
Image Captions as Paraphrases

Master's Thesis Defence

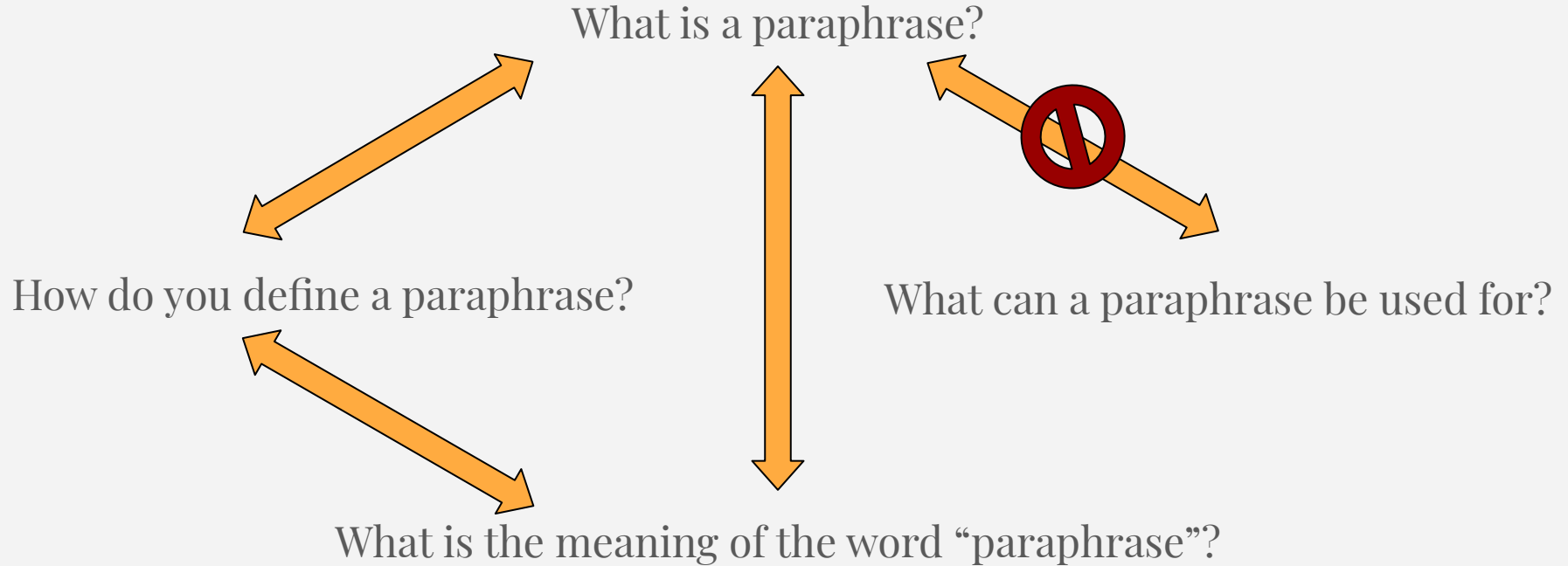
Marcel Gohsen
15th March 2021

Contents

1. Introduction
2. Formal Definition of Paraphrases and Textual Entailment
3. Paraphrase Acquisition Approach
4. Experiment and Evaluation
5. Conclusion

1. Introduction

Paraphrase Definition



Paraphrase Applications

NLP applications that exploited paraphrase resources:

- Textual Entailment e.g., Izadinia et al., 2015, Marelli et al., 2014
- Semantic Similarity Analysis e.g., Agirre et al., 2015, Li and Srikumar, 2016
- Machine Translation e.g., Seraj et al., 2015
- Question Answering e.g., Fader et al., 2013
- Sentence Compression e.g., Cordeiro et al., 2007
- Text Summarization e.g., Cordeiro et al., 2007
- ...

Existing Paraphrase Corpora

Corpus	Reference	Example Pairs	Paraphrases	Acquisition Methodology	Granularity
PPDB:Eng 1.0	Ganitkevitch et al., 2013	221,400,000	169,600,000	automatically extracted	phrases (max 4-grams)
SICK	Marelli et al., 2014	9,840	4,366	automatically extracted	sentences
PAWS	Zhang et al., 2019	108,463	35,789	automatically generated	sentences
ParaNMT-50M	Wieting et al., 2018	50,000,000	? (no labels)	automatically generated	sentences
MSRPC	Dolan et al., 2005	5,801	3,900	manually annotated	sentences
Webis-CPC-11	Burrows et al., 2013	7,859	4,067	crowd-sourced	passages

2. Formal Paraphrase Definition

Paraphrase Definition

$t_{1,2}$: “Formal representations of texts”

α : “Common domain or world knowledge”

$\Phi_{1,2}$: “Sets of deducible statements”

$$(t_1 \wedge \alpha) \models \Phi_1 \quad \Leftrightarrow \quad (t_2 \wedge \alpha) \models \Phi_2$$

$$\Phi_1 \cap \Phi_2 \neq \emptyset$$

Textual Entailment

Is a statement entailed in a text?

Example:

“U.S. president Barack Obama enjoys playing golf” **entails** that “Barack Obama is the president of the United states”.

Can we find statements that are entailed in both texts t_1, t_2 ?

Forward/Reverse Entailment

$t_{1,2}$: “Formal representations of texts”
 $t_1 \sqsubset t_2$: “ t_1 entails t_2 ”
 $\Phi_1 \subset \Phi_2$: “Every statement that can be inferred from $t_1 \wedge \alpha$ can also be inferred from $t_2 \wedge \alpha$ ”

$$\begin{aligned}
 t_1 \sqsubset t_2 &\Leftrightarrow \Phi_1 \supset \Phi_2 \\
 t_1 \supset t_2 &\Leftrightarrow \Phi_1 \subset \Phi_2
 \end{aligned}$$

t_1 : “Barack Obama is the president of the United States”

t_2 : “The 44th president of the United States is Barack Obama”

$$\Rightarrow t_2 \sqsubset t_1$$

Semantic Relations

Equivalence: $t_1 \equiv t_2 \iff \Phi_1 = \Phi_2$

Entailment: $t_1 \sqsubset t_2 \iff \Phi_1 \supset \Phi_2$

Exclusion: $t_1 \mid t_2 \iff \Phi_1 \cap \Phi_2 = \emptyset$

Negation: $t_1 \leftrightarrow \neg t_2 \iff \exists \varphi_1, \varphi_2 : \varphi_1 \leftrightarrow \neg \varphi_2$

Intersection:

$$t_1 \sqcap t_2 \iff \Phi_1 \cap \Phi_2 \neq \emptyset \wedge (\Phi_1 \not\sqsubset \Phi_2) \wedge (\Phi_2 \not\sqsubset \Phi_1)$$

3. Paraphrase Acquisition

Example

British Airways

From Wikipedia, the free encyclopedia



British Airways and Iberia merged in January 2011, forming International Airlines Group, one of the world's largest airlines.

https://en.wikipedia.org/wiki/British_Airways

History of British Airways

From Wikipedia, the free encyclopedia

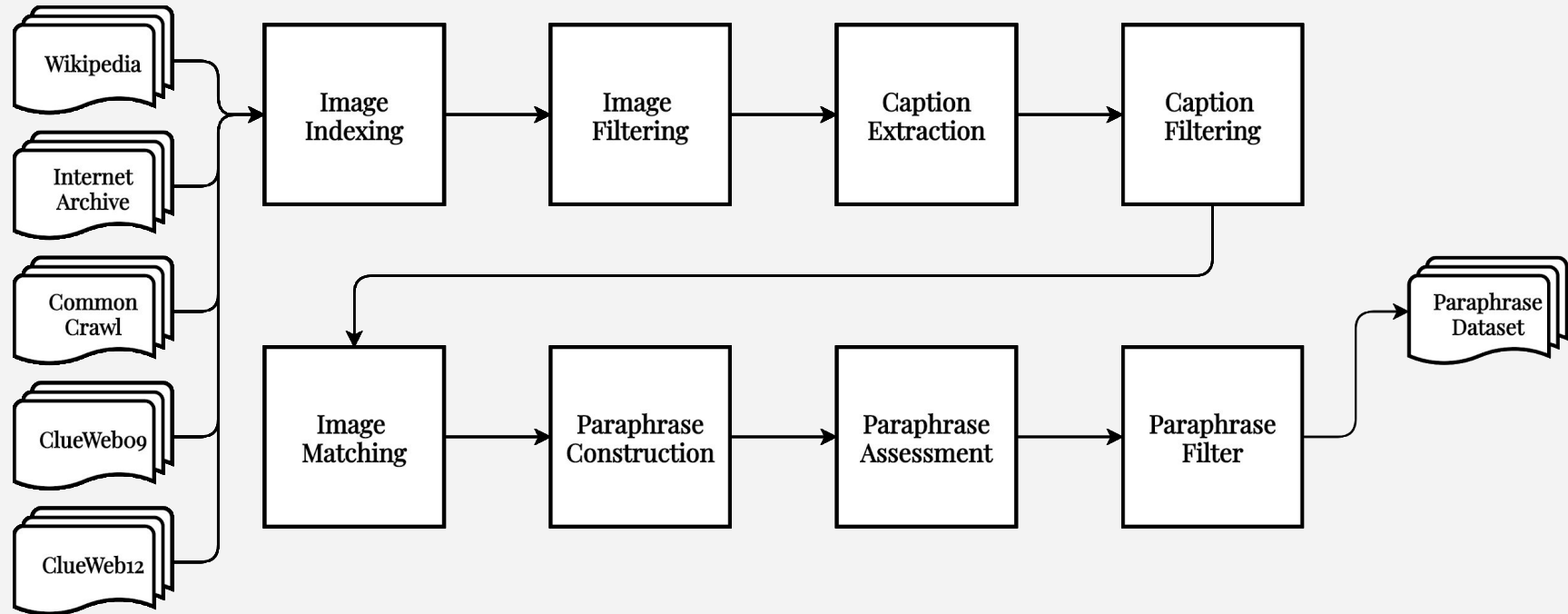


British Airways and Iberia announced their merger in April 2010, creating the **International Airlines Group**

https://en.wikipedia.org/wiki/History_of_British_Airways



Paraphrase Acquisition Pipeline



Web Resources



Wikimedia dumps

Web pages:	20,707,605
Unique images:	3,106,419 (upper bound)
Image references:	5,745,681 (1.8 per image)
Captions:	3,389,952 (59% of references)

High textual quality manually observed from 500 randomly drawn captions.

Web Resources



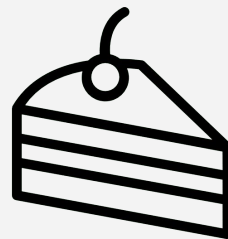
Common Crawl

Web pages:	2,886,236,237
Unique images:	8,419,862,945 (upper bound)
Image references:	67,615,864,465 (8 per image)
Captions:	34,078,395,690 (~50% of references) (lower bound)

Varying text quality manually observed from 500 randomly drawn captions.

Image Filter

Goal: Exclude images which potentially don't have an explanatory caption



...

Find distinguishable characteristics which separate positive and negative examples

- Randomly draw 5000 images balanced by size from Common Crawl
- Annotate images whether the annotator expects an explanatory caption

(0x0, 32x32], (32x32, 256x256], (256x256, 2048x2048), (2048x2048, 4096x4096], (4096x4096, inf]

Image Filter

Manually derived heuristics for positive examples:

$$resolution(img) \in [0.003, 1] \text{ Mpx}$$

$$\wedge file_size(img) > 3000 \text{ bytes}$$

$$\wedge aspect_ratio(img) \in [1, 3]$$

$$\wedge transparent_pixels(img) = 0$$

$$\wedge num_references(img) < 10$$

Caption Extraction from Wikipedia

Wikitext Image Tag Syntax

```
[ [File:Name|Type|Border|...|alt=Alt text|...|Caption] ]
```

“Alternative text is text associated with an image that serves the same purpose and conveys the same essential information as the image.” ^[1]

Caption Extraction from Common Crawl

HTML

```

```

Caption specification is not standardized in HTML

“[...] the alt attribute must be specified and its value must not be empty; the value must be an appropriate replacement for the image” [1]

Caption Extraction in Common Crawl

```
<figure>
  
  <figcaption>Caption</figcaption>
</figure>
```

```
<table>
  <tr><td></td></tr>
  <tr><td class="caption">Caption</td></tr>
</table>
```

```
<div>
  
  <div>Caption</div>
</div>
```

...

Caption Filter

Goal: Maintain high textual quality of captions

Filter heuristics for a quality caption:

$$\text{num_words}(\text{caption}) > 10$$

$$\wedge \text{num_profane_words}(\text{caption}) = 0$$

$$\wedge \text{language}(\text{caption}) = \text{'English'}$$

$$\wedge \text{is_proper_sentence}(\text{caption}) = \text{true}$$

Proper Sentence Detection

Goal: Retain sentences and exclude sentence fragments

Procedure:

1. Sample 500 and 100 captions for training and test sets
2. Annotate both sets whether they are sentences or fragments
3. Manually derive POS-based rules for sentences

Statistics:

Training set: 500 (154 sentences)

Test set: 100 (19 sentences)

Sentence Detection

Rule	Premise	Pattern
1	. * MD . *	. * MD RB? VB . *
2	. * (WDT WP WRB) . *	$[\neg(\text{WDT} \text{WP} \text{WRB})]^* (\text{VBP} \text{VBZ} \text{VBD}) . *$
3	. * IN . *	$[\neg \text{IN}]^* (\text{VBP} \text{VBZ} \text{VBD}) . *$
4	\perp	. * (VBP VBZ VBD) . *

	Precision	Recall
Test set	0.94	0.79

Image Equivalence

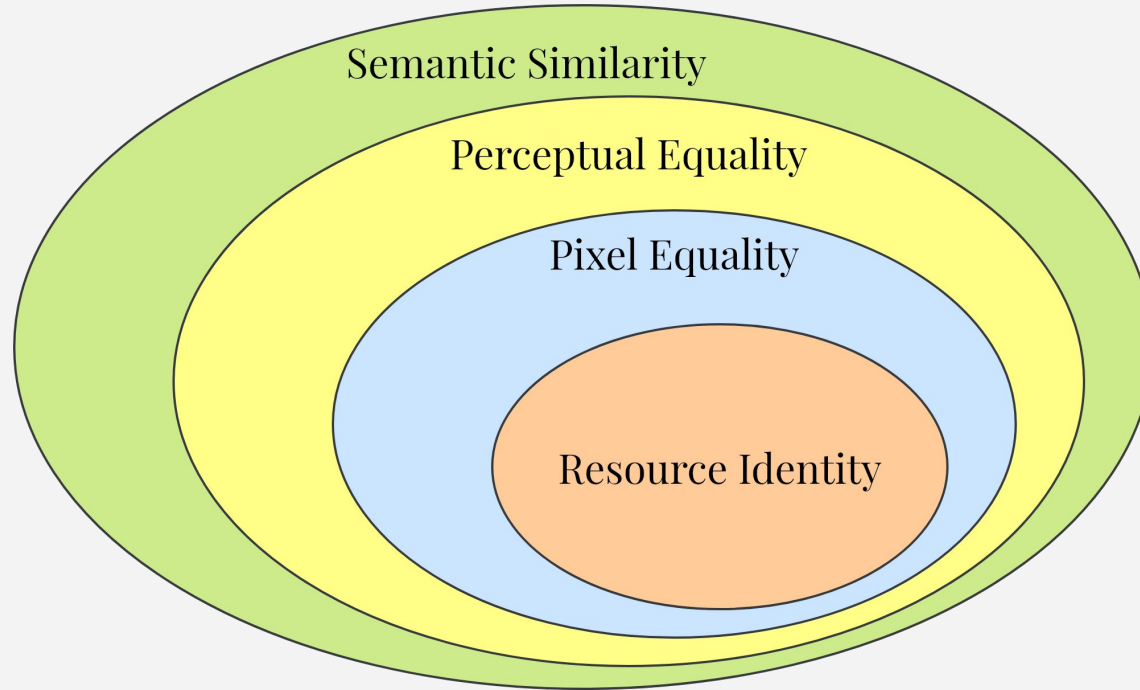


Image Equivalence

Perceptual Equality: Equivalent images are near perceptual duplicates



https://upload.wikimedia.org/wikipedia/commons/5/51/Schneegl%C3%B6ckchen_%28Galanthus%29_%2812714600103%29.jpg



https://upload.wikimedia.org/wikipedia/commons/thumb/5/51/Schneegl%C3%B6ckchen_%28Galanthus%29_%2812714600103%29.jpg/320px-Schneegl%C3%B6ckchen_%28Galanthus%29_%2812714600103%29.jpg

Image Equivalence

Semantic Similarity: The content of equivalent images are semantically similar



https://upload.wikimedia.org/wikipedia/commons/5/51/Schneegl%C3%B6ckchen_%28Galanthus%29_%2812714600103%29.jpg



https://images.immediate.co.uk/production/volatile/sites/18/2017/02/II_2011-12_Galanthus-092_JasonIngram-75b346a.jpg?quality=45&resize=556,556

Paraphrase Construction



“White snowdrops in the woods” - Alt text

“White snowdrops in the woods” - Alt text

“In the forest, many snowdrops can be found during this season” - Figcaption

“In this season, many beautiful snowdrops can be found in the woods” - Figcaption

Paraphrase Assessment

Word-based Levenshtein distance

Accumulation of minimum costs to transform a string to another

n-Gram Overlap

$$sim_o(a_n, b_n, N) = \frac{1}{N} \sum_{n=1}^N \frac{|a_n \cap b_n|}{\min(|a_n|, |b_n|)}$$

LCP n-Gram Overlap

$$sim_{exo}(a_n, b_n, N) = \max_{n \in 1, \dots, N} \frac{|a_n \cap^{LCP} b_n|}{\min(|a_n|, |b_n|)}$$

Paraphrase Assessment

BLEU (adaption for paraphrases by Cordeiro et al., 2007)

$$bleu(a_n, b_n, N) = \exp \left(\sum_{n=1}^N w_n \cdot \log \frac{|a_n \cap b_n|}{\min(|a_n|, |b_n|)} \right)$$

$$bleu(a_n, b_n, N) = \sqrt[N]{\prod_{n=1}^N \frac{|a_n \cap b_n|}{\min(|a_n|, |b_n|)}}$$

Paraphrase Assessment

Sumo (proposed by Cordeiro et al., 2007)

$$sumo(a_1, b_1, \lambda) = \begin{cases} S(a_1, b_1, \lambda) & S(a_1, b_1, \lambda) < 1 \\ e^{-k \cdot S(a_1, b_1, \lambda)} & \text{otherwise} \end{cases}$$

$$S(a_1, b_1, \lambda) = \alpha \log_2 \left(\frac{\max(|a_1|, |b_1|)}{\lambda} \right) + \beta \log_2 \left(\frac{\min(|a_1|, |b_1|)}{\lambda} \right)$$

“Barack Obama is the first African-American president of the United States”

“The 44th president of the United States is Barack Obama”

4. Experiment and Evaluation

Experiment

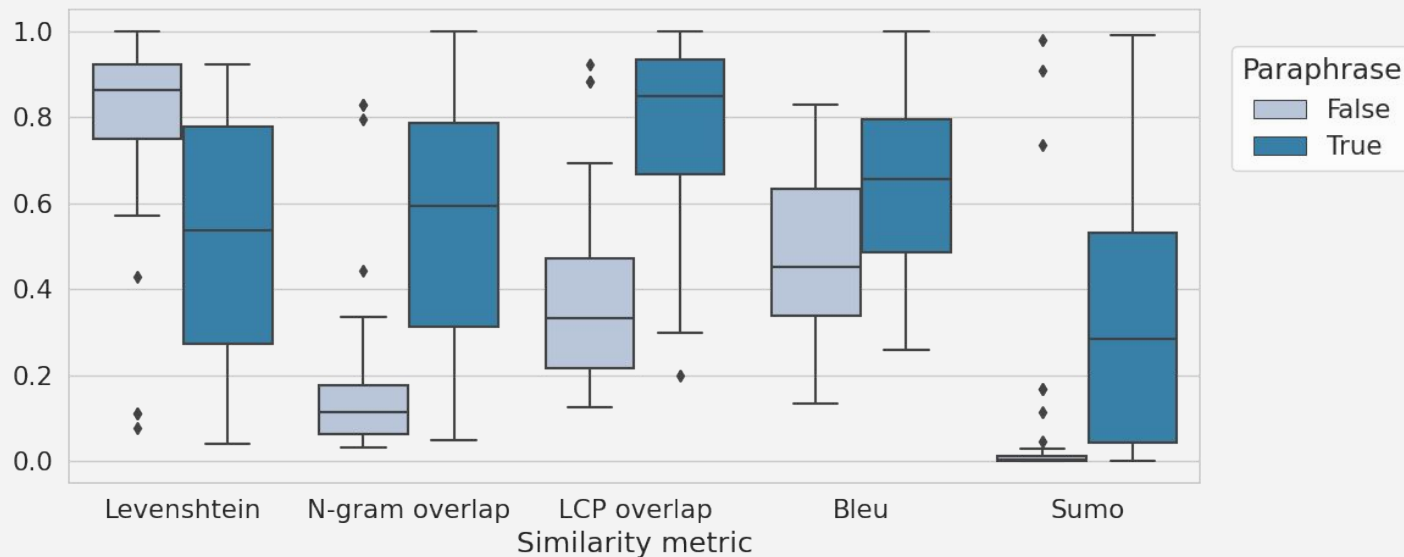
Application of the paraphrase acquisition pipeline on the Wikimedia dumps:

→ 23,944 paraphrase candidates

Manual annotation of 100 randomly drawn paraphrase candidates:

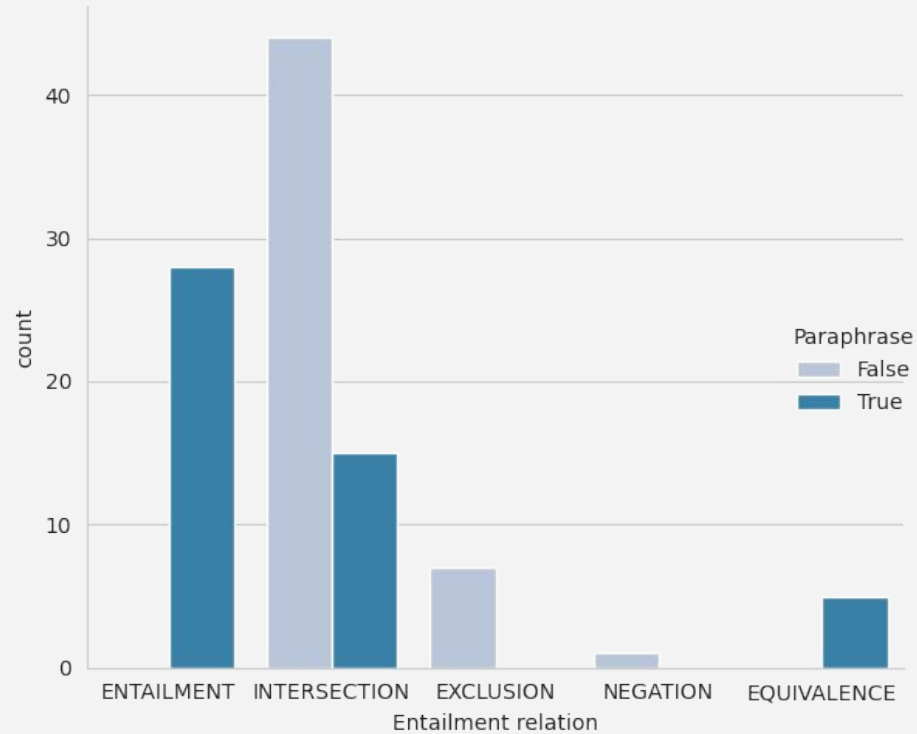
- Paraphrase judgements
- Entailment relations

Evaluation



100 candidates: 49 paraphrases / 51 non-paraphrases

Evaluation



5. Conclusion

Contributions

- Formalize paraphrase definition
- Connect paraphrasing and textual entailment
- Develop novel paraphrase acquisition approach
- Create prototype paraphrase corpora
- Show potential of image captions as paraphrases

Future Work

- Conduct experiments on larger web crawls (e.g., Common Crawl, Web Archive)
- Compare textual diversity and quality to existing paraphrase corpora
- Train classifier for paraphrase labels
- Develop paraphrase model to evaluate potential for language modeling

Thank you for listening!



I acknowledge your attention!