LIMSI PhD Day

ANR Project - ILES Group - ED STIC

June 6th, 2019

A priori knowledge and domain adaptation for building word embeddings in specialized domains

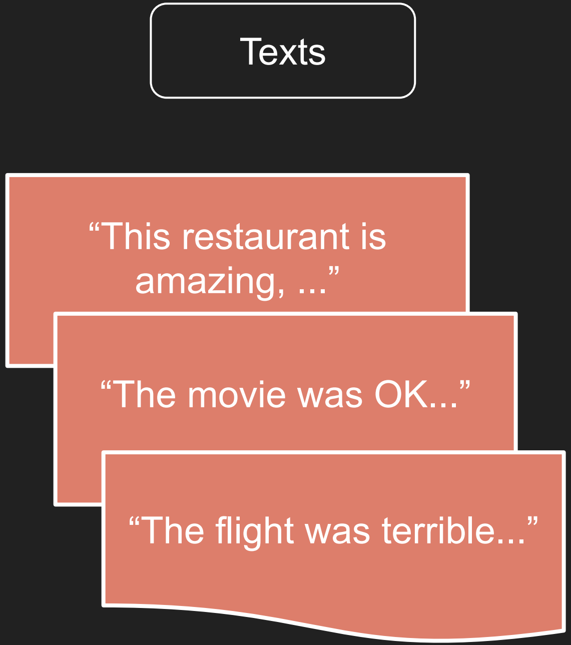
PhD Student

[LIMSI] **Hicham EL BOUKKOURI**

Supervisors

[LIMSI] **Pierre ZWEIGENBAUM**
[LIMSI] **Thomas LAVERGNE**
[CEA-LIST] **Olivier FERRET**

NLP: Natural Language Processing



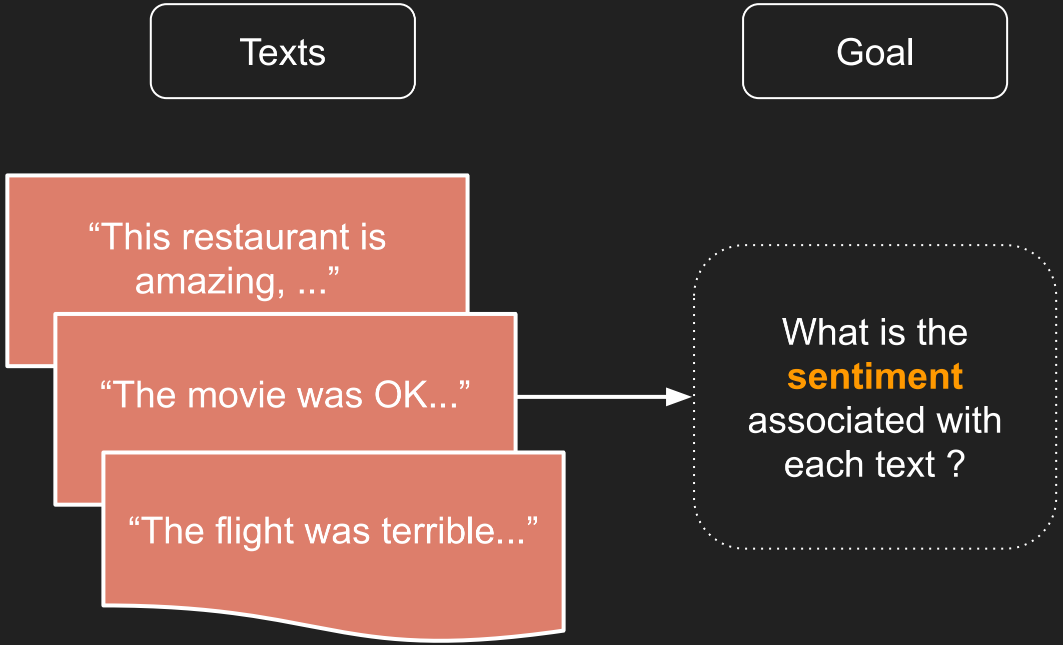
Texts

“This restaurant is amazing, ...”

“The movie was OK...”

“The flight was terrible...”

NLP: Natural Language Processing



Texts

Goal

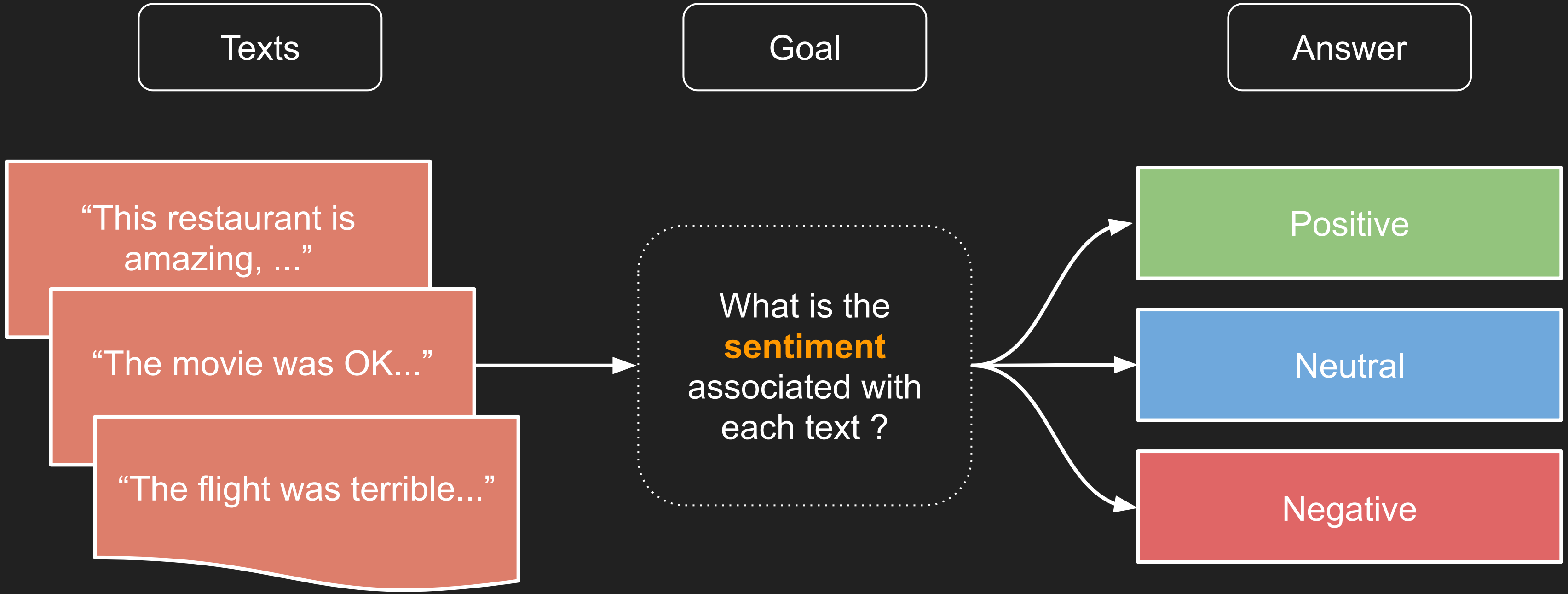
“This restaurant is amazing, ...”

“The movie was OK...”

“The flight was terrible...”

What is the **sentiment** associated with each text ?

NLP: Natural Language Processing



Texts

Goal

Answer

“This restaurant is amazing, ...”

“The movie was OK...”

“The flight was terrible...”

What is the **sentiment** associated with each text ?

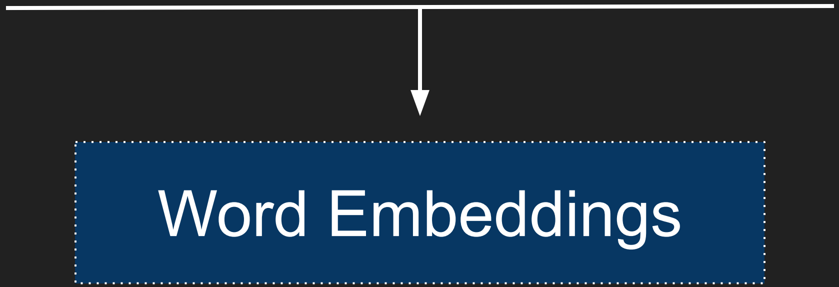
Positive

Neutral

Negative

Word Embeddings

We need to get the algorithm to "understand" the meaning of words.



Word Embeddings

Word Embeddings: General Approach

Contexts in a collection of documents

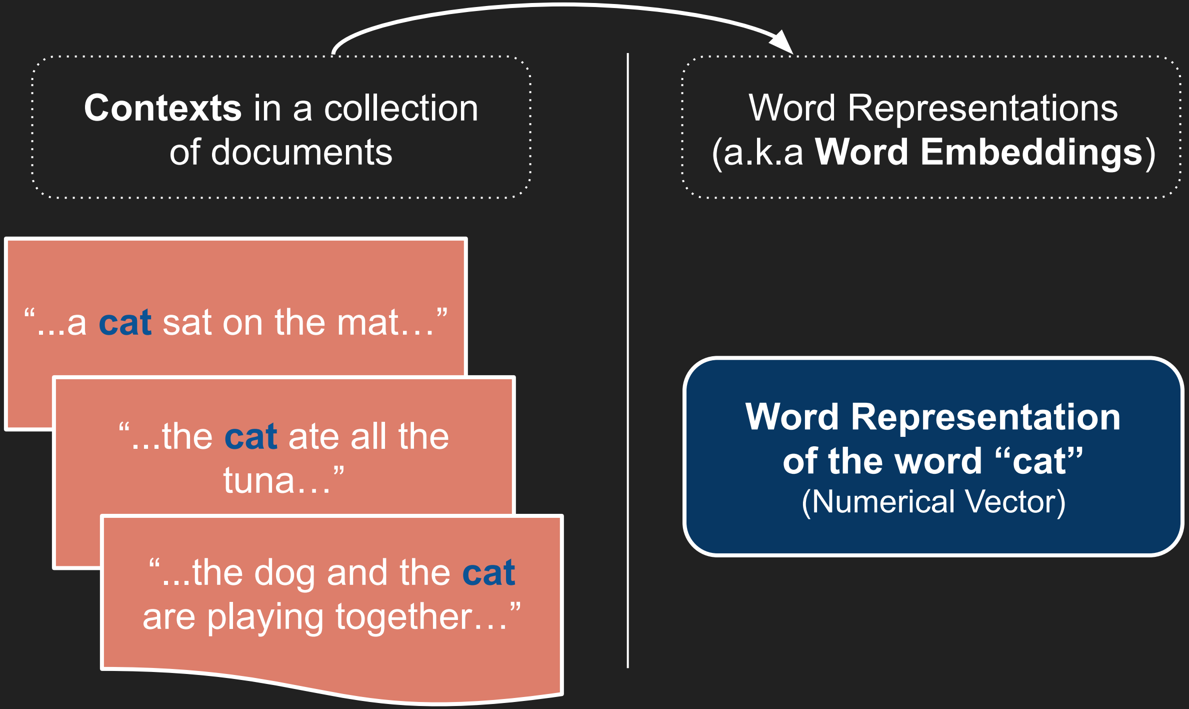
“...a **cat** sat on the mat…”

“...the **cat** ate all the tuna…”

“...the dog and the **cat** are playing together…”

Word Embeddings: General Approach

Step 1: Learn Embeddings



Contexts in a collection of documents

"...a **cat** sat on the mat…"

"...the **cat** ate all the tuna…"

"...the dog and the **cat** are playing together…"

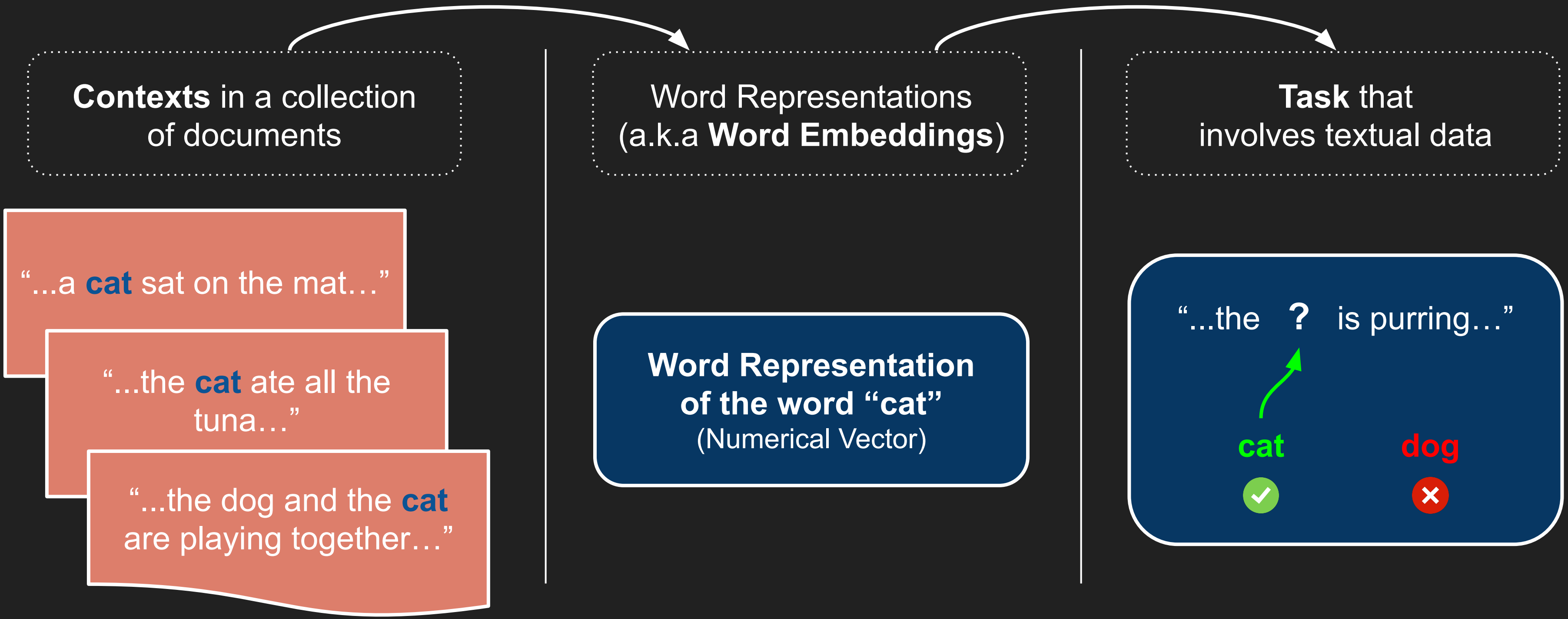
Word Representations (a.k.a **Word Embeddings**)

Word Representation of the word "cat"
(Numerical Vector)

Word Embeddings: General Approach

Step 1: Learn Embeddings

Step 2: Solve a given Task



Contexts in a collection of documents

Word Representations (a.k.a **Word Embeddings**)

Task that involves textual data

"...a **cat** sat on the mat..."

"...the **cat** ate all the tuna..."

"...the dog and the **cat** are playing together..."

Word Representation of the word "cat"
(Numerical Vector)

"...the **?** is purring..."

cat ✓

dog ✗

Case of Specialized Domains (e.g Medical Domain)

Case of Specialized Domains (e.g Medical Domain)

Medical Concept Detection*

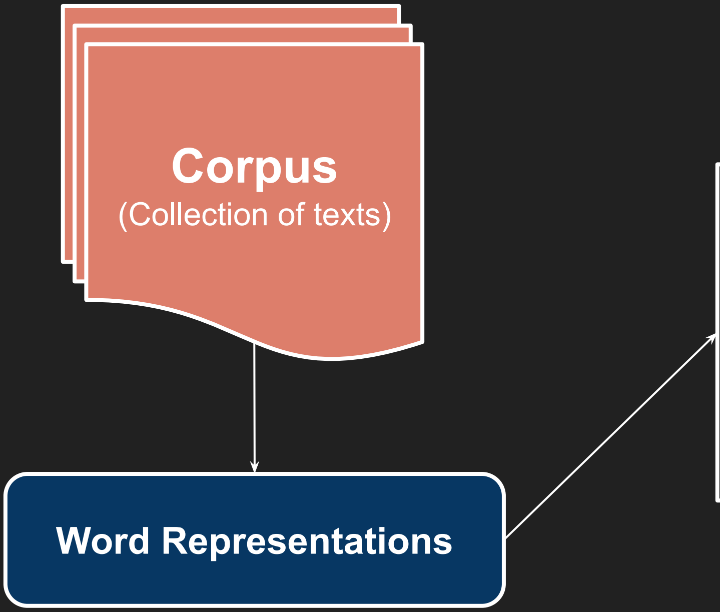
The patient had **headache** that was relieved only with **oxycodone** . A **CT scan of the head** showed **microvascular ischemic changes** . A **followup MRI** which also showed **similar changes** . This was most likely due to **her multiple myeloma** with **hyperviscosity** .

Medical Concept Types

Problem Treatment Medical Test

* i2b2/VA 2010 Concept Detection Task (**Uzuner et al., 2011**)

Case of Specialized Domains (e.g Medical Domain)



Corpus

(Collection of texts)

Word Representations

Medical Concept Detection*

The patient had **headache** that was relieved only with **oxycodone** . A **CT scan of the head** showed **microvascular ischemic changes** . A **followup MRI** which also showed **similar changes** . This was most likely due to **her multiple myeloma** with **hyperviscosity** .

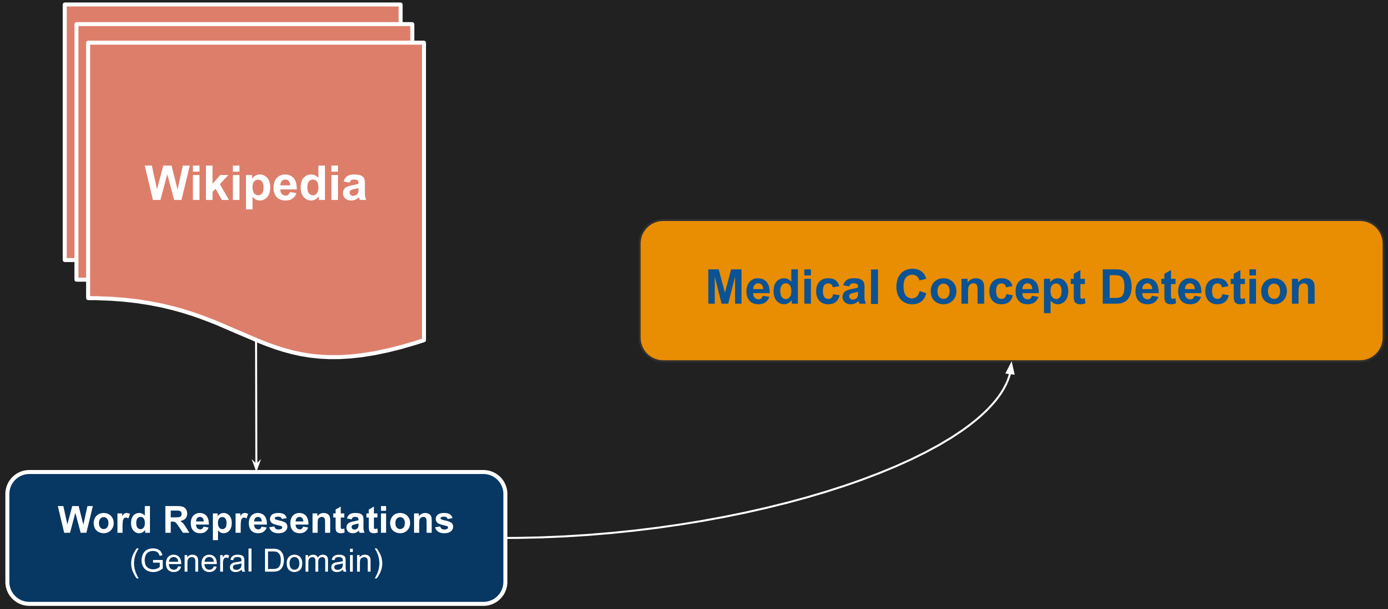
Medical Concept Types

🟥 Problem 🟩 Treatment 🟦 Medical Test

* i2b2/VA 2010 Concept Detection Task (**Uzuner et al., 2011**)

Using General Domain Corpora

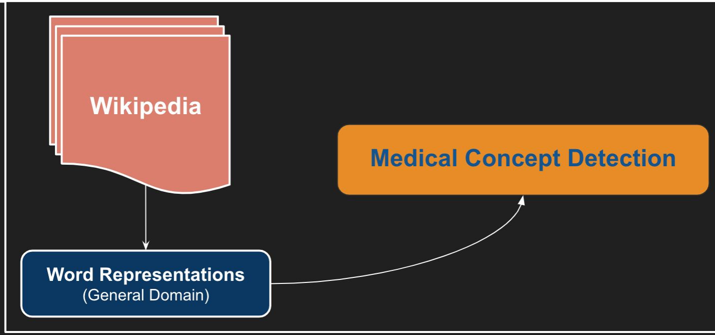
Using General Domain Corpora



Using General Domain Corpora [Issues]

Issues with using a corpus like Wikipedia

Many medical terms **never appear** in the corpus
⇒ Can’t learn their word representations

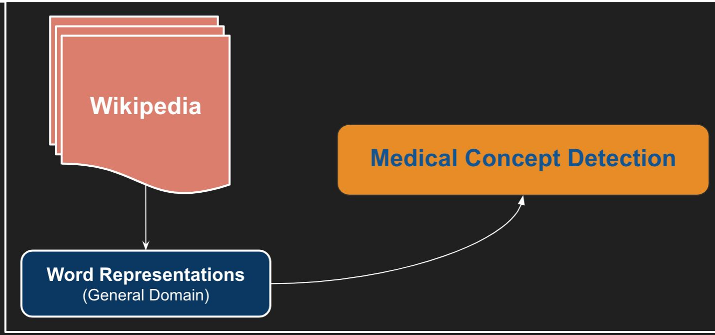


Using General Domain Corpora [Issues]

Issues with using a corpus like Wikipedia

Many medical terms **never appear** in the corpus
⇒ Can’t learn their word representations

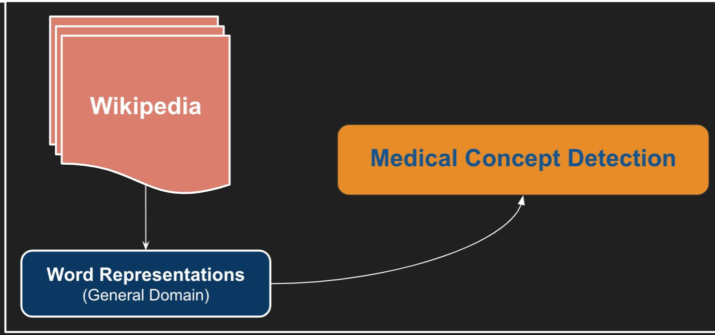
Those who do appear, do not appear **frequently enough**
⇒ Not enough contexts to precisely deduce their meaning



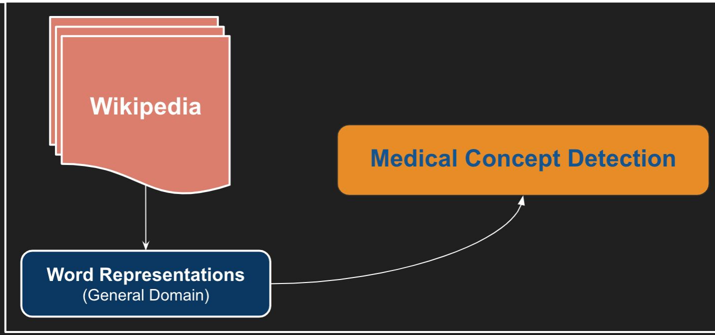
Using General Domain Corpora [Solutions]

Possible Solutions

Use a **large** collection of **specialized texts**
(e.g. scientific medical articles)
⇒ **Not always possible**



Using General Domain Corpora [Solutions]



Possible Solutions

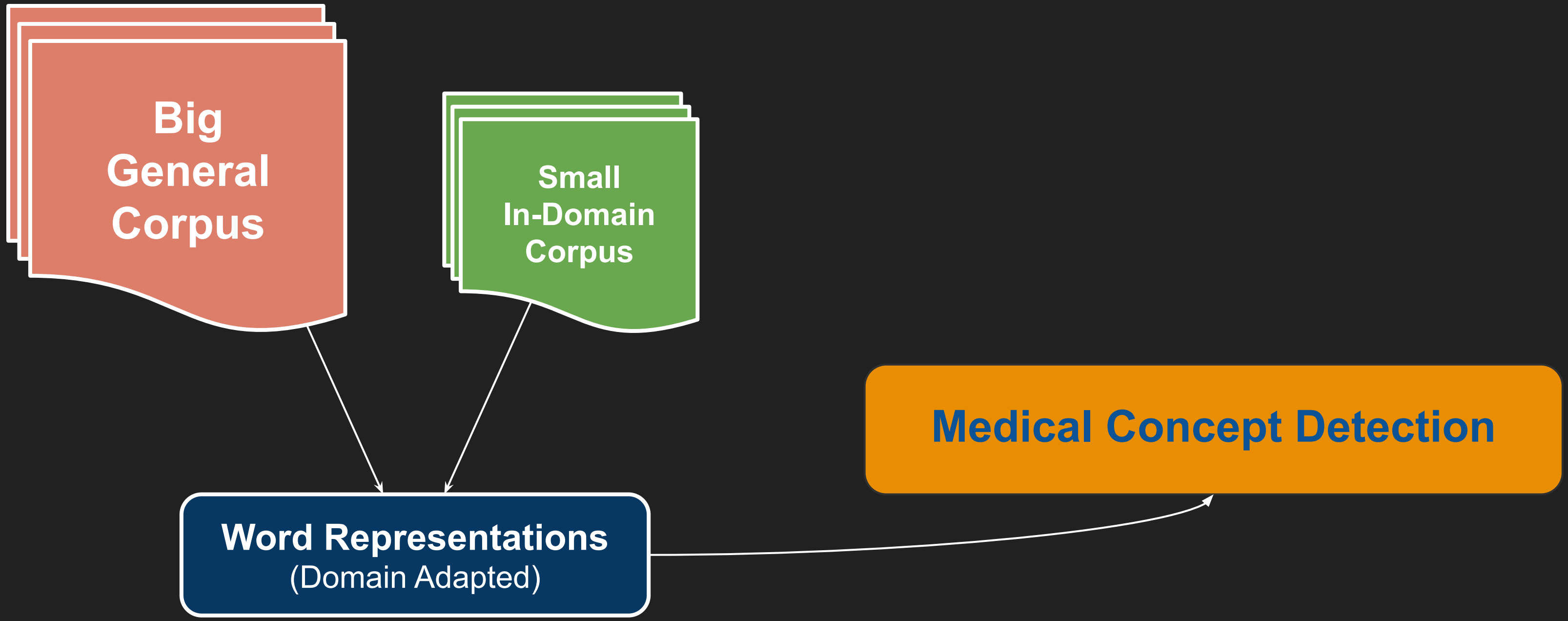
Use a **large** collection of **specialized texts**
(e.g. scientific medical articles)
⇒ **Not always possible**

Adapt general domain embeddings
using **domain knowledge**

Domain Adaptation w/ Domain Knowledge: explored so far

Domain Adaptation w/ Domain Knowledge: explored so far

- Mix general-domain embeddings with representations learned on small in-domain data



Big General Corpus

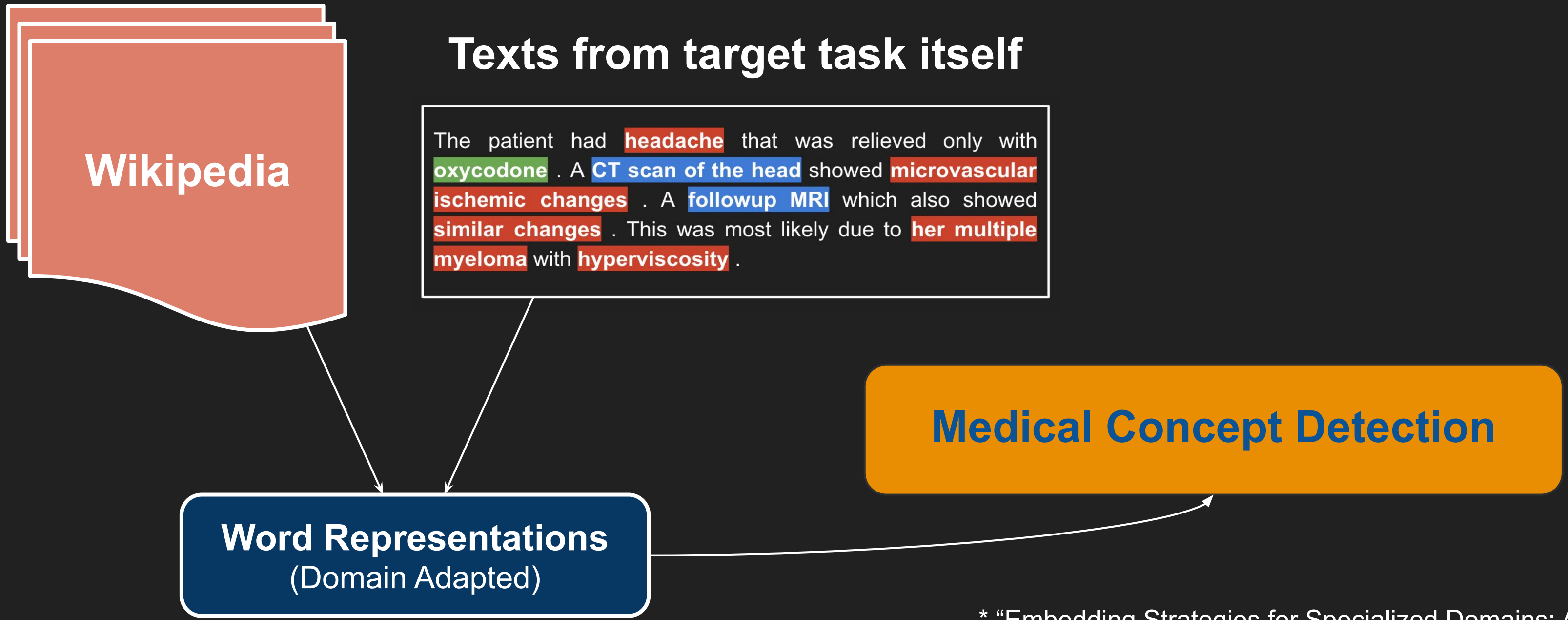
Small In-Domain Corpus

Word Representations (Domain Adapted)

Medical Concept Detection

Domain Adaptation w/ Domain Knowledge: explored so far

- Mix general-domain embeddings with representations learned on the task data itself*



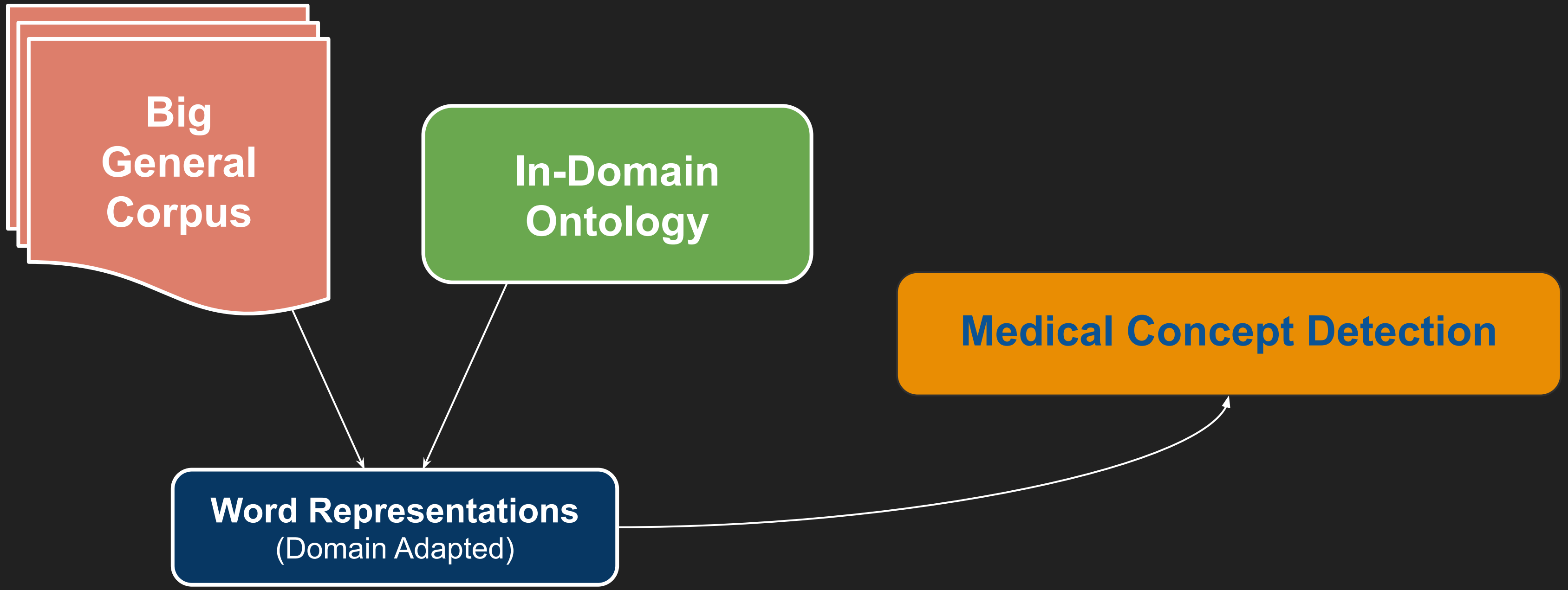
The patient had **headache** that was relieved only with **oxycodone** . A **CT scan of the head** showed **microvascular ischemic changes** . A **followup MRI** which also showed **similar changes** . This was most likely due to **her multiple myeloma** with **hyperviscosity** .

* “Embedding Strategies for Specialized Domains: Application to Clinical Entity Recognition” (**Paper accepted @ACL SRW**)

Domain Adaptation w/ Domain Knowledge: yet to explore

Domain Adaptation w/ Domain Knowledge: yet to explore

- Adapt general-domain embeddings using **ontologies** from the domain of interest



Big General Corpus

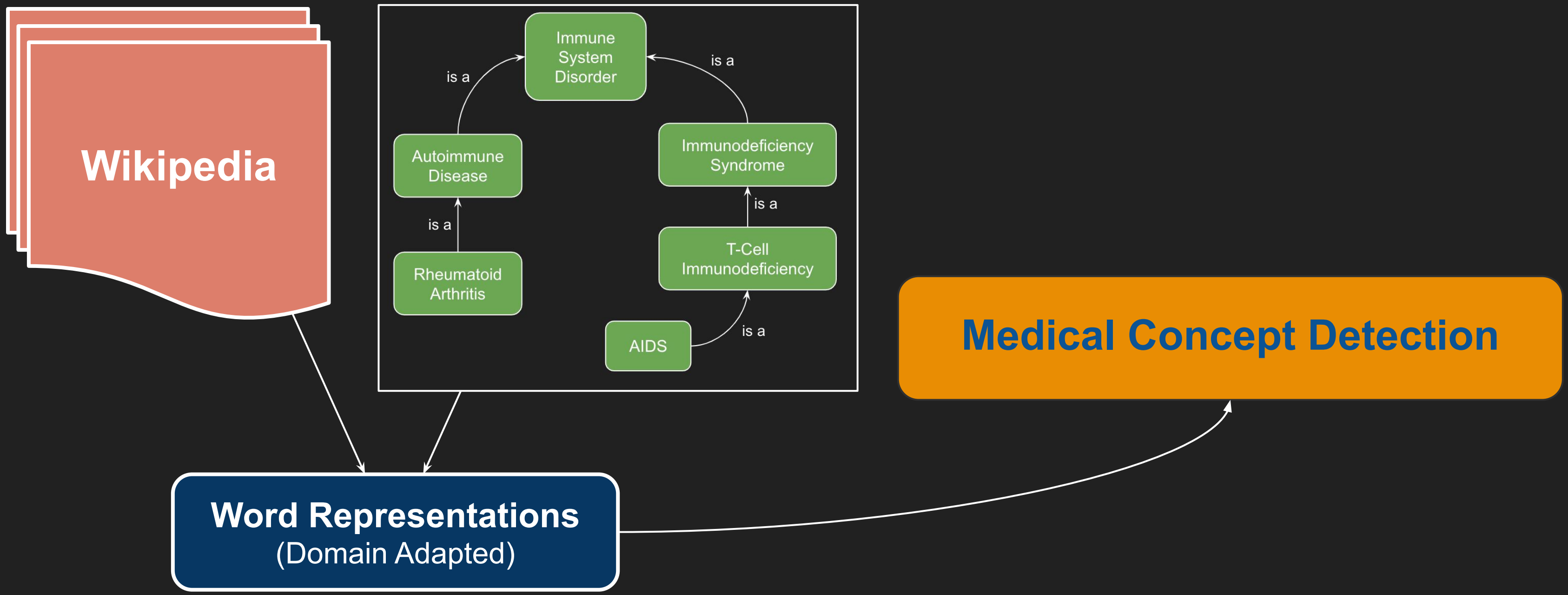
In-Domain Ontology

Medical Concept Detection

Word Representations
(Domain Adapted)

Domain Adaptation w/ Domain Knowledge: yet to explore

- Adapt general-domain embeddings using **ontologies** from the domain of interest



Domain Adaptation w/ Domain Knowledge: yet to explore

- Something else...

?

Thanks for your attention

Thanks for your attention

Any questions ?