Marcel Gohsen[1] Matthias Hagen[2] Martin Potthast[3] Benno Stein[1]

[1]Bauhaus-Universität Weimar [2]Friedrich-Schiller-Universität Jena [3]Leipzig University and ScaDS.AI

Task-Oriented Paraphrase Analytics

Paraphrase Definitions

level of semantic similarity

"convey the **same meaning** but use different words" [Bhagat and Hovy, 2013]

sem. equivalent

Sentence Compression
"generate a shorter paraphrase of a sentence"
[Filippova et al., 2013]

Original: *The future of the nation is in your hands.*
Paraphrase: *The nation's future is in your hands.*

"restatements with **approximately the same meaning**" [Wang et al., 2019]

Reframing
"shift in perspective without distorting the underlying meaning"
[Ziems et al., 2022]

Original: *This was a bland dish.*
Paraphrase: *I've made dishes that are much tastier than this one.*

sem. similar

"talking about the same situation **in a different way**" [Hirst, 2003]

Tasks require texts to be *semantically equivalent* or *similar*.

Paraphrasing Task Taxonomy

Find paraphrasing tasks from the literature that are

1) explicitly defined as paraphrasing tasks and/or
2) align with common paraphrase definitions

Found: **25 paraphrasing tasks**

Paraphrase Generation Tasks

Semantically Equivalent Paraphrasing

- **Copy editing**
  - Improvement of coherence
  - Text simplification
  - Sentence compression
  - Sentence expansion
- **Data augmentation**
  - Adversarial example generation
- **Linguistic steganography**
  - Acrostification
  - Natural language watermarking
- **Style adjustment**
  - Author obfuscation
  - Plagiarizing
  - Style transfer

Semantically Similar Paraphrasing

- **Context change**
  - Image recaptioning
  - Positive reframing
  - Text localization
- **Conversational interaction**
  - Argument repetition
  - Question dodging
  - Rogerian rhetoric application
  - Utterance clarification
- **Textual entailment generation**
- **Information disguise**
- **Query suggestion**

Human Paraphrase Classification

Considered Tasks & Datasets

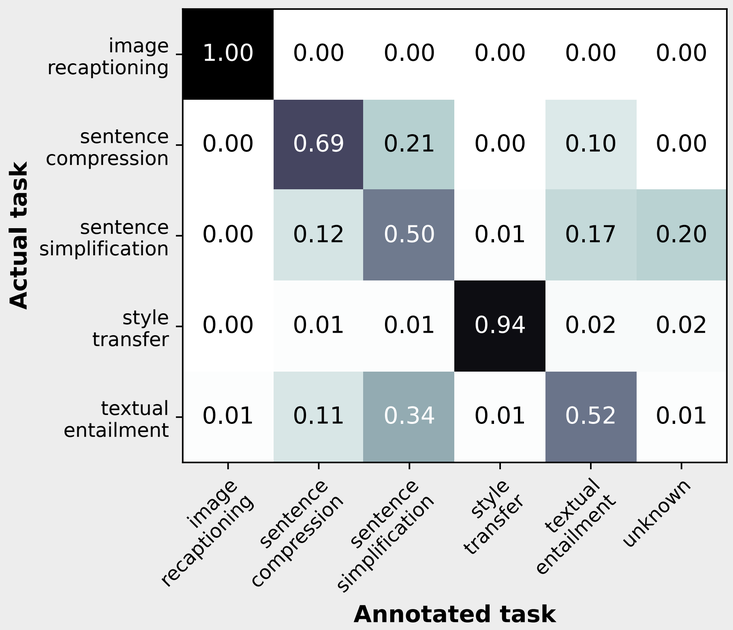
Sentence compression
- Google Compressions [Filippova and Altun, 2013]
- Microsoft Compressions [Toutanova et al., 2016]

Sentence simplification
- TurkCorpus [Xu et al., 2016]
- WikiLarge [Zhang and Lapata, 2017]

Style transfer
- ParaDetox [Logacheva et al., 2022]
- Bible style transfer [Carlson et al., 2018]

Image recaptioning
- MSCOCO [Lin et al., 2015]
- VizWiz [Gurari et al., 2020]

Textual entailment generation
- SciTail [Bowman et al., 2015]
- HELP [Yanaka et al., 2019]



Sample: 50 random paraphrases per dataset.
Performance: F1 = 0.73.

Machine Paraphrase Classification

Paraphrase Classifier

Training data
- 4,000 random paraphrases per dataset (40,000 in total)
- Paraphrases from five tasks

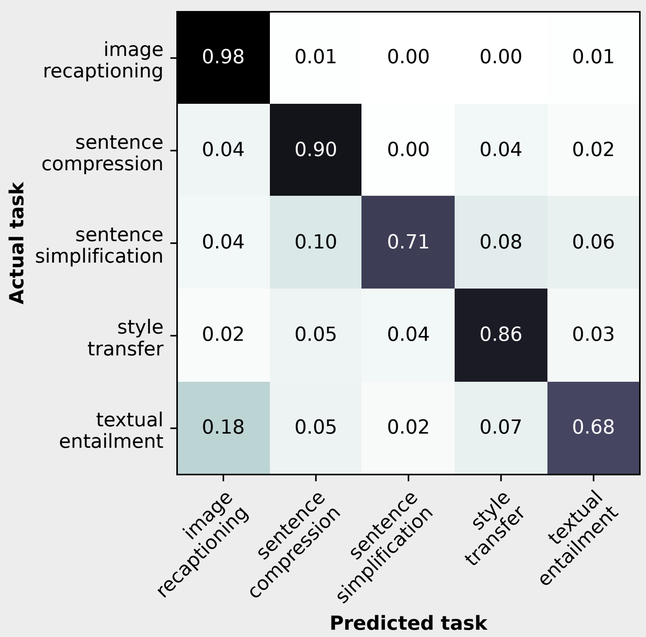
Random Forest classifier
- Max. depth of 15

Features

Lexical similarity:
- ROUGE-1
- BLEU

Syntactic features:
- POS n-gram frequencies

Semantic similarity:
- Sentence-BERT



Sample: 1,000 random paraphrases per dataset.
Performance: F1 = 0.82.

Inhomogeneity of Common Paraphrase Datasets

Paraphrase Dataset	Image Recaptioning		Sentence Compression		Sentence Simplification		Style Transfer		Textual Entailment		Total
MSRPC	6.7%	390	32.0%	1,858	38.6%	2,241	11.4%	653	11.4%	659	5,801
PAWS	5.2%	3,367	24.7%	16,194	**62.7%**	41,004	3.7%	2,442	3.6%	2,394	65,401
TaPaCo	1.8%	4,140	8.4%	18,949	1.0%	2,141	**76.8%**	172,718	**12.0%**	26,877	224,825
Wikipedia-IPC$_{silver}$	**16.3%**	37,489	**62.0%**	142,492	19.8%	45,535	0.2%	427	1.7%	3,934	229,877
Total	8.6%	45,386	34.1%	179,493	17.3%	90,921	33.5%	176,240	6.4%	33,864	525,904

Code

github.com/webis-de/LREC-COLING-24