

# ParaGraph: Mapping Wikidata Tail Entities to Wikipedia Paragraphs

**Natalia Ostapuk**, Djellel Difallah and Philippe Cudré-Mauroux

IEEE BigData 2022, Osaka, Japan

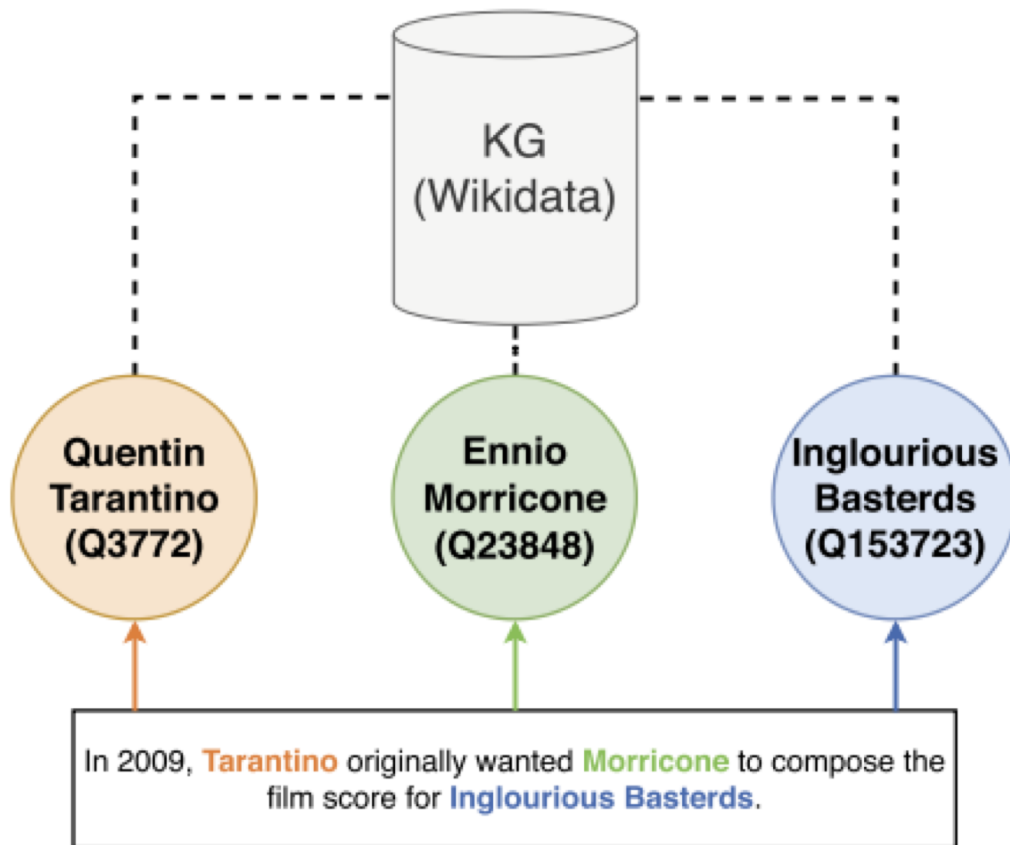
# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

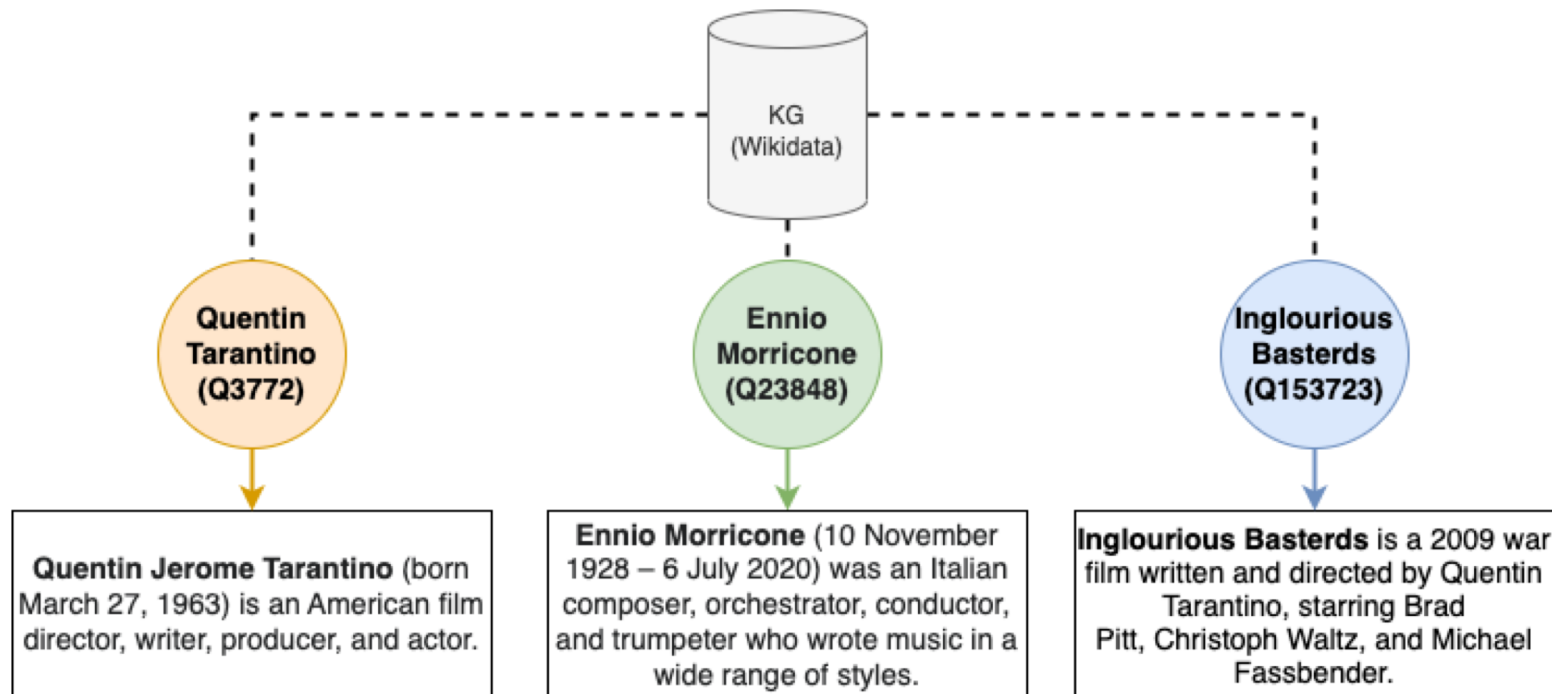
# Introduction: Problem Definition



- **Entity linking:** the task of linking **textual mentions** of named entities with their corresponding **entities in a knowledge graph**.

# Introduction: Problem Definition

**Entity Mapping:** given a **knowledge graph entity** and a collection of **paragraphs**, find the paragraph that best describes the input entity.

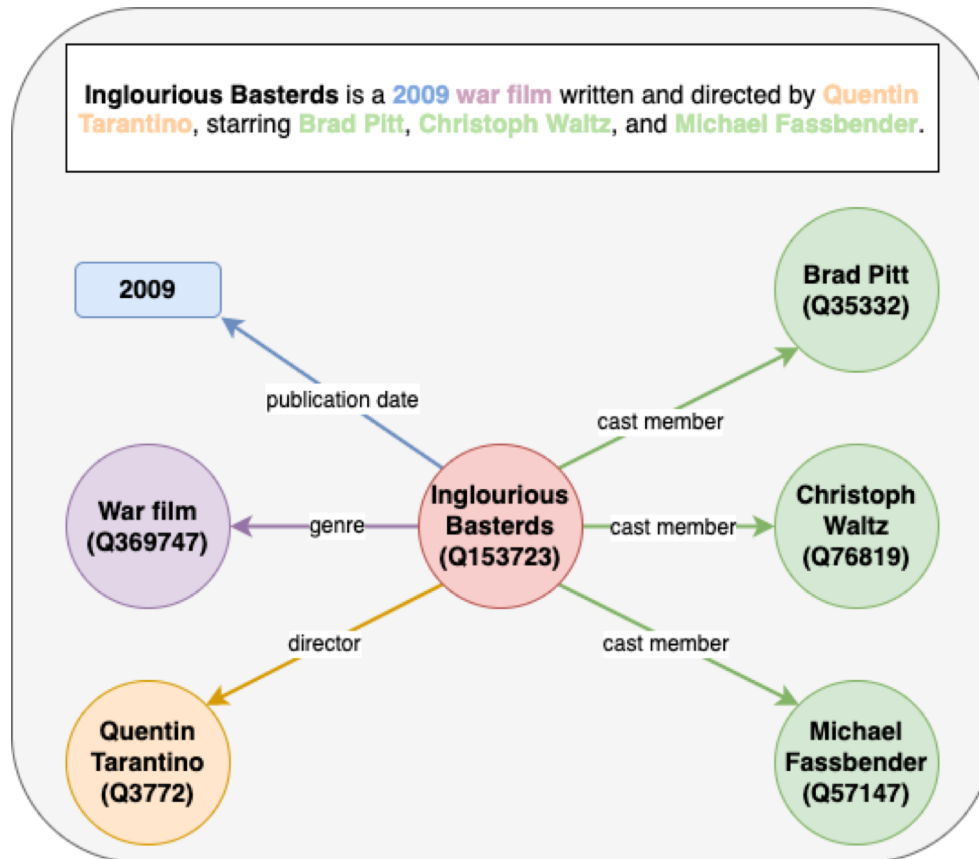


# Contents

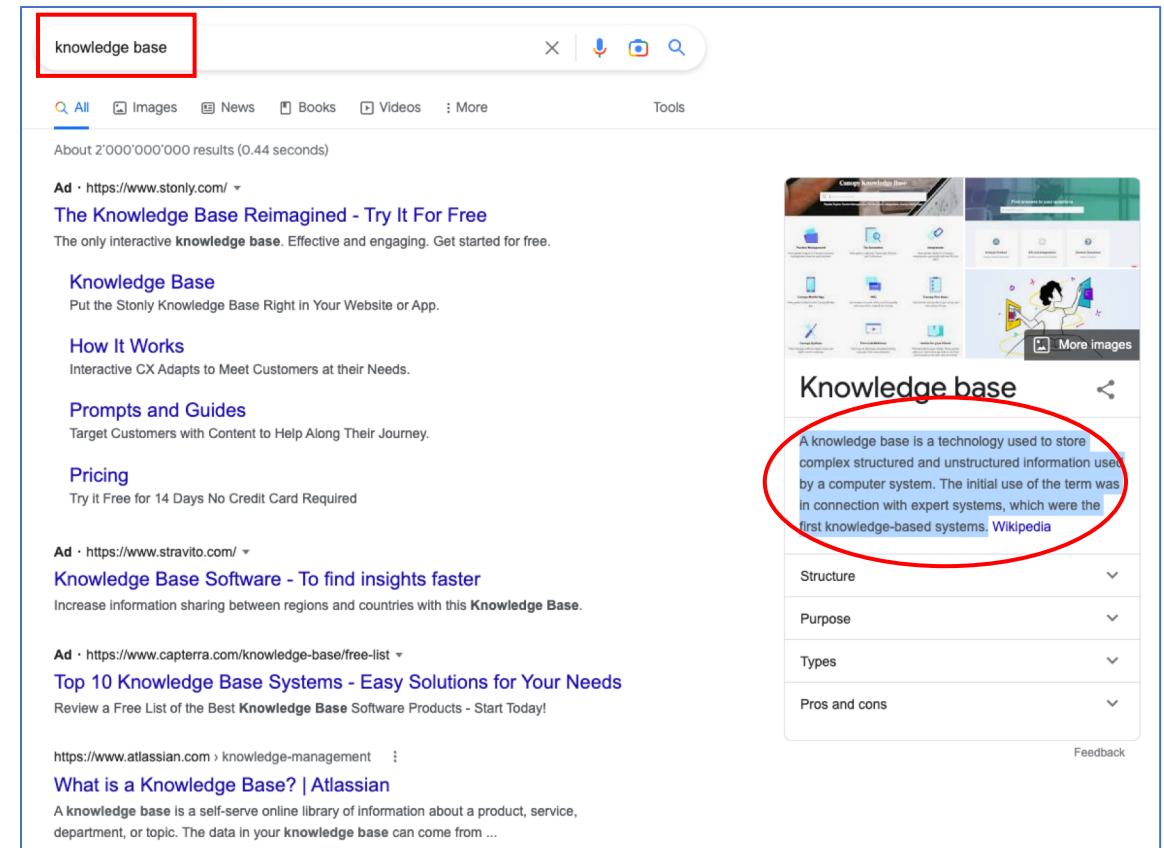
- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# Introduction: Motivation

## 1. Knowledge Graph Augmentation



## 2. Information Retrieval



The screenshot shows a search engine results page for the query "knowledge base". The search bar at the top contains the text "knowledge base". Below the search bar, there are navigation tabs for "All", "Images", "News", "Books", "Videos", and "More". The search results show approximately 2,000,000,000 results in 0.44 seconds.

The first result is an advertisement from stonly.com for "The Knowledge Base Reimagined - Try It For Free". The ad describes it as "The only interactive knowledge base. Effective and engaging. Get started for free." and includes links for "Knowledge Base", "How It Works", "Prompts and Guides", and "Pricing".

The second result is an advertisement from stravito.com for "Knowledge Base Software - To find insights faster". The ad describes it as "Increase information sharing between regions and countries with this Knowledge Base."

The third result is an advertisement from capterra.com for "Top 10 Knowledge Base Systems - Easy Solutions for Your Needs". The ad describes it as "Review a Free List of the Best Knowledge Base Software Products - Start Today!"

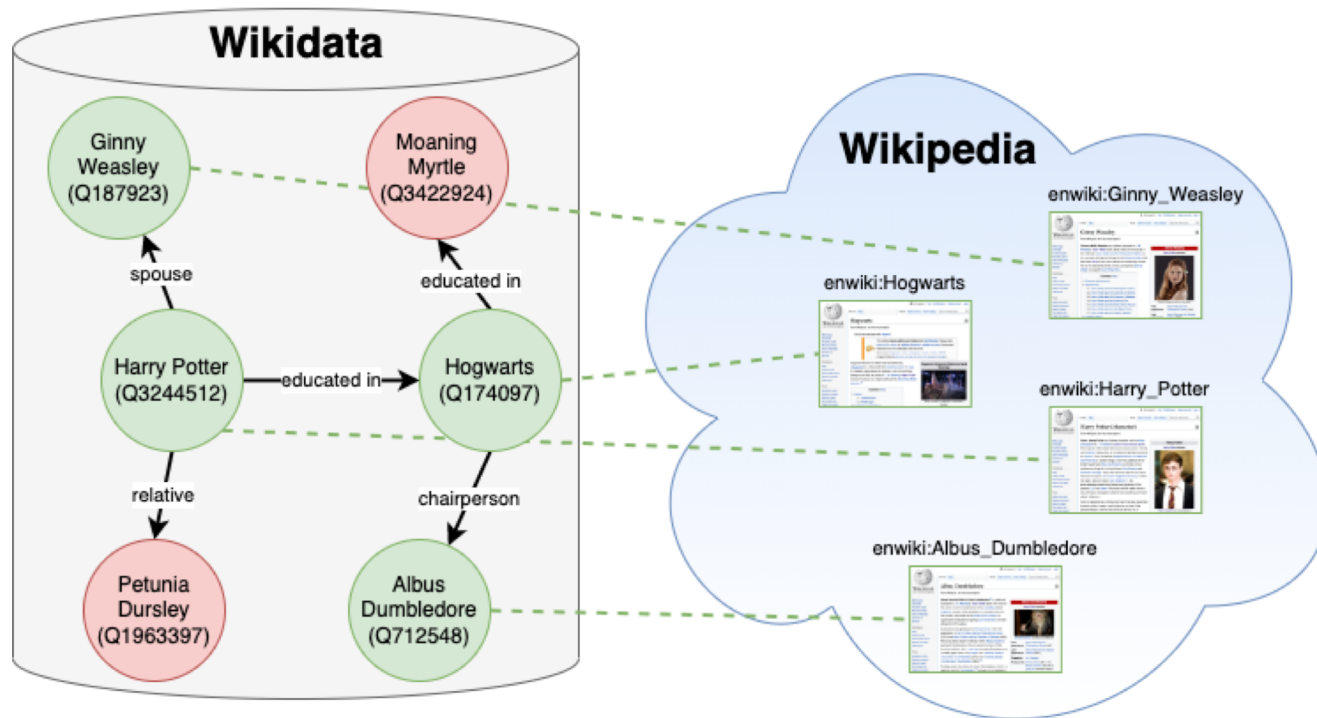
The fourth result is from atlassian.com for "What is a Knowledge Base? | Atlassian". The ad describes it as "A knowledge base is a self-serve online library of information about a product, service, department, or topic. The data in your knowledge base can come from ..."

On the right side of the search results, there is a "More images" section showing a grid of images related to knowledge bases. Below the images, there is a "Knowledge base" section with a definition: "A knowledge base is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems, which were the first knowledge-based systems. Wikipedia". This definition is circled in red. Below the definition, there are expandable sections for "Structure", "Purpose", "Types", and "Pros and cons".

# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# Introduction: Wikidata-Wikipedia Mapping

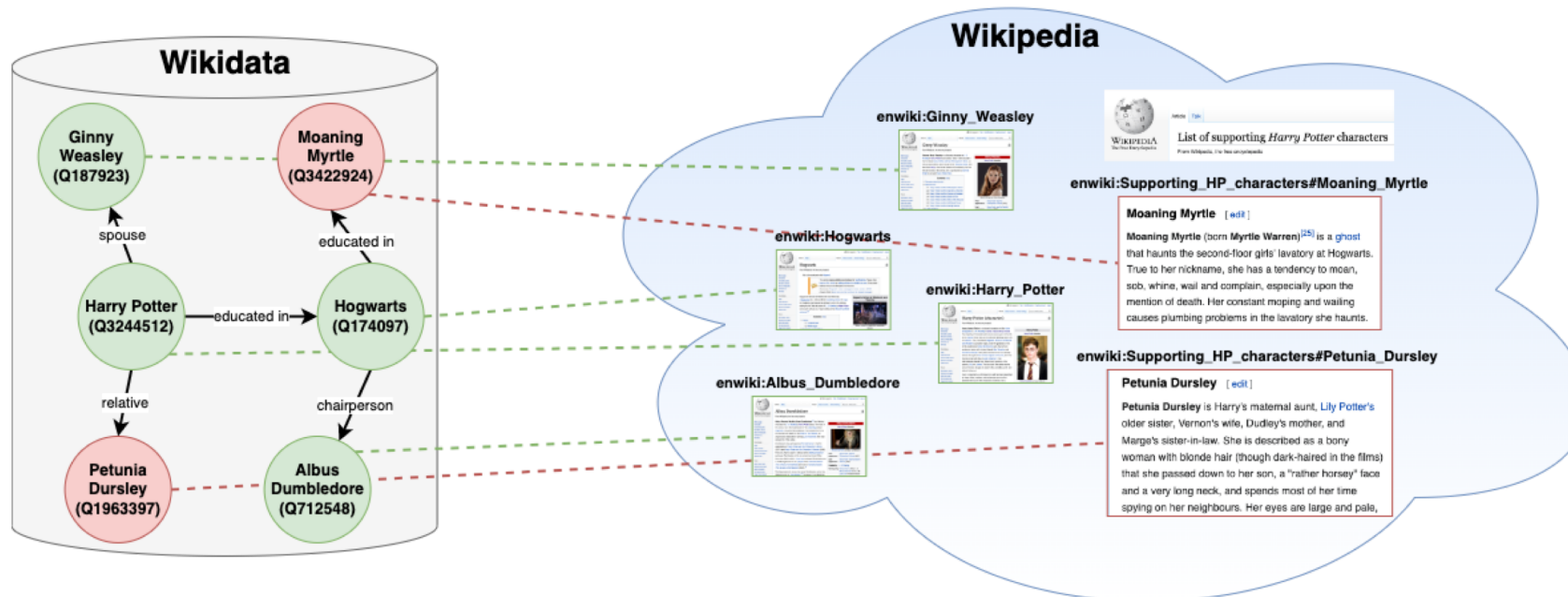


- Wikidata: multilingual knowledge graph hosted by the Wikimedia Foundation.
- Every Wikipedia article has a corresponding Wikidata entity.
- **NOT** every Wikidata entity has a corresponding Wikipedia article.
- *Orphans*: Wikidata entities not linked to a Wikipedia article.

# Introduction: Wikidata-Wikipedia Mapping

Observation: *orphan* entities are often described within existing Wikipedia articles in the form of sections, subsections, and paragraphs of a more generic concept or fact.

Task: establish a fine-grained mapping between Wikidata **orphan entities** and Wikipedia **(sub-) sections**.

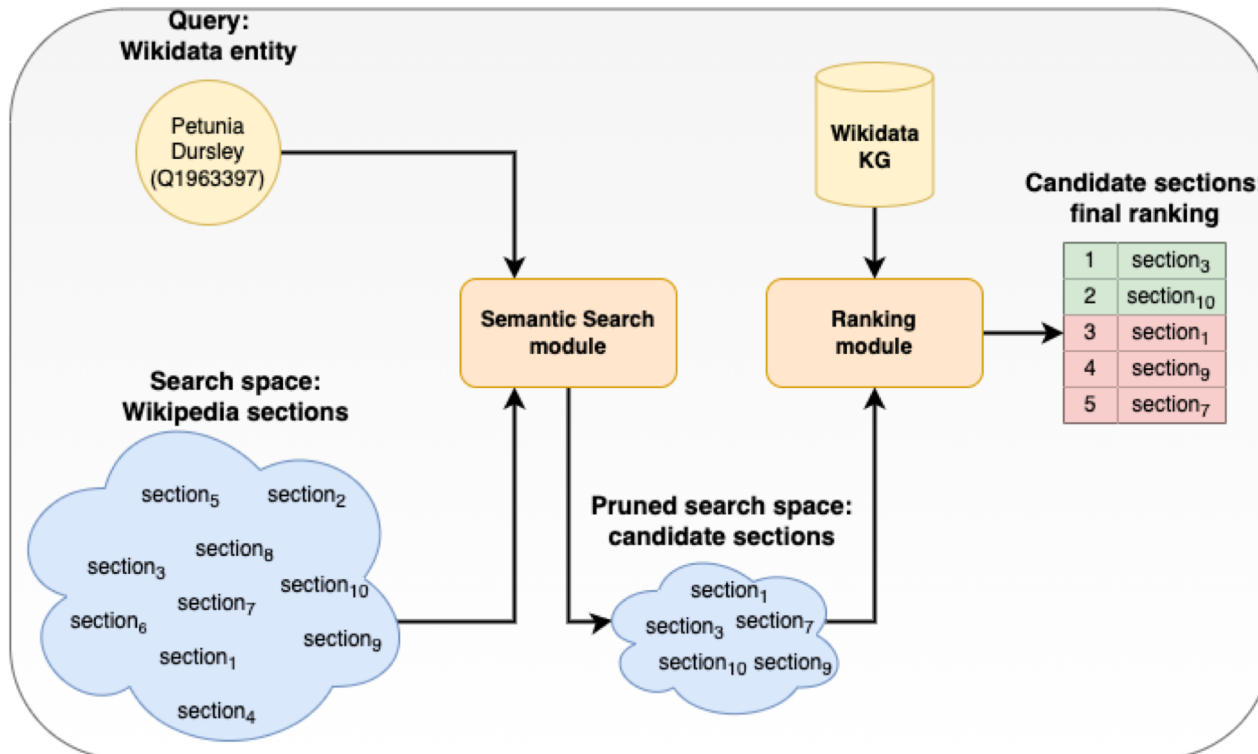


# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# ParaGraph: Overview

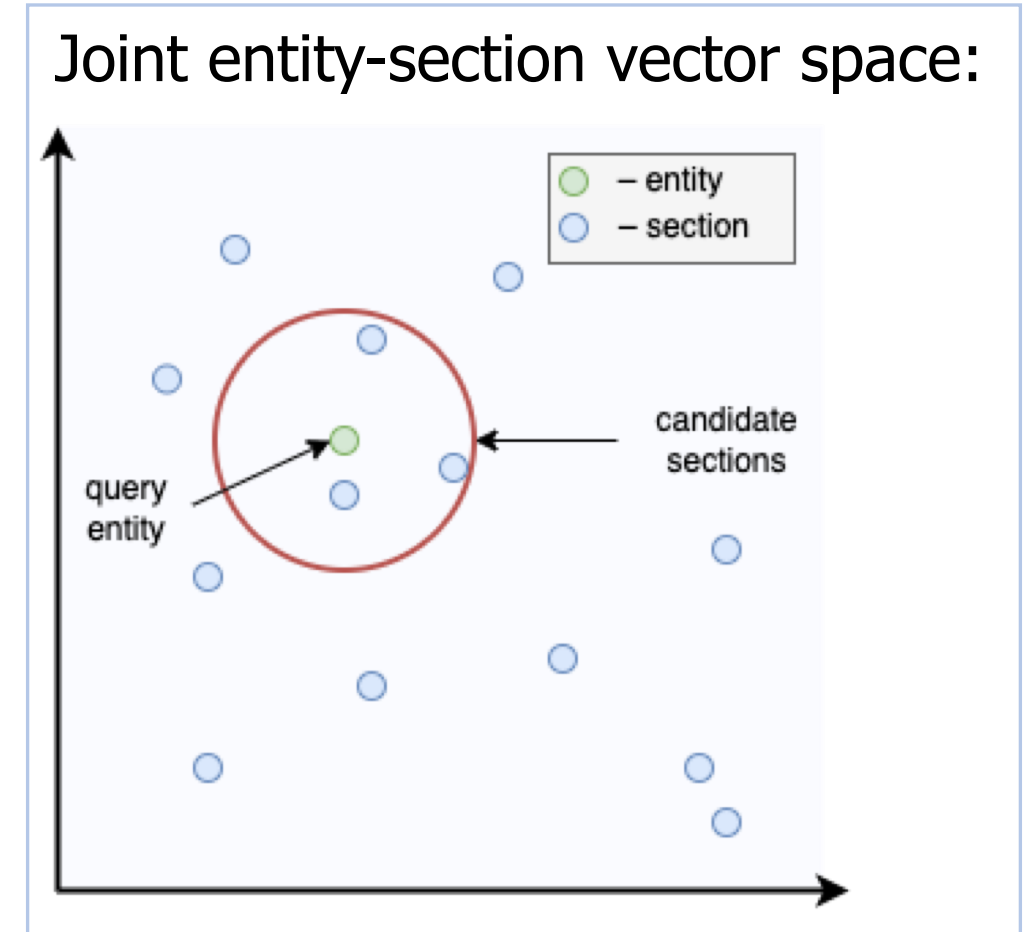
## ParaGraph framework:



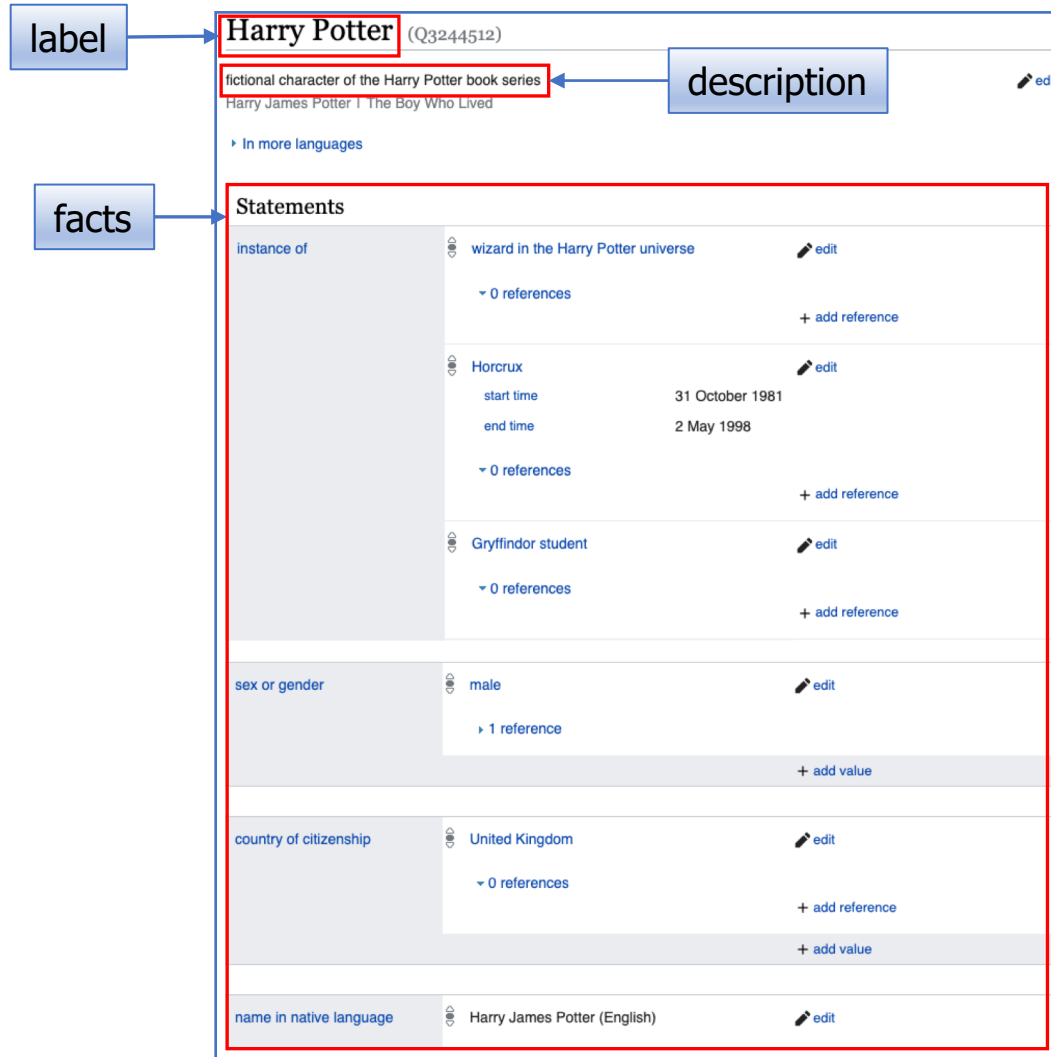
- **Input:**
  - Wikidata entity (query)
  - Wikipedia sections (search space)
- **Semantic Search module:**
  - Pre-select a subset of candidate sections
  - Search space pruning
- **Ranking module:**
  - Re-rank candidate sections
- **Output:**
  - Top sections most relevant to a query entity

# ParaGraph: Semantic Search Module

- General idea:
  - Project entities and sections onto the same vector space.
  - Select K nearest sections as candidates.
- Approaches:
  - Text-to-entity
  - **Entity-to-text**
- How to represent an entity as a text? 🤔



# ParaGraph: Entity Lexicalization



**label** → **Harry Potter** (Q3244512)

**description** → fictional character of the Harry Potter book series

**facts** → **Statements**

- instance of: wizard in the Harry Potter universe
- Horcrux: 31 October 1981 – 2 May 1998
- Gryffindor student
- sex or gender: male
- country of citizenship: United Kingdom
- name in native language: Harry James Potter (English)

Entity: **Q1963397**

Entity textual representation:

- Label: *Harry Potter*
  - Character? Movie series? Book series?
- Label + description: *Harry Potter, fictional character of the Harry Potter book series*
  - Better, but...

# ParaGraph: Entity Lexicalization

Entity (textual form): *Harry Potter, fictional character of the Harry Potter book series*

Relevant paragraph:

Harry James Potter is a fictional character and the titular protagonist in J. K. Rowling's series of eponymous novels. On his eleventh birthday he learns that he is a wizard. Thus, he attends Hogwarts School of Witchcraft and Wizardry to practise magic under the guidance of the kindly headmaster Albus Dumbledore along with his best friends Ron Weasley and Hermione Granger.

Irrelevant paragraph:

Harry Potter is a fictional character who is more alive than any of us! To all Potterheads, Potter is a famous half-blood prince who reincarnates, again and again, to emerge as one of the most famous wizards of all time!

# ParaGraph: Entity Lexicalization

Entity textual representation: **label + description + facts:**

*Harry Potter* IS A *fictional character of the Harry Potter book series.*

*Instance of Gryffindor student.*

*Sex or gender male.*

*Country of citizenship United Kingdom.*

*Date of birth 31 July 1980.*

*Father James Potter.*

*Mother Lily Potter.*

*Native language British English.*

*Educated at Hogwarts.*

...

# ParaGraph: Entity Lexicalization

- Entity textual representation: **label + description + facts.**
- Not all facts are equally useful.

+	<i>Harry Potter IS A fictional character of the Harry Potter book series.</i>
+	<i>Instance of Gryffindor student.</i>
–	<i>Sex or gender male.</i>
?	<i>Country of citizenship United Kingdom.</i>
+	<i>Date of birth 31 July 1980.</i>
?	<i>Father James Potter.</i>
?	<i>Mother Lily Potter.</i>
–	<i>Native language British English.</i>
+	<i>Educated at Hogwarts.</i>

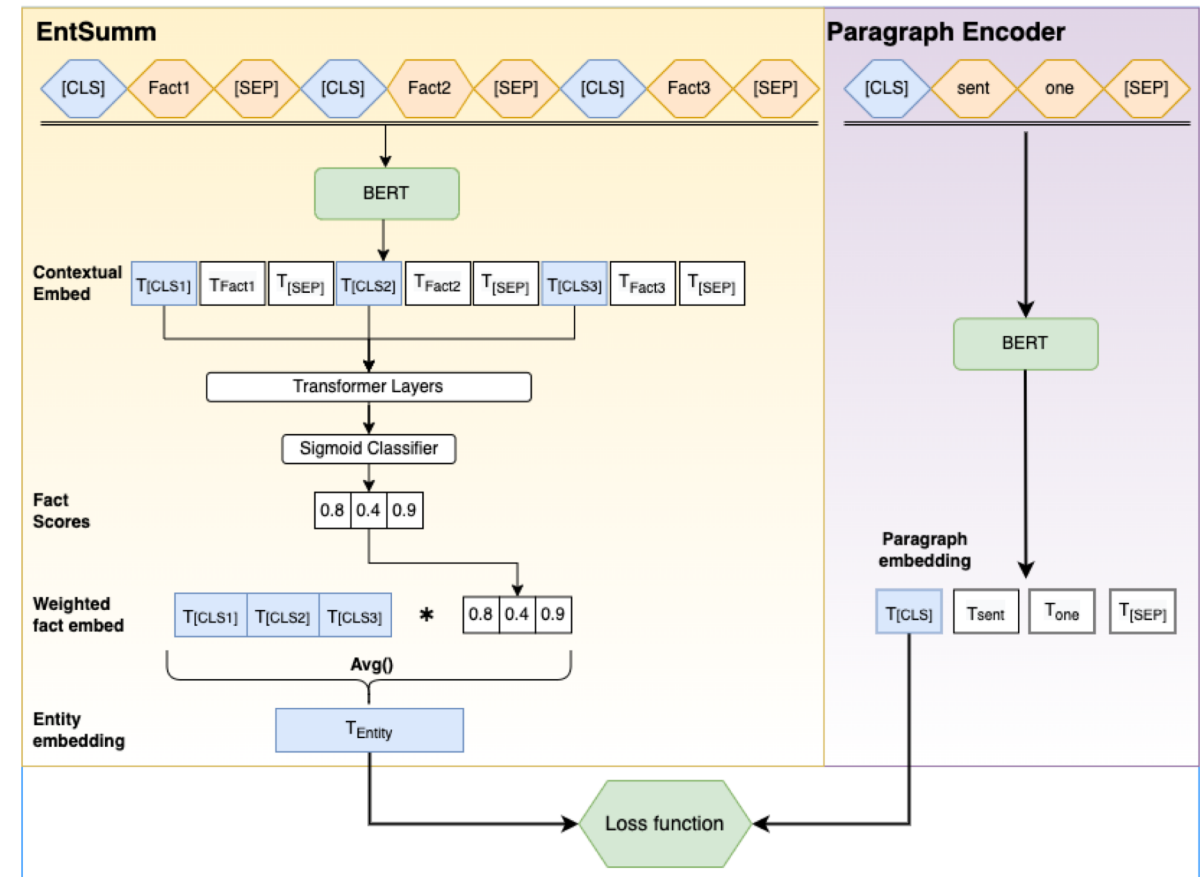
# ParaGraph: Entity Summarization

## • EntSumm:

- Select facts that summarize an entity in the best way.
- Objective: generate an entity representation similar to a description in natural language.

## • Algorithm:

- Fact-level encoder based on BERT
- Express the semantics of a entity and obtain representations for its facts.
- Based on BERTSumExt [1].



# ParaGraph: Semantic Search Module

## Semantic Search Algorithm

1. Preprocessing phase: represent an entity with its label, description and concatenation of all facts.
2. Training phase: jointly train an entity summarizer (EntSumm) and a paragraph encoder.
3. Inference phase: generate an entity summary embedding and paragraphs embeddings; select top-K nearest paragraphs as a candidate set.

# ParaGraph: Ranking Module

- Learning to Rank (LTR) algorithm – **LambdaRank**.
- Feature set:
  - **SimScore** – semantic similarity between entity and paragraph (from the previous step).
  - **TitleScore** – semantic similarity between the entity label and the section title.
  - **PathScore** – score of the path between the query entity and the article entity.

# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# Experiments: Dataset Collection

## 1. Mappings between Wikidata entities and Wikipedia articles.

Minerva McGonagall

From Wikipedia, the free encyclopedia

**Professor Minerva McGonagall** is a fictional character in J. K. Rowling's *Harry Potter* series. Professor McGonagall is a professor at *Hogwarts School for Witchcraft and Wizardry*, the head of *Gryffindor House*, the professor of *Transfiguration*, the *Deputy Headmistress* under *Albus Dumbledore* and a member of the *Order of the Phoenix*. Following *Lord Voldemort*'s defeat at the hands of her student *Harry Potter* and the deaths of Headmasters *Albus Dumbledore* and *Severus Snape*, McGonagall takes the position of Headmistress. Professor McGonagall was portrayed in the film adaptations by actress *Maggie Smith*, and later by *Fiona Glascott* in the *Fantastic Beasts* prequel films *The Crimes of Grindelwald* and *The Secrets of Dumbledore*.<sup>[1]</sup>

Contents [hide]

1 Fictional character biography

1.1 Education and employment at Hogwarts

## 3. Links between sections and full articles.

Balls

Main article: [Tennis ball](#)

Tennis balls were originally made of cloth strips stitched together with the optic yellow in the latter part of the 20th century to allow for improved visibility. The official diameter as 65.41–68.58 mm (2.575–2.700 in). Balls must weigh 56–59.4 g (2.0–2.1 oz). As they remained virtually unchanged for the past 100 years, the majority of manufacturers must use balls that are approved by the International Tennis Federation.

## 2. Interlingual Wikipedia.

Download as PDF

Printable version

Languages

Català

Deutsch

Español

فارسی

Français

Polski

中文

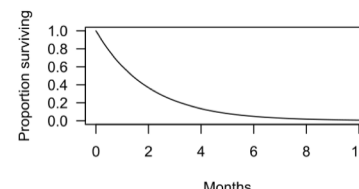
3 more

Edit links

Examples of survival functions [edit]

The graphs below show examples of hypothetical survival functions beyond time t.

Survival function 1



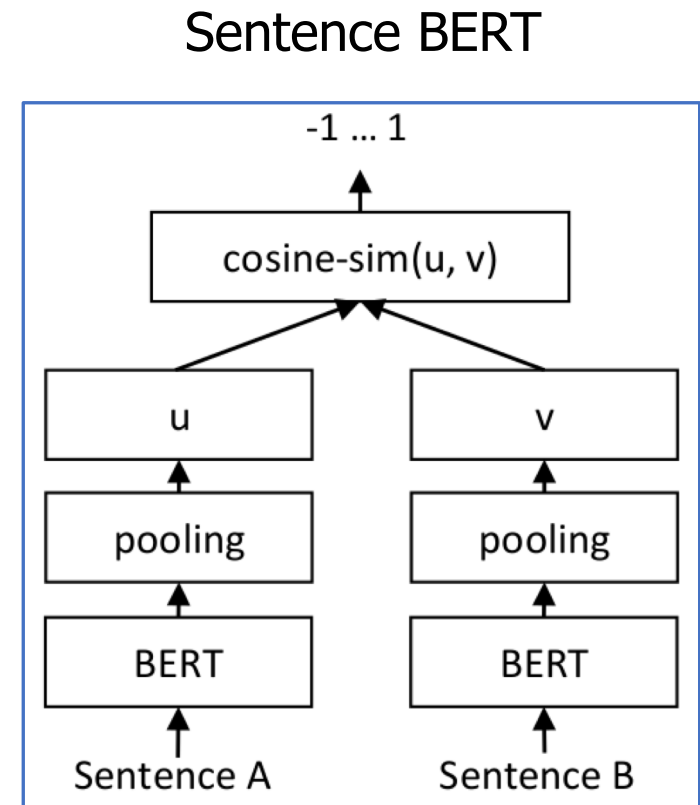
[https://en.wikipedia.org/wiki/Survival\\_function](https://en.wikipedia.org/wiki/Survival_function)

[https://fr.wikipedia.org/wiki/Analyse\\_de\\_survie#Fonction\\_de\\_survie](https://fr.wikipedia.org/wiki/Analyse_de_survie#Fonction_de_survie)

# Experiments: Semantic Search Baseline

- Semantic Textual Similarity (STS) task.
- Fine-tuned Sentence BERT model.

Entity representation:	
Label:	$SBERT_{label}$
Label + description:	$SBERT_{label\_descr}$
Label + description + bag-of-facts:	$SBERT_{bof}$



<https://www.sbert.net/examples/training/sts/README.html>

# Experiments: Semantic Search Results

Entity2Section dataset (min. 2 facts, avg. 10 facts):

	Hits@1	Hits@10	Hits@50
SBERT <sub>label</sub>	39.73	70.76	82.16
SBERT <sub>label_descr</sub>	46.81	76.42	86.61
SBERT <sub>bof</sub>	47.40	<b>80.02</b>	90.28
ParaGraph (EntSumm)	<b>48.79</b>	<b>80.02</b>	<b>90.75</b>

RichEntity2Section dataset (min. 10 facts, avg. 20 facts):

	Hits@1	Hits@10	Hits@50
SBERT <sub>bof</sub>	51.57	81.43	90.82
ParaGraph (EntSumm)	<b>57.17</b>	<b>83.75</b>	<b>92.10</b>

# Experiments: Information Retrieval Baselines

- ParaGraph: Semantic Search + Ranking
- Information Retrieval models:
  - BM25
  - ColBERT
  - ANCE
- Entity representation:
  - Label: (i)
  - Label + description: (ii)

# Experiments: ParaGraph Results

	Hits@1	Hits@10	Hits@50
BM25 (i)	46.94	85.60	<b>92.36</b>
BM25 (ii)	43.10	83.30	91.75
ColBERT (i)	41.95	77.27	90.46
ColBERT (ii)	46.87	80.12	92.18
ANCE (i)	39.25	68.92	82.11
ANCE (ii)	41.61	72.84	85.78
ParaGraph	<b>70.27</b>	<b>87.34</b>	90.75

# Contents

- Introduction
  - Problem definition
  - Motivation
  - Wikidata to Wikipedia
- ParaGraph
- Experiments and Results
- Conclusion and future work

# Conclusion and Future Work

- Introduce **entity mapping** task.
- Present **ParaGraph** – a two stage framework for mapping **Wikidata entities** onto **Wikipedia sections**.
- Demonstrate the effectiveness of **ParaGraph** compared to **STS** and **IR** baselines.

## Future work:

- Unsupervised entity summarization.
- Extension beyond Wikidata-Wikipedia scenarios.

# Thank you!



<https://exascale.info>