvement
- Conclusion
- Future Work
- Mentor Suggestions

Motivation

What are the **diseases** associated with **BRCA1**?

Motivation

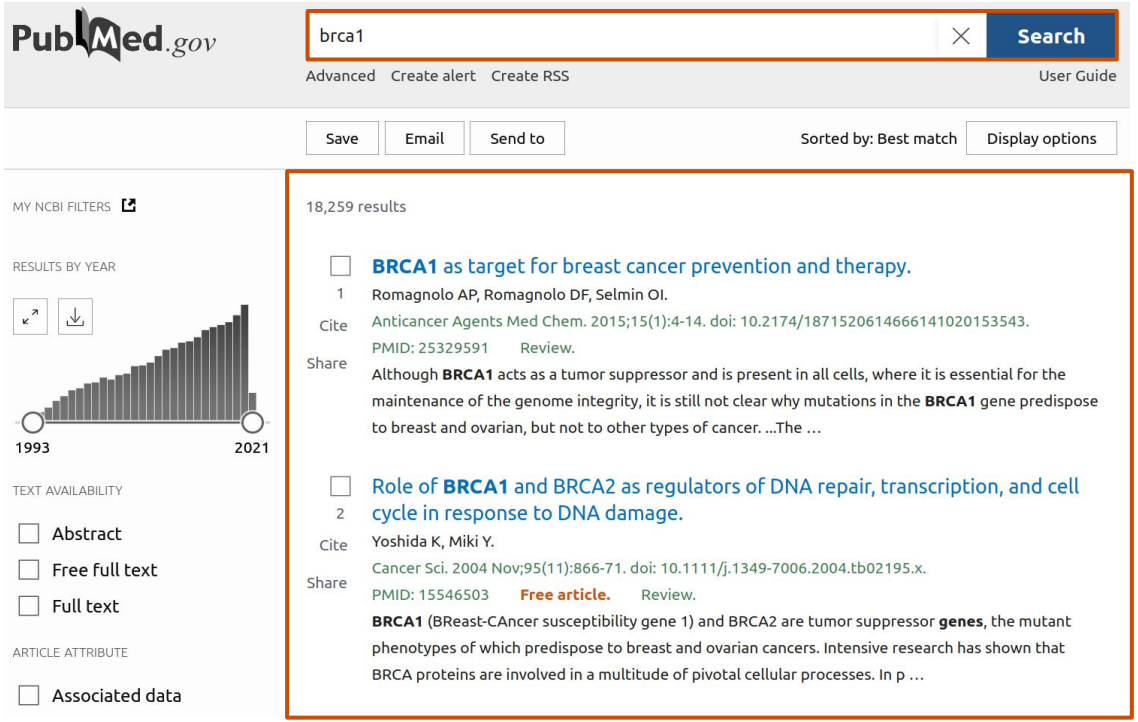
What are the **diseases** associated with **BRCA1**?

Biomedical Literature

- breast cancer
- ovarian cancer
- **others?**

Online Information
PubMed

Motivation



What are the **diseases** associated with **BRCA1**?

- breast cancer
- ovarian cancer
- **prostate cancer**
- **others?**

PubMed

32 million citations

- ❑ unstructured textual information
- ❑ highly heterogeneous
- ❑ time-consuming task

Motivation

What are the **diseases** associated with **BRCA1**?

- breast cancer
- ovarian cancer
- prostate cancer
- **Alzheimer's disease**

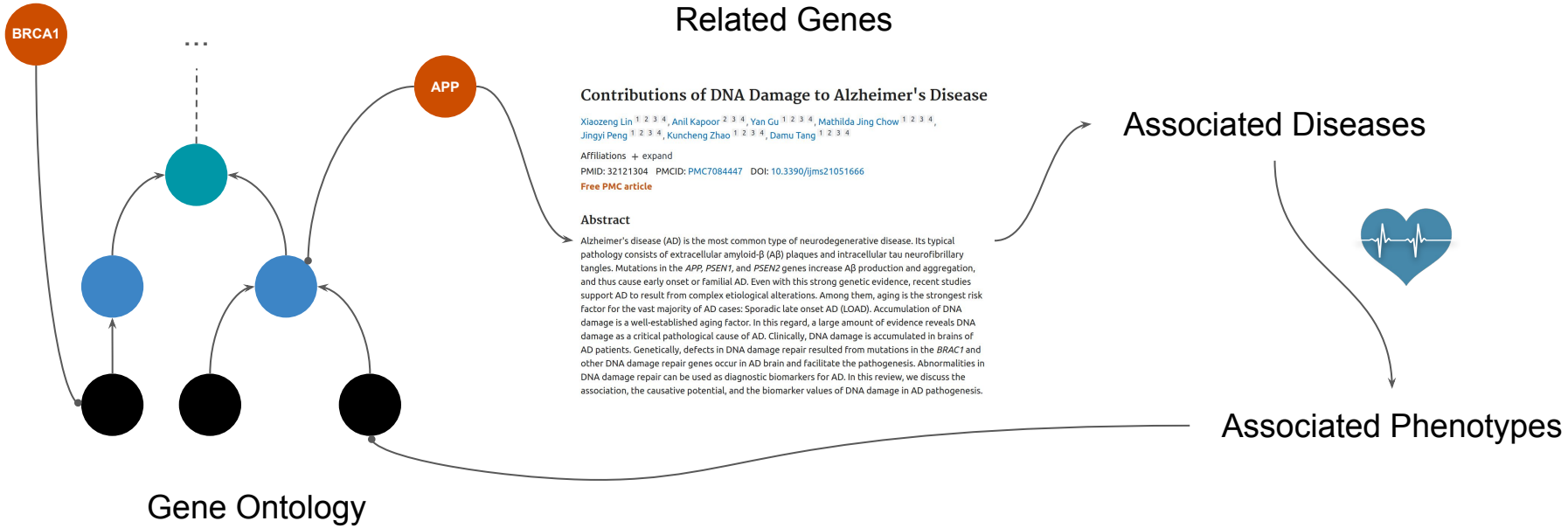
What are other **genes** associated with **Alzheimer's disease**?

What are the **human phenotypes** associated with **Alzheimer's disease**?

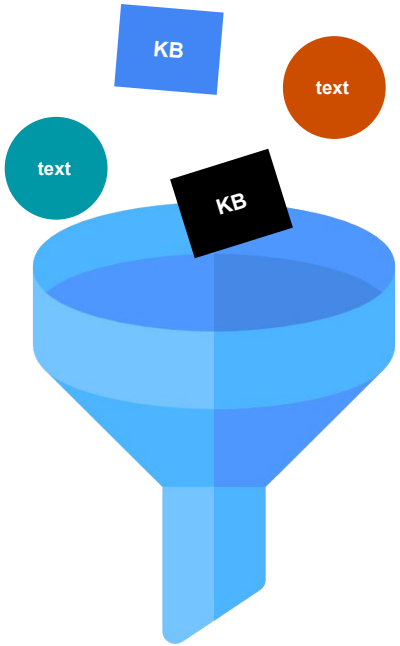
How to capture semantic information about **BRCA1**?

Motivation

How to capture semantic information about **BRCA1**?



Background



Knowledge distillation

Text Mining

BRCA1 relates to **Alzheimer's disease**

Alzheimer's disease relates to **APP**

APP has a common ancestor with **BRCA1**

APP and **BRCA1** common ancestor is related to the **phenotype aging**

Background

*Mutations in the **APP** gene increase Aβ production and aggregation, and thus cause early onset or familial **Alzheimer's disease**.*

Text Mining Tasks

- Named-Entity Recognition (NER) **APP**, Offset 17-20 | **Alzheimer's disease**, Offset 105-124
- Named-Entity Linking (NEL) **APP**, 351 (NCBI) | **Alzheimer's disease**, DOID:10652 (DO)
- **Relation Extraction (RE)** <**APP**, **Alzheimer's disease**, *true*>

Background

Approaches for Relation Extraction

Co-occurrence

Pattern/Rule-based

Supervised Machine Learning

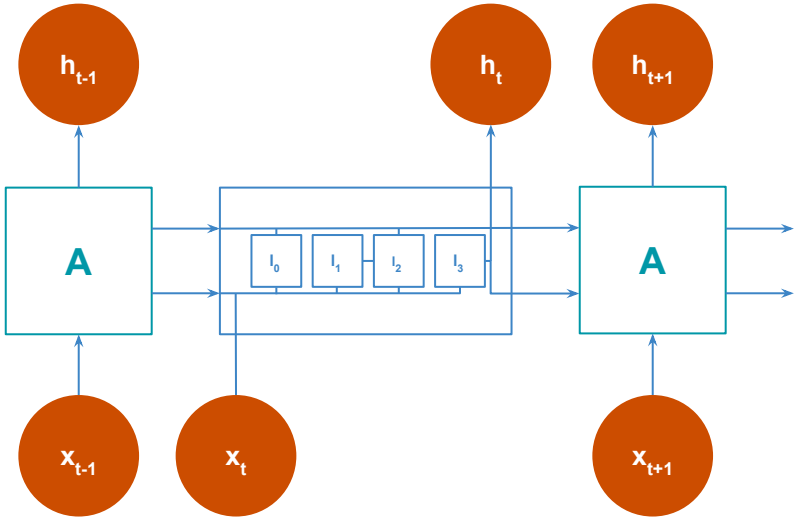
output is known

predict class or value label

State-of-the-art

Approaches for Relation Extraction

Deep Learning - Long Short-Term Memory (LSTM)



long and descriptive sentences

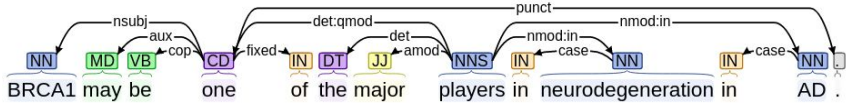
bidirectional LSTMs

State-of-the-art

Approaches for Relation Extraction

Data Representations

BRCA1 *may be one of the major players in neurodegeneration in* ***AD****.*



Shortest Dependency Path (SDP)

- Part-of-speech (POS)

  BRCA1$_{NP.e1}$ - be$_{VB}$ - one$_{NP}$ - of$_{PP}$ - players$_{NP}$ - in$_{PP}$ - **AD**$_{NP.e2}$

- Word Embeddings

  BRCA1 - be - one - of - players - in - **AD**

lookup table

	S1	S2	S3
BRCA1	1	0	1
	0	...	3
AD	2	0	1

word vector

sentence vector

State-of-the-art

Evaluation

Supervised Machine Learning

70%

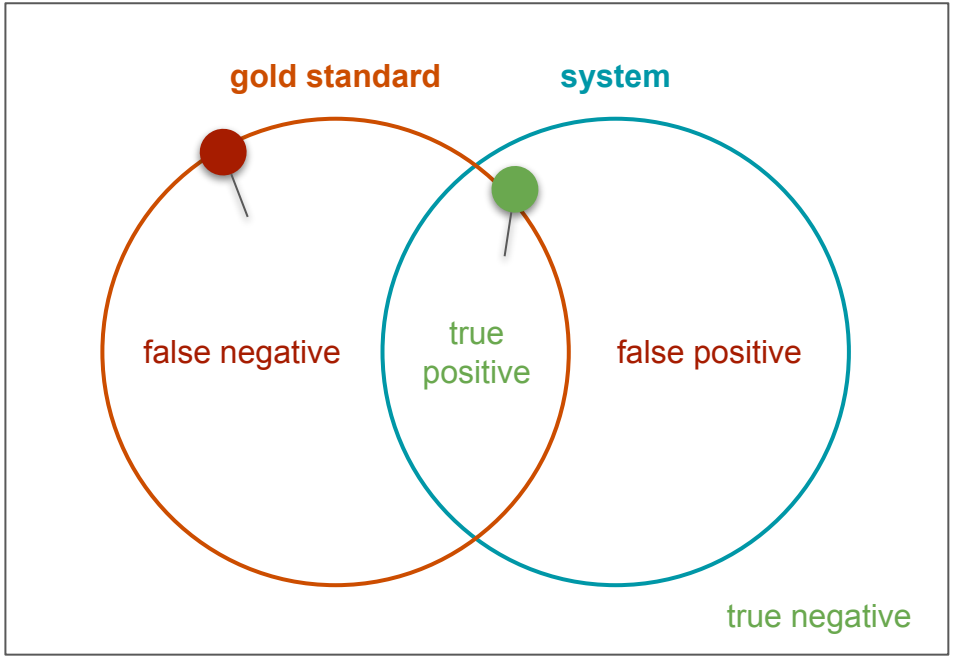
training set + validation set

30%

test set

State-of-the-art

Evaluation



1st scenario

gold standard
<**BRCA1**, **Alzheimer's disease**, *true*>

system
<**BRCA1**, **Alzheimer's disease**, *true*>

2nd scenario

gold standard
<**BRCA1**, **Alzheimer's disease**, *true*>

system
<**BRCA1**, **Alzheimer's disease**, *false*>

State-of-the-art

Evaluation

number of correct results identified

$$Recall = \frac{TP}{TP + FN}$$

how **often** the results are correct

$$Precision = \frac{TP}{TP + FP}$$

harmonic mean of precision and recall to express overall performance

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

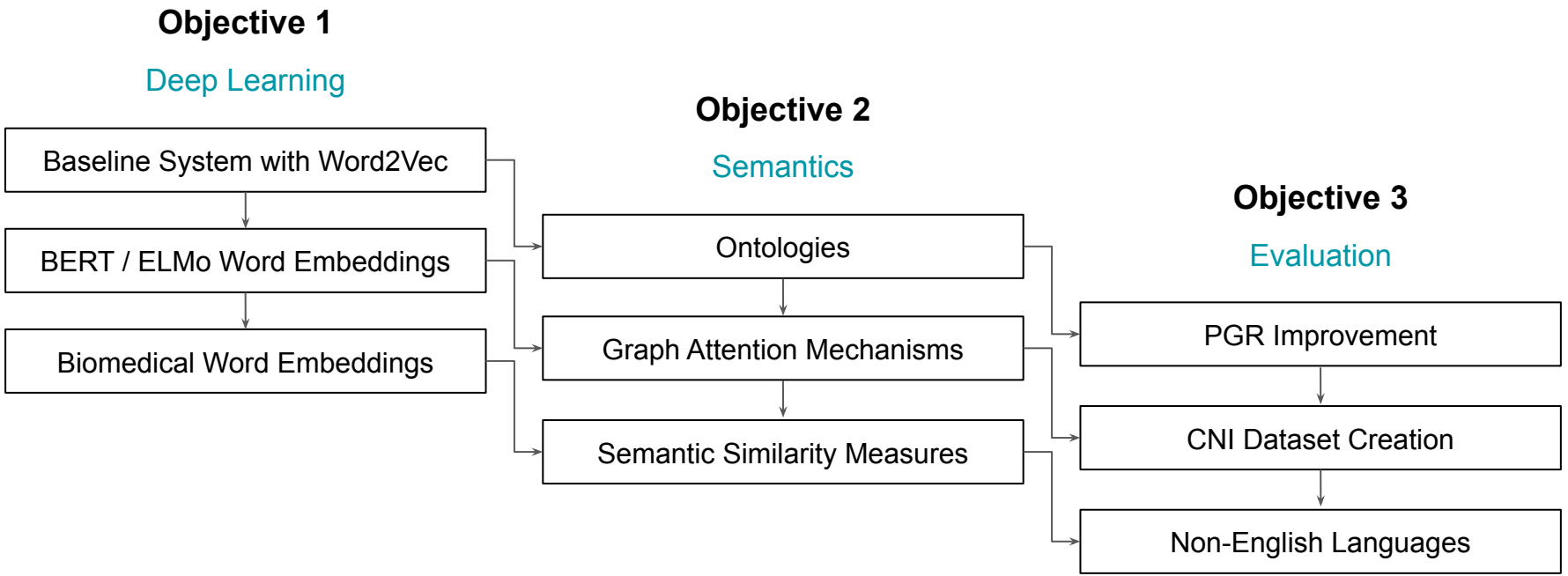
Problem Statement

1. Biomedical relation extraction (RE) systems are **scarce** and **limited**
2. Usually, do not resort to external sources of knowledge
3. Lack of annotated datasets (in English and other languages)

Research Questions

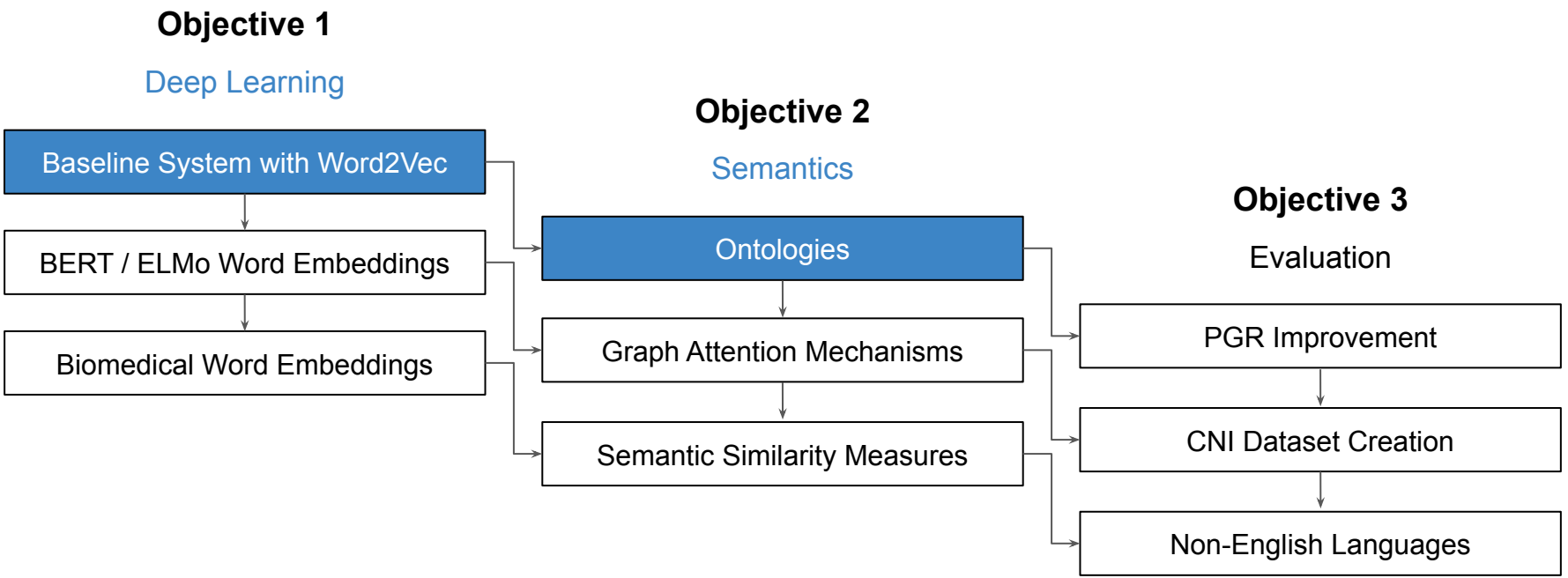
- Can the latest advances in language representations be used to create a state-of-the-art biomedical RE deep learning system?
- Can we use biomedical semantics as an add-on for RE systems?
- How can we evaluate RE systems regarding the biomedical domain in English and non-English languages?

Thesis Outline



PGR - Phenotype-Gene Relations Corpus
CNI - Cancer-Nutrition Interactions

Methodology



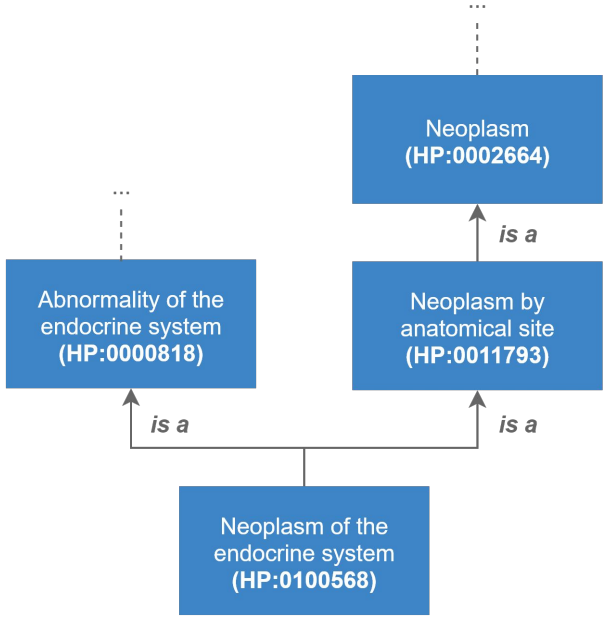
PGR - Phenotype-Gene Relations Corpus
CNI - Cancer-Nutrition Interactions

Methodology

Baseline System with Word2Vec → Ontologies

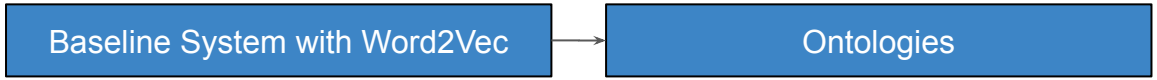
Ontologies

structured way of providing a common vocabulary in which shared knowledge is represented



Human Phenotype Ontology

Methodology



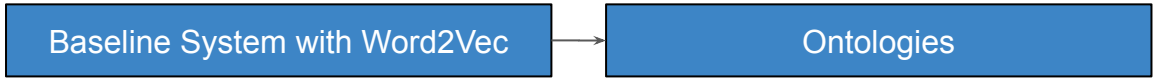
BR-LSTM System (Entity + Annotation Vector) - *Xu et al.* [2018]

Word Embeddings + WordNet Hypernyms + Unified Medical Language System (UMLS) concepts

3 channels
of information

limited to drug-drug interactions

Methodology



BiOnt System (Entity + Annotation Vector)

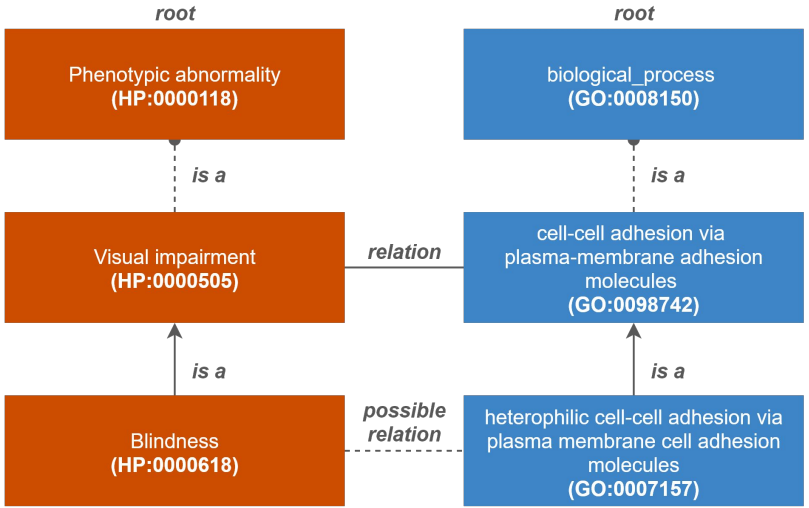
Word Embeddings + WordNet Hypernyms + **Ontology Embeddings**

4 channels of information

Methodology

Baseline System with Word2Vec → Ontologies

*The **CRB1** gene is a key target in the fight against **blindness**.*

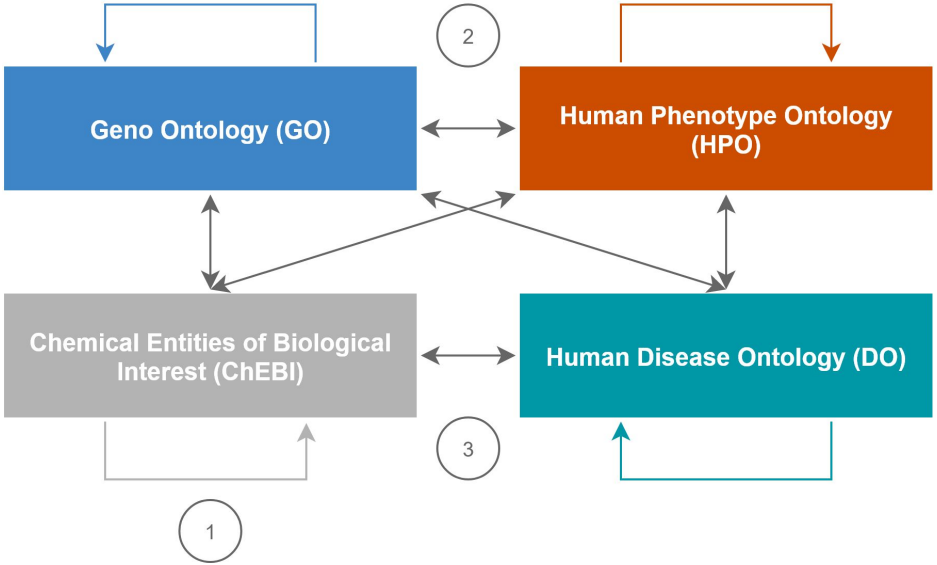


heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules **(GO:0007157)**

most representative GO term for the CRB1 gene

Methodology

Baseline System with Word2Vec → Ontologies



3 out of 10 possible combinations

1. Drug-Drug Interactions
2. Phenotype-Gene Relations
3. Chemical-Induced Disease Relations

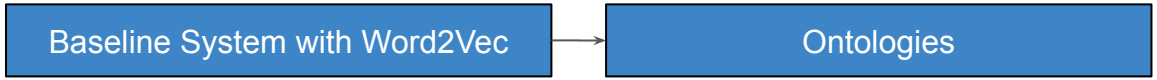
Results

Baseline System with Word2Vec → Ontologies

Dataset	System	Precision	Recall	F-Measure
DDI Corpus	BR-LSTM*	0.7134	0.6410	0.6753
	BiOnt	0.6784	0.7775	**0.7246**
PGR Corpus	BR-LSTM*	0.8421	0.6666	0.7442
	BiOnt	0.8438	0.7500	**0.7941**
BC5CDR Corpus	BR-LSTM*	0.5371	0.7264	0.6175
	BiOnt	0.5770	0.7173	**0.6396**

*without UMLS concepts

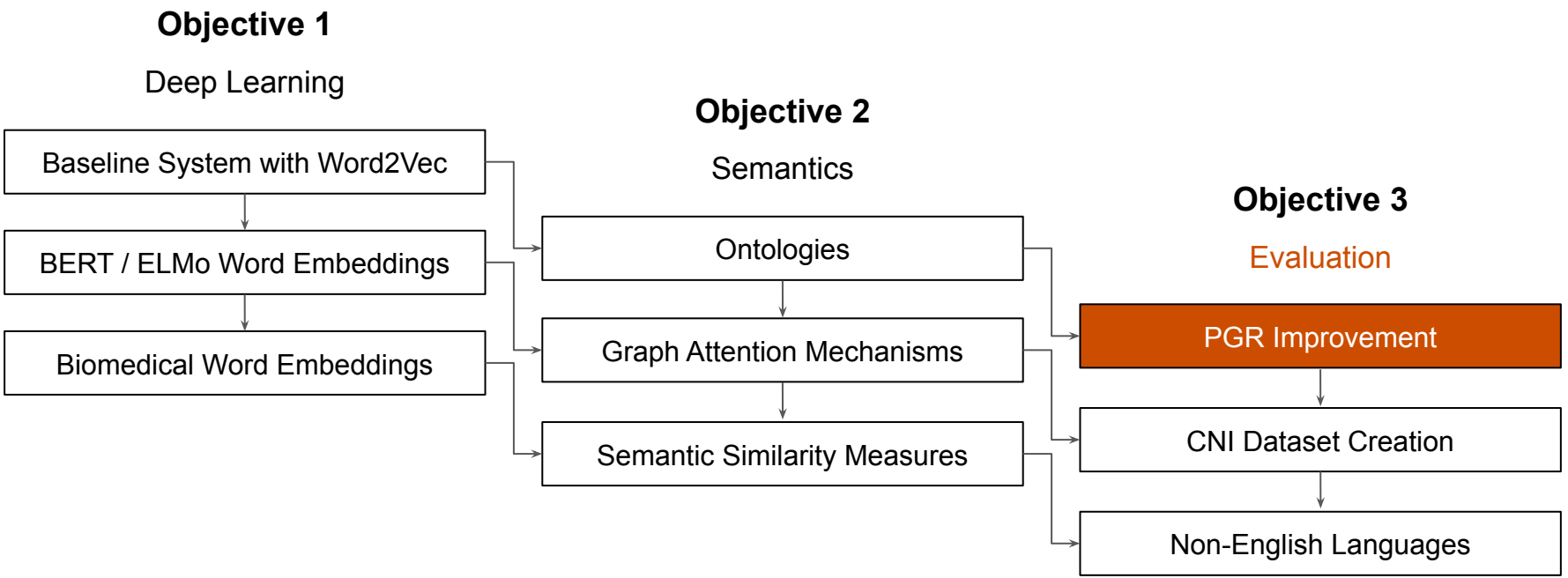
Main Takeaways



- Biomedical ontologies improve the performance of deep text mining systems
- The **BiOnt system** can be used to populate knowledge bases regarding gold standard relations
- Exploration of new experimental hypotheses can provide evidence about possible unknown associations between biomedical entities

Conference (Core A): BiOnt: Deep Learning Using Multiple Biomedical Ontologies for Relation Extraction (**Diana Sousa** and Francisco M. Couto) in the Proceedings of the 42nd European Conference on Information Retrieval, 2020

Methodology



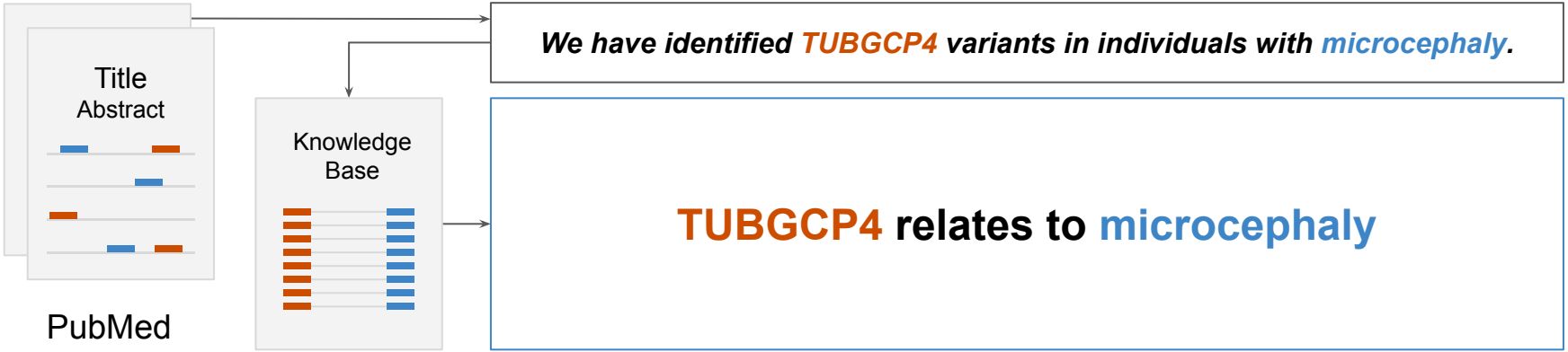
PGR - Phenotype-Gene Relations Corpus
CNI - Cancer-Nutrition Interactions

Methodology

PGR Improvement

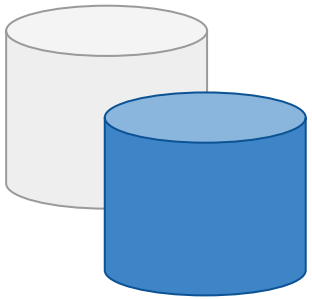
Silver Standard Dataset

Phenotype-Gene Relations (PGR)



Methodology

PGR Improvement



Knowledge Bases

Biomedical data retrieved from the **biomedical literature**

- Domain Expert Annotators
- Crowdsourcing Platforms
- Distantly Supervised Techniques

Methodology

PGR Improvement

- **Domain Expert Annotators** → More reliable
- Crowdsourcing Platforms → 
- **Distantly Supervised Techniques** → Less reliable
  - Still needs to be reviewed
  - Still time- and resource- consuming

Methodology

PGR Improvement

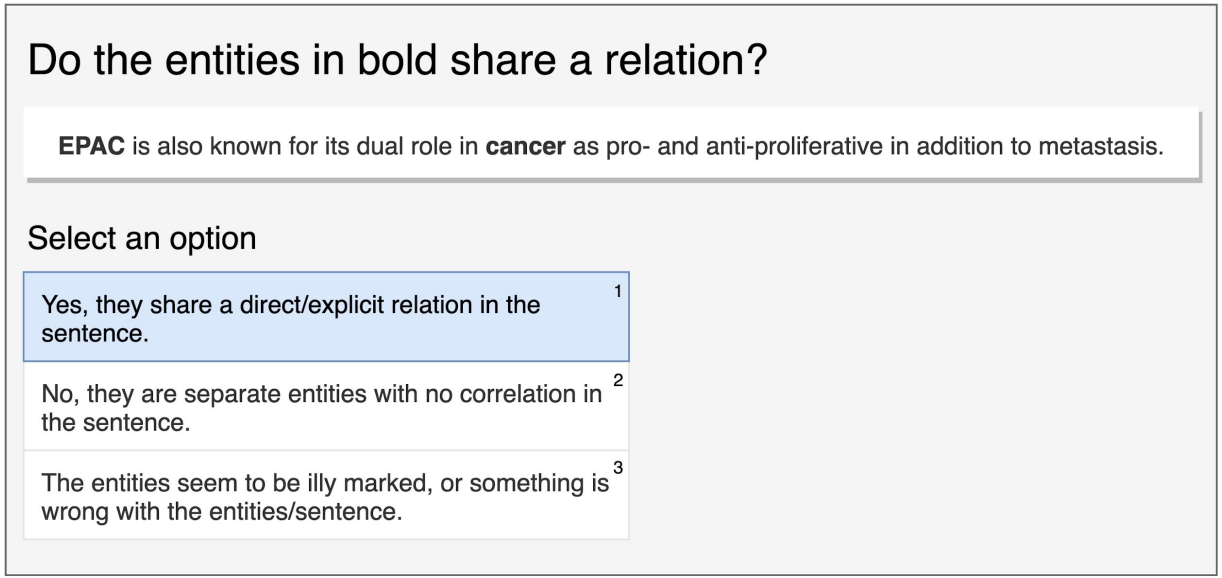
Distantly Supervised Techniques
PGR (7963 relations)

Crowdsourcing Platforms

amazon mechanical turk

Methodology

PGR Improvement



On 30%

\+ extra-rater on-site
\+ domain expert

Results

PGR Improvement

Evaluation on two different deep learning systems: **BiOnt and BioBERT**

Method		Precision	Recall	F-measure	Accuracy
BiOnt	PGR original	0.8140	0.3070	0.4459	0.4821
	Amazon (train) + PGR (test)	0.7000	**0.9825**	0.8175	0.7024
	Amazon (train) + Amazon / extra-rater consensus (test)	0.6810	0.9670	0.7992	0.6726
	Amazon (train) + Expert (test)	**0.8142**	0.9721	**0.8861**	**0.7989**
BioBERT	PGR original	**0.8542**	0.3445	0.4910	0.5143
	Amazon (train) + PGR (test)	0.6744	0.9856	0.8000	0.6775
	Amazon (train) + Amazon / extra-rater consensus (test)	0.6700	0.9763	0.7946	0.6680
	Amazon (train) + Expert (test)	0.8103	**0.9906**	**0.8915**	**0.8096**

Main Takeaways

PGR Improvement

- ~16% of relations were excluded by MTurk workers
- Extra rater's work had a two times superior cost than the revisions done by MTurk workers
- Crowdsourcing takes into account the low availability of experts for some domains
- The PGR improvement had a significant impact on model training: the BiOnt and BioBERT systems had an increase in average F-measure of 0.3494 percentual points

Journal (Q1 Scimago): A Hybrid Approach towards Biomedical Relation Extraction Training Corpora: Combining Distant Supervision with Crowdsourcing (**Diana Sousa**, Andre Lamurias, and Francisco M. Couto) in Database: The Journal of Biological Databases and Curation, 2020

Conclusions

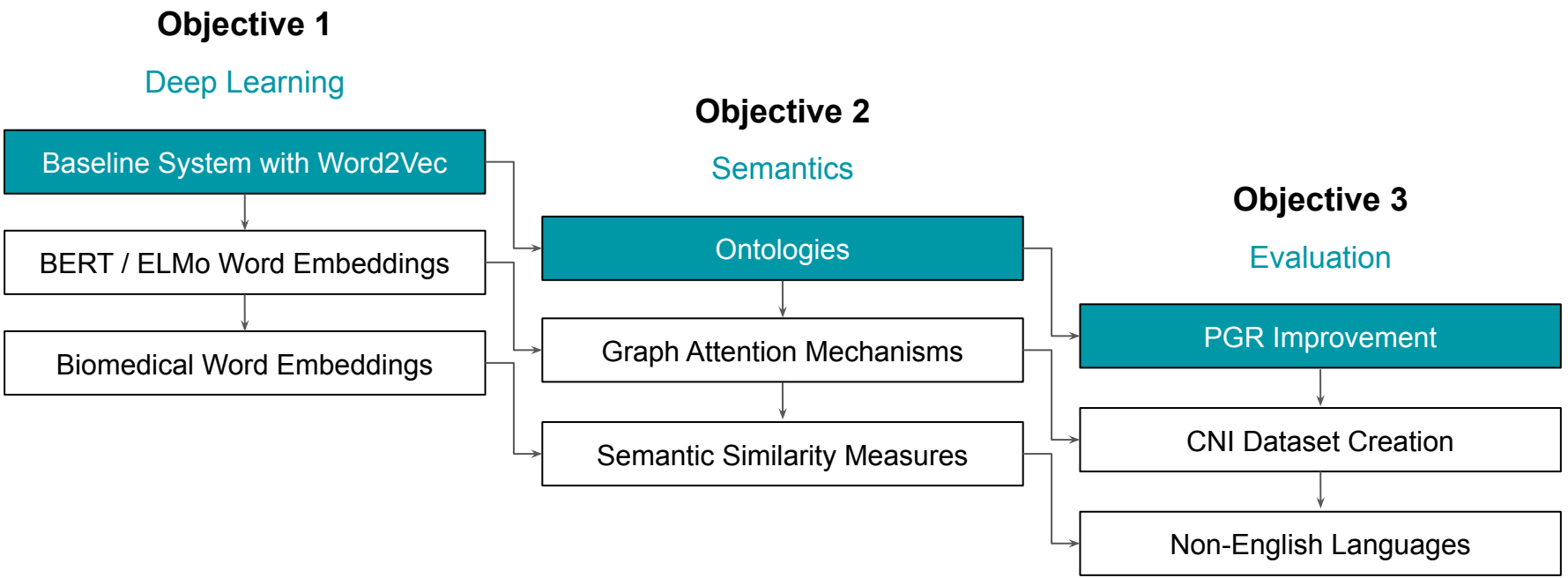
- The **knowledge encoded in biomedical ontologies** plays a vital part in the development of learning systems

- Ultimately, it can be used to explore new experimental hypotheses providing evidence to researchers and clinicians about possible unknown associations between biomedical entities

---

- There is potential in using **distant supervision allied with crowdsourcing** to produce gold standard datasets with which we can train viable models and detect relevant biomedical relations

Future Work



PGR - Phenotype-Gene Relations Corpus
CNI - Cancer-Nutrition Interactions

Mentor Suggestions

Professor Karin Verspoor

- The importance of generalization of ML algorithms. What type of results do we want?
  - Memorization versus Generalization
  - Think datasets to get the rare cases
  - Test for generalization
- Full integration of knowledge graphs in the neural network
- Integration of graph neural networks knowledge

Contacts

DeST: Deep Semantic Tagger

http://dest.rd.ciencias.ulisboa.pt/

https://github.com/lasigeBioTM

dfsousa@lasige.di.fc.ul.pt

Thank you!

List of Contributions

Chapter (Q3 Scimago): Using Neural Networks for Relation Extraction from Biomedical Literature for the book Artificial Neural Networks: Methods and Applications (**Diana Sousa**, Andre Lamurias, and Francisco M. Couto) in the Springer “Methods in Molecular Biology” series, 2020

Conference (Core A): BiOnt: Deep Learning Using Multiple Biomedical Ontologies for Relation Extraction (**Diana Sousa** and Francisco M. Couto) in the Proceedings of the 42nd European Conference on Information Retrieval, 2020

Journal (*unassigned*): Improving accessibility and distinction between negative results in biomedical relation extraction (**Diana Sousa**, Andre Lamurias, and Francisco M. Couto) in Genomics & Informatics, 2020

Journal (Q1 Scimago): A Hybrid Approach towards Biomedical Relation Extraction Training Corpora: Combining Distant Supervision with Crowdsourcing (**Diana Sousa**, Andre Lamurias, and Francisco M. Couto) in Database: The Journal of Biological Databases and Curation, 2020

co-authored:

Journal (Q1 Scimago): Generating Biomedical Question Answering Corpora From Q&A Forums (Andre Lamurias, **Diana Sousa**, and Francisco M. Couto) in IEEE Access, 2020

Workshop: COVID-19: A Semantic-Based Pipeline for Recommending Biomedical Entities (Márcia Barros, Andre Lamurias, **Diana Sousa**, Pedro Ruas, and Francisco M. Couto) in the Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, 2020

Extra Links

- https://shreyachopra711.medium.com/how-to-gather-information-1972617ea539 (Figure 1 - Slide 2 and 3)
- https://pubmed.ncbi.nlm.nih.gov/?term=brca1 (Figure 2 - Slide 5)
- https://www.logo.wine/logo/PubMed (Figure 3 - Slide 5)
- https://pubmed.ncbi.nlm.nih.gov/32121304/ (Figure 4 - Slide 7)
- https://www.flaticon.com/premium-icon/funnel_460326 (Figure 5 - Slide 8)
- https://corenlp.run/ (Figure 6 - Slide 11)
- https://www.searchenginejournal.com/blogging-challenges/321109/ (Figure 7 - Slide 16)
- https://www.civhc.org/partner-with-us/question-mark/ (Figure 8 - Slide 17)
- https://commons.wikimedia.org/wiki/File:Eo_circle_orange_arrow-up-right.svg (Figure 9 - Slide 25)
- https://twitter.com/amazonmturk (Figure 10 - Slide 30)