ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Iprova
INVENT FIRST

From Relation Extraction to Knowledge Graphs

Master Thesis Project – IC School

Diego Antognini — **Current PhD Student (LIA)**

diegoantognini@gmail.com

Dr. Jean-Cédric Chappelier — **EPFL Supervisor**

Bernard Maccari — **Iprova Supervisor**

Outline

1. Problem Statement
2. Our Method
   - Relation Classification Models
   - Building Knowledge Graphs
3. Conclusion & Future Work

Our Objectives

1. Given a sentence, extract concepts and find the relationship among them if such one exists.

   EEG measures brain activities

2. Given a corpus, build Knowledge Graphs of concepts, favoring precision over recall.

- What are concepts ?
- What are the relationships ?
- What are Knowledge Graphs ? Why use them ?

$\Rightarrow$ Come back later !

Concepts

- Short phrases made of adjectives and nouns
  - Gyroscope ✓
  - Rotational motion ✓
  - Brain electrical activity ✓
  - The new model S developed by Tesla ✗
  - Galaxy S8 of Samsung ✗

Relations

Directed relations Iprova is interested in

Relation	Example
Cause	Those cancers were caused by radiation exposures.
Contain	My apartment has a large kitchen.
Measure	EEG measures brain activities.
Produce	A factory manufactures suits.
TypeOf	NoSQL databases such as MongoDB.
Use	Bluetooth is used in audio equipment.
Other	A misty ridge uprises from the surge.

Relations

Directed relations Iprova is interested in

Relation	Example
Cause	Those cancers were caused by radiation exposures.
Contain	My apartment has a large kitchen.
Measure	EEG measures brain activities.
Produce	A factory manufactures suits.
TypeOf	NoSQL databases such as MongoDB.
Use	Bluetooth is used in audio equipment.
Other	A misty ridge uprises from the surge.

2 First Approaches

1. **Relation Extraction**
   - Named entities (Location, Organization, Person, etc.)
   - Specified relations (e.g. CoFounder, BornIn)
   - **Need a lot of data**
2. **Open Information Extraction**
   - Named entities/nominals (nouns/base noun phrases)
   - No specified relations

     ⇒ **find a mapping to “ontology”**

     e.g. was included in ⇒ Contain(e2,e1)

Relation/Open Information Extraction

Bill Gates, **Microsoft co-founder**, stepped down as **CEO** in January 2000. **Gates** was included in the **Forbes wealthiest list** since 1987 and **was the wealthiest** from 1995 to 2007...

It was announced that **IBM** would buy **Ciao** for an undisclosed amount. The **CEO**, **MacLorrance** has occupied the **corner office of the Hopkinton**, company

The company's storage business is also threatened by new, born-on-the Web could providers like Dropbox and Box, and …

RE →

Co-founder(Bill Gates, Microsoft)
Director-of (MacLorrance, Ciao)
Employee-of (MacLorrance, Ciao)
…

Open IE →

(Bill Gate, be, Microsoft co-founder)
(Bill Gates, stepped down as, CEO)
(Bill Gates, was included in, the Forbes wealthiest list)
(Bill Gates, was, the wealthiest)
(IBM, would buy, Ciao)
(MacLorrance, has occupied, the corner office of the Hopkinton)
…

	RE	**Open IE**
Input	Sentences + Labeled relations	Sentences
Relation	Specified relations in advance	Free discovery
Extractor	Specified relations	Independent-relations

Image from Vo and Bagheri, 2016

Chosen approach

3. **Relation Classification**

- Specified relations
- Named entities/Nominals/Concepts
  - **These are given with the sentence**
  - How to find concepts ?
    $\Rightarrow$ **Using existing concept extraction system**
    (ADJ)* (NOUN)+

Our Method

Relation Classification Models

Building Knowledge Graphs

Relation Classification

Input:

"The [factory]$_{e1}$'s products have included flower pots, Finnish rooster-whistles, pans, [trays]$_{e2}$, tea pots, ash trays and air moisturisers."

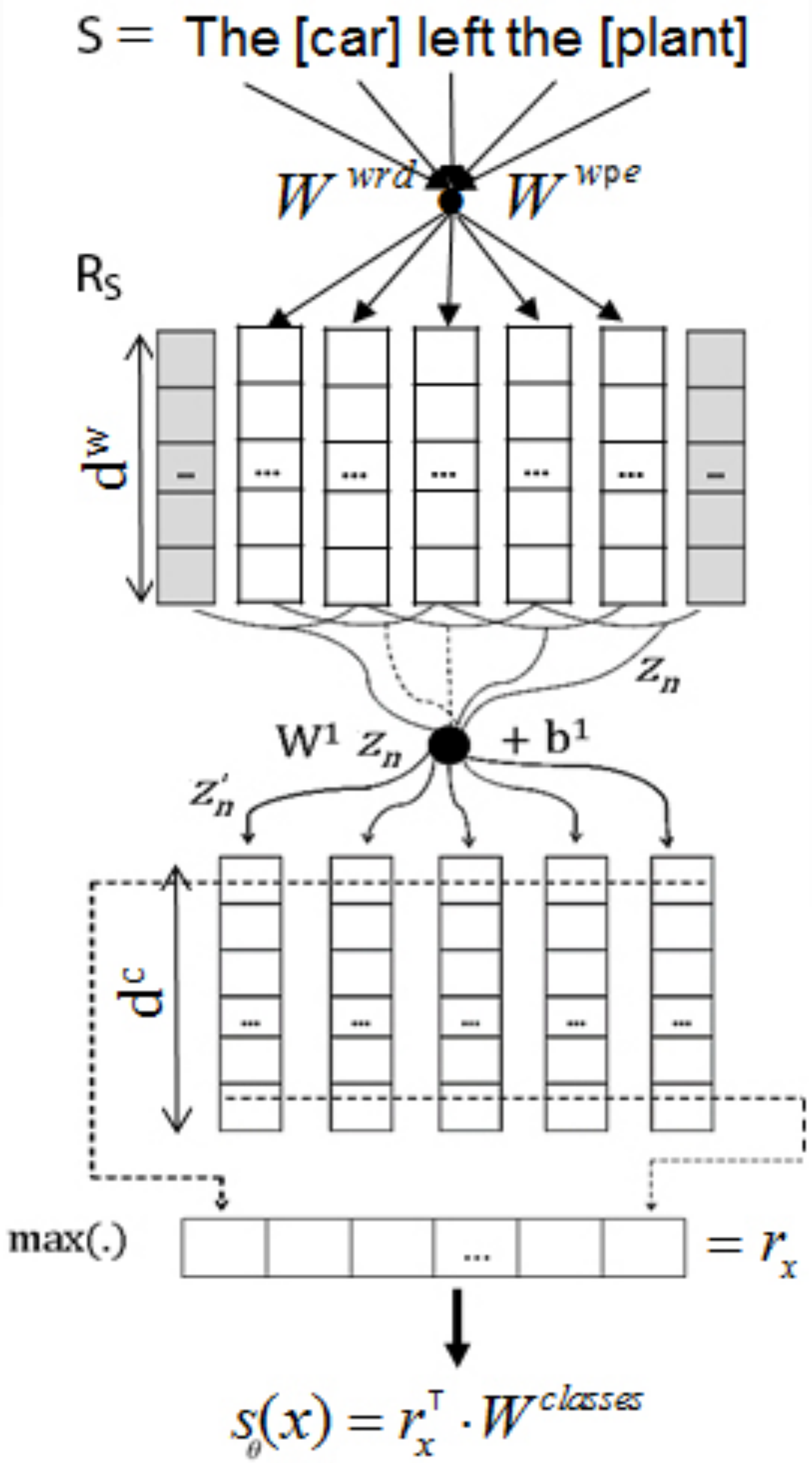
Output:

- The directed relationship among factory and trays
  $\Rightarrow$ **Produce(factory, trays)**

Developed models

- 2 Models for Relation Classification Task:
  - CR-CNN: Convolutional neural networks
  - BRCNN: Recurrent & Convolutional neural networks
- These have been shown to be efficient architecture for RC

Model 1: CR-CNN



- State of the art for 2015
- Convolutional Neural Network
- Simple features: word embeddings and relative distance
- Omit “Other” class embeddings
- Pairwise ranking loss function

$$L = log(1 + exp(\gamma(m^+ - s_\theta(x)_{y^+}))$$
$$+ log(1 + exp(\gamma(m^- + s_\theta(x)_{c^-}))$$

Model 2: BR-CNN

- State of the art since 2016
- Bi-Recurrent Convolutional Neural Network
- Use shortest dependency path

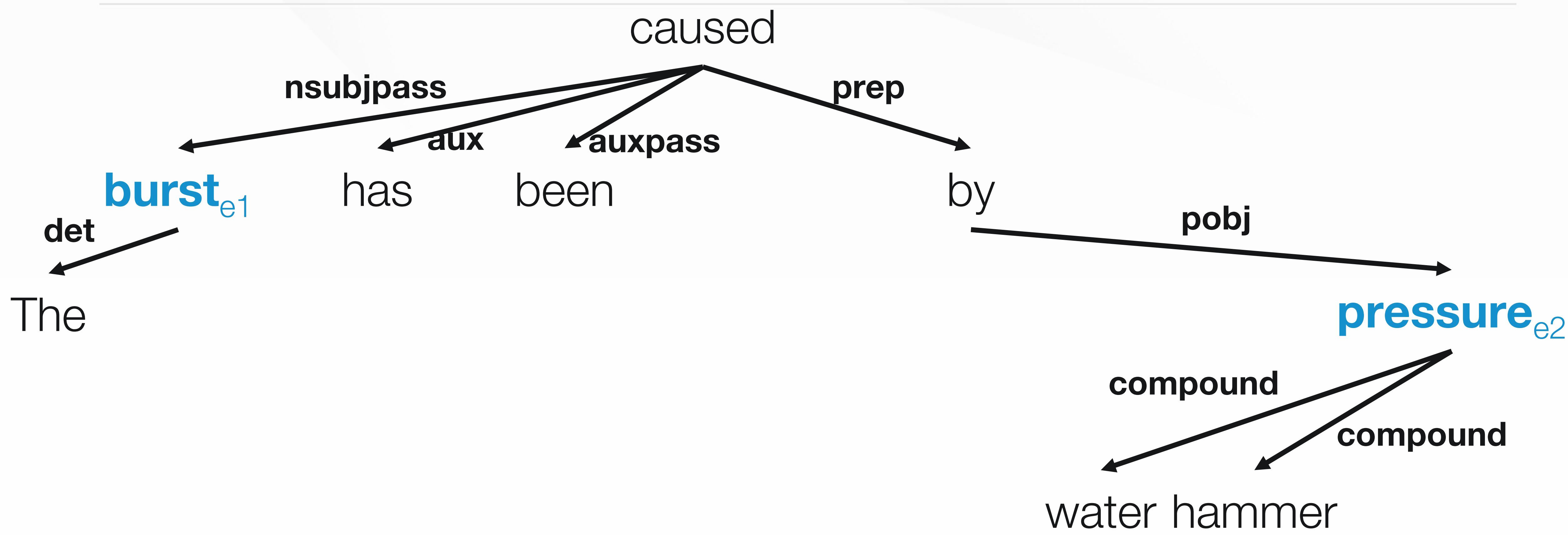
BRCNN_1
- Word embeddings, dependency tag embeddings

BRCNN_2
+ POS tags, NER tags and WordNet hypernyms

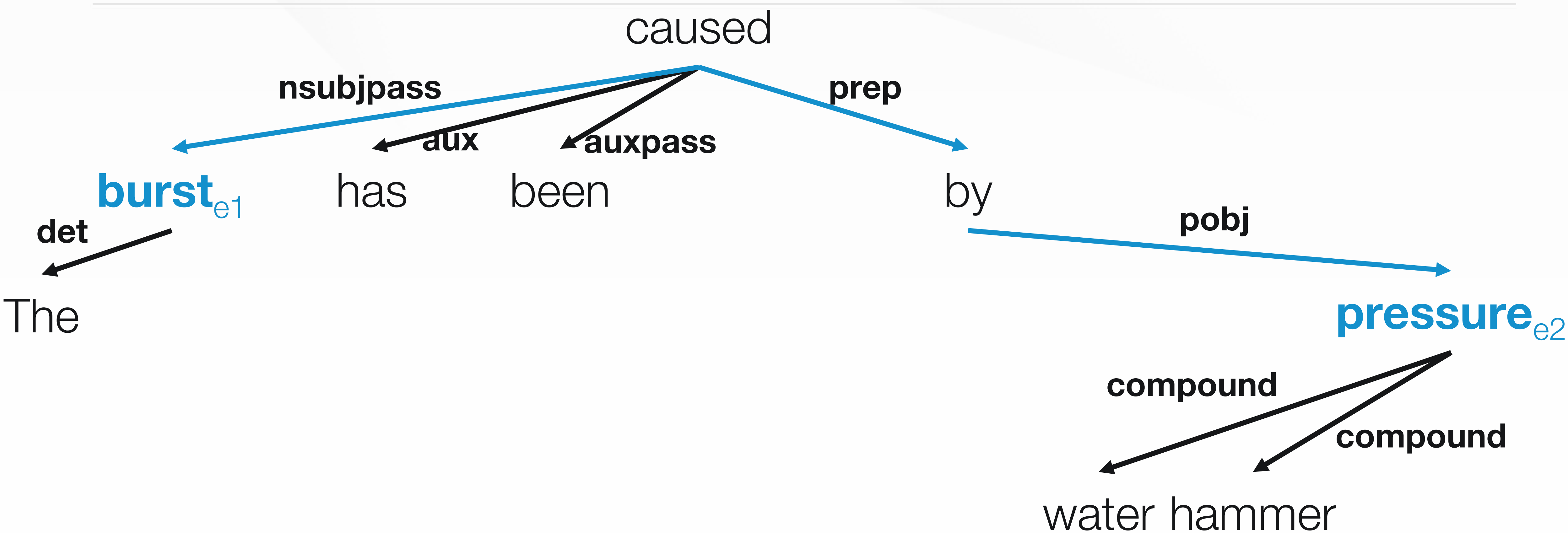
- Training objective: cross entropy

$$J(x_i) = -\sum_{i=1}^{2K+1} \overrightarrow{t}_i \log \overrightarrow{y}_i - \sum_{i=1}^{2K+1} \overleftarrow{t}_i \log \overleftarrow{y}_i - \sum_{i=1}^{K+1} t_i \log y_i$$

Shortest Dependency Path

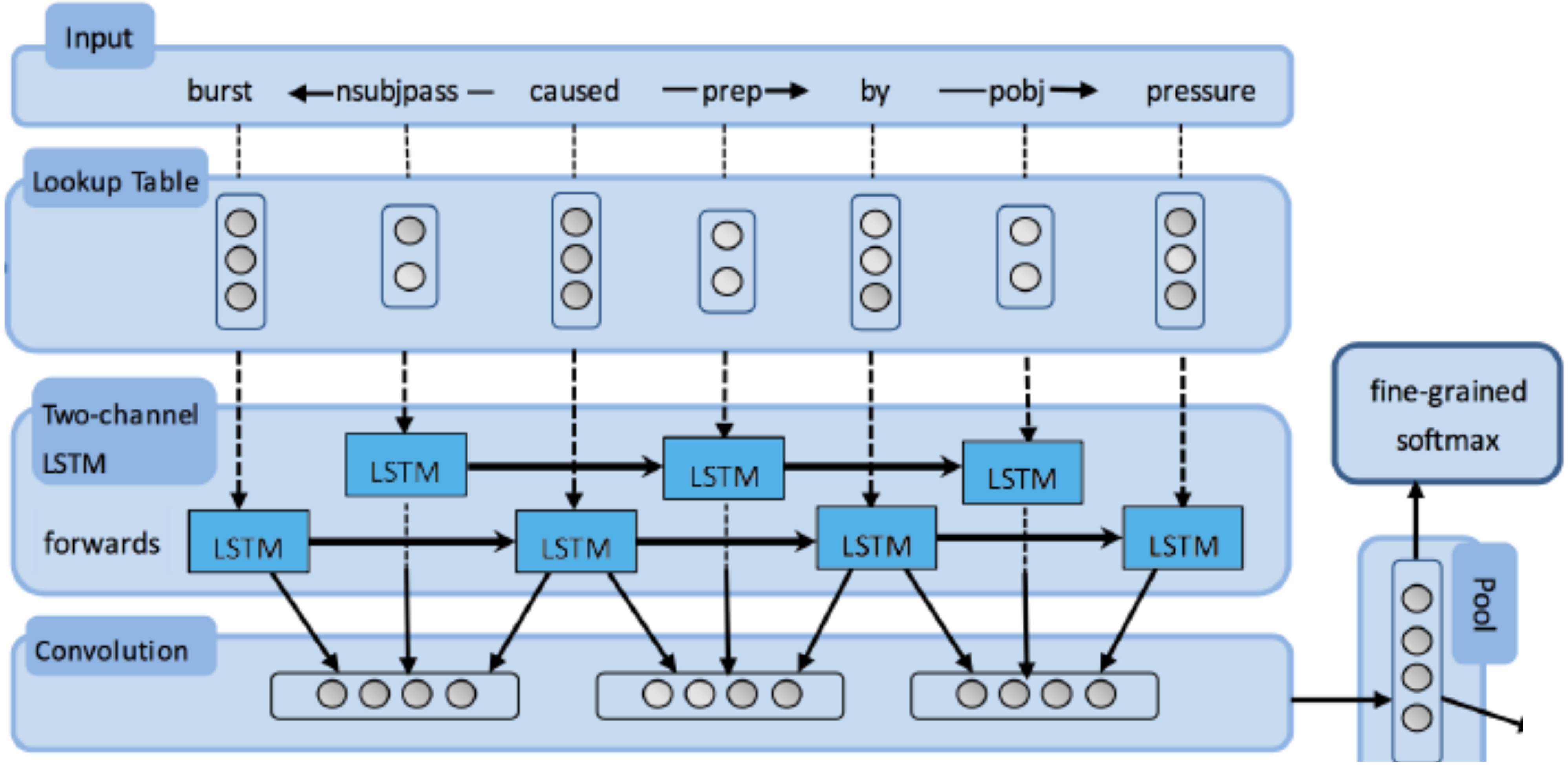


Shortest Dependency Path



Shortest Dependency Path
burst $\xleftarrow{\text{nsubjpass}}$ caused $\xrightarrow{\text{prep}}$ by $\xrightarrow{\text{pobj}}$ pressure
pressure $\xleftarrow{\text{pobj}}$ by $\xleftarrow{\text{prep}}$ caused $\xrightarrow{\text{nsubjpass}}$ burst

Model 2: BR-CNN



Experiments

- Run on the 3 datasets 5 times (mean + stdev)
- Compute macro F_1-Score excluding class Other
- Tune on validation set and evaluate on test set
- Comparison with 3 baselines
  - **UTD**: Support Vector Machine with lexical features
  - **SPTree**: bi-Recurrent Neural Networks
  - **DRNN**: deep Recurrent Neural Networks

Datasets

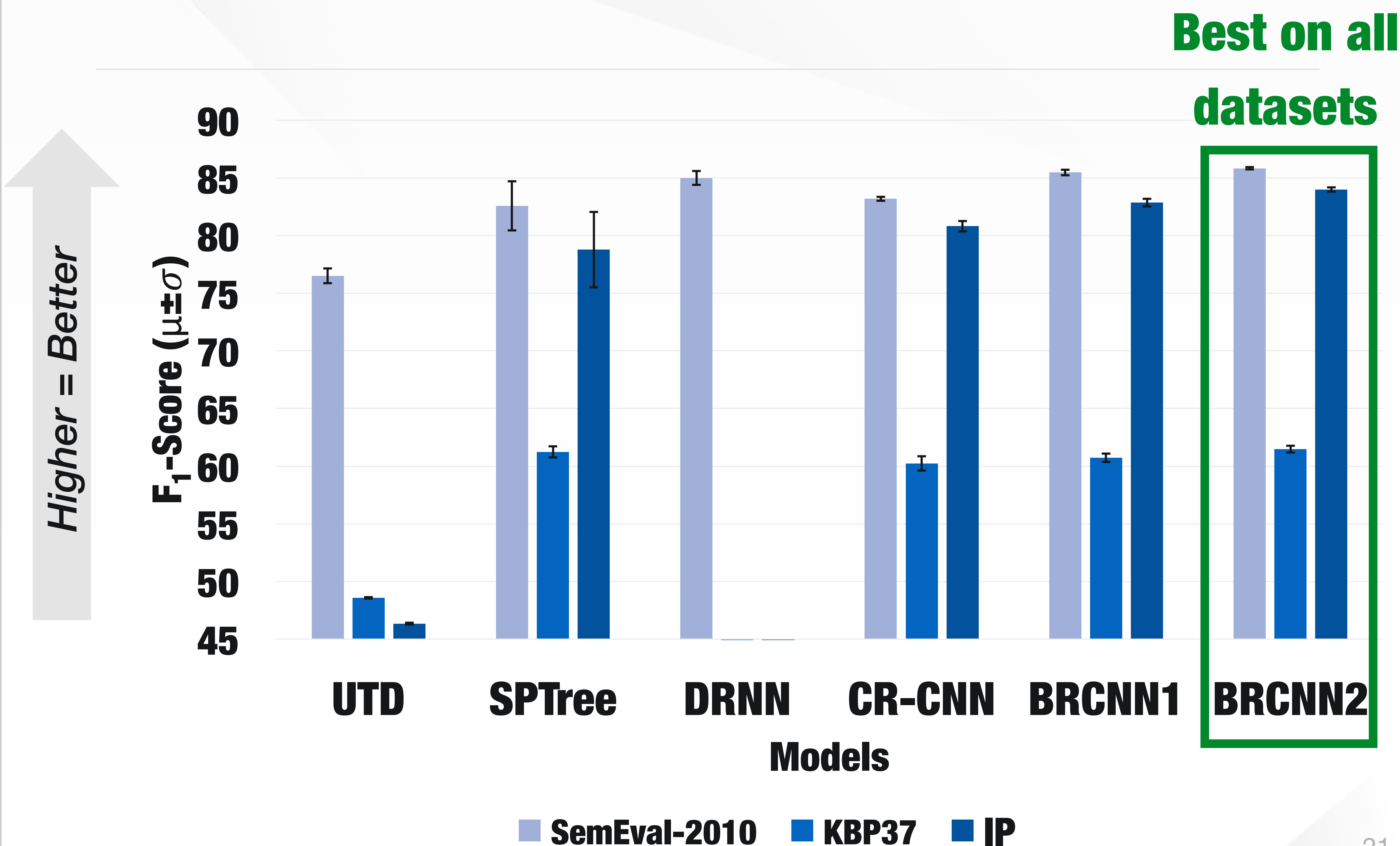
1. SemEval-2010 Task 8
   - Established benchmark for Relation Classification
   - Most of the sentences are either short or average
   - 2x9 relations + 1 Other ⇒ **19 relations**
2. KBP37
   - Named entities
   - Longer sentences
   - 2x18 relations + 1 Other ⇒ **37 relations**

Datasets

3. **IP**

- Training set partially based on SemEval-2007/2010
- Manually gathered sentences from various websites & manual searches on the Internet
- Most of the sentences are either short or average
- Relations of interest for Iprova
  2x6 relations + 1 Other ⇒ **13 relations**

Results


Best on all datasets
Higher = Better
F1-Score (μ±σ)
90
85
80
75
70
65
60
55
50
45
UTD
SPTree
DRNN
CR-CNN
BRCNN1
BRCNN2
Models
SemEval-2010
KBP37
IP

Improvements

Gum disease rates were highest in $[\text{males}]_{e_1}$, Mexican Americans, adults with less than a high school education, adults below the $[\text{poverty line}]_{e_2}$ and current smokers.

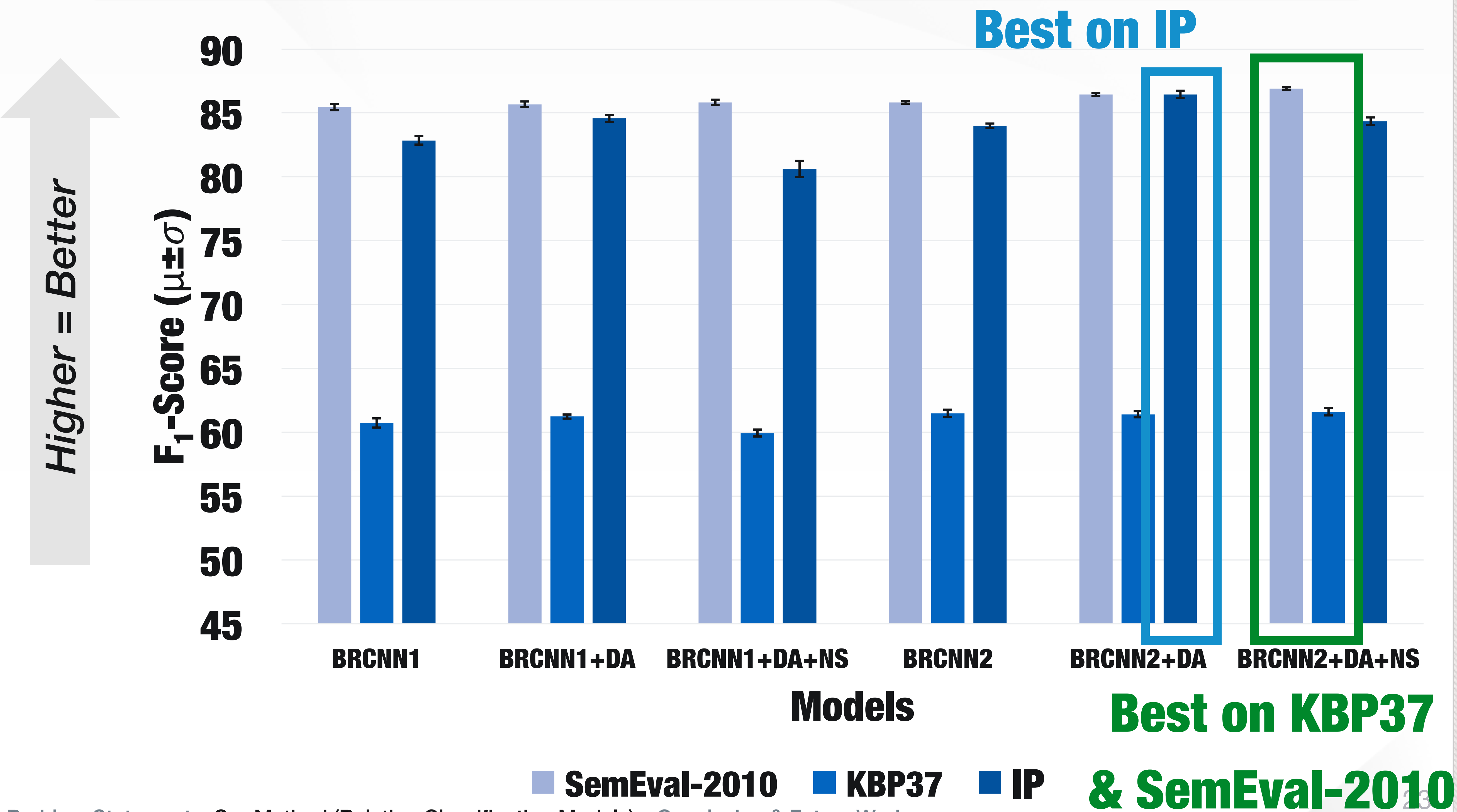
- Data augmentation
  - Replace some words with neighbors in Word2Vec space

Gum infection rates were highest in $[\text{males}]_{e_1}$, Peruvian Americans, adults with less than a junior-senior school education, adults below the $[\text{poverty line}]_{e_2}$ and former nonsmokers.

- Negative Sampling
  - Assign tags $[]_{e1}$, $[]_{e2}$ to other words

Gum disease rates were highest in males, Mexican Americans, adults with less than a high school education, $[\text{adults}]_{e_1}$ below the poverty line and current $[\text{smokers}]_{e_2}$.

Results



Best on IP
Higher = Better
F1-Score (μ±σ)
90
85
80
75
70
65
60
55
50
45
BRCNN1
BRCNN1+DA
BRCNN1+DA+NS
BRCNN2
BRCNN2+DA
BRCNN2+DA+NS
Models
Best on KBP37
SemEval-2010
KBP37
IP
& SemEval-2010

Recap

- Incorporating linguistic information in network's architecture is still important and beneficial
- Data Augmentation & Negative Sampling techniques help to strengthen classifiers
- $BRCNN_2$+DA+NS outperforms all models on Sem & KBP
- $BRCNN_2$+DA outperforms all models on IP dataset

  ⇒ **Will be used to build Knowledge Graphs**

Our Method

Relation Classification Models

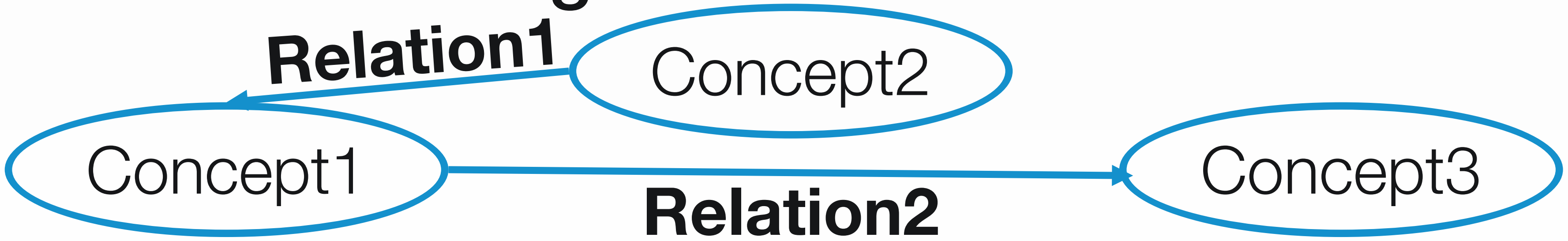
Building Knowledge Graphs

What, Why, How ?

- **What:** structured representation of semantic knowledge and relations among nodes

What, Why, How ?

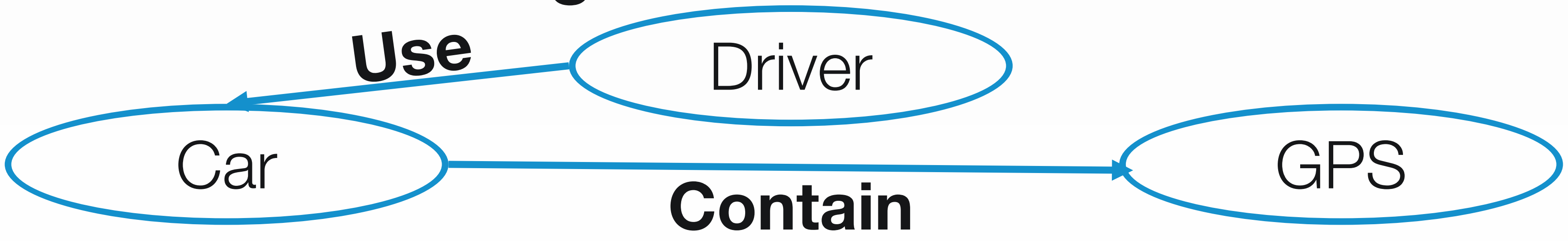
- **What:** structured representation of semantic knowledge and relations among nodes



- **Why:** model domains of interest, infer new relations, basis for a Question-Answering system, etc.

What, Why, How ?

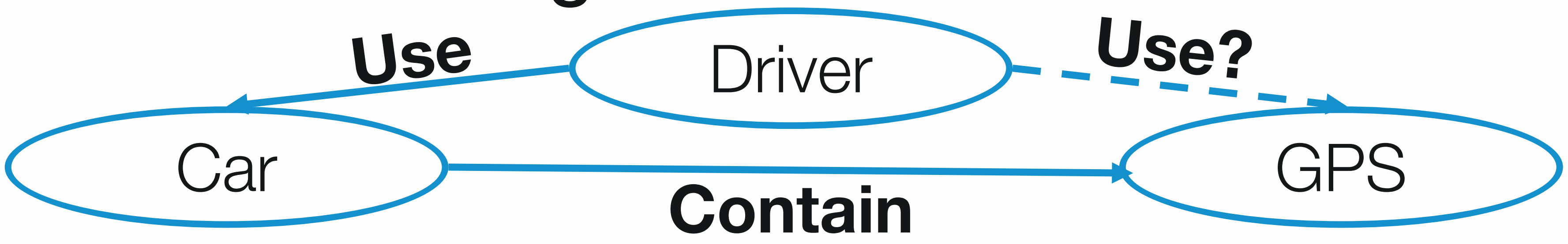
- **What:** structured representation of semantic knowledge and relations among nodes



- **Why: model domains of interest**, infer new relations, basis for a Question-Answering system, etc.

What, Why, How ?

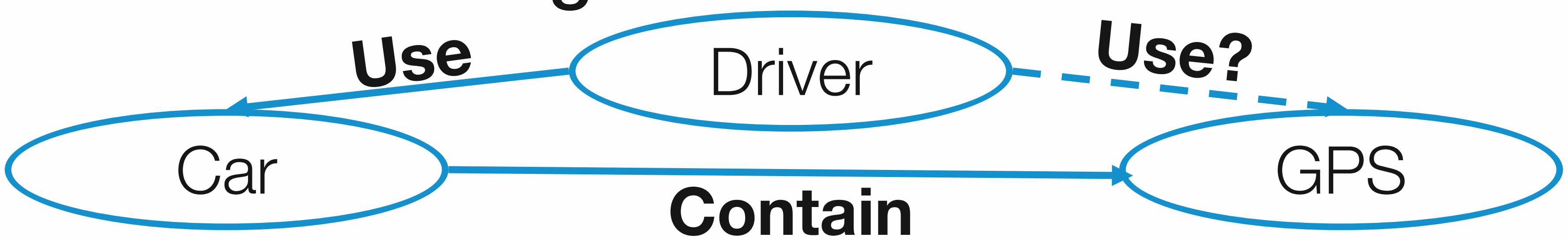
- **What:** structured representation of semantic knowledge and relations among nodes



- **Why:** model domains of interest, **infer new relations**, basis for a Question-Answering system, etc.

What, Why, How ?

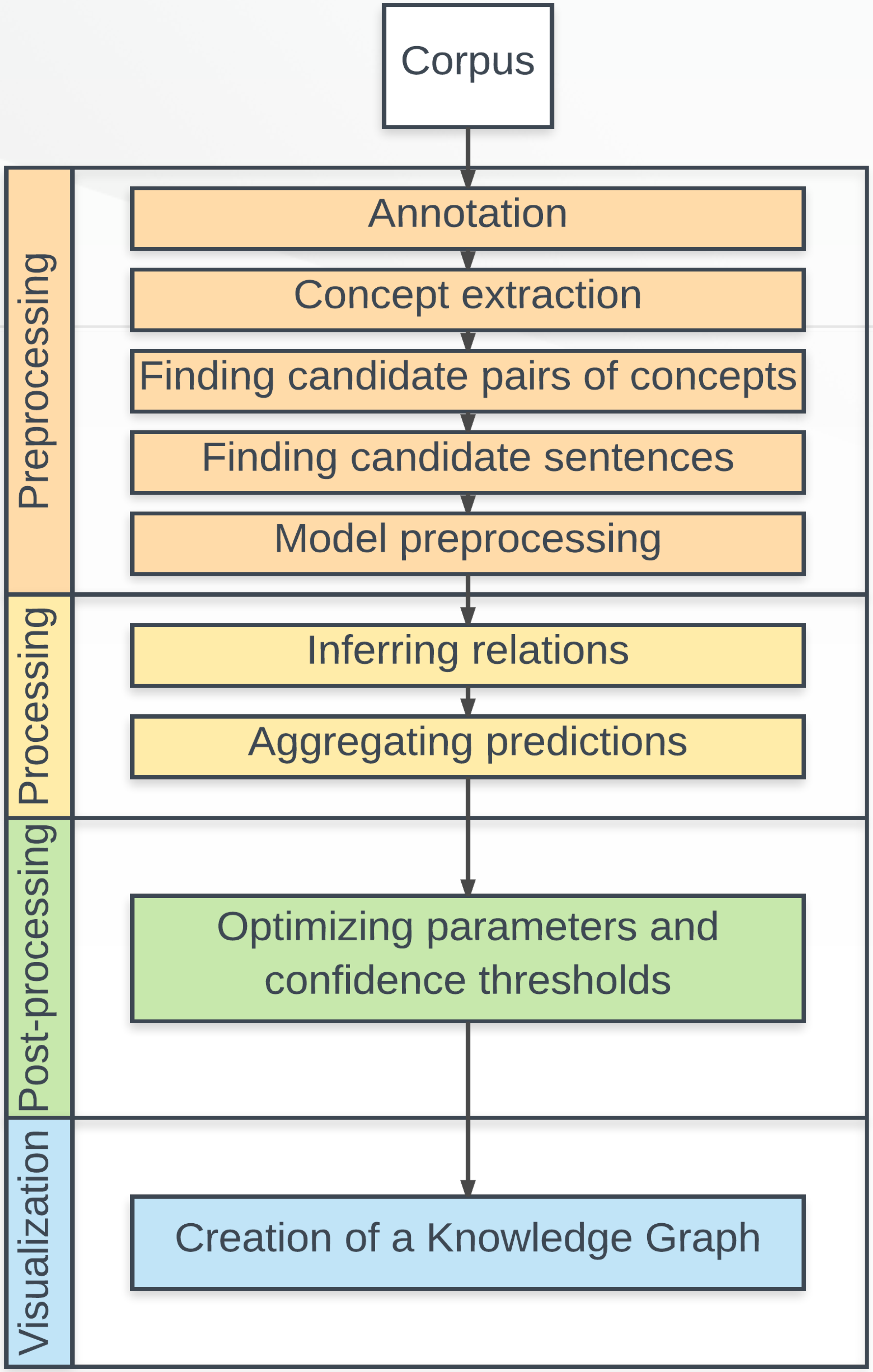
- **What:** structured representation of semantic knowledge and relations among nodes



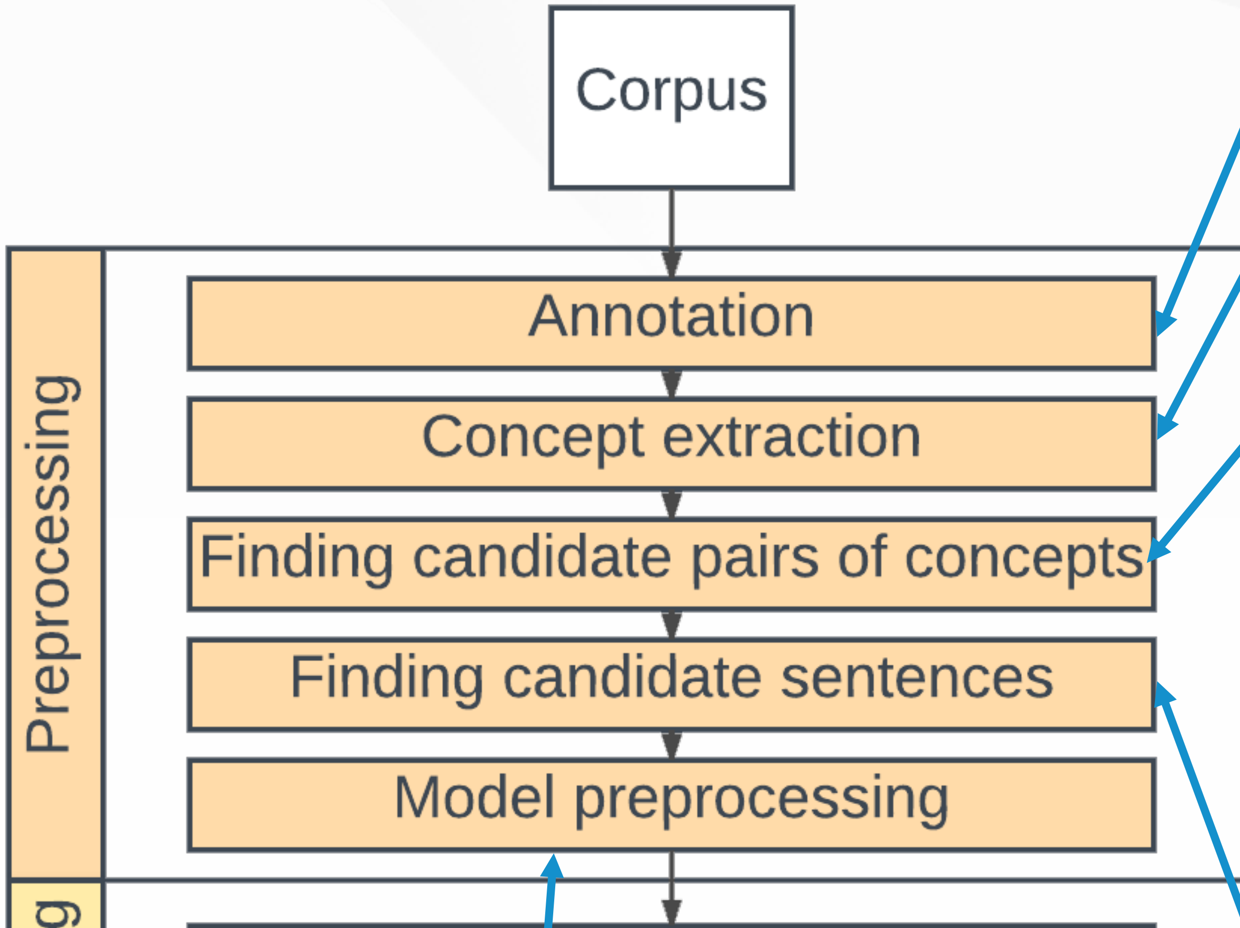
- **Why:** model domains of interest, infer new relations, basis for a Question-Answering system, etc.

- **How:**
  - Extracting pairs of concepts from large corpora
  - Infer relations with best model on IP dataset: $BRCNN_2$+DA

Pipeline


Corpus
Preprocessing
Annotation
Concept extraction
Finding candidate pairs of concepts
Finding candidate sentences
Model preprocessing
Processing
Inferring relations
Aggregating predictions
Post-processing
Optimizing parameters and confidence thresholds
Visualization
Creation of a Knowledge Graph

Preprocessing



Stanford's tools

Iprova's concept extracter

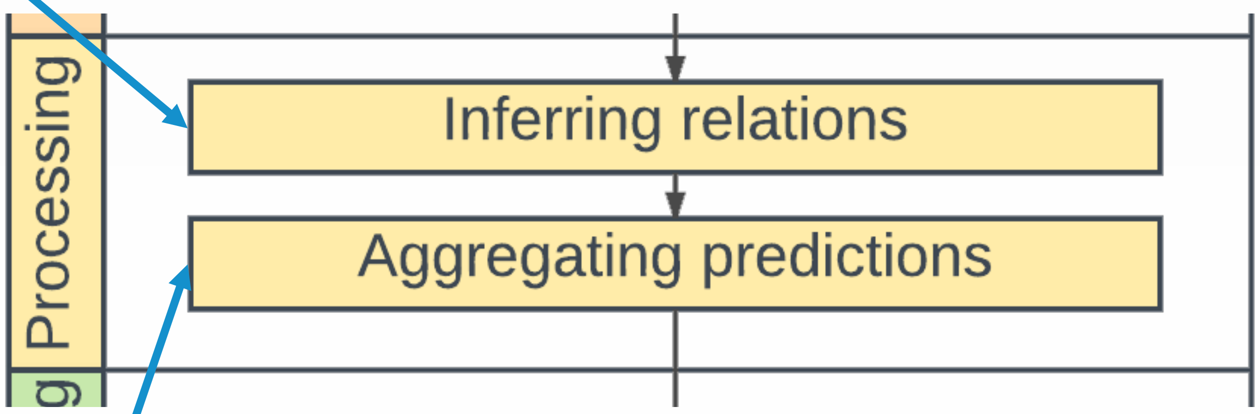
- At least 2 sentences
- Concepts not too far away

- Containing both concepts
  A concept might be part of bigger concept e.g.
  diabetes $\in$ type 2 diabetes
  schedule $\in$ rotating schedule

Preprocessing of **BRCNN$_2$+DA** model

Processing

Using model BRCNN2+DA



What if R1(Concept1, Concept2) & R2(Concept1, Concept2) ?

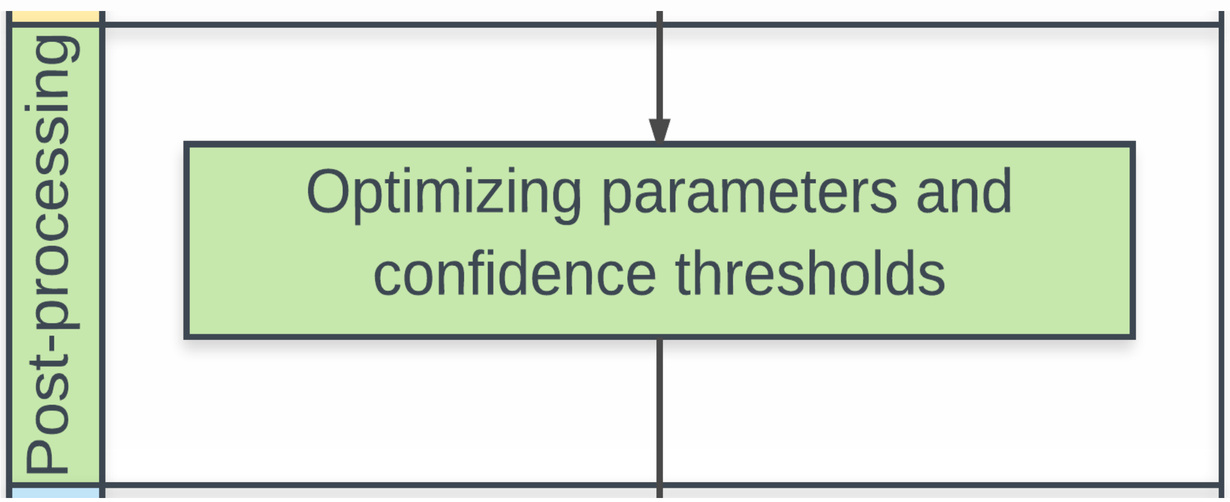
⇒ **Aggregate probability distribution vectors by class-label**

Pair	Conf. R1	Conf. R2	Conf. R3
(c1,c2)	0.5	0.3	0.2
(c1,c2)	0.8	0.2	0.0
(c1,c2)	0.2	0.1	0.7

Median

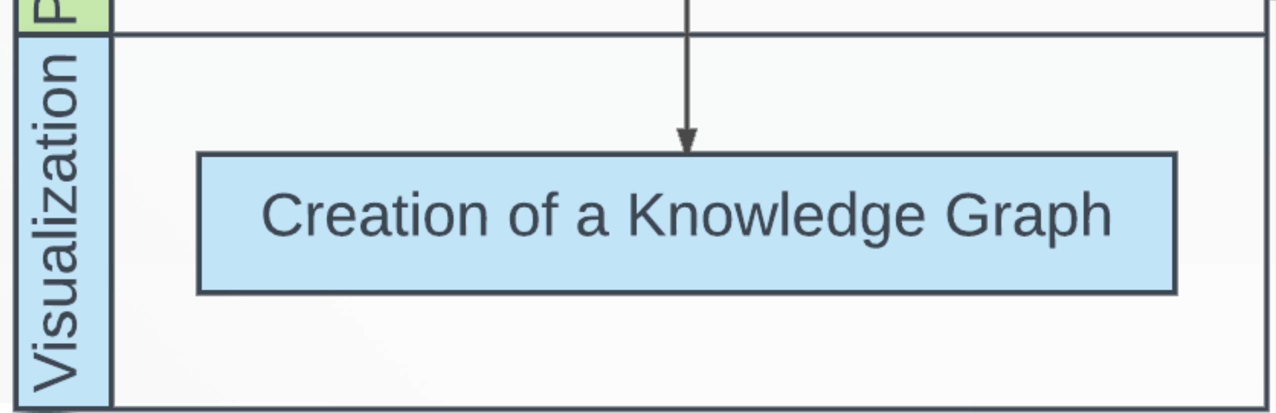
Pair	Conf. R1	Conf. R2	Conf. R3
(c1,c2)	0.5	0.2	0.2

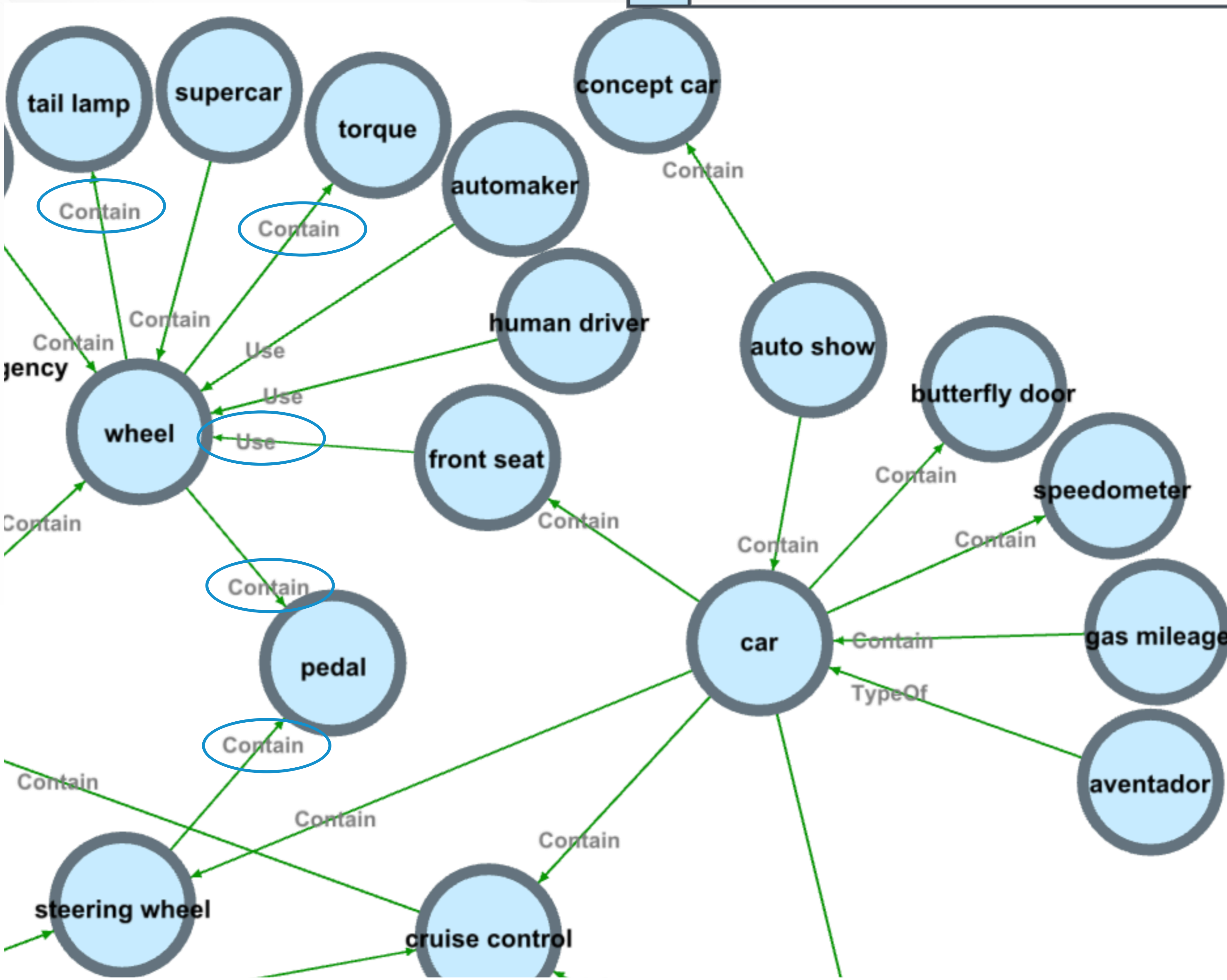
Post-Processing



- **Goal:** filter out noise
- **Parameters:** free to setup during visualization by Iprova
- **Confidence Thresholds**
  Confidence threshold for each class
  - >= threshold ⇒ keep relation
  - < threshold ⇒ Other

Visualization





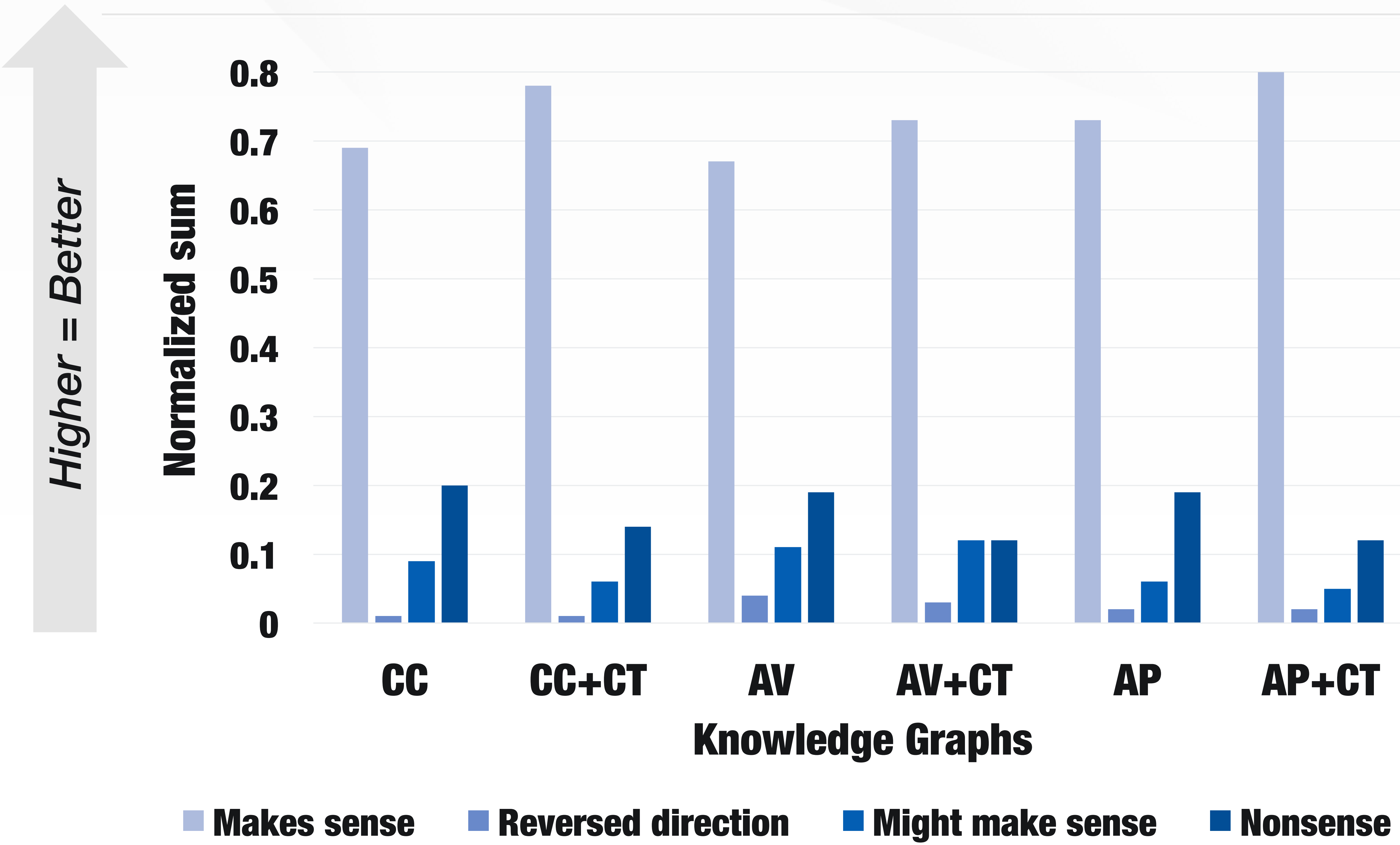
Qualtitative Evaluation

- Build representative Knowledge Graphs from 3 corpora with and without confidence thresholds (CT)
- **Manually assess** quality of the predictions on 2% of each KG
- Classify each sample in one of the four classes:
  1. **Makes sense**, e.g. Contain(car, wheels)
  2. **Reversed direction**, e.g. Contain(wheels, car)
  3. **Might make sense**, e.g. Use(racing, drivers)
  4. **Nonsense**, e.g. TypeOf(neck, tail)

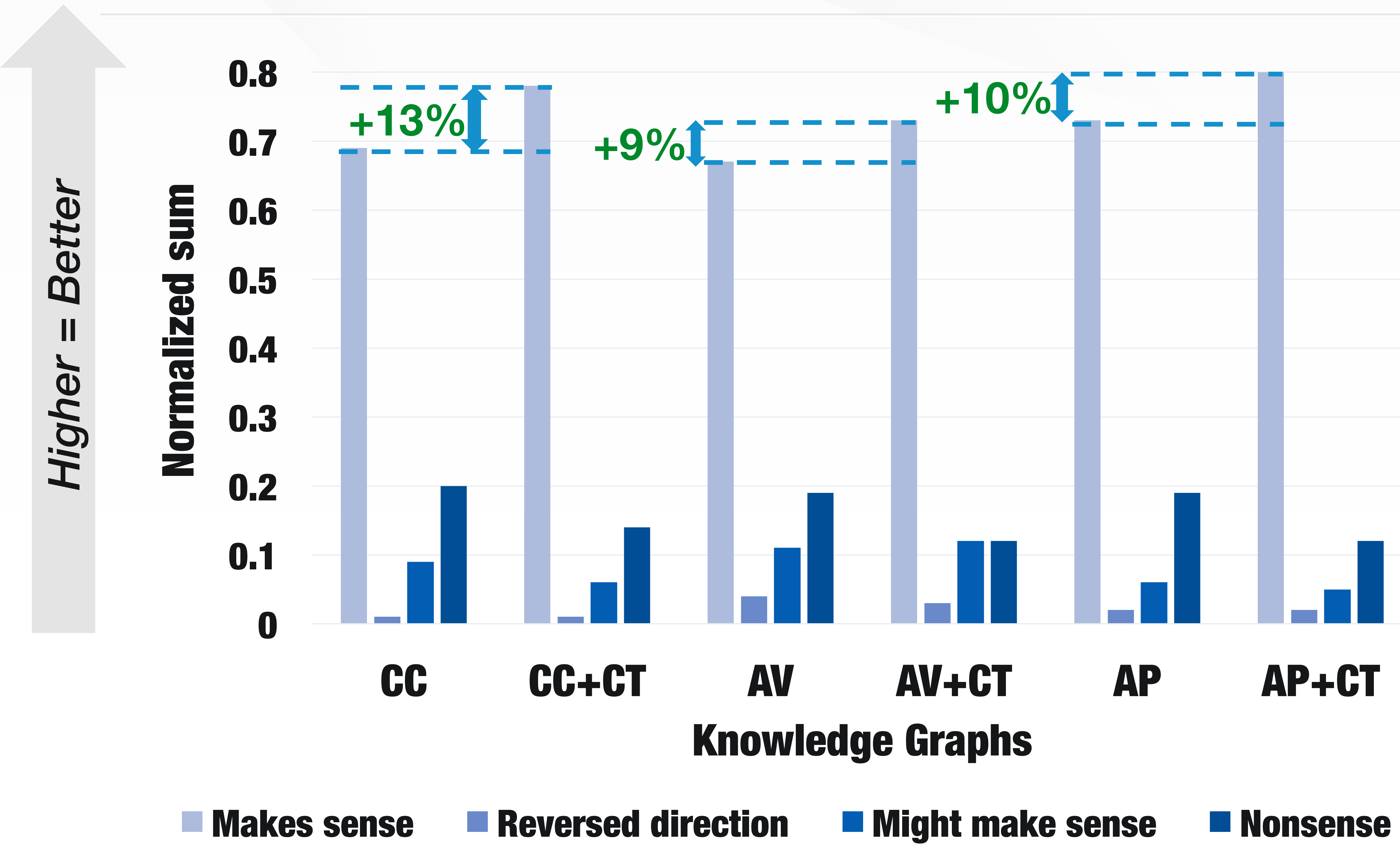
Corpora

1. **Common Crawl**
   - based on ScienceDaily and Phys.org
   - ~20 millions sentences
2. **Autonomous Vehicles** documents
   - ~ $^1/_2$ million sentences
3. **Air Purifier** documents
   - ~1 million sentences

Results

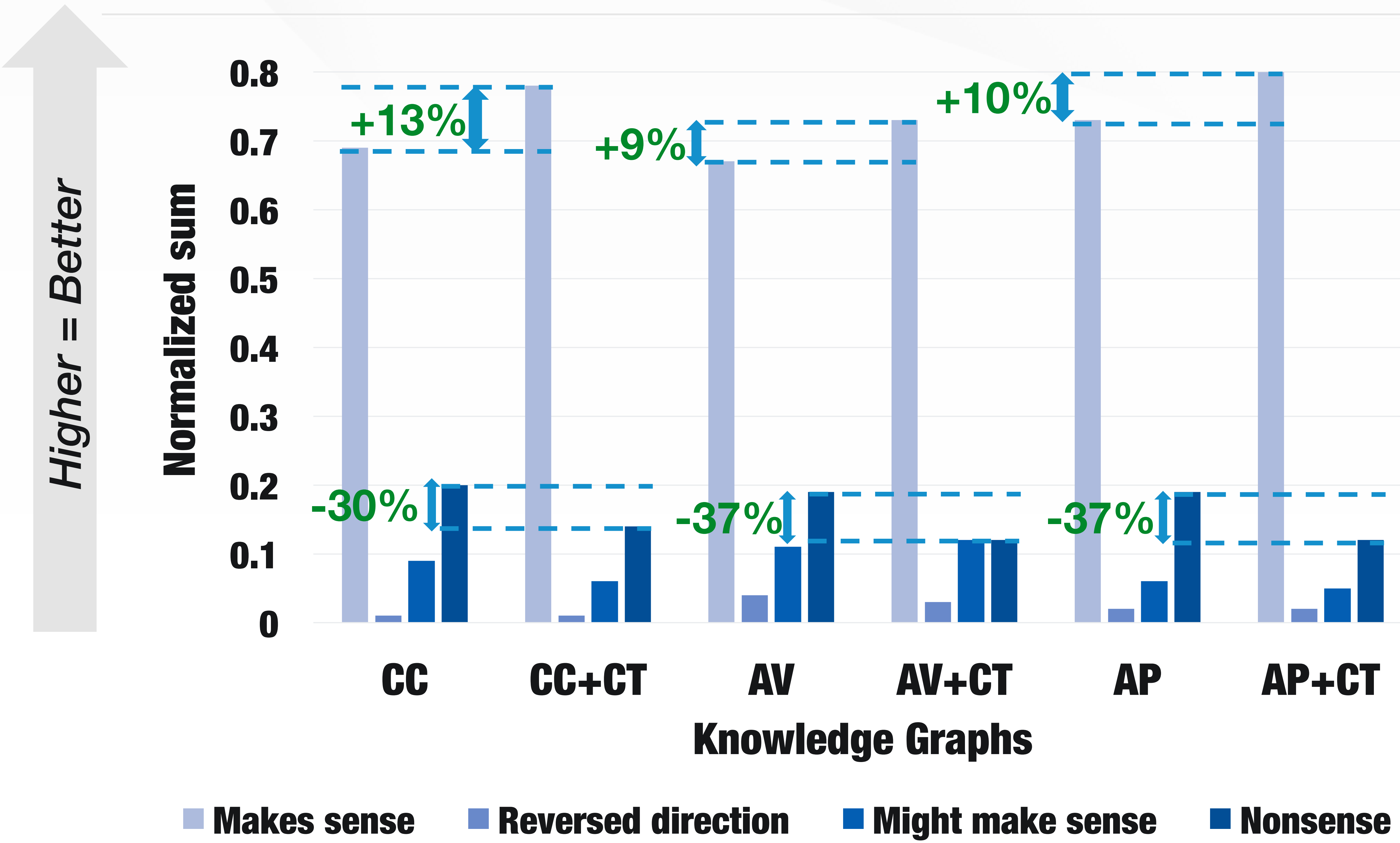

Higher = Better
Normalized sum
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CC
CC+CT
AV
AV+CT
AP
AP+CT
Knowledge Graphs
Makes sense
Reversed direction
Might make sense
Nonsense

Results

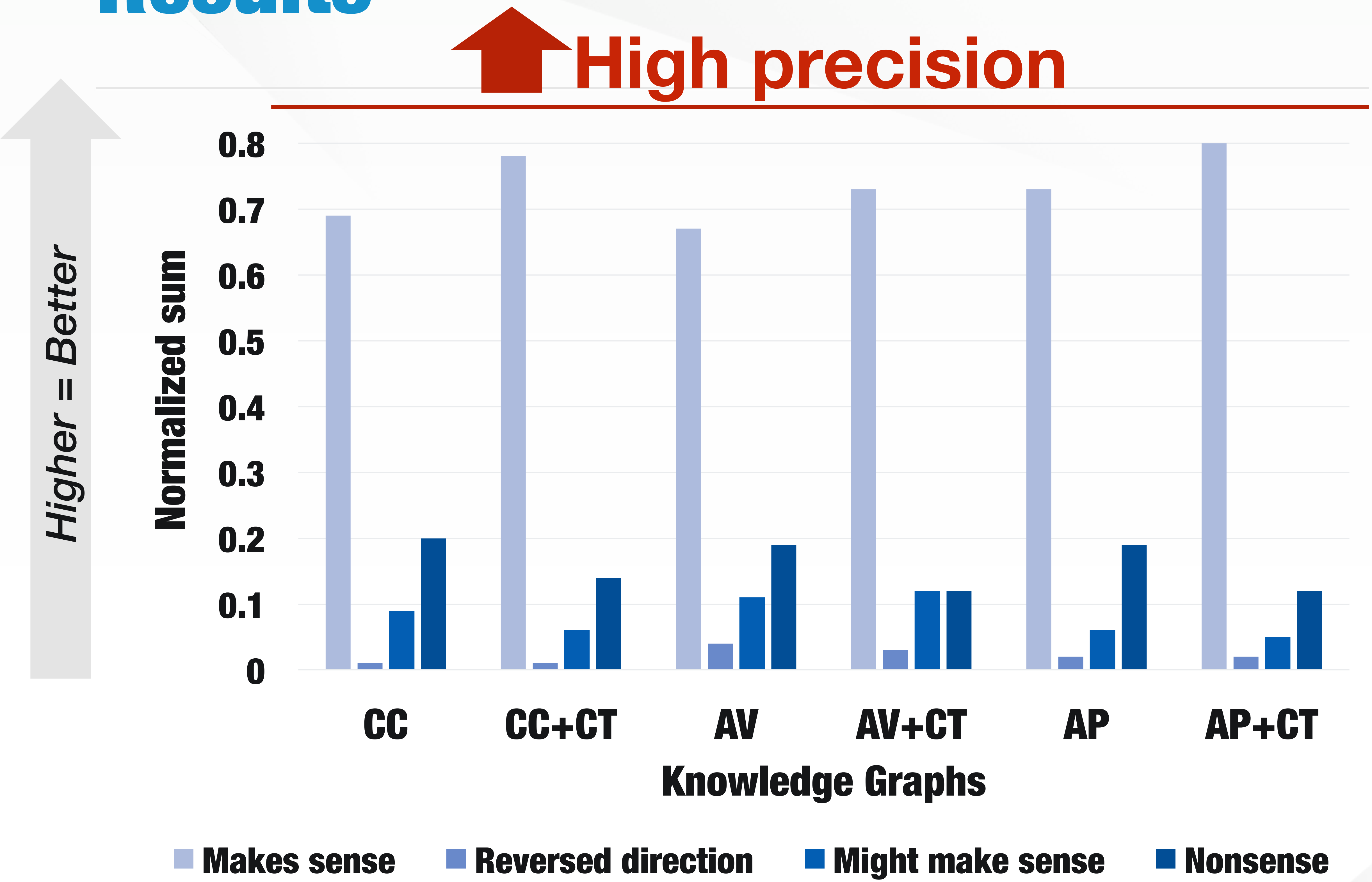


Higher = Better
Normalized sum
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
+13%
+9%
+10%
CC
CC+CT
AV
AV+CT
AP
AP+CT
Knowledge Graphs
Makes sense
Reversed direction
Might make sense
Nonsense

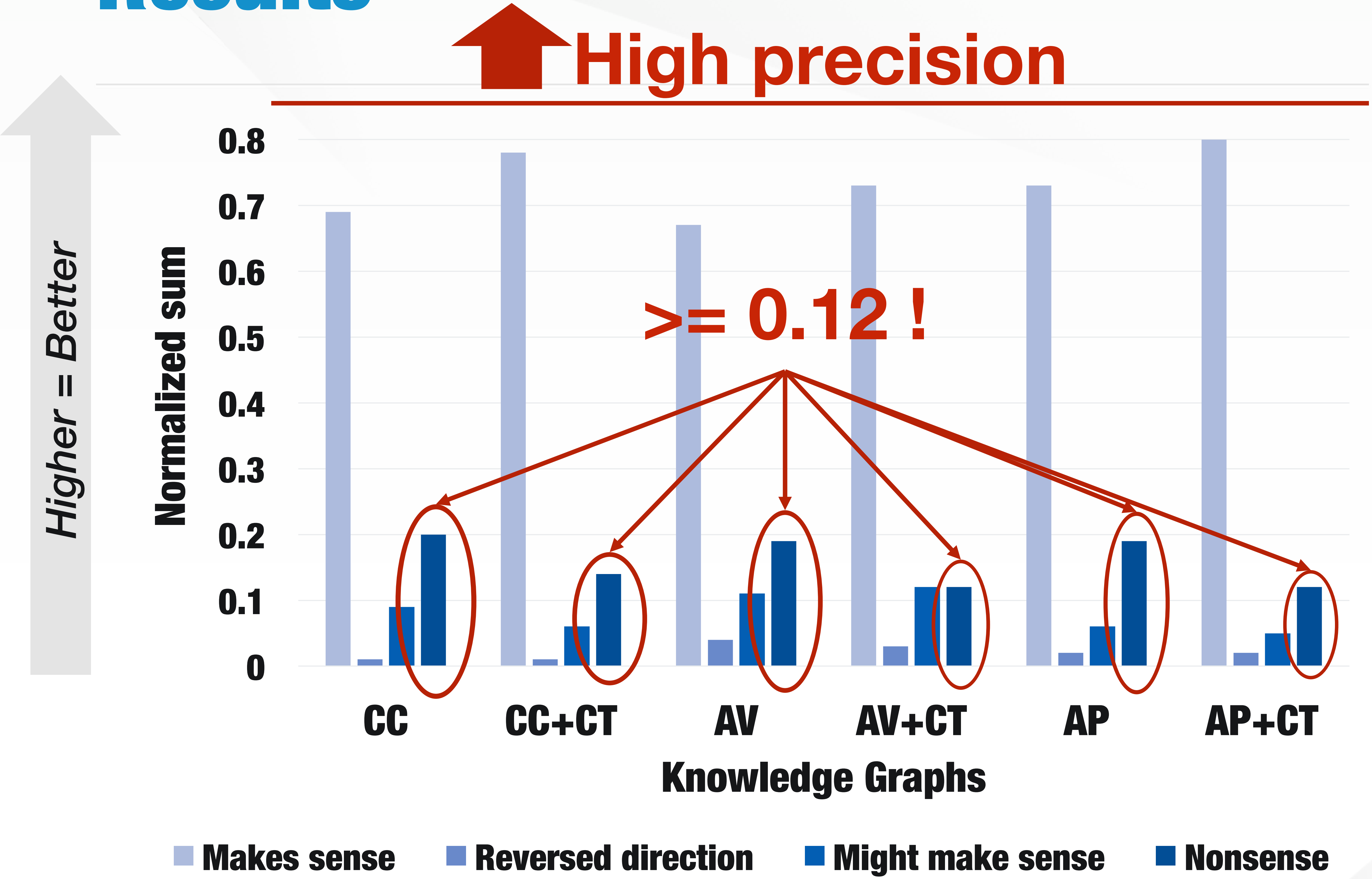
Results


Higher = Better
Normalized sum
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
+13%
+9%
+10%
-30%
-37%
-37%
CC
CC+CT
AV
AV+CT
AP
AP+CT
Knowledge Graphs
Makes sense
Reversed direction
Might make sense
Nonsense

Results


High precision
Higher = Better
Normalized sum
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CC
CC+CT
AV
AV+CT
AP
AP+CT
Knowledge Graphs
Makes sense
Reversed direction
Might make sense
Nonsense

Results


High precision
Higher = Better
Normalized sum
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
>= 0.12 !
CC
CC+CT
AV
AV+CT
AP
AP+CT
Knowledge Graphs
Makes sense
Reversed direction
Might make sense
Nonsense

Limitations

- **Potential overlaps among relations**
  e.g. Use(laptop, processor) & Contain(laptop, processor)
- **Delimitation of the relations**
  e.g. Contain(mouse, genes)
- **Mixture of semantic meanings**
  e.g. Contain(mouse, brain) & TypeOf(mouse, device)
- **Hypothetic relations**
  e.g. Contain(artery, clot)

Conclusion & Future Work

Conclusion

- State of the art model for Relation Classification task
  - Linguistic information features and +/- sampling help !
- Create a dataset fitting Iprova's needs and build KGs
  - Precision not high enough yet
  - Has some limitations
  - Can be used as an help for humans
- This kind of Knowledge Graphs doesn't exist

  ⇒ **We provide a tool to model domains of interest**

Future Work

- Inferring new relations by using prior knowledge from KG
- Training pair-words embeddings
- Use pairwise ranking loss function
- Better filtering for Knowledge Graphs
- Improve concept extraction system

Questions ?