

### 43<sup>rd</sup> 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





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



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

#### **Biomedical Literature**

- breast cancer
- ovarian cancer
- others?



Online Information
PubMed



Pub Med.gov	brca1 × Search Advanced Create alert Create RSS User Guide
	Save         Email         Send to         Sorted by: Best match         Display options
MY NCBI FILTERS	18,259 results
RESULTS BY YEAR	BRCA1 as target for breast cancer prevention and therapy.         1       Romagnolo AP, Romagnolo DF, Selmin OI.         Cite       Anticancer Agents Med Chem. 2015;15(1):4-14. doi: 10.2174/1871520614666141020153543.         PMID: 25329591       Review.         Although BRCA1 acts as a tumor suppressor and is present in all cells, where it is essential for the maintenance of the genome integrity, it is still not clear why mutations in the BRCA1 gene predispose to breast and ovarian, but not to other types of cancerThe
TEXT AVAILABILITY         Abstract         Free full text         Full text         ARTICLE ATTRIBUTE         Associated data	<ul> <li>Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell</li> <li>cycle in response to DNA damage.</li> <li>Yoshida K, Miki Y.</li> <li>Cancer Sci. 2004 Nov;95(11):866-71. doi: 10.1111/j.1349-7006.2004.tb02195.x.</li> <li>PMID: 15546503 Free article. Review.</li> <li>BRCA1 (BReast-CAncer susceptibility gene 1) and BRCA2 are tumor suppressor genes, the mutant phenotypes of which predispose to breast and ovarian cancers. Intensive research has shown that BRCA proteins are involved in a multitude of pivotal cellular processes. In p</li> </ul>

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

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

### Publed

## **32 million citations**

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

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

- breast cancer
  - prostate cancer
- ovarian 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**?

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



#### **Related Genes**

#### Contributions of DNA Damage to Alzheimer's Disease

Xiaozeng Lin <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>, Anil Kapoor <sup>2</sup> <sup>3</sup> <sup>4</sup>, Yan Gu <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>, Mathilda Jing Chow <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>, Jingyi Peng <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>, Kuncheng Zhao <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>, Damu Tang <sup>1</sup> <sup>2</sup> <sup>3</sup> <sup>4</sup>

Affiliations + expand PMID: 32121304 PMCID: PMC7084447 DOI: 10.3390/ijms21051666 Free PMC article

Alzheimer's disease (AD) is the most common type of neurodegenerative disease. Its typical pathology consist of extracellular anyolicife (AB) plaques and intracellular tain eurorithrillary tangles. Mutations in the APP, PSEN1, and PSEN2 genes increase AB production and aggregation, and thus cause early onset or familial AD. Even with this strong genetic evidence, recent studies support AD to result from complex etiological alteriations. Among them, aging is the strongest risk factor for the vast majority of AD cases: Sporadic late onset AD (LOAD). Accumulation of DNA damage is a velicetablished aging factor. In this regard, a large amount of evidence reveals DNA damage as a critical pathological cause of AD. Clinically, DNA damage is accumulated in brains of AD patients. Genetically, defects in DNA damage repair resulted from mutations in the BRAC1 and other DNA damage repair genes occur in AD brain and facilitate the pathogenesis. Abnormalities in DNA damage repair can be used as diagnostic biomarkers for AD. In this review, we discuss the association, the causative potential, and the biomarkers for AD. In this review, we discuss the

#### Associated Diseases



#### **Associated Phenotypes**

Gene Ontology

BRCA1





#### 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



Mutations in the **APP** gene increase  $A\beta$  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>



#### **Approaches for Relation Extraction**



#### **Approaches for Relation Extraction**

Deep Learning - Long Short-Term Memory (LSTM)





#### **Approaches for Relation Extraction**

**Data Representations** 

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





### **Evaluation**

Supervised Machine Learning



#### **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>

### **Evaluation**

number of correct results identified

how often the results are correct

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

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**

#### **Objective 1** Deep Learning **Objective 2** Baseline System with Word2Vec **Semantics Objective 3** Ontologies Evaluation BERT / ELMo Word Embeddings PGR Improvement Graph Attention Mechanisms **Biomedical Word Embeddings CNI** Dataset Creation Semantic Similarity Measures Non-English Languages

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





Human Phenotype Ontology



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

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



limited to drug-drug interactions



#### **BiOnt System** (Entity + Annotation Vector)

#### Word Embeddings + WordNet Hypernyms + Ontology Embeddings

4 channels of information







common concatenation ancestors of ancestors



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





### 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

#### **Objective 1** Deep Learning **Objective 2** Baseline System with Word2Vec Semantics **Objective 3** Ontologies Evaluation BERT / ELMo Word Embeddings PGR Improvement Graph Attention Mechanisms **Biomedical Word Embeddings CNI** Dataset Creation Semantic Similarity Measures Non-English Languages



**PGR** Improvement

#### **Silver Standard Dataset**

Phenotype-Gene Relations (PGR)





**PGR Improvement** 



#### **Knowledge Bases**

- **Domain Expert Annotators** •
- **Crowdsourcing Platforms** •
- **Distantly Supervised Techniques** •

**PGR** Improvement

• Domain Expert Annotators



#### More reliable

• Crowdsourcing Platforms





- Distantly Supervised Techniques
  - O Still needs to be reviewed
  - O Still time- and resource- consuming



Less reliable



**PGR** Improvement

### **Distantly Supervised Techniques**

PGR (7963 relations)



#### **PGR** Improvement

#### Do the entities in bold share a relation?

EPAC is also known for its dual role in cancer as pro- and anti-proliferative in addition to metastasis.

#### Select an option

Yes, they share a direct/explicit relation in the sentence.

No, they are separate entities with no correlation in the sentence.

The entities seem to be illy marked, or something is <sup>3</sup> wrong with the entities/sentence.

### 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**



## **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



### **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**

- <u>https://shreyachopra711.medium.com/how-to-gather-information-1972617ea539</u> (Figure 1 Slide 2 and 3)
- <u>https://pubmed.ncbi.nlm.nih.gov/?term=brca1</u> (Figure 2 Slide 5)
- <u>https://www.logo.wine/logo/PubMed</u> (Figure 3 Slide 5)
- <u>https://pubmed.ncbi.nlm.nih.gov/32121304/</u> (Figure 4 Slide 7)
- <u>https://www.flaticon.com/premium-icon/funnel\_460326</u> (Figure 5 Slide 8)
- <u>https://corenlp.run/</u> (Figure 6 Slide 11)
- <u>https://www.searchenginejournal.com/blogging-challenges/321109/</u> (Figure 7 Slide 16)
- <u>https://www.civhc.org/partner-with-us/question-mark/</u> (Figure 8 Slide 17)
- <u>https://commons.wikimedia.org/wiki/File:Eo\_circle\_orange\_arrow-up-right.svg</u> (Figure 9 Slide 25)
- <u>https://twitter.com/amazonmturk</u> (Figure 10 Slide 30)