Zero-shot learning for multilingual discourse relation classification

Eleni Metheniti, Philippe Muller, Chloé Braud, Margarita Hernández-Casas

UT3 - IRIT - CNRS

firstname.lastname@irit.fr

Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

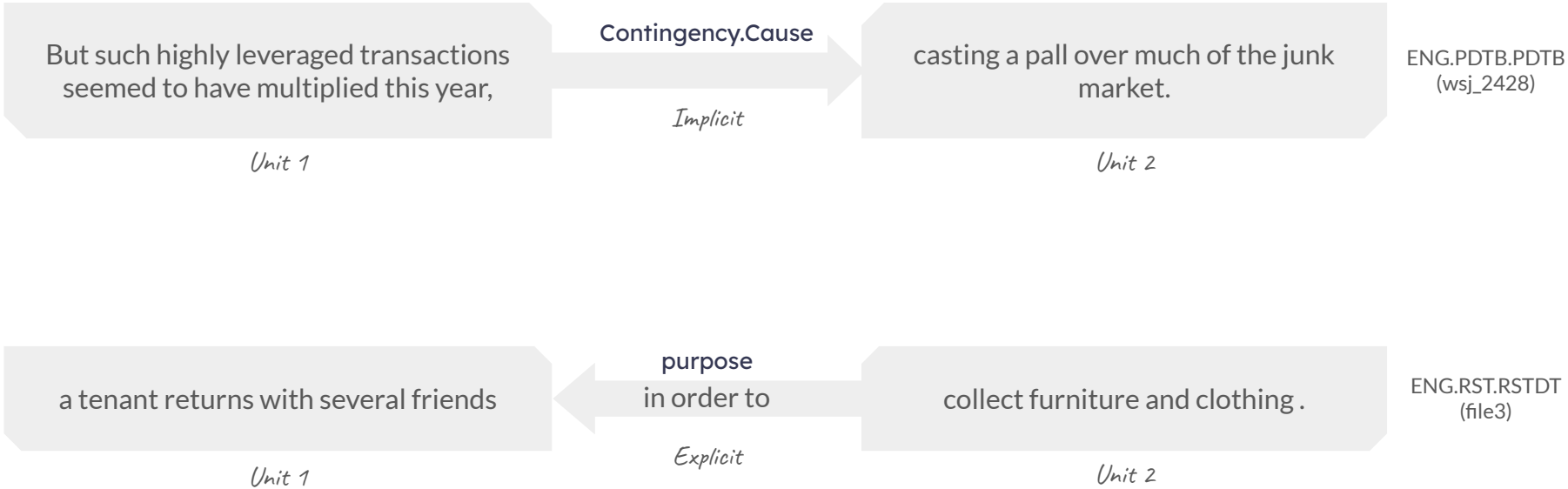
cnrs dépasser les frontières

ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

KEEP CALM AND ANDIAMO AVANTI

source image
Andiamo

Discourse relations



Main questions

Is **zero-shot learning** possible for **discourse relation classification**?

Across languages & frameworks?

Best zero-shot setup?

Datasets

- DISRPT 2023: 26 datasets, 13 languages, 4 frameworks
- Submitted systems:
  - 🥇 HITS: monolingual or framework-based, large models
  - 🥇 DisCoDisCo (2021): monolingual, features, direction annotation
  - 🥈 DiscReT: **multilingual only**, **label harmonization**, switching units for direction
  - 🥉 DiscoFLAN: generative models, **prediction filtering**

Methodology

Methodology

- Classifiers built with **mBERT**, DistilmBERT, XLM-RoBERTa

Methodology

- Classifiers built with **mBERT**, DistilmBERT, XLM-RoBERTa
- Label harmonization (163 → 128 labels)

(PDTB) expansion.conjunction
(PDTB) conjunction
} conjunction

Methodology

- Classifiers built with **mBERT**, DistilmBERT, XLM-RoBERTa
- Label harmonization (163 → 128 labels)

  (PDTB) expansion.conjunction
  (PDTB) conjunction
  } conjunction

- Switching units for unified relation directions
  - `1>2:[CLS] Unit 1 [SEP] Unit 2`
  - `1<2:[CLS] Unit 2 [SEP] Unit 1`

Methodology

- Classifiers built with **mBERT**, DistilmBERT, XLM-RoBERTa
- Label harmonization (163 → 128 labels)

  (PDTB) expansion.conjunction
  (PDTB) conjunction
  } conjunction

- Switching units for unified relation directions
  - `1>2:[CLS] Unit 1 [SEP] Unit 2`
  - `1<2:[CLS] Unit 2 [SEP] Unit 1`
- Label filtering for model predictions

  PDTB: `{expansion.conjunction: 0.2, ~~joint: 0.25~~, expansion.disjunction: 0.02 …}`

Zero-Shot Methodology

- Train zero-shot classifier with mBERT
- Compare accuracy with monolingual mBERT classifier

Language Families

- Germanic
- Romance

Frameworks

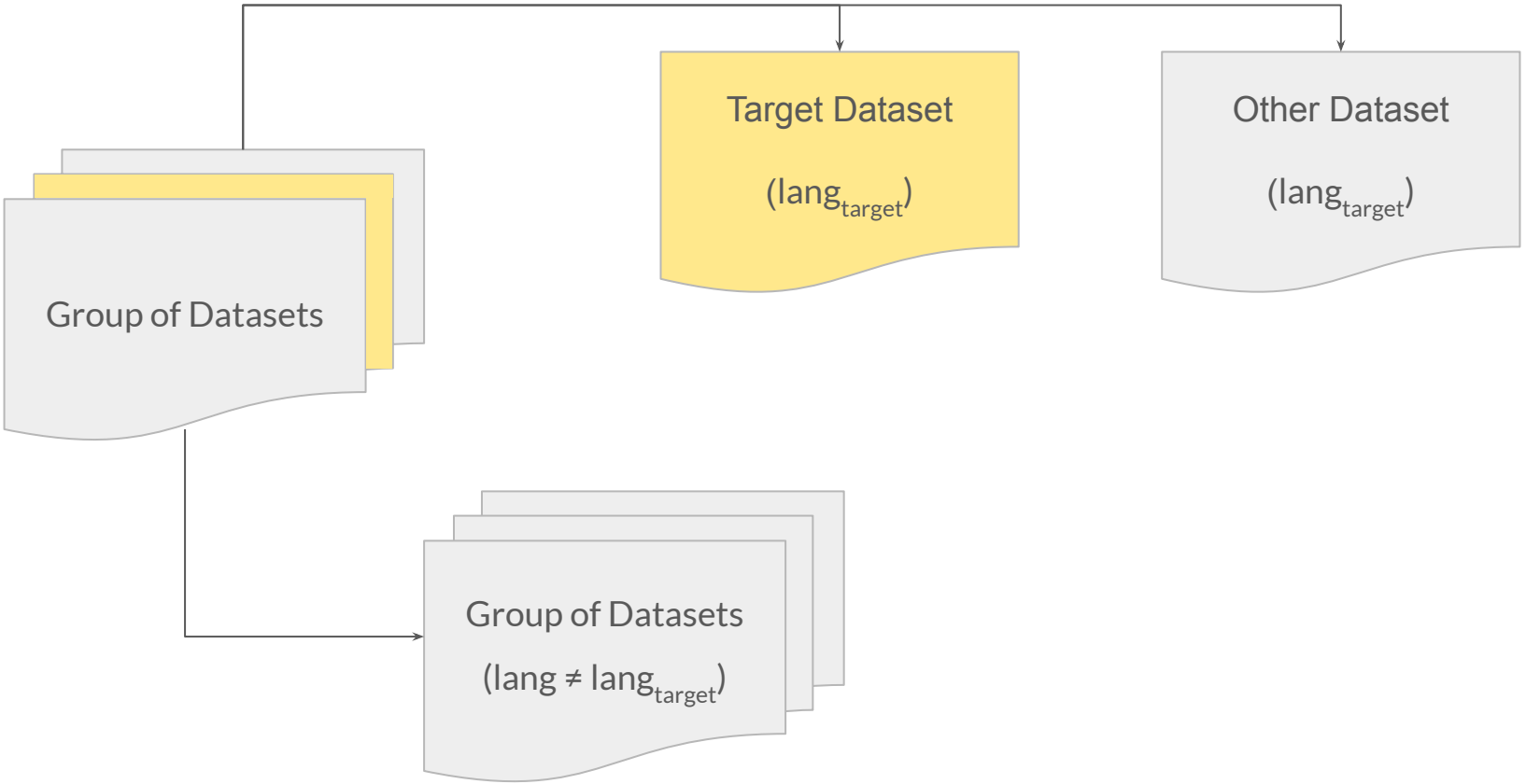
- RST
- PDTB
- SDRT
- DEP

Jaccard Similarity

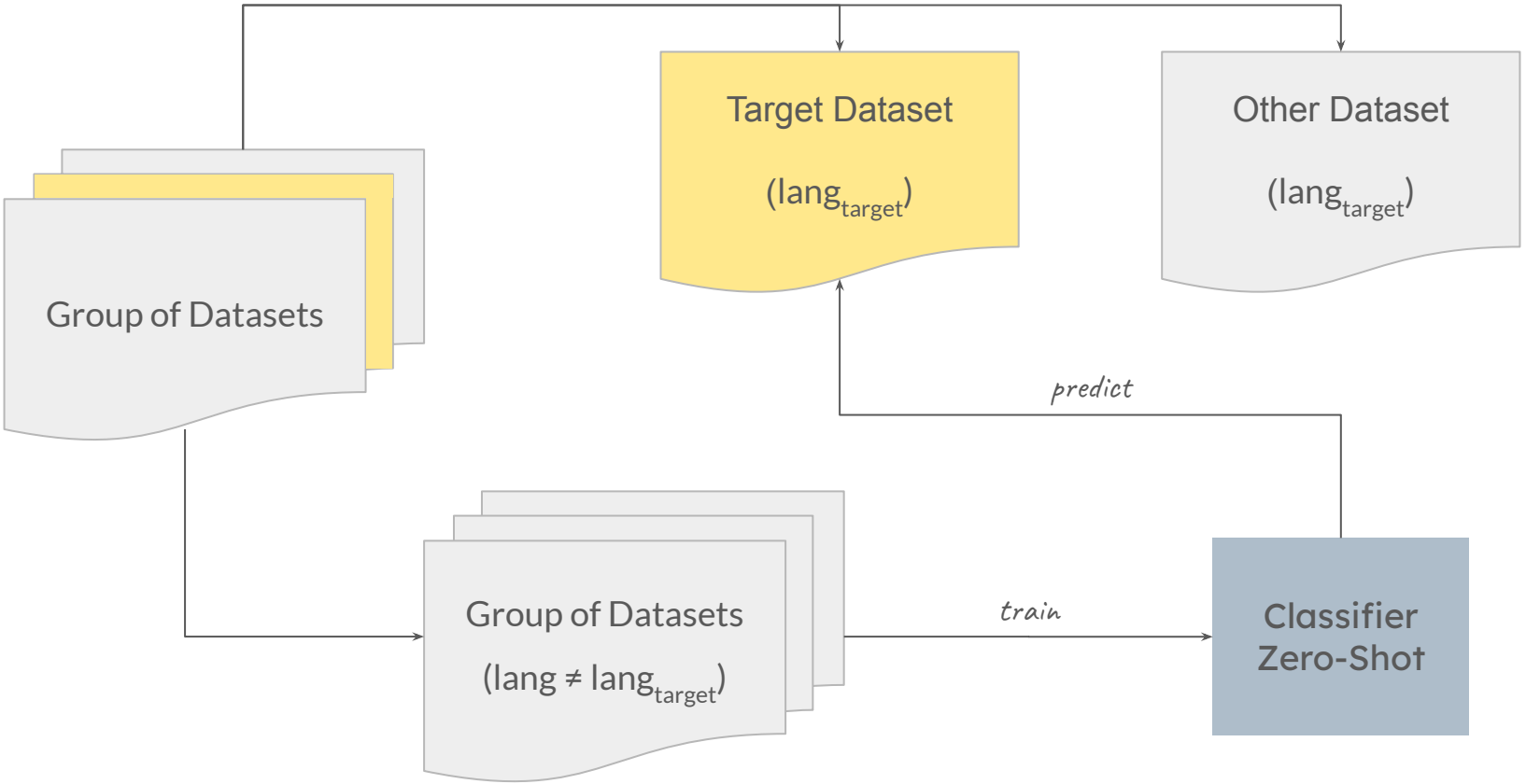
- Computed on label sets of each dataset
- Groups of pair similarity > 0.4

Zero-Shot Methodology

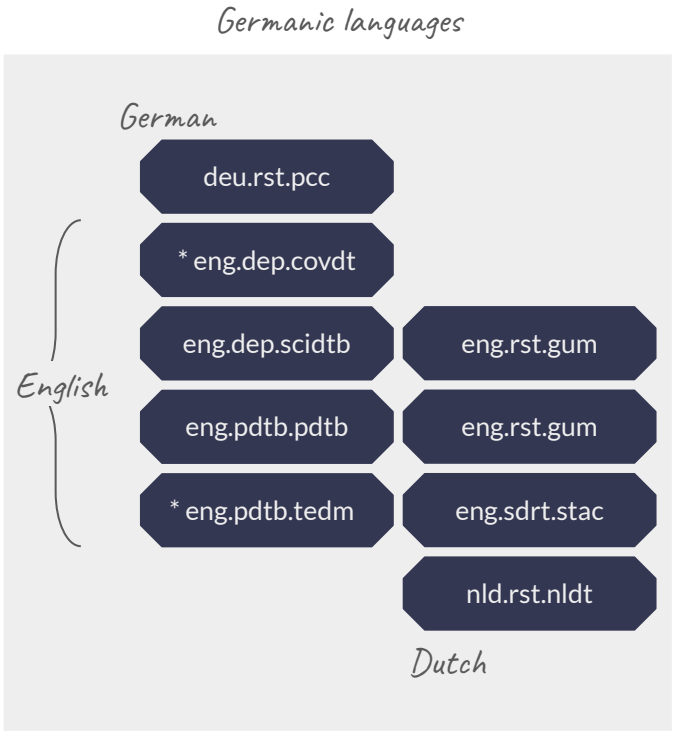
Zero-Shot Methodology



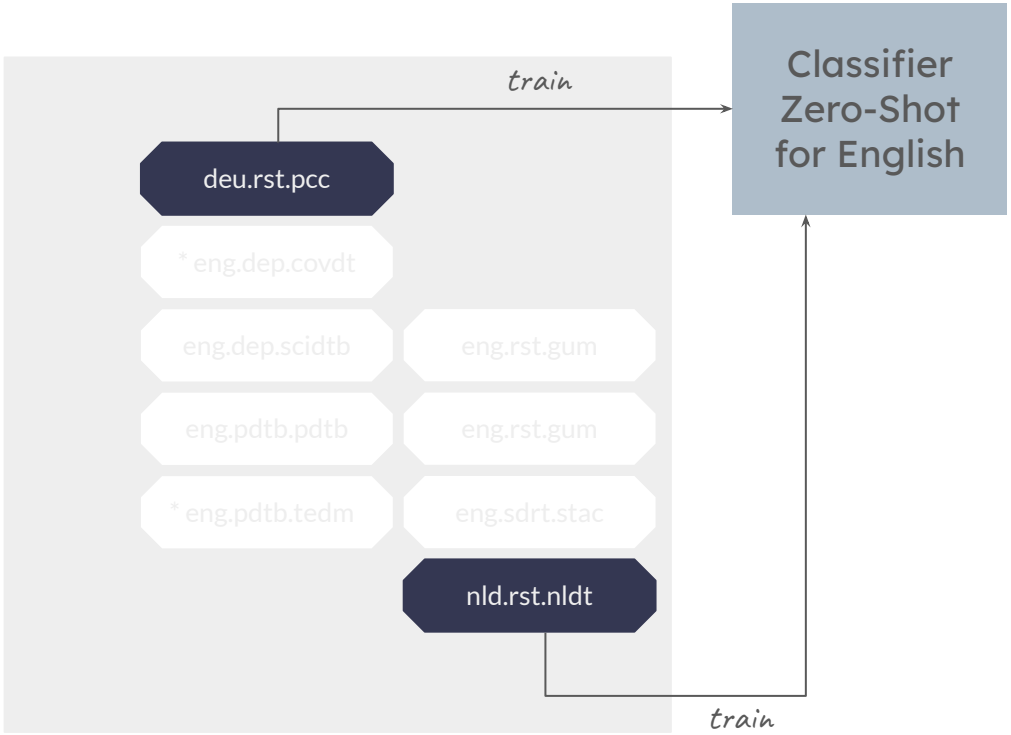
Zero-Shot Methodology



Zero-Shot: Language families (example)

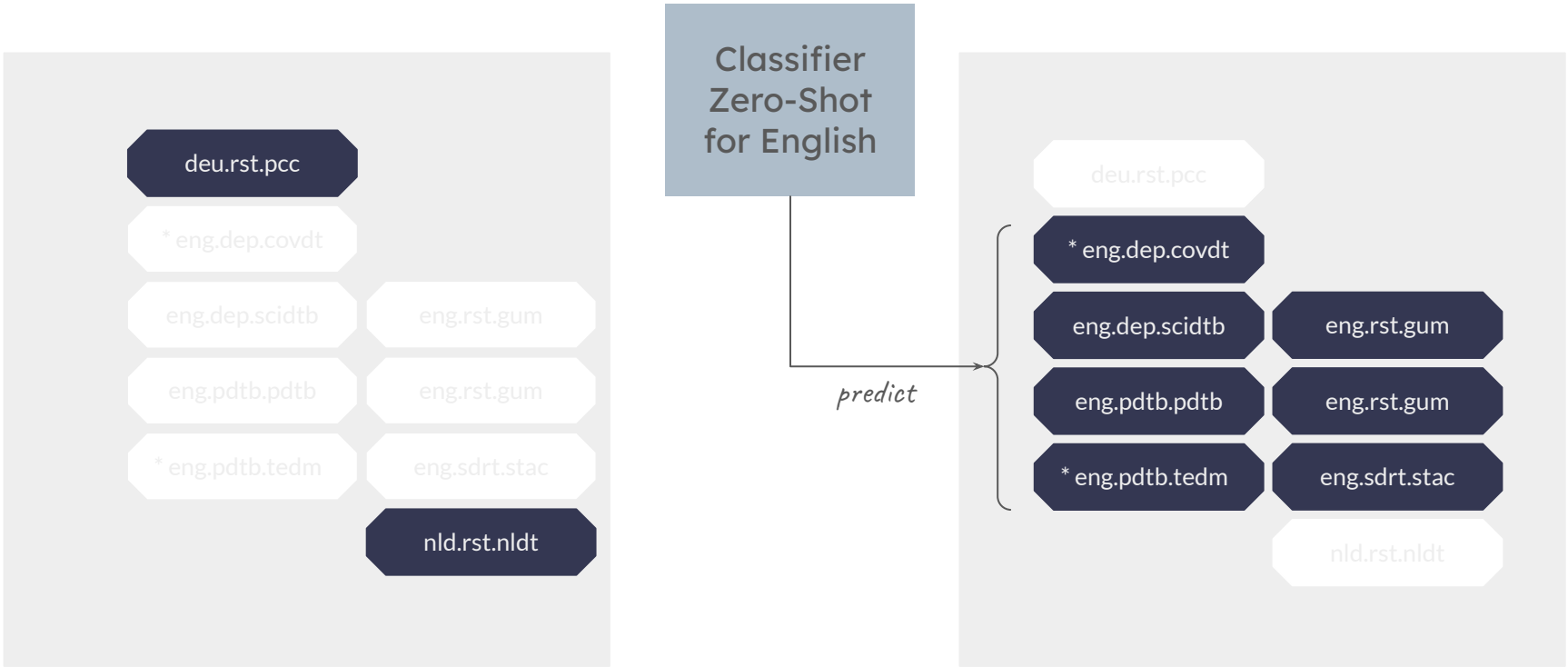


Zero-Shot: Language families (example)



Classifier Zero-Shot for English

Zero-Shot: Language families (example)



Jaccard similarity

spa.rst.rstdt

preparation
list
motivation
circumstance
elaboration
...

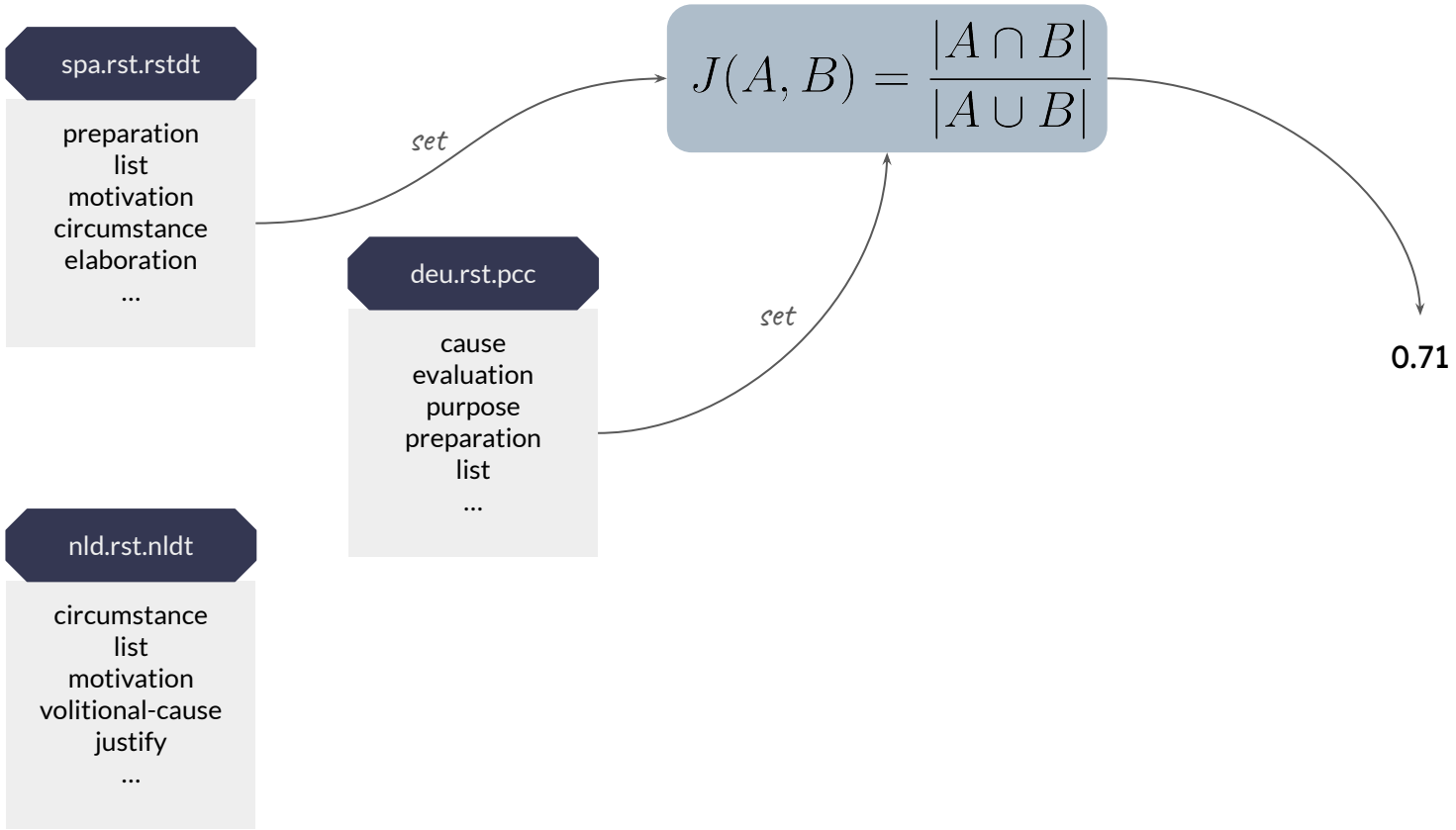
deu.rst.pcc

cause
evaluation
purpose
preparation
list
...

nld.rst.nldt

circumstance
list
motivation
volitional-cause
justify
...

Jaccard similarity



Jaccard similarity

spa.rst.rstdt

preparation
list
motivation
circumstance
elaboration
...

deu.rst.pcc

cause
evaluation
purpose
preparation
list
...

nld.rst.nldt

circumstance
list
motivation
volitional-cause
justify
...

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

	spa.rst.rstdt	deu.rst.pcc	nld.rst.nldt
spa.rst.rstdt		0.71	0.73
deu.rst.pcc	0.71		0.56
nld.rst.nldt	0.73	0.56	

Results

Results - Language Families

Germanic languages	Baseline	Zero-Shot
deu.rst.pcc	0.32	0.15
***eng.dep.covdtb**	*0.63	0.52
eng.dep.scidtb	0.72	0.06
eng.pdtb.pdtb	0.73	0.03
***eng.pdtb.tedm**	*0.52	0.02
eng.rst.gum	0.54	0.05
eng.rst.rstdt	0.64	0.40
eng.sdrt.stac	0.62	0.09
nld.rst.nldt	0.43	0.26

Romance languages	Baseline	Zero-Shot
fra.sdrt.annodis	0.46	0.23
ita.pdtb.luna	0.52	0.20
por.pdtb.crpc	0.66	0.04
***por.pdtb.tedm**	*0.44	0.05
por.rst.cstn	0.57	0.29
spa.rst.rststb	0.56	0.25
spa.rst.sctb	0.43	0.35

Baseline: **mBERT** trained only with target dataset

Results - Language Families

- Zero-shot is difficult!
- Zero-shot with groups of **language families**:
  - Steep drop in accuracy for most corpora, low similarity of label sets
  - OOD did well
  - Some English corpora, Portuguese ≈ zero accuracy (unique label sets)

Results - Frameworks

PDTB framework	Baseline	Zero-Shot
eng.pdtb.pdtb	0.73	0.55
***eng.pdtb.tedm**	*0.52	0.55
ita.pdtb.luna	0.52	0.42
por.pdtb.crpc	0.66	0.48
***por.pdtb.tedm**	*0.44	0.45
tha.pdtb.tdtb	0.94	0.57
tur.pdtb.tdb	0.41	0.37
***tur.pdtb.tedm**	*0.35	0.40
zho.pdtb.cdtb	0.83	0.47

DEP framework	Baseline	Zero-Shot
***eng.dep.covdtb**	*0.63	0.11
eng.dep.scidtb	0.72	0.35
zho.dep.scidtb	0.55	0.41

Baseline: **mBERT** trained only with target dataset

RST framework	Baseline	Zero-Shot
deu.rst.pcc	0.32	0.20
eng.rst.gum	0.54	0.10
eng.rst.rstdt	0.64	0.42
eus.rst.ert	0.42	0.33
fas.rst.prstc	0.52	0.40
nld.rst.nldt	0.43	0.30
por.rst.cstn	0.57	0.49
rus.rst.rrt	0.59	0.40
spa.rst.rststb	0.56	0.46
spa.rst.sctb	0.43	0.60
zho.rst.gcdt	0.6	0.01
zho.rst.sctb	0.46	0.48

SDRT framework	Baseline	Zero-Shot
eng.sdrt.stac	0.62	0.19
fra.sdrt.annodis	0.46	0.24

Results - Frameworks

- Zero-shot is difficult!
- Zero-shot with **frameworks**:
  - Lower accuracy overall
  - Significant drop for Thai (only explicit relations)
  - Significant drop for some English and Chinese datasets (unique labels)
  - OOD datasets in Turkish and Portuguese: improvement with PDTB-only classifier!
  - Improvement for Spanish datasets
  - Low accuracy for SDRT- and DEP-only classifiers (small training sets)

Results - Jaccard similarity

PDTB-adjacent	Baseline	Zero-Shot
eng.pdtb.pdtb	0.73	0.55
***eng.pdtb.tedm**	*0.52	0.55
por.pdtb.crpc	0.66	0.47
***por.pdtb.tedm**	*0.44	0.46
tha.pdtb.tdtb	0.94	0.58
tur.pdtb.tdb	0.41	0.38
***tur.pdtb.tedm**	*0.35	0.42

DEP-RST-adjacent	Baseline	Zero-Shot
*eng.dep.covdtb	*0.63	0.21
eng.dep.scidtb	0.72	0.4
eng.rst.rstdt	0.64	0.37
fas.rst.prstc	0.52	0.46
zho.dep.scidtb	0.55	0.43

RST-adjacent	Baseline	Zero-Shot
deu.rst.pcc	0.32	0.18
eus.rst.ert	0.42	0.36
nld.rst.nldt	0.43	0.31
por.rst.cstn	0.57	0.46
rus.rst.rrt	0.59	0.31
spa.rst.rststb	0.56	0.49
spa.rst.sctb	0.43	0.61
zho.rst.sctb	0.46	0.51

Baseline: **mBERT** trained only with target dataset

Results - Jaccard similarity

- Zero-shot is difficult!
- Zero-shot with **Jaccard similarity groups**:
- First group: most PDTB corpora
  - Significant drop for Thai (only explicit relations)
  - OOD datasets in Turkish and Portuguese: improvement
- Second group: most RST corpora
  - Better performance for Spanish and Chinese
  - Worse performance for German, Dutch, Russian
- Third group: RST+DEP corpora
  - All accuracies low (lower similarities)

Conclusion

- Zero-shot is difficult… but discourse relation classification even harder!

- Data-hungry classifiers vs. small datasets
- OOD datasets liked zero-shot!
- But… realistic and not impossible!

Gitlab repository:

https://gitlab.irit.fr/melodi/andiamo/discret-zero-shot

Thank you for your attention!