Deep Semantic Entity Linking

Pedro Ruas

PhD student in Informatics

LASIGE-FCUL

Faculdade de Ciências, Universidade de Lisboa, Portugal

Mentor: Professor Wen Hua

ECIR 2021 Doctoral Consortium

28th March, 2021

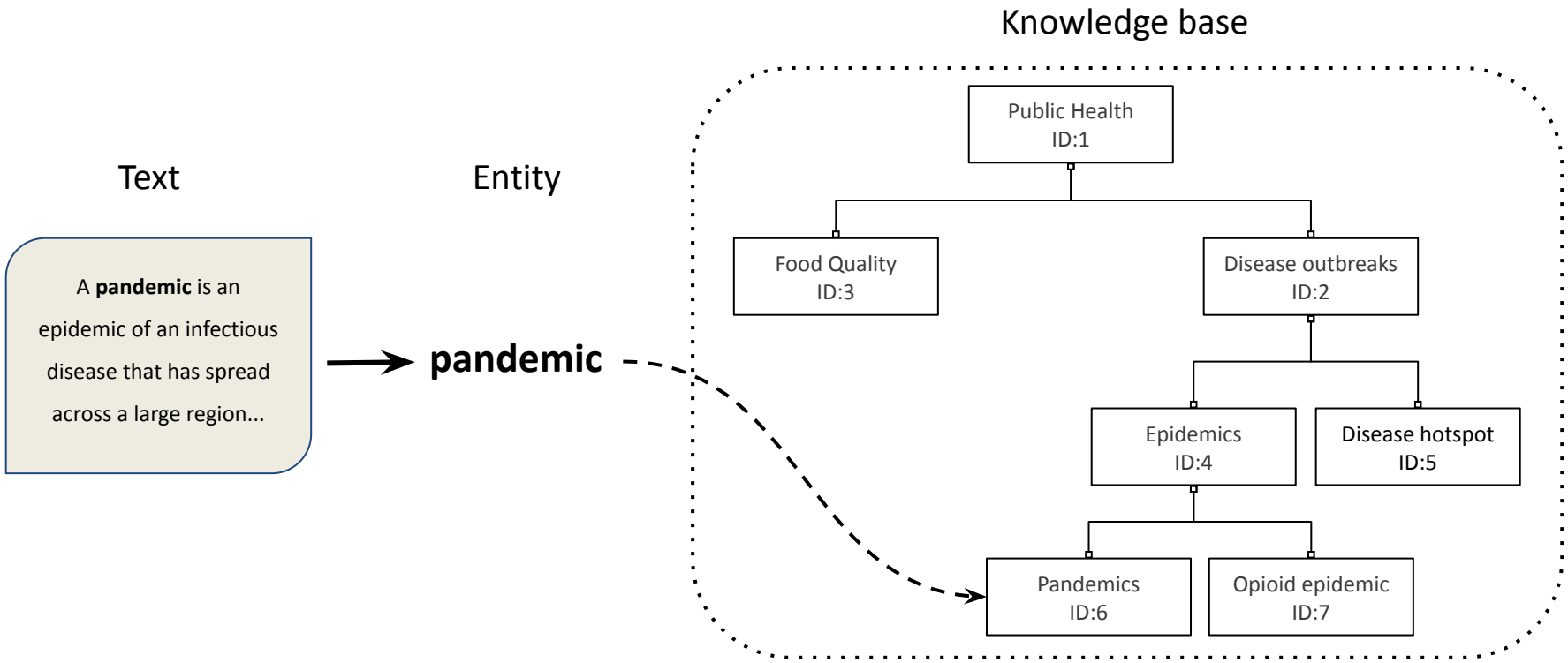
Motivation

Text Mining

Transformation of **unstructured text data** into **machine-readable** formats

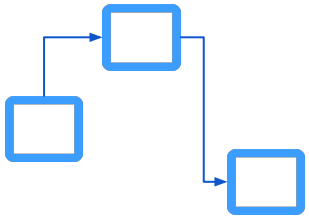
Named Entity Linking (NEL)

To link **entities** in text to **concepts** of Knowledge Bases (KBs) that describe their meaning



Motivation

Why are NEL systems important?



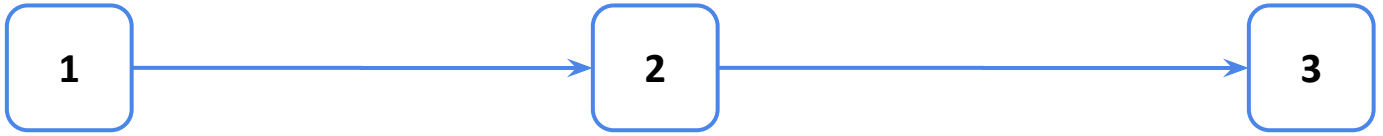
Essential component in
Text Mining pipelines

Performance of
search engines

Knowledge Base
curation/building

Motivation

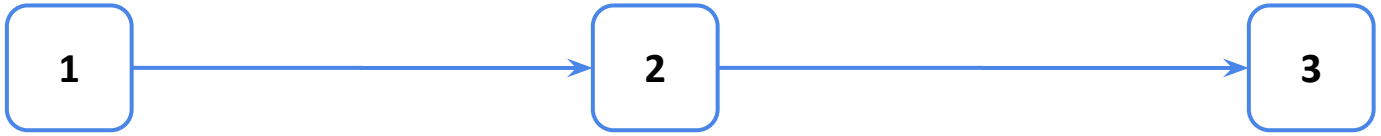
Typical NEL System



Candidate generation **Candidate ranking** **Candidate disambiguation**

Motivation

Typical NEL System



Candidate generation **Candidate ranking** **Candidate disambiguation**

Limitations

- lack of context
- incomplete knowledge bases
- multilingual performance

In the biomedical domain

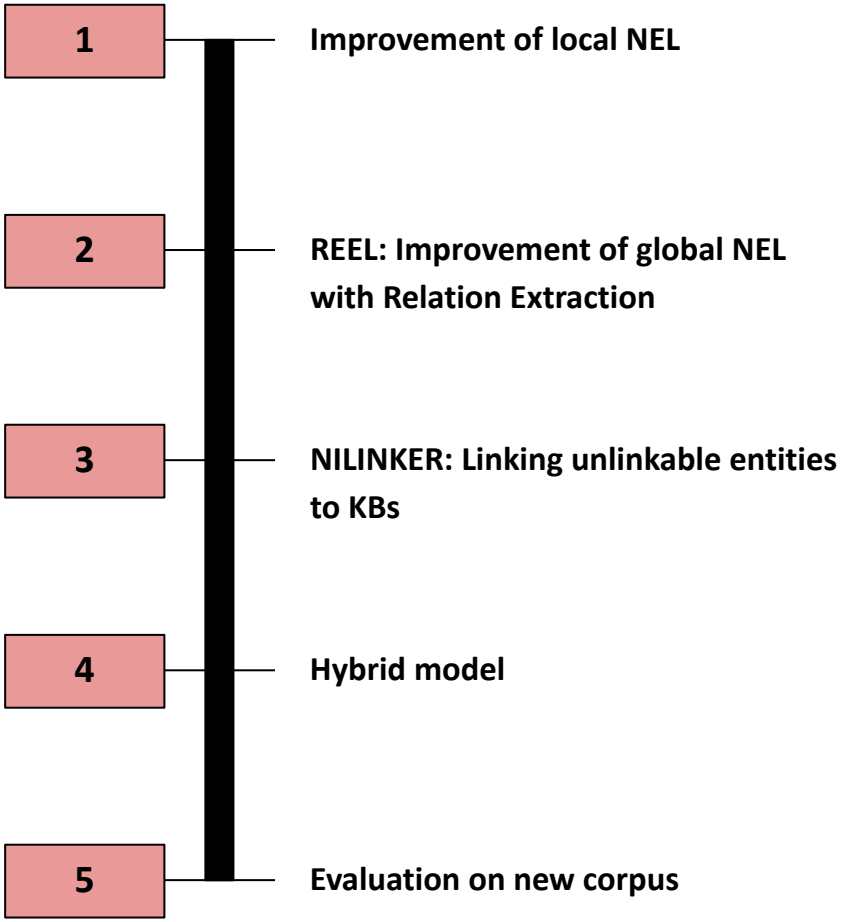
- specialized language
- lack of annotated datasets

Goals

Main goal

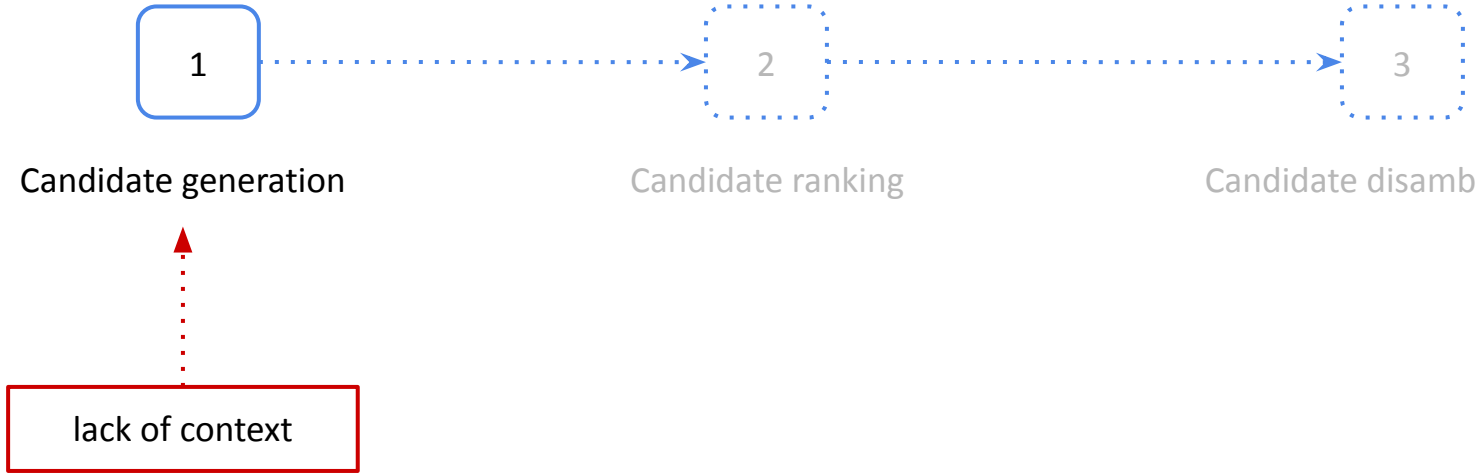
To tackle limitations of NEL models to achieve SOTA in the biomedical domain

Specific goals



1. Improvement of local NEL

Typical NEL System



An entity has always the same candidates list

1. Improvement of local NEL

Goal

➔ To use **pre-trained language models** to improve the determination of local similarity between entity and KB candidates

Approach

→ 1. Large **KB candidates list** for each entity:

- abbreviation expansion
- string matching
- synonyms lookup

→ 2. **Contextualised embeddings** for entities + candidates (BERT, ClinicalBERT, BioBERT)

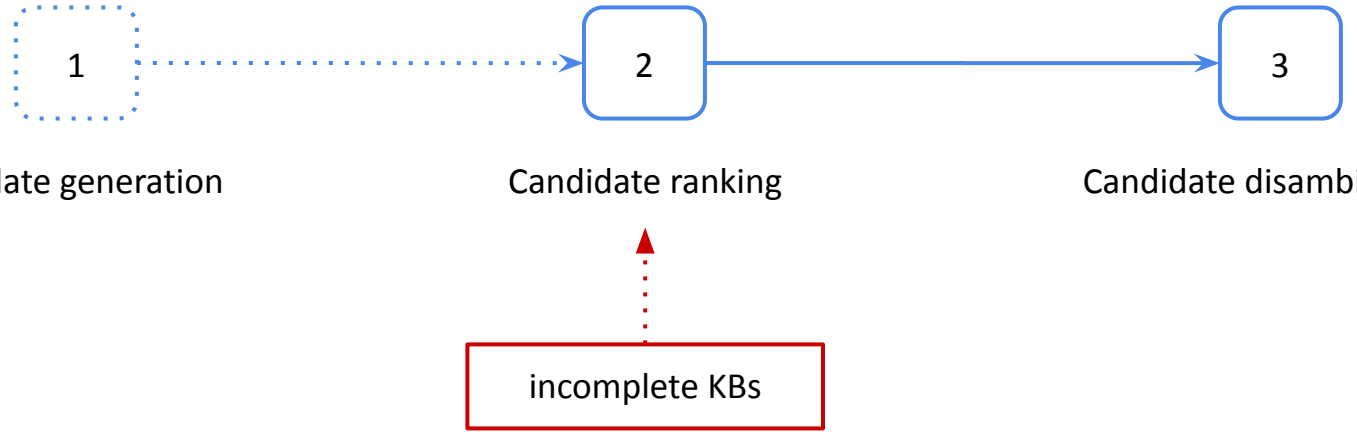
→ 3. Similarity score (entity/candidate pairs)

multi-layer perceptron → conditional probabilities for each pair mention-candidate

→ 4. **Filter out** less relevant candidates

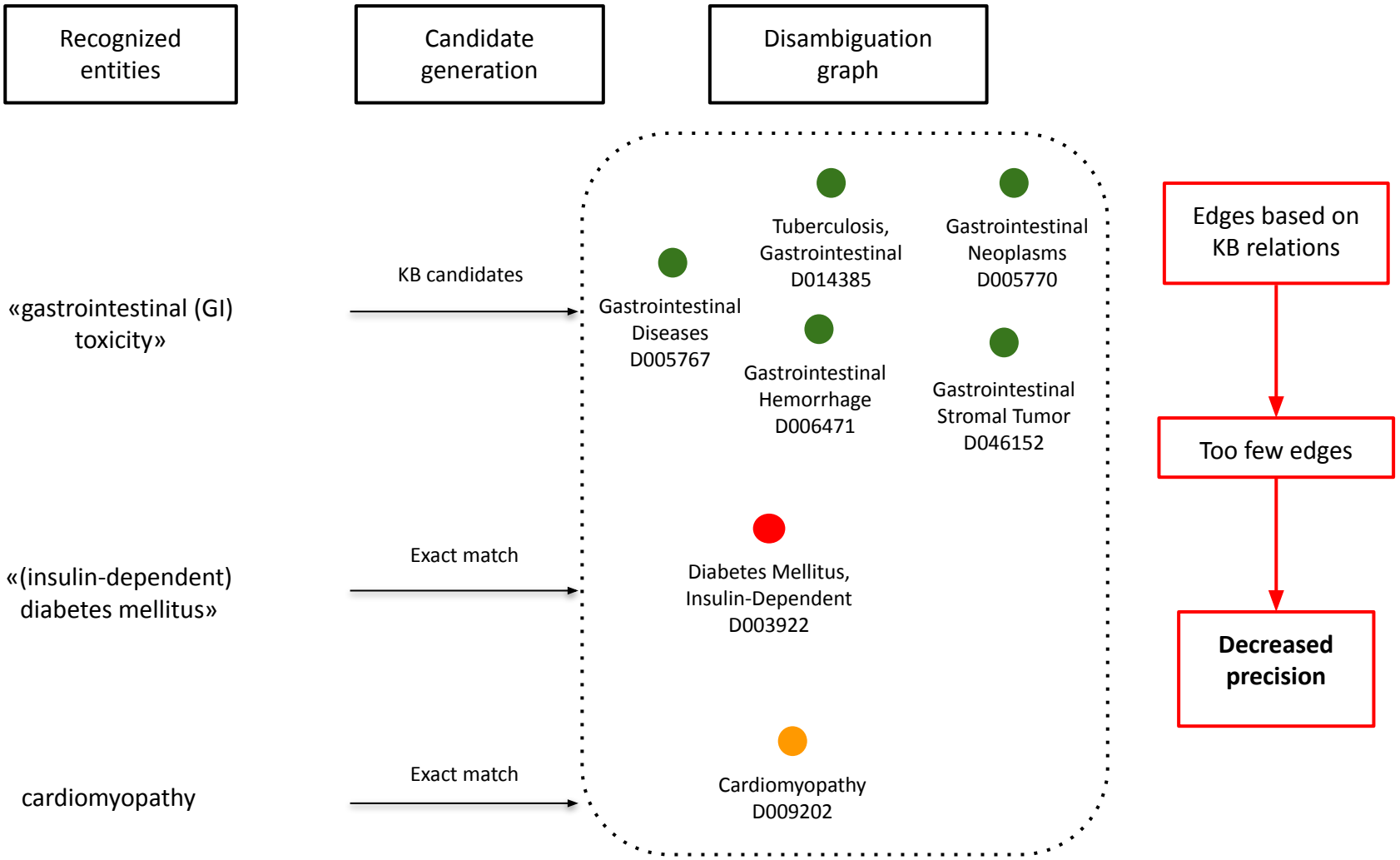
2. REEL: Improvement of global NEL with Relation Extraction

Typical NEL System



2. REEL: Improvement of global NEL with Relation Extraction

Graph-based systems build a disambiguation graph with candidates for all entities in a given document



2. REEL: Improvement of global NEL with Relation Extraction

Hypothesis

➔ **Relation Extraction** systems increase the available semantic information for NEL systems, which in turn increases their **precision**

Challenges

- Availability of Relation Extraction systems in biomedical and life sciences domains

2. REEL: Improvement of global NEL with Relation Extraction

Solution: to apply BO-LSTM (a deep learning RE tool) to complete disambiguation graphs

1. Entities recognized in scientific literature + normalization with MER tool

- «**Gastrointestinal diseases** can be associated with **diabetes mellitus**»
- «**Cardiomyopathy** is defined by the existence of abnormal myocardial structure in individuals with **diabetes mellitus**»

Gastrointestinal diseases → **Gastrointestinal diseases D005767**

diabetes mellitus → **Diabetes Mellitus, Insulin-Dependent D003922**

Cardiomyopathy → **Cardiomyopathy D009202**

2. Extraction of relations between entities

D005767: Gastrointestinal diseases ↔ D003922: diabetes mellitus

D009202: cardiomyopathy ↔ D003922: diabetes mellitus

2. REEL: Improvement of global NEL with Relation Extraction

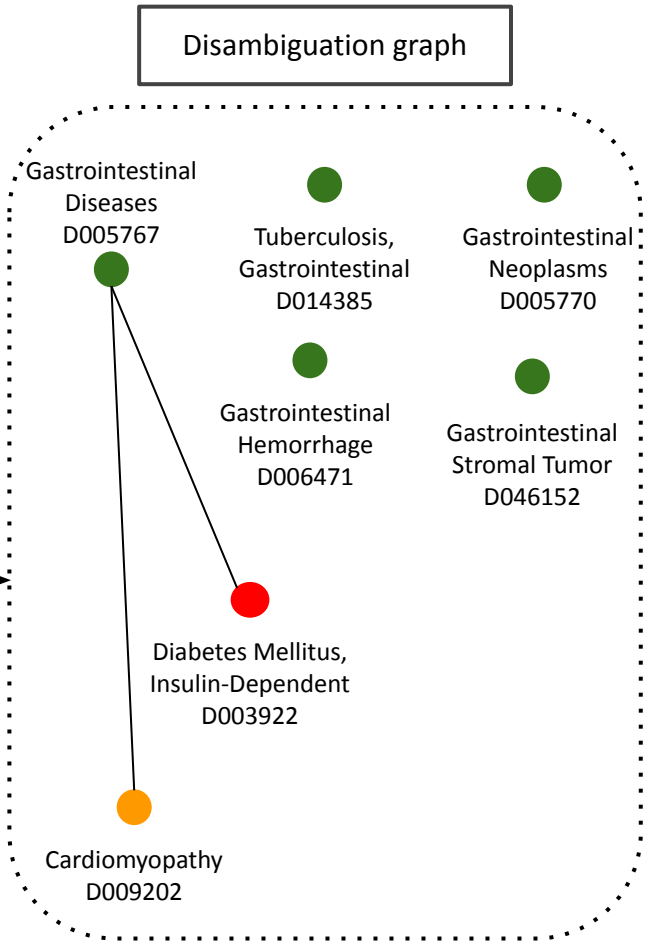
Solution: to apply BO-LSTM (a deep learning RE tool) to complete disambiguation graphs

2. Extraction of relations between entities

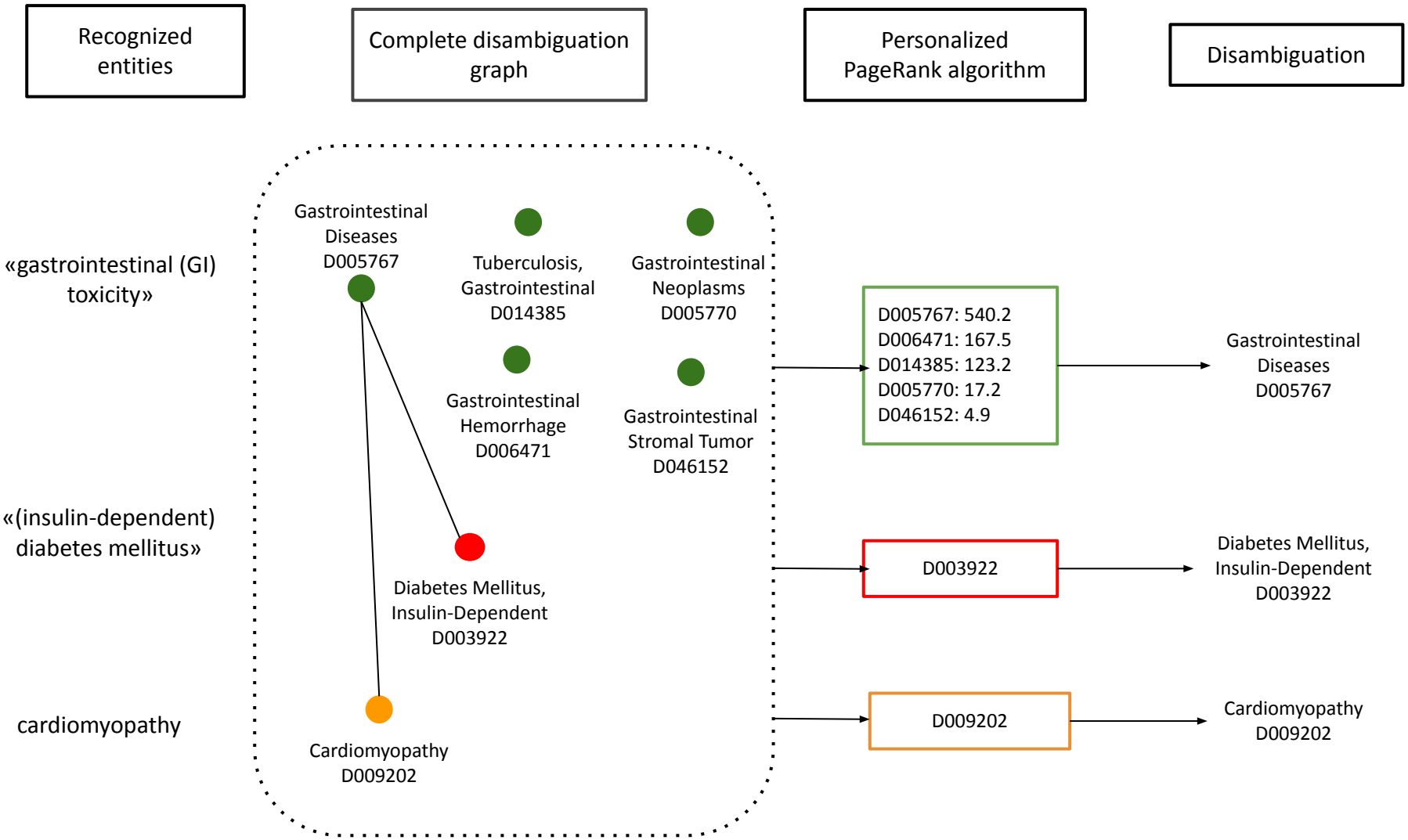
D005767: Gastrointestinal diseases ↔ D003922: diabetes mellitus

D009202: cardiomyopathy ↔ D003922: diabetes mellitus

3. Add relations to disambiguation graph

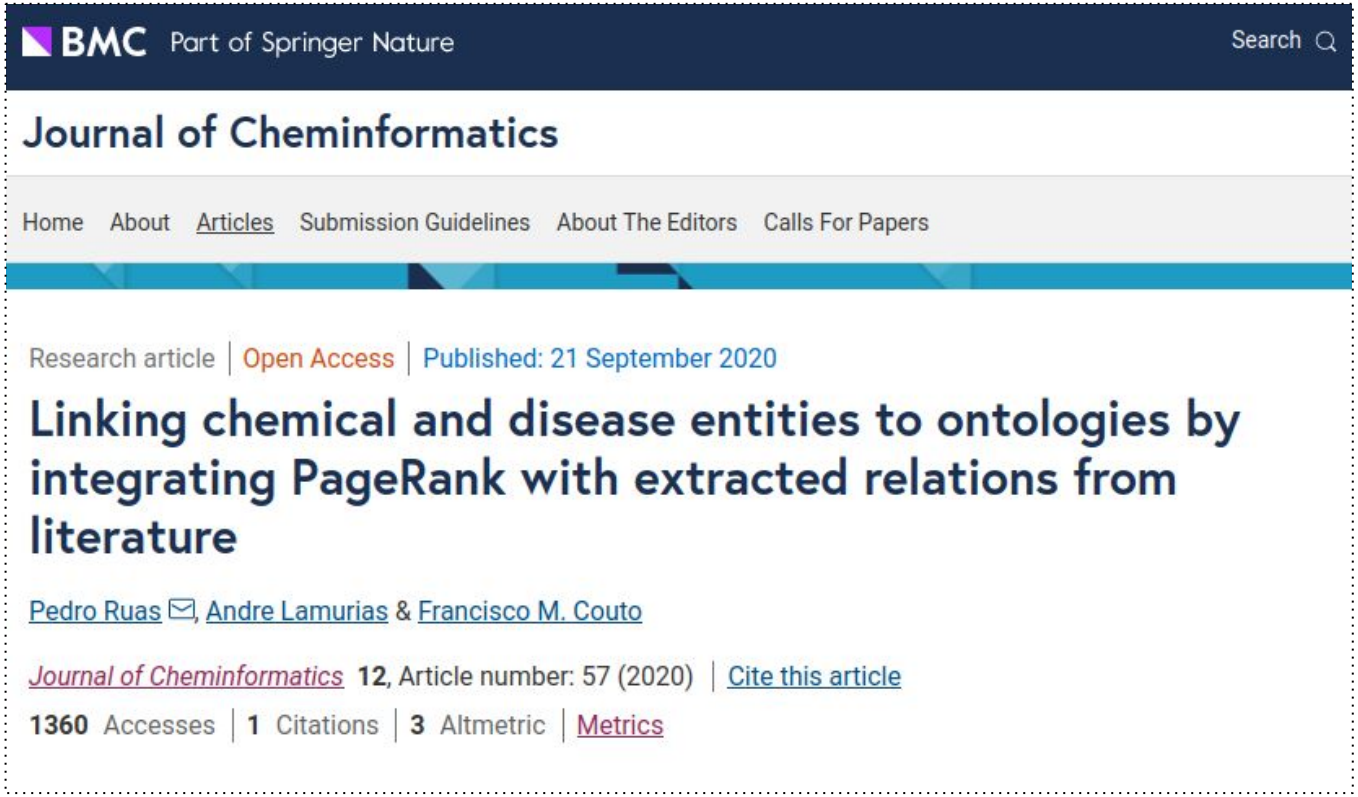


2. REEL: Improvement of global NEL with Relation Extraction



2. REEL: Improvement of global NEL with Relation Extraction

Work published in a journal article:



BMC Part of Springer Nature

Search

Journal of Cheminformatics

Home About Articles Submission Guidelines About The Editors Calls For Papers

Research article | Open Access | Published: 21 September 2020

Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature

Pedro Ruas, Andre Lamurias & Francisco M. Couto

Journal of Cheminformatics **12**, Article number: 57 (2020) | Cite this article

1360 Accesses | **1** Citations | **3** Altmetric | Metrics

2. REEL: Improvement of global NEL with Relation Extraction

Results on ChEBI annotations from CRAFT corpus

- Relations extracted by BO-LSTM on the same corpus

baselines

Model	CRAFT-ChEBI		
	P	R	F1
String matching	77.8	78.0	77.9
PPR-IC	87.1	79.9	83.3
REEL(Corpus)	*91.3*	*80.9*	*85.8*
REEL(KB+Corpus)	*91.3*	*80.9*	*85.8*

2. REEL: Improvement of global NEL with Relation Extraction

BC5CDR corpus: Gold labeled relations of the corpus included in the disambiguation graph

Results on disease annotations

BC5CDR-Diseases			
Model	**All**		
	P	**R**	**F1**
String matching	82.7	74.7	78.5
PPR-IC	83.4	74.9	78.9
REEL(Corpus)	*86.9*	*75.6*	*80.9*
REEL(KB+Corpus)	86.6	75.5	80.7

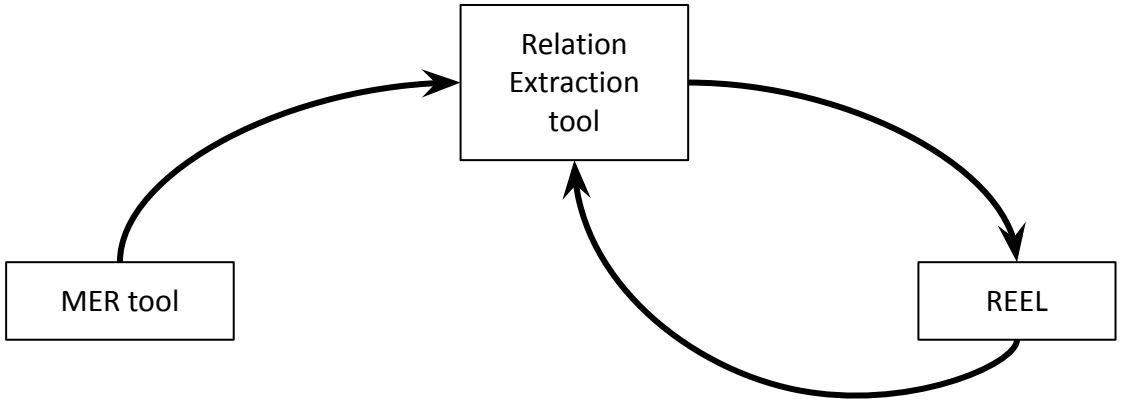
baselines

Results on chemical annotations

BC5CDR-Chemicals			
Model	**All**		
	P	**R**	**F1**
String matching	94.9	84.1	89.2
PPR-IC	96.7	84.4	90.1
REEL(Corpus)	*97.0*	*84.4*	*90.3*
REEL(KB+Corpus)	*97.0*	*84.4*	*90.3*

baselines

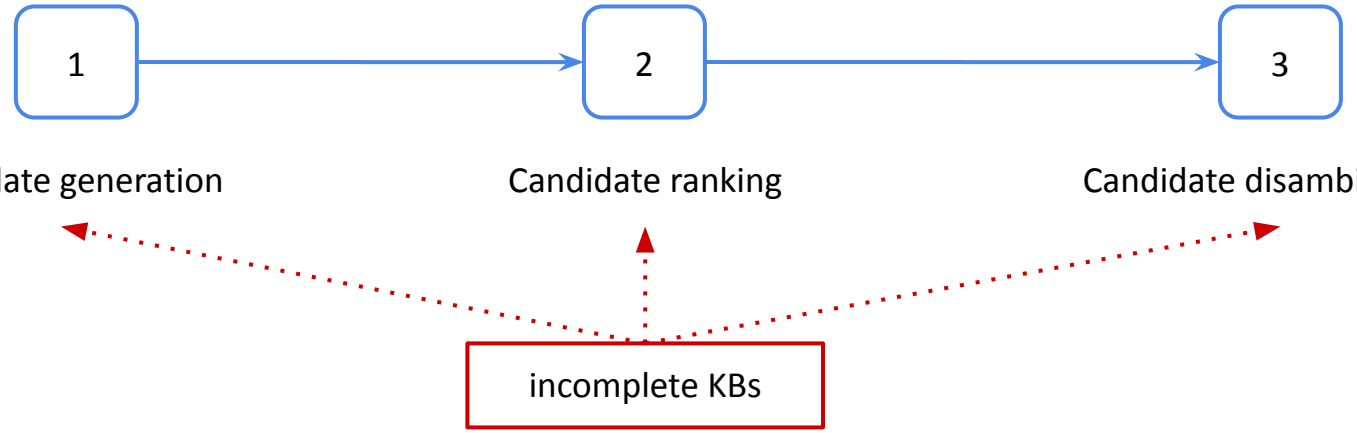
2. REEL: Improvement of global NEL with Relation Extraction



Feedback loop between **Relation Extraction** and **NEL**

3. NILINKER: Linking unlinkable entities to KBs

Typical NEL System



A **NIL entity is unlinkable** when there is no concept in the KB that describe it

Recognized entity

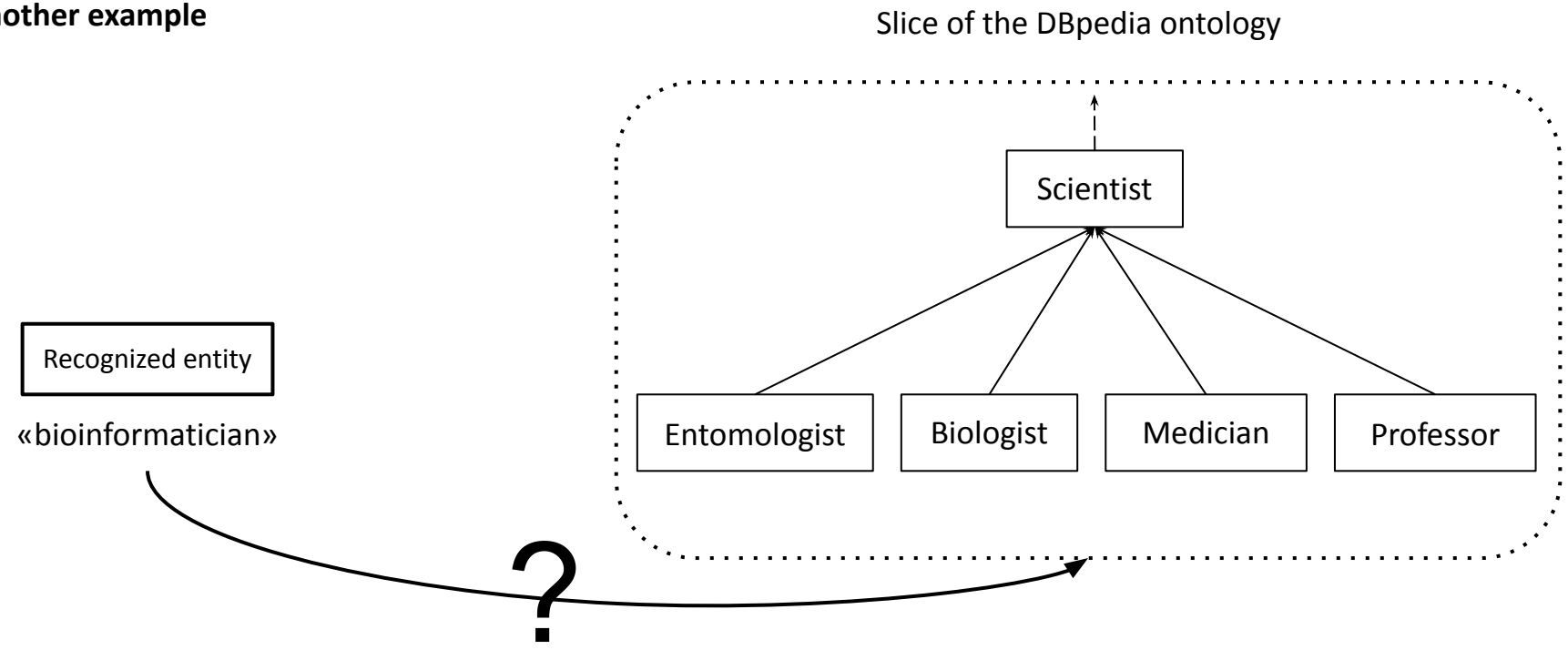
«Arrhythmogenic cardiomyopathy» →

- Empty candidates list
- Correct KB candidate not on the list
- Score below defined threshold

→ **Decreased NEL performance**

3. NILINKER: Linking unlinkable entities to KBs

Another example



What concept do we choose to describe «bioinformatician»?

3. NILINKER: Linking unlinkable entities to KBs

Hypotheses

➔ It is possible to **partially link NIL or unlinkable** entities to knowledge bases (i.e. **NIL entity Linking**)

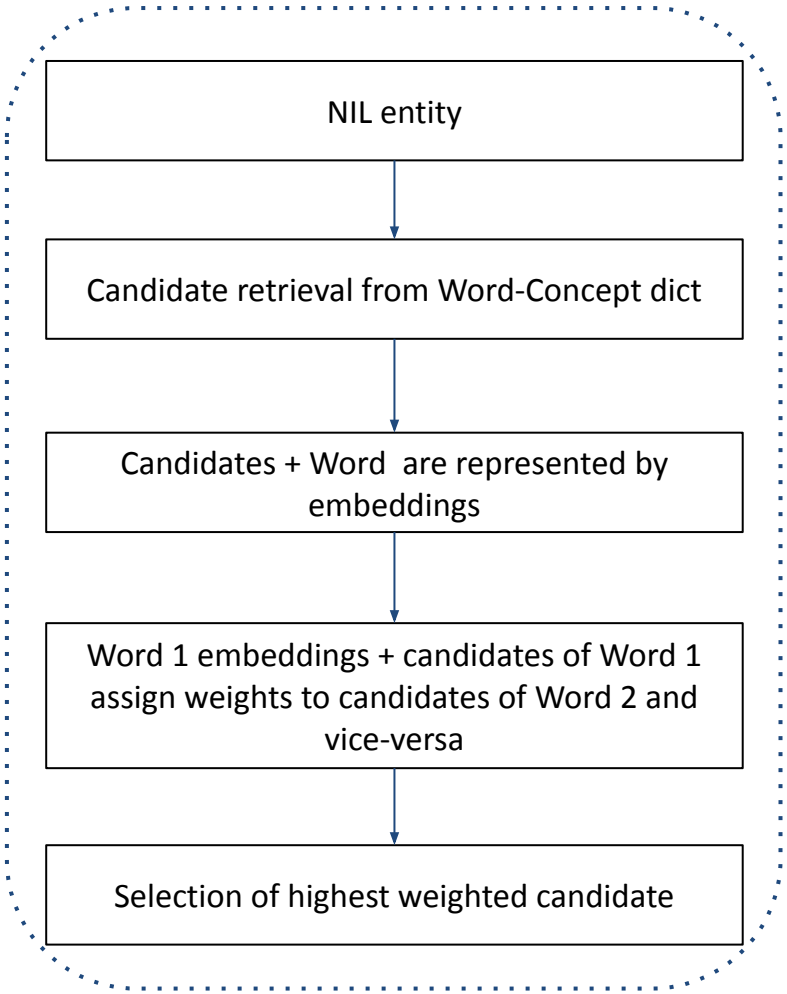
➔ NIL entity linking improves the **performance** of NEL models

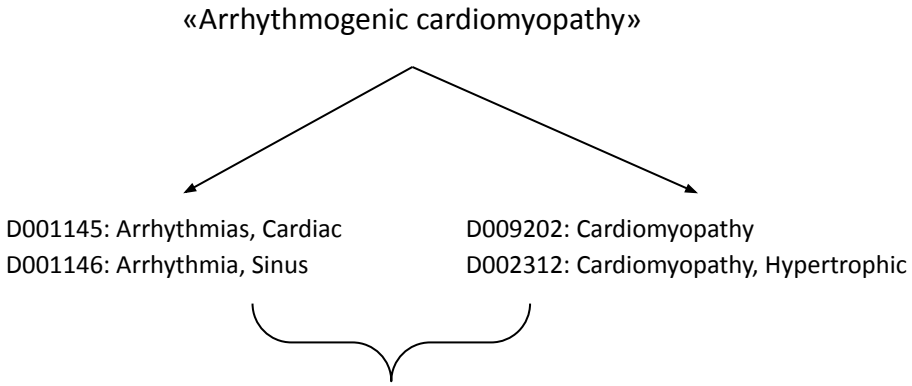
Challenges

- Lack of evaluation dataset
- Lack of baseline approaches

3. NILINKER: Linking unlinkable entities to KBs

Attention-based model to select the most relevant candidate for a **NIL entity**

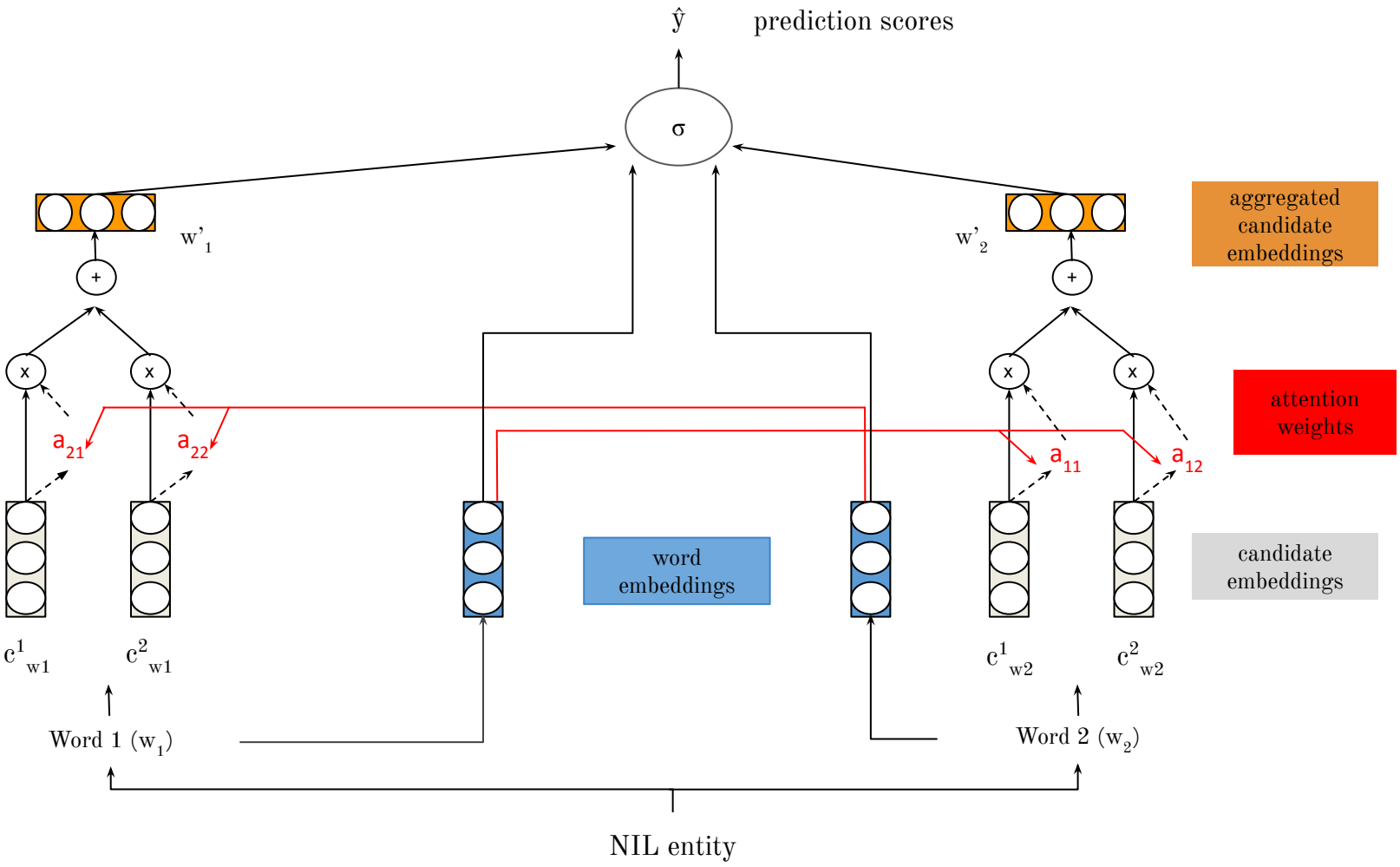




«Arrhythmogenic cardiomyopathy» can be added to the KB as a **child concept** of "Cardiomyopathy"

3. NILINKER: Linking unlinkable entities to KBs

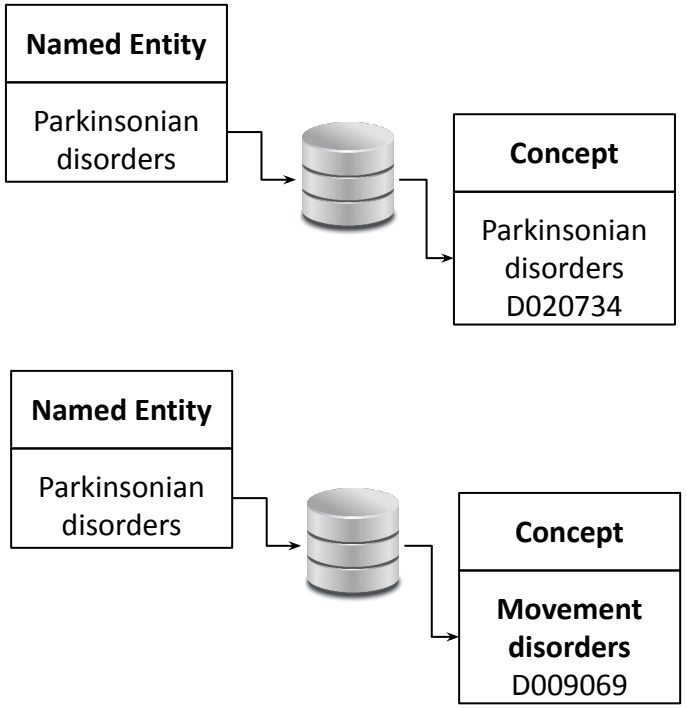
Attention-based model to select the most relevant candidate for a **NIL entity**



3. NILINKER: Linking unlinkable entities to KBs

Two different evaluation approaches

1. **Generation of new dataset or adaptation of existing corpora**

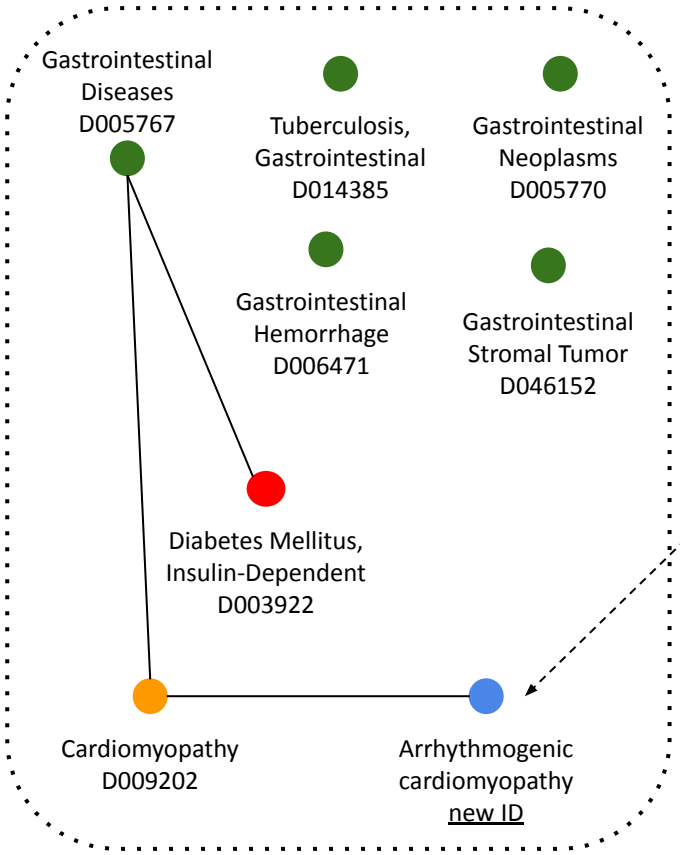


Assumption: If there was not a disambiguating concept for an entity, it should be linked to its **direct ancestor**

3. NILINKER: Linking unlinkable entities to KBs

Two different evaluation approaches

2. Performance of the graph-based NEL model on existing corpora



- Convert 30 % of the annotations into NIL entities
- Application of NILINKER
- Complete the **disambiguation graph** with the output

Evaluation corpora

- NCBI disease
- BC5CDR

4. Hybrid model

Hypothesis

➔ The integration of the the previous modules achieves state-of-the-art performance in the Biomedical and Life Sciences domains

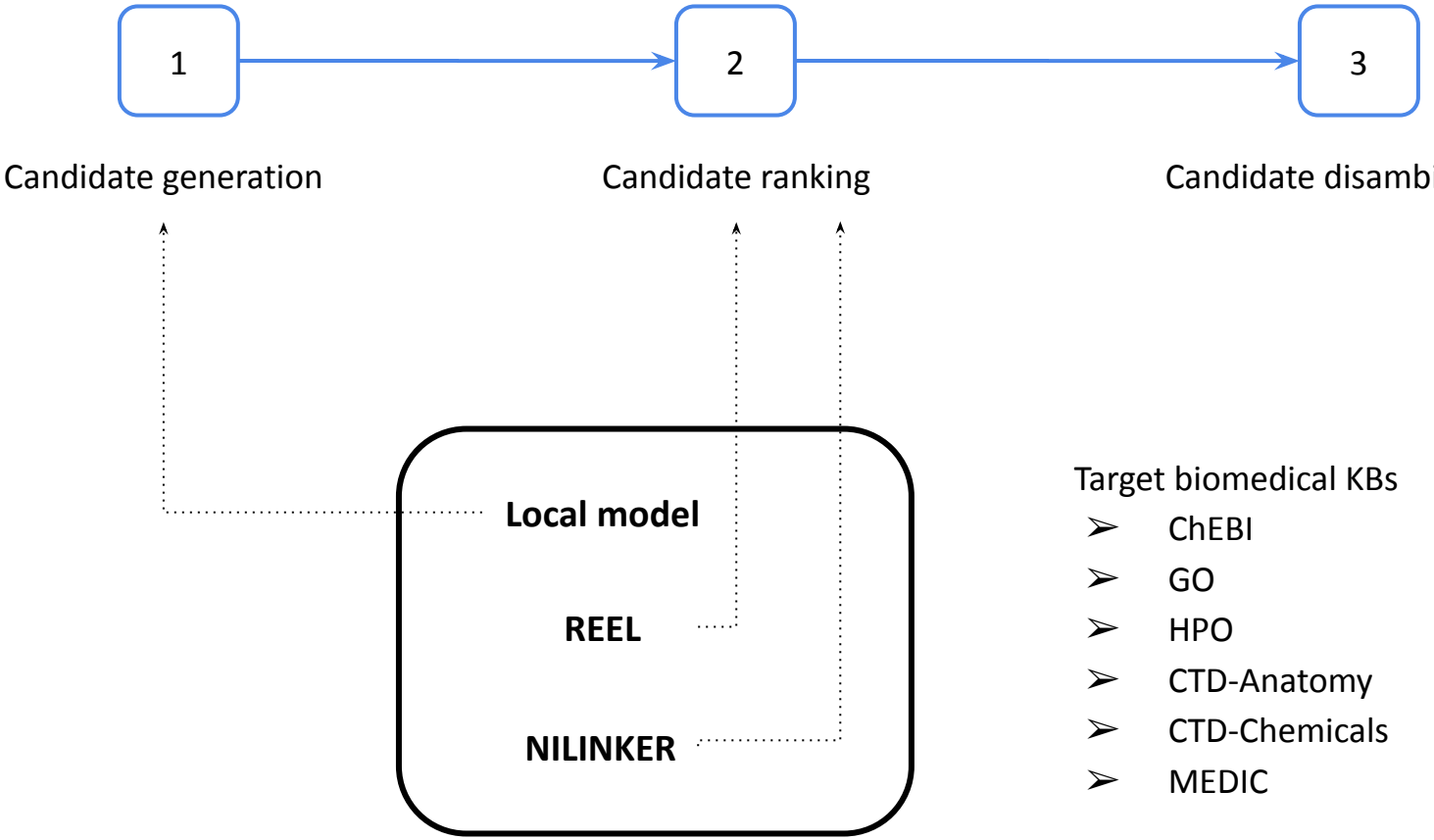
Challenges

- Combination of different modules into a single model

Evaluation corpora

- NCBI disease
- BC5CDR
- (new) MultiNEL: Portuguese, English and Spanish biomedical corpus

4. Hybrid NEL model



Target biomedical KBs

- ChEBI
- GO
- HPO
- CTD-Anatomy
- CTD-Chemicals
- MEDIC

5. Evaluation on new corpus

MultiNEL: Portuguese, English and Spanish biomedical corpus

Document retrieval (SciELO, PubMED)

Automatic entity annotation (DeCS vocabulary, ICD10-CM)

next step

Annotation revision by humans

Preliminary work: Short paper published in Proceedings of the SIIRH 2020 Workshop (ECIR 2020)

Towards a multilingual corpus for Named Entity Linking evaluation in the clinical domain *

Pedro Ruas[1][0000−0002−1293−4199], André Lamúrias[1][0000−0001−7965−6536], and Francisco M Couto[3][0000−0003−0627−1496]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
ps_ruas@fc.ul.pt

Abstract. We propose a new multilingual, parallel corpus for Named Entity Linking benchmarking which comprises English, Portuguese and Spanish clinical case reports[1]. The medical diagnostic entities in the reports were annotated with the respective code of the International Classification of Diseases 10 - Clinical Modification (ICD10-CM) terminology and its Portuguese and Spanish versions. The result is a preliminary annotation set, which will be further validated and expanded by humans. Additionally, the ICD10-CM codes in the annotations will be mapped to the respective Medical Subject Headings (MeSH) identifiers when possible.

Keywords: Text Mining · Multilingual clinical case reports · Named Entity Linking · Information retrieval · Named Entity Recognition

	English	Portuguese	Spanish
Abstracts retrieved	639	639	639
Abstracts with annotations	217	197	199
Ratio of annotated abstracts	0.340	0.308	0.314
Entity mentions	533	432	465
Entity mentions per annotated abstract	2.456	2.193	2.340
Linked entity mentions	463	432	389
Linked entity mentions per annotated abstract	2.134	2.193	1.955
Ratio of linked entity mentions	0.867	1.000	0.837

Suggestions by Prof. Wen Hua

- To focus the presentation on biomedical challenges
- To improve the presentation by:
  - talking about the local model first
  - presenting the results of the published article
- To improve the explanation of the interaction between NEL/RE in the REEL work
- To provide more examples of NIL entities, specially in the biomedical domain
- To improve the evaluation of the NILINKER model

Deep Semantic Entity Linking

Pedro Ruas

PhD student in Informatics

LASIGE-FCUL

Faculdade de Ciências, Universidade de Lisboa, Portugal

Mentor: Professor Wen Hua

ECIR 2021 Doctoral Consortium

28th March, 2021

Funding

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017, LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020) and by FCT through funding of PhD Scholarship ref.2020.05393.BD

FCT Fundação para a Ciência e a Tecnologia

U LISBOA | UNIVERSIDADE DE LISBOA

Ciências ULisboa

LASIGE driven by excellence