Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training

Merel Scholman[a], Valentina Pyatkin[b], Frances Yung[a],
Ido Dagan[b], Reut Tsarfaty[b], Vera Demberg[a]

[a] *Saarland University, Germany*
[b] *Bar Ilan University, Israel*

LREC 2022

Introduction

- Discourse relations (DR) are logical links between segments of text
- Annotating DRs is difficult, even for experts (Spooren & Degand, 2010)

Example:

*I love dogs. [***But/Specifically***] I think poodles are the best.*

- CONCESSION, SPECIFICATION?

Introduction

- Discourse relations (DR) are logical links between segments of text
- Annotating DRs is difficult, even for experts (Spooren & Degand, 2010)

- Traditional annotation is time- and cost-intensive
- Crowdsourcing can provide solution, but crowdsourcing tasks require adaptations:
  - Task design (Yung et al, 2019; Pyatkin et al., 2020)
  - Worker selection and training (current contribution)

Example:

*I love dogs. [**But/Specifically**] I think poodles are the best.*

- CONCESSION, SPECIFICATION?

Introduction

Controlled crowdsourcing annotation protocols and learning curricula effective in other fields:

- Controlled crowd annotation protocols: (Nangia et al., 2021; Roit et al., 2020) crowd-wide recruitment round → screening → training → production
- Annotation curricula: gradually train workers by ordering items from easier examples to more difficult ones (Lee et al., 2021; Tauchmann et al., 2020)

Introduction

Controlled crowdsourcing annotation protocols and learning curricula effective in other fields:

- Controlled crowd annotation protocols: (Nangia et al., 2021; Roit et al., 2020) crowd-wide recruitment round → screening → training → production
- Annotation curricula: gradually train workers by ordering items from easier examples to more difficult ones (Lee et al., 2021; Tauchmann et al., 2020)

Current contribution:

- Study trade-off between resources and reliability of crowdsourced DR annotation, across two independent annotation methods
  - Study 1: No worker selection or training
  - Study 2: Selection-and-training
  - Study 3: Selection-only

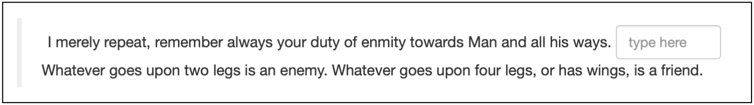
Table of Contents

1. Annotation methods: DC and QA
2. Data
3. Study 1: No selection or training
4. Study 2: Selection-and-training
5. Study 3: Selection-only
6. Conclusion

Method: Discourse Connectives (DC)

Two-step DC method:

1. Freely insert connective to express relation

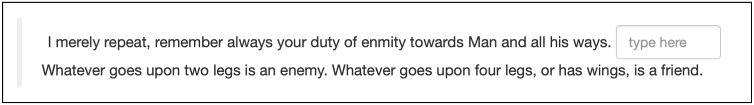


I merely repeat, remember always your duty of enmity towards Man and all his ways. type here
Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

Method: Discourse Connectives (DC)

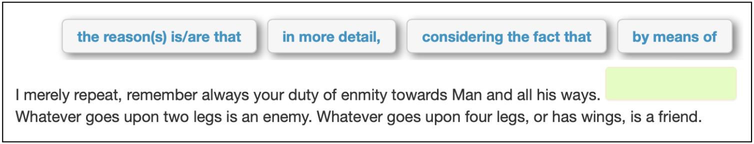
Two-step DC method:

1. Freely insert connective to express relation



I merely repeat, remember always your duty of enmity towards Man and all his ways. Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

2. Choose from automatically provided list to disambiguate



I merely repeat, remember always your duty of enmity towards Man and all his ways. Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

Mapping between connectives and PDTB relation labels: a connective bank created for this method

Yung, Scholman & Demberg (2019), *LAW*.

Method: Question-Answer (QA)

Relate two clauses with a QA pair:

Lucie is feeling tired. She is going to a party.

1. Choose a Question Prefix from a predefined set of question starts:
    - **Despite what**

Method: Question-Answer (QA)

Relate two clauses with a QA pair:

Lucie is feeling tired. She is going to a party.

1. Choose a Question Prefix from a predefined set of question starts:
   - **Despite what**
2. Complete the question with text from either one of the two clauses:
   - **Despite what** is she going to a party?
3. The other clause should answer the created question:
   - **Despite what** is she going to a party?
   - Lucie is feeling tired.

Mapping between QAs and PDTB labels: one-to-one mapping from question prefixes + clause order to labels

Pyatkin, Klein, Tsarfaty & Dagan (2020), *EMNLP*.

Table of Contents

1. Annotation methods: DC and QA
2. Data
3. Study 1: No selection or training
4. Study 2: Selection-and-training
5. Study 3: Selection-only
6. Conclusion

Method: Data

- Implicit relations from Wikipedia and Blog Authorship Corpus
- Gold labels provided by three expert annotators

Method: Data

- Implicit relations from Wikipedia and Blog Authorship Corpus
- Gold labels provided by three expert annotators

- Same texts used across the studies:
  - Study 1: No worker selection or training
  - Study 2: Selection-and-training
  - Study 3: Selection-only

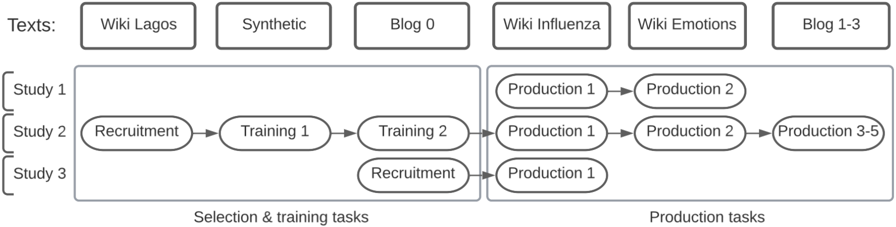


Table of Contents

1. Annotation methods: DC and QA
2. Data
3. Study 1: No selection or training
4. Study 2: Selection-and-training
5. Study 3: Selection-only
6. Conclusion

Study 1: No selection or training

- Aim: establish baseline for agreement using the DC and QA methods
- Prolific workers (n=10) annotated one text with DC method and other with QA method

Study 1: No selection or training

- Aim: establish baseline for agreement using the DC and QA methods
- Prolific workers (n=10) annotated one text with DC method and other with QA method

Results:

Task	DC		QA	
	κ	Agree gold-maj	κ	Agree gold-maj
Influenza	.27	45	.18	18
Emotions	.20	28	.09	17

Table: κ: Cohen's kappa agreement between the gold and majority label per item; *Agree gold-maj*: percent agreement between the gold label and majority label.

Study 1: No selection or training

- Aim: establish baseline for agreement using the DC and QA methods
- Prolific workers (n=10) annotated one text with DC method and other with QA method

Results:

Task	DC		QA	
	κ	Agree gold-maj	κ	Agree gold-maj
Influenza	.27	45	.18	18
Emotions	.20	28	.09	17

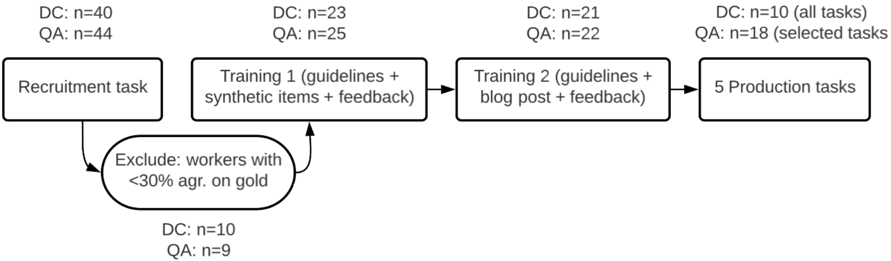
Table: κ: Cohen's kappa agreement between the gold and majority label per item; *Agree gold-maj*: percent agreement between the gold label and majority label.

- Much room for improvement
- Discrepancy with original results of both methods due to alterations (inter-sentential implicit relations, different relational classes, etc.)

Table of Contents

1. Annotation methods: DC and QA
2. Data
3. Study 1: No selection or training
4. Study 2: Selection-and-training
5. Study 3: Selection-only
6. Conclusion

Study 2: Selection-and-training



- Recruitment task to exclude poorest performers
- For training, workers were provided with PDF guidelines to explain task
- Training item selection corresponded to a learning curriculum
- During training: immediate feedback

Study 2: Selection-and-training – results

	DC		QA	
	κ	% Agree gold-maj	κ	% Agree gold-maj
Recruitment	.61	67	.53	61
Training	.97	97	.85	84
Production	.7	74	.56	62

Study 2: Selection-and-training – results

	DC		QA	
	κ	% Agree gold-maj	κ	% Agree gold-maj
Recruitment	.61	67	.53	61
Training	.97	97	.85	84
Production	.7	74	.56	62

- Agreement high on training texts → task and methods are feasible
- All agreement metrics are higher after training than before training

Study 2: Selection-and-training – results

	DC		QA	
	κ	% Agree gold-maj	κ	% Agree gold-maj
Study 1 Influenza	.27	45	.18	18
Study 2 Influenza	.61	73	.47	64
Study 1 Emotions	.20	28	.09	17
Study 2 Emotions	.64	72	.35	44

- Agreement high on training texts → task and methods are feasible
- All agreement metrics are higher after training than before training
- Performance on Influenza & Emotions texts: Clear boost between the untrained group in Study 1 (κs $<$.27) and the trained group in Study 2

→ Selection-and-training yields more reliable annotations for both methods

Study 2: Selection-and-training – drawback

- Drawback: proportion of the trained workers would not return to new tasks
  → Training investment misspent & data collection slowed
- Selection-and-training method might not be optimal for certain research efforts, given the available resources

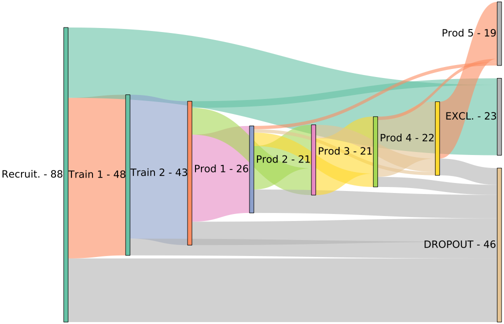
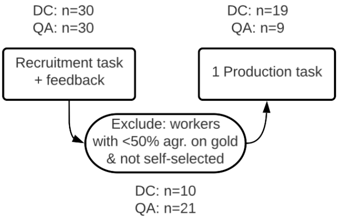


Figure: Illustration of Study 2's pipeline for both methods combined.

Table of Contents

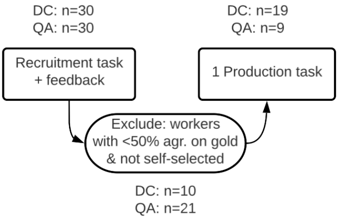
1. Annotation methods: DC and QA
2. Data
3. Study 1: No selection or training
4. Study 2: Selection-and-training
5. Study 3: Selection-only
6. Conclusion

Study 3: Selection-only


DC: n=30
QA: n=30
Recruitment task + feedback
DC: n=19
QA: n=9
1 Production task
Exclude: workers with <50% agr. on gold & not self-selected
DC: n=10
QA: n=21

- Engaged a larger pool of workers with a recruitment task; used more stringent selection criteria to create subpool of “talented” workers

Study 3: Selection-only


DC: n=30
QA: n=30

Recruitment task
+ feedback

Exclude: workers
with <50% agr. on gold
& not self-selected

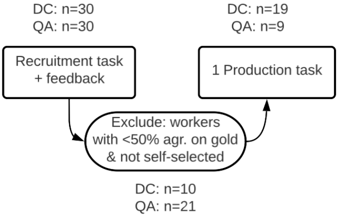
DC: n=10
QA: n=21

DC: n=19
QA: n=9

1 Production task

- Engaged a larger pool of workers with a recruitment task;
  used more stringent selection criteria to create subpool of "talented" workers
  - Cost-efficient: no training investment, so more workers can be recruited
  - Time-effective: tasks completed faster because of larger subpool

Study 3: Selection-only



- Engaged a larger pool of workers with a recruitment task;
  used more stringent selection criteria to create subpool of "talented" workers
  - Cost-efficient: no training investment, so more workers can be recruited
  - Time-effective: tasks completed faster because of larger subpool
- Recruitment task: training 2 text, including the feedback component
- More stringent pre-selection (workers must have completed university)
  and post-selection (including self-selection)

Study 3: Selection-only – results recruitment task

Study	Participant type	Invested cost GBP	DC κ	DC % Agree gold-maj	QA κ	QA % Agree gold-maj
2	Trained part.	10.10	**.92**	94	**.84**	85
3	All recruit.	0	.84	89	.77	83
3	Final selection	0	**.85**	89	**.7**	61

Study 3: Selection-only – results recruitment task

			DC		QA	
		Invested	κ	% Agree	κ	% Agree
Study	Participant type	cost GBP		gold-maj		gold-maj
2	Trained part.	10.10	**.92**	94	**.84**	85
3	All recruit.	0	.84	89	.77	83
3	Final selection	0	**.85**	89	**.7**	61

- Results show promise considering study 3 workers have less experience

Study 3: Selection-only – results influenza task

			DC		QA	
Study	Participant type	Invested cost GBP	κ	% Agree gold-maj	κ	% Agree gold-maj
1	Untrained	0	.27	45	.18	18
2	Trained	11.98	.61	73	.47	64
3	All selected	1.88	.41	68	.28	41
3	Decent selected	1.88	.58	77	.45	55

Study 3: Selection-only – results influenza task

			DC		QA	
Study	Participant type	Invested cost GBP	κ	% Agree gold-maj	κ	% Agree gold-maj
1	Untrained	0	.27	45	.18	18
2	Trained	11.98	.61	73	.47	64
3	All selected	1.88	.41	68	.28	41
3	Decent selected	1.88	.58	77	.45	55

- With continuous quality monitoring, κs similar to trained participants can be obtained
- Selection-only method appears to be an attractive alternative to the selection-and-training method

Table of Contents

1 Annotation methods: DC and QA

2 Data

3 Study 1: No selection or training

4 Study 2: Selection-and-training

5 Study 3: Selection-only

6 Conclusion

Resource comparison

Worker selection entails trade-off between resources and annotation quality:

- Study 1: Quick and cheap, but lowest-quality data
- Study 2: High quality data, but slow and expensive due to dropout (52%)
- Study 3: Relatively quick, data quality comparable to Study 2, but 77% decrease in cost investment compared to Study 2

Resource comparison

Worker selection entails trade-off between resources and annotation quality:

- Study 1: Quick and cheap, but lowest-quality data
- Study 2: High quality data, but slow and expensive due to dropout (52%)
- Study 3: Relatively quick, data quality comparable to Study 2, but 77% decrease in cost investment compared to Study 2

Relevant considerations:

- Continuous quality monitoring is necessary, even with “talented” workers. E.g., bonuses, accuracy check reminders, intermediate quality checks

Conclusion

- Training leads to more reliable annotated data, but this comes at a high cost (time and money)
- Selection-only approach more viable for certain projects in terms of resources

Conclusion

- Training leads to more reliable annotated data, but this comes at a high cost (time and money)
- Selection-only approach more viable for certain projects in terms of resources

- First step in a larger project: study how design choices for discourse annotation tasks shape research results
- Future work: detailed comparison between the obtained annotations from different methods

Conclusion

- Training leads to more reliable annotated data, but this comes at a high cost (time and money)
- Selection-only approach more viable for certain projects in terms of resources
- First step in a larger project: study how design choices for discourse annotation tasks shape research results
- Future work: detailed comparison between the obtained annotations from different methods

Thank you for your attention!

IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

Bar-Ilan
University
אוניברסיטת בר־אילן