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

Bauhaus-Universität Weimar[1] Friedrich-Schiller-Universität Jena[2] Leipzig University[3]

Paraphrase Acquisition from Image Captions

Motivation

- Paraphrasing is an important NLP task
- Large sets of paraphrases are useful for training or fine-tuning paraphrasing models
- Manual paraphrase acquisition is expensive to scale

Contributions

- Method for paraphrase acquisition from image captions
- Wikipedia-IPC: Image caption paraphrase dataset consisting of
  - 30.237 gold-quality paraphrases
  - 229.877 silver-quality paraphrases
  - 656.560 bronze-quality paraphrases
- Paraphrase "sophistication" metric $\Delta_{\text{sem,syn}}$

Easter Bunny

From Wikipedia, the free encyclopedia

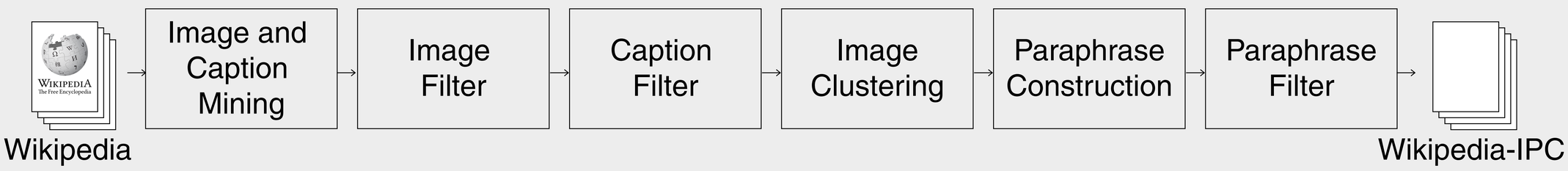
A 1907 postcard featuring the Easter Bunny

Ēostre

From Wikipedia, the free encyclopedia

An Easter postcard from 1907 depicting a rabbit

Caption-based Paraphrase Acquisition



Caption Mining
- Captions
- Alternative texts

Image Filter
- Number of references in [2, 10]

Caption Filter
- Captions with at least six words
- Captions that are sentences

Image Clustering
- Equivalent images have same URL

Paraphrase Construction
- Combine captions and alt texts within an image cluster

Paraphrase Filter
- Caption pairs with significant difference

Quantitative Similarity Analysis

Similarity Metrics

Lexical and Syntactic Similarity
- ROUGE-1
- ROUGE-L
- BLEU

Semantic Similarity
- Word Mover Distance (WMS)
- BERTScore
- Sentence Transformer (ST)

Paraphrase Sophistication $\Delta_{\text{sem,syn}}$
- Semantically similar
- Syntactically dissimilar
- Delta between semantic and syntactic similarity

Average Similarity of Paraphrases

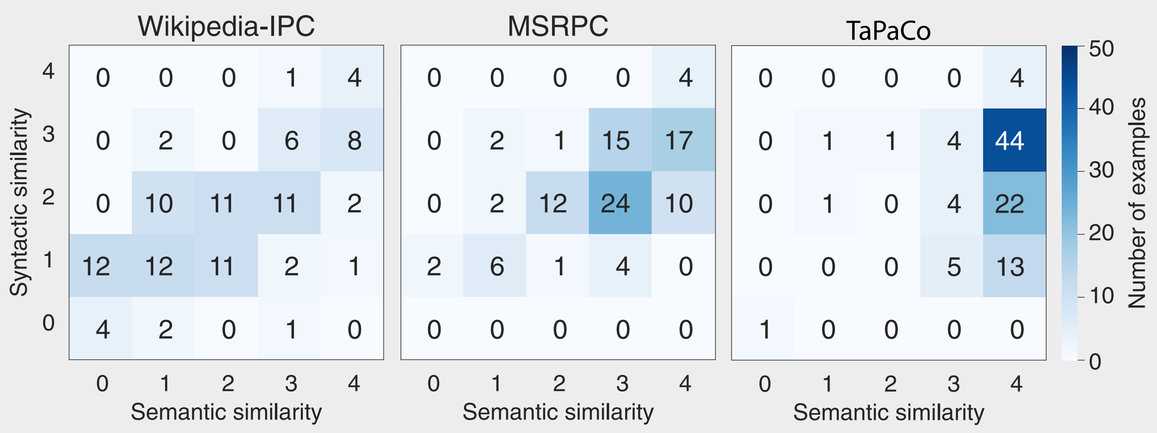
		Syntactic similarity				Semantic similarity				
Corpus	**Acquisition**	ROUGE-1	ROUGE-L	BLEU	Avg.	WMS	BERT	ST	Avg.	$\Delta_{\text{sem,syn}}$
Wikipedia-IPC$_{\text{gold}}$	Caption	0.74	0.71	0.56	0.67	**0.83**	0.69	0.91	0.81	0.14
Wikipedia-IPC$_{\text{silver}}$	Caption	0.71	0.67	0.52	0.63	0.63	0.81	0.90	0.78	0.15
Flickr8k	Caption	0.53	0.48	**0.22**	0.41	0.59	0.73	0.86	0.73	**0.32**
MS-COCO	Caption	**0.51**	**0.45**	**0.22**	**0.39**	0.57	0.71	0.86	0.71	**0.32**
PASCAL	Caption	**0.51**	0.47	**0.22**	0.40	0.59	0.72	0.86	0.73	**0.32**
ParaNMT-5m	Generated	0.63	0.60	0.33	0.52	0.60	0.75	0.87	0.74	0.22
PAWS	Generated	0.94	0.79	0.69	0.81	0.82	**0.96**	**0.97**	**0.92**	0.11
MSRPC	Dist. supervision	0.73	0.69	0.54	0.65	0.72	0.82	0.90	0.82	0.16
PPDB 2.0	Dist. supervision	0.64	0.63	0.32	0.53	0.64	0.63	0.89	0.72	0.19
TaPaCo	Dist. supervision	0.65	0.63	0.30	0.53	0.78	0.79	0.91	0.83	0.30

Qualitative Similarity Analysis

Manual Annotation
- Semantic/Syntactic similarity
- 5-point Likert scale
- Three datasets
- 100 paraphrases per dataset

Correlation Analysis

Syntax:	ROUGE-1	ROUGE-L	BLEU	Average
r	0.78	0.77	0.70	0.79
Semantics:	**WMS**	**BERT**	**ST**	**Average**
r	0.59	0.70	0.78	0.76



Code and Data

github.com/webis-de/EACL-23