Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science for Digital Media

Image Captions as Paraphrases

Master's Thesis

Marcel Gohsen

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Volker Rodehorst

Submission date: February 8, 2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, February 8, 2021

.....
Marcel Gohsen

Abstract

Paraphrase resources have been exploited for various problems in the domain of natural language processing. Although there are large paraphrase datasets available, many of them lack textual diversity or differ from human-written paraphrases conceptually. Phrasal paraphrases also limit the usefulness of a dataset for language modeling applications. To overcome these shortcomings, we propose a novel paraphrase acquisition approach that has the potential to extract vast amounts of paraphrases through distant supervision. To show its effectiveness and point out distinctive characteristics of the resulting paraphrase candidates, we build two prototype corpora and evaluate its textual properties. Further, we emphasize the connection between paraphrasing and recognition of textual entailment to distinguish semantic relations that can hold between paraphrases.

Contents

Acknowledgements

First and foremost, I would like to thank Prof. Dr. Benno Stein, Prof. Dr. Matthias Hagen, and Prof. Dr. Martin Potthast for their excellent supervision and support. They were very dedicated and interested in the success of this thesis. Im grateful that they shared their advices, experiences, and insights with me to help me to grow as a researcher.

Further, I'd like to thank Magdalena Wolska for her competent support in all linguistic questions. She helped me to build the sentence detection rules and always offered her help when linguistic expertise was necessary.

Last but not least, I would like to thank Vanya Gercheva. Without her emotional support and care, I would have never been able to finish my Master's degree. She made me who I am today, and I really appreciate it.

Chapter 1

Introduction

Loosely speaking, paraphrases are texts which convey exactly the same meaning. However, linguistic literature argues that paraphrases are not limited to strict synonymous expressions [Bhagat and Hovy, 2013]. It is more common that paraphrases incorporate a similar meaning rather than an equivalent interpretation. Hirst [2003] even goes beyond this by defining "talking about the same situation in a different way" as paraphrasing. We define paraphrases here as text pairs with a significant amount of information that can be deduced from both texts.

A variety of natural language processing (NLP) applications such as textual entailment [Izadinia et al., 2015, Marelli et al., 2014], semantic similarity [Agirre et al., 2015, Li and Srikumar, 2016], machine translation [Seraj et al., 2015], and question answering [Fader et al., 2013] exploit paraphrase resources [Lan et al., 2017] in the one or other form. Similarly, the use of broadly defined paraphrases and an analysis of textual entailment relations enhances usefulness of sentence compression and text summarization [Cordeiro et al., 2007b].

Paraphrasing has gained popularity in the NLP community [Scherrer, 2020], and thus various datasets of paraphrases have been created over the years. The largest corpora have been (semi-) automatically generated and they hence differ conceptually from human-written paraphrases. In contrast, manually created or annotated paraphrases are costly to produce, which leads to small datasets that often lack negative examples. However, Zhang et al. [2019] strongly emphasize the necessity of negative examples for training paraphrasing models. Effective learning-to-paraphrase models also require at least sentence-level granularity, which limits the usefulness of the datasets with phrasal- or even word-level paraphrases. Another common issue of existing corpora is their limited textual diversity within the dataset, a fact which was shown to be crucial for language modeling [Gao et al., 2020].

1.1 Contributions

In Chapter 2 we will present related work with regard to the creation of paraphrase corpora and show their similar construction methodology, which is based on a 20 year old approach that require parallel corpora. Additionally, we show that prior works have considered media captions as paraphrases before and assumed its suitability. Finally, we point out that the literature emphasizes the connection between paraphrasing and recognizing textual entailment.

As part of this thesis, we formalize the definition of paraphrases and establish a connection between paraphrasing and textual entailment. We distinguish 5 levels of semantic relations that can hold between a pair of texts. These contributions can be found in Chapter 3.

In this thesis, we propose a novel paraphrase acquisition approach that incorporates a level of indirection, and that has the potential of creating a vast amount of paraphrase candidates through distant supervision. With this method, we are capable of creating paraphrase datasets of human-written passage-level paraphrases. The details of all the steps of the paraphrase acquisition pipeline can be found in Chapter 4.

To point out distinctive characteristics of paraphrases that can be created with the proposed paraphrasing approach, we develop two different prototype datasets from different sources and analyze their properties in Chapter 5.

In Chapter 6, we analyze the effectiveness of our rule-based sentence detection algorithm and show its promising potential. Further, we evaluate the usefulness of perceptual hashing for detecting equivalent images of transformed image duplicates. These technologies are essential building blocks of our paraphrasing approach.

Chapter 2

Related Work

In this chapter, we shed light on related literature in terms of paraphrase corpora. Additionally, we show that media captions have been considered as useful for paraphrasing. Finally, prior works are examined for the connection between textual entailment and paraphrasing.

2.1 Paraphrase Corpora

The creation of huge state-of-the-art paraphrase datasets relies on the existence of parallel corpora, up to the present day. Barzilay and McKeown [2001] first proposed the method of extracting paraphrases from a monolingual parallel corpus of aligned sentences, which originated from multiple English translations of classic literature. Dolan and Brockett [2005] used this method of paraphrase extraction by creating monolingual parallel corpora from news clusters and therewith created the Microsoft Research Paraphrase Corpus (MSRPC). Similar sentences within these clusters were aligned using word-based edit distance and a heuristic strategy, that pairs initial sentences [Dolan et al., 2004]. Additionally, monolingual machine translation generated sentence pairs from these news clusters [Quirk et al., 2004], which extended the base data for the creation of MSRPC. From this data, 5801 sentence pairs were selected and manually annotated from which 3900 were considered paraphrases. With the goal of sentence compression, Knight and Marcu [2002] automatically extracted 1067 sentence pairs from parallel corpora in the form of newspaper articles and their corresponding abstracts from which many qualify as paraphrases.

More recent research base their data acquisition on multilingual parallel corpora and machine translation to create much bigger datasets. First, Bannard and Callison-Burch [2005] used bilingual parallel corpora to develop an approach for automatic paraphrasing by searching and aligning English phrases that translate into the same phrase in a foreign pivot language. Based on

this method, PPDB 1.0 [Ganitkevitch et al., 2013], one of the biggest collection of paraphrases, was created for the English and Spanish language. The English portion of this paraphrase database consists of 7.6 million lexical (i.e., synonymous) and 68.4 million phrasal paraphrases. An extension of the PPDB [Ganitkevitch and Callison-Burch, 2014] was released one year later to add 23 additional languages to this data set. Besides PPDB, the pivot language translation technique of Bannard and Callison-Burch [2005] was applied to create Opusparcus [Creutz, 2018], a paraphrase corpus for six European languages build from translated movie and TV subtitles. The training set of Opusparcus was automatically ranked and filtered such that the English portion contained 7 million sentence and phrase pairs, which the authors claimed to be "good" or "mostly good" paraphrases. The dedicated development and test set consists of 3088 examples which were manually annotated with quality labels from which 1997 can be considered paraphrases. Another example of a dataset that follows the idea of language pivoting is TaPaCo [Scherrer, 2020]. Their multilingual source data belongs to Tatoeba[1], which is a crowdsourcing-based collection of sentences and translations. This sentence-aligned data was transformed into equivalence graphs in order to create clusters of paraphrases in 73 different languages. The English part of TaPaCo contains 158 thousand sentences that are clustered in 62 thousand paraphrase sets. An alternative to the pivot language technique in multilingual corpora is neural machine translation with bilingual parallel data. ParaNMT-50M [Wieting and Gimpel, 2018] is a collection of automatically generated paraphrases from translating Czech sentences to English in English-Czech sentence pairs from parallel corpora. With this approach, the creators of ParaNMT-50M managed to form a set of 50 million sentence pairs from which they estimated 30 million to be "strong" paraphrases.

To overcome the need of parallel corpora, paraphrase datasets can be generated with automatic or semi-automatic paraphrasing methods. Zhang et al. [2019] applied a mixture of word scrambling and back translation to sentences drawn from the Quora Question Pairs corpus [Iyer et al., 2017] and Wikipedia to create a balanced dataset of paraphrases and non-paraphrases with high lexical overlap. In total, 108463 sentence pairs were manually labeled, and additionally, a dataset of 656 thousand pairs were assigned silver labels by considering examples from word swapping as non-paraphrases and from back translation as paraphrases.

Crowdsourcing has been proven to be useful to create high quality human-written paraphrase corpora. Crowdworkers were advised to "rewrite the original text [...] so that the rewritten version has the same meaning, but a com-

[1] `https://tatoeba.org/eng/about`

pletely different wording and phrasing" to build Webis-CPC-11 [Burrows et al., 2013]. The original texts were 4096 excerpts randomly drawn from literature that should be paraphrased, which resulted in 7859 pairs of automatically classified positive and negative examples. An important characteristic of this dataset is, that this is the only passage-level paraphrase corpus to our knowledge. Xu et al. [2014] made use of crowdsourcing to select candidate sentences with the same meaning as a shown original sentence, which were extracted from Twitter's trending topics and their associated tweets. PIT-2015 is the result of this procedure and consists of 18862 sentence pairs from which 5641 are paraphrases.

The lack of parallel corpora motivated Lan et al. [2017] to build their Twitter News URL Corpus. They introduced a level of indirection by linking tweets that contained the same shared URL and use them as paraphrase candidates. They labeled 51524 sentence pairs through crowdsourcing and claimed to be the largest human-labeled paraphrase corpus to that time.

2.2 Paraphrases and Captions

In order to assess the quality of automatic image captioning models, Vinyals et al. [2015] showed high `BLEU` scores, a metric from machine translation that is often used to assess paraphrase quality, between human-written image captions in the Pascal VOC 2008 [Farhadi et al., 2010], the Flickr 30k [Young et al., 2014], and the Flickr 8k [Rashtchian et al., 2010] datasets. The high `BLEU` scores indicate suitability of image captions as paraphrases [Prakash et al., 2016]. SICK [Marelli et al., 2014] is a collection of 9840 paraphrase candidates that were extracted from the Flickr 8k and the MSR Video Paraphrase Corpus [Chen and Dolan, 2011]. The Flickr 8k dataset provides 5 descriptive captions for each of the 8000 images acquired through crowdsourcing. The MSR Video Paraphrase Corpus comprises 85 thousand English descriptions of 2000 videos which were obtained by asking crowdworkers to describe within one sentence what can be seen in a short video clip.

Moreover, MSCOCO [Lin et al., 2014], a dataset of 120 thousand images with 5 human-annotated captions each, was used to train a neural paraphrase generation model [Prakash et al., 2016]. By evaluating their model, they showed that paraphrasing can be learned effectively on image captions.

2.3 Paraphrases and Entailment

In prior works, paraphrases were vaguely defined as a pair of texts that approximately mean the same thing [Pavlick et al., 2015]. However, more semantic

entailment relationships can be expressed and has been used to characterize paraphrases.

With the goal of sentence compression, Cordeiro et al. [2007b] distinguished symmetrical and asymmetrical paraphrases, where symmetry describes a bi-directional entailment relation. Even more fine-grained semantic relations can be inferred from a pair of texts. An extension to natural logic was build in order to define and express 7 distinct entailment relations [MacCartney and Manning, 2009]. Based on these, Pavlick et al. [2015] annotated 5 adapted entailment relations in PPDB and used them for automatic classification of entailment. Furthermore, SICK also encodes three different semantic relations [Marelli et al., 2014]. These annotations had been used and extended with hypotheses to create a corpus for natural language inference [Bowman et al., 2015], which focuses on the extraction of semantic relations like entailment and contradiction.

In conclusion, Madnani and Dorr [2010] stated that there is a strong relation between paraphrasing and recognizing textual entailment. Some entailment recognition systems benefit from paraphrasing approaches while entailment recognition helps analyzing paraphrases qualitatively.

Chapter 3

Modeling and Hypothesis

In this chapter, we formalize the problem of paraphrasing and contrast this formalization to former definitions. Further, we describe the concept of textual entailment and distinguish 5 different semantic relations that can hold between a pair of texts.

3.1 Paraphrases

Related literature in the field of paraphrasing define paraphrases as a pair of divergent texts in the same language that approximately have the same meaning. We want to refine this definition and connect paraphrasing with textual entailment in order to determine semantic properties of paraphrases. This formalization is an extension to the logical definition of paraphrases introduced by Burrows et al. [2013].

Let a text t be a concept in a semantic ontology. Let α be a common domain or world knowledge in the form of relations within this ontology. A pair of texts t_1, t_2 as concepts in an ontology allow semantic deduction of individual information sets Φ_1, Φ_2.

$$(t_1 \wedge \alpha) \models \Phi_1 \quad \Leftrightarrow \quad (t_2 \wedge \alpha) \models \Phi_2 \tag{3.1}$$

The common definition of paraphrases constrain the inferred information sets to $\Phi_1 \approx \Phi_2$. However, this limits the lexical diversity of paraphrases to be word-wise permutations of each other. We weaken this constraint to allow various semantic relations between paraphrases. The necessary condition for paraphrases models that a pair of texts needs to contain shared information.

$$\Phi_1 \cap \Phi_2 \neq \emptyset \tag{3.2}$$

With this formalization it would be sufficient to infer a single fact from both texts to consider them a paraphrase. However, it may be that the non-shared

information cover distinct topics. For instance, "*President Barack Obama enjoys playing golf*" and "*President Barack Obama traveled to Germany*" both entail the fact that Barack Obama is a president. Although it satisfies the necessary condition, it is not a paraphrase. Hence, we state that the amount of information that can be inferred from both texts needs to be a significant portion related to all information entailed in these texts.

3.2 Entailment relations

Textual entailment describes the relation whether a hypothesis can be inferred from a text. In case of paraphrases, a pair of texts is given and we search for hypotheses which can be inferred from both. Based on the deduced information, we define 5 distinct levels of entailment, which are adaptations of the basic semantic relations presented by MacCartney and Manning [2009]. We follow the syntax of description logic to formalize entailment relations since we consider texts to be concepts in an ontology.

3.2.1 Equivalence

Two texts are semantically equivalent if their deduced information sets are identical. This can be interpreted as bidirectional entailment.

$$t_1 \equiv t_2 \quad :\Leftrightarrow \quad \Phi_1 = \Phi_2 \tag{3.3}$$

In the above equation $:\Leftrightarrow$ is defined as logically equivalent and it hence is interpreted as "t_1 is semantically equivalent to t_2 which is logically equivalent to Φ_1 equals Φ_2". As stated, semantic equivalence is usually considered the only relation that holds between two texts to be a paraphrase. However, paraphrases that satisfy semantic equivalence can only be permutations or synonym replacements of a source sentence with significant text reuse. "*Barack Obama is the president of the United States*" paraphrases "*The president of the U.S. is Barack Obama*" which are semantically identical. Note that changing the first sentence to "*Barack Obama is the president*" would make it more general because we can't deduce which country Barack Obama is president of, and thus it would not be semantically equivalent to the second sentence anymore. However, this modified example still would qualify as a paraphrase. All examples that fulfill semantic equivalence are in fact paraphrases.

3.2.2 Forward/Reverse Entailment

Forward and reverse entailment are unidirectional entailment relations where all information from one text can be inferred from the other, but one sentence

is more general or specific than the other.

$$t_1 \sqsubset t_2 \quad :\Leftrightarrow \quad \Phi_1 \supset \Phi_2 \tag{3.4}$$

$$t_1 \sqsupset t_2 \quad :\Leftrightarrow \quad \Phi_1 \subset \Phi_2 \tag{3.5}$$

Since entailment is a directional relation and there is no ordering of texts in a paraphrase, we merged forward and reverse entailment into a single class. In the literature, paraphrases that have this entailment property are denoted as "*asymmetric paraphrases*" [Cordeiro et al., 2007b]. For instance, let t_1="*The Space Shuttle Endeavour docked with the International Space Station*" and t_2="*Space Shuttle Endeavour docked to the ISS on STS-134*". In this example t_1 is entailed by t_2 ($t_1 \sqsupset t_2$), since all information that can be inferred from t_1 can be inferred from t_2, but we also know from t_2 the additional fact that during the mission STS-124 Edeavour docked to the ISS. This mean that Φ_1 is a subset of Φ_2. The shown example would classify as a paraphrase, but others which fulfill the entailment relation might be debatable, and their classification depends on the proportion of shared versus unshared information.

3.2.3 Intersection

Semantic intersection of a text pair models the fact that there is information shared between the two but both contain additional facts that are not entailed in the other one.

$$t_1 \sqcap t_2 \quad :\Leftrightarrow \quad \Phi_1 \cap \Phi_2 \neq \emptyset \wedge (\Phi_1 \not\subseteq \Phi_2) \wedge (\Phi_2 \not\subseteq \Phi_1) \tag{3.6}$$

An example of semantically intersecting paraphrases is the pair "*Thein Sein meets US President Barack Obama*" and "*President Obama meets with the President of Myanmar, Thein Sein*". From the first sentence we can infer that Obamas first name is Barack and he is the president of the United States, a fact which we can not deduce from the second sentence. From the second sentence however, we know that Thein Sein is the president of Myanmar, which is not entailed in the first sentence. From both texts we know that Thein Sein meets president Obama. Although this example represents a paraphrase, semantically intersecting text pairs may not necessarily be paraphrases, which depends on the portion of shared information.

3.2.4 Exclusion

A pair of texts that mutually exclude their semantics will be assigned with the semantic relation of exclusion. To fulfill this relation, it is required that there is not a single fact that can be derived from both texts.

$$t_1 \mid t_2 \quad :\Leftrightarrow \quad \Phi_1 \cap \Phi_2 = \emptyset \tag{3.7}$$

Semantically excluding text pairs appear completely unrelated to each other. "*The repeating decimal continues infinitely*" and "*This is identical to one*" do not share information and therefore do not qualify as a paraphrase. Since it violates the necessary condition of paraphrases (Equation 3.2), text pairs with this property can never be paraphrases.

3.2.5 Negation

With the semantic relation of negation we encode that there is at least one fact in a text that contradicts a fact in another text. Let $\varphi_1 \in \Phi_1$ and $\varphi_2 \in \Phi_2$.

$$t_1 \leftrightarrow \neg t_2 \quad :\Leftrightarrow \quad \exists \varphi_1, \varphi_2 : \varphi_1 \leftrightarrow \neg \varphi_2 \tag{3.8}$$

For instance, "*Brad Keselowski finished 10 points behind Martin Truex Jr. in fourth place*" and "*Brad Keselowski won the race*" contradict each other. It might be that two different races are referred to in these texts but, as it is not further specified, it is a semantic contradiction. Texts whose semantic relation is expressed through negation defeat the main idea of paraphrases and are therefore not classified as such.

Chapter 4

Paraphrase Pipeline

The key concept of our paraphrase acquisition pipeline is that different captions of the same image potentially paraphrase each other. Hence, we link equivalent images in web crawls and consider their divergent captions as paraphrase candidates. In this chapter, we will give a detailed explanation of every step in the acquisition pipeline. An overview of these steps can be seen in Figure 4.1.

4.1 Web Resources

Our paraphrase acquisition approach can be flexibly applied to nearly all kinds of web crawls. The only requirement is that a dataset is significantly large because of the necessity of "aggressive" filter heuristics. Therefore, we base our approach on some of the largest web crawls available whose image-related statistics can be observed in Table 4.1. The sum of all occurrences of image references in a web page (e.g., through an `img` tag in HTML) will be called

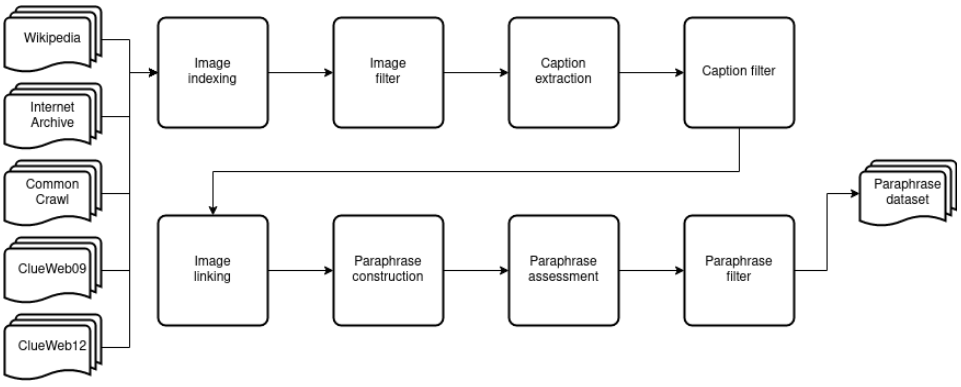


Figure 4.1: Overview of the paraphrase acquisition pipeline.

number of image URIs. Labeled as unique URIs, we count how many different image URIs are encountered either as image reference or as a header for a physical image payload within a WARC file. WARC denotes the "Web ARChive" format that allows to store all kinds of digital resources together with useful metadata which is the corresponding file format for all the considered web crawls despite the Wikimedia dumps. The number of images corresponds to "physical" images in the web crawl, i.e., images for which the data is present.

Crawl	#URIs	#Unique URIs	#Images
ClueWeb09[1]	21,808,796,671	1,983,184,678	4,052
ClueWeb12[2]	16,938,833,429	1,000,751,805	1,948
Common Crawl[3]	67,615,864,465	8,419,862,945	4,302,467
Wikimedia dump[4]	5,745,684	3,133,150	0

Table 4.1: Image statistics of web crawls.

ClueWeb09 [Callan et al., 2009] and ClueWeb12 [Callan et al., 2013] are sufficiently large web crawls, which are designed for research in the field of information retrieval and comprise 9TB and 6TB respectively of compressed data. ClueWeb09 consists of more than 1 billion web pages in 10 different languages. ClueWeb12 contains only English web pages, which sum up to approximately 700 million. We found that on average a web page in ClueWeb09 contains 21 images and, that a page in ClueWeb12 contains 23 images. These number appear to be unreasonably large at first sight. However, considering that the HTML image tag is the 5th most-frequently used tag in the web in 2020 and the 3rd most-frequently used in 2005[5], it becomes plausible again. Both datasets did not intentionally contain physical images. However, we still find image instances from which almost none qualifies for the paraphrase acquisition. One type of data that we found were server side includes with a wrong content type declaration. Others are indeed images from which some show strings of numbers and letters which presumably are part of captchas. Another type of images within these crawls are monochrome and very small images which may be used design elements in the web page. Thus, these physical images can be neglected in the paraphrase acquisition pipeline, but the web pages from these crawls are still useful resources. Of the approximately 22 and 17 billion image URIs, 8 and 7 billion have a non-empty caption in form of

[1] `https://lemurproject.org/clueweb09/`
[2] `https://lemurproject.org/clueweb12/`
[3] `https://commoncrawl.org/`
[4] `https://dumps.wikimedia.org/enwiki/20201101/`
[5] `https://almanac.httparchive.org/en/2020/`

an alt-text in ClueWeb09 and ClueWeb12 respectively. These are ratios of 36% and 41%, which indicates their high potential for our paraphrase acquisition approach. Another beneficial fact is, that the ratio of URIs to unique URIs is sufficiently high, which indicates frequent image reuse. This means that on average an image was reused almost 17 times in the ClueWeb09 and close to 11 times in ClueWeb12. The more image reuse is present in a web crawl, the more paraphrase candidates can be generated.

The Common Crawl comprises a collection of monthly crawls, that sum up to many petabytes of data gathered over 12 years. The portion of this large collection that we used is the 2020-16 crawl, which we use because this is the newest release we have downloaded so far and verified with MD5 checksum. This part of the Common Crawl contains 2.8 billion web pages and media files. In contrast to the ClueWebs, the crawl methodology of this dataset included to crawl all referenced media from the web pages, too. Thus, the compressed data size accumulates to approximately 69TB. From Table 4.1 we can observe, that the amount of image URIs is staggering. On average, a web page from the Common Crawl contains more than 23 images. Since the creator of the Common Crawl provide statistics about their crawls we can perform a sanity check if our computation is trustworthy. The number of images we calculated is within an error margin of 0.8%. Thus, we consider these numbers to be reliable. A strong selling point of the Common Crawl is, that more than 50% of all image tags have a non-empty alt tag. However, the ratio of image URIs to unique URIs is lower compared to both ClueWebs with an image being referenced around 8 times on average.

The Web Archive[6] is the largest collection of web pages of all the considered crawls. Currently, more than 525 billion web pages has been stored over time. A representative portion of this collection has been crawled, which accumulates to a size of more than 500TB of compressed data [Völske et al., 2020]. Due to technical difficulties while handling this amount of data, we are not able to present the same statistics as for the other crawls. However, this crawl contains billions of images and is therefore the resource, that may allows us to acquire the most paraphrases from.

Wikipedia[7] is an excellent source for human-written texts of good quality. Thus, we also considered the Wikimedia dumps as source for paraphrases. The Wikimedia dump without page edit history from the 1st November 2020 comprises more than 20 million pages. Around 28% of the pages contain images, and 59% of all images have an extractable caption. This is the highest image-to-caption ratio of all the considered web crawls, which encouraged us to use it with the paraphrase acquisition pipeline. However, this dump does not

[6] `https://web.archive.org/`

[7] `https://en.wikipedia.org/`

contain physical images, which limits us in terms of equality determination. Fortunately, Wikipedia embeds images from a common image database such that equal images can be found by their database URI.

4.2 Image Indexing

To reduce the amount of data that needs to be maintained within the pipeline, we index all images and image references that we encounter in the crawls. These are grouped by their associated image URLs. With this grouping scheme, we know for each image how often and where it is referenced. Furthermore, we can derive meta-information from an image and associate it with all corresponding image references.

4.3 Image Filter

A requirement for an image to be suitable for our paraphrase acquisition approach is the existence of an "explanatory caption". In order to reduce the image search space, the goal of the image selection step is to exclude not promising candidates such as icons, symbols, and design elements of a web page. These images typically do not have an explanatory caption and will be discarded in a first-pass filtering step. If an image is discarded, all corresponding image references will be discarded too.

4.3.1 Image Annotation

To find reasonable filter criteria to distinguish between potentially useful and less useful images, we annotate a set of images which we sample from the Common Crawl. Our sampling procedure randomly selects images balanced between the following size categories.

- `Icon` $(0 \times 0,\ 32 \times 32]$
- `Small` $(32 \times 32,\ 256 \times 256]$
- `Medium` $(256 \times 256,\ 2048 \times 2048]$
- `Large` $(2048 \times 2048,\ 4096 \times 4096]$
- `Giant` $(4096 \times 4096, \infty]$

Target size of the sampling procedure are 5000 images. However, balancing between these size categories are not possible because there are not enough

Property	Label	Mean	Min	Max	Std
Image size	Likely	91,109	3,000	1,000,000	118,992
Image size	Unlikely	141,799	1	19,738,400	617,840
File size	Likely	5,983	579	11,664	2,694
File size	Unlikely	3,432	52	12,082	3,027
Aspect ratio	Likely	1.81	1.00	8.44	0.98
Aspect ratio	Unlikely	2.72	1.00	480.00	14.72
Transparent pixels	Likely	0	0	24	2
Transparent pixels	Unlikely	20,297	0	19,738,400	468,885

Table 4.2: Characteristics of the annotated images by label. Likely translates to "*likely to have a caption*" and unlikely analogously.

images that are large enough to fall in `Large` or `Giant` in the Common Crawl. Thus, a total of 3010 images were finally annotated from which 1000 are within the boundaries of each `Icon`, `Small`, and `Medium`, and 10 in the category `Large`.

Our annotation guidelines are as follows. We asked an expert annotator to decide if he would expect an image to have an explanatory caption by just looking at the image. Although this task depends on the impression of the annotator, we can be sure that symbols, logos or similar images will not be selected, and hence suitable criteria for the exclusion of these can be found.

In favor of a fast annotation procedure, we developed an image annotation tool which allows the selection or deselection of images through a graphical user interface. The tool is implemented in Java based on Java Swing and provides us with an excellent annotation efficiency.

From the annotated images, we tried to find characteristics that distinguishes images that may have an explanatory caption and these that presumably have none. From the 3010 images, 210 are labeled as likely and 2800 as unlikely to have a caption. The observed image characteristics were image size (in pixels), file size (in bytes), aspect ratio, and the number of transparent pixels. The aspect ratio is computed as the ratio between the longer and the shorter dimension of an image.

Table 4.2 presents the computed characteristics grouped by annotation label. In terms of image size, images that are likely to have a caption tend to be in a specific range while images that most-likely don't have a caption are much bigger or smaller than that. There are even images with a resolution of 1x1. Considering the file size in bytes, no clear trend is visible. Although the lower bound of images, that most-likely do not have a caption, is lower, it is

hard to define a specific range to discard these. With respect to the aspect ratio, we can easily define an upper bound to eliminate unwanted images.

4.3.2 Filter heuristics

From the found results, we can derive the following filter heuristics. An image needs to be at least 3000 and at max 1,000,000 pixels large. We also lower bound the file size of an image to 3000 bytes. Moreover, the aspect ratio should be at least 1 and at max 3. Further, we do not allow images to have any transparent pixels.

A property that clearly divides logos and icons from other images in the web is their number of references. Logos or icons like e.g., arrows or magnifying glasses are frequently reused within a domain and appear on almost every web page. Images that occur more than 10 times in a crawl are discarded.

4.4 Caption Extraction

The method of extracting image captions from a web page depends on the markup language a page is written in. While every web page comes down to the Hypertext Markup Language (HTML), MediaWiki built Wikitext[8] on top of HTML to ease unified formatting of Wikipedia articles written by multiple authors.

4.4.1 HTML

A creator of a web page has numerous options to specify a caption for an image in HTML. Unfortunately, there is no standardized way for the specification. Thus, we need to find suitable candidates for caption approximations that can be extracted from an HTML page.

The purpose of the `alt` attribute of an image tag in HTML is to present an explanatory text to a reader of a web page, if the associated image can't be displayed properly. In theory, these texts are meant to describe the content of an image and hence can be considered as image caption. Since the content of alt-texts are not controlled or restricted by some mechanism, selecting appropriate captions require filter criteria that exclude search engine optimization terms, Twitter hash-tags [Sharma et al., 2018], etc.. To give an alt-text to an image is mandatory, but does not prevent people from using empty strings and therefore needs to influence the decision for appropriate filter criteria.

[8] `https://en.wikipedia.org/wiki/Help:Wikitext`

HTML5 provides the `figure` tag to include images as floating blocks within a web page and allows to specify an associated caption with the `figcaption` tag. Although this is a convenient method to define captions, HTML5 features are often not supported by old browser versions. Therefore, other tags to annotate captions to an image are often preferred by web content authors.

Another common practice for image caption realizations is to define a single-column two-cell table. The image is stored in the top cell while the caption is written in the bottom one. In contrast to the other caption extraction methods, extracting texts in a cell below an image can yield texts that are not intended to be captions and might not have a strong relation to an image.

The most difficult task to find a caption for an arbitrary image is to extract it from the surrounding text. Achieving high precision in selecting explanatory paragraphs from the context requires a semantic analysis between the image content and the caption candidates but would most-likely produce high-quality passages.

4.4.2 Wikitext

Wikitext provides its own syntax for embedding images in Wikipedia articles. This syntax includes an optional specification of a displayed caption as well as an alternative text that will be shown if an image can't be displayed. Thus, extracting captions from Wikitext is easy and reliable, if a caption exists at all.

4.5 Caption Filter

The textual quality of image captions depends on the source from which a caption is collected as well as on the chosen extraction strategy. Hence, different filter heuristics have to be developed to ensure that only "qualitatively satisfying" captions remain for the paraphrase construction.

4.5.1 Preprocessing

The preprocessing step is applied to a caption before it is passed to the actual filtering step. A preprocessed caption replaces its original version and persists through all further steps in the pipeline.

Within the first preprocessing step, we remove all line breaks, which increases legibility and lowers the chance of incomprehensible tokenization. As a next step, non-printable unicode characters (e.g., `\x00`) are deleted. Finally, we normalize whitespaces such that multiple spaces are reduced to one. Legibility also benefits from this step and prevents empty tokens during tokenization.

After these normalizations the texts are enriched with part-of-speech (POS) tags and word tokenization annotations. Both information are obtained with the Stanford CoreNLP Toolkit [Manning et al., 2014].

4.5.2 Filter Heuristics

A simple but still important filter criterion is the caption length. Empty captions can occur due to missing restrictions of the `alt` attribute in HTML or as a result of the preprocessing pipeline and are excluded from further processing. With the goal of passage-level paraphrases, longer captions are preferred, and a caption should contain at least 10 words.

The amount of pornographic or offensive content in the web is not neglectable. Some crawls are already cleaned from such content, but the Web Archive crawl, for instance, is not. There are many detectors for sexual content in images but just few sophisticated methods for texts. Swearword filters mainly work with comprehensive word blacklists, and thus we follow a similar approach. We use a list of more than 1300 offensive or profane words [von Ahn, 2019] and exclude those captions that include one of them.

Since we are interested in English paraphrases in this work, we perform language detection on captions from multilingual web crawls with Lingua[9]. Lingua distinguishes 74 languages and claims to be the most accurate language detection application that outperforms even Apache Tika[10] and OpenNLP[11]. Only captions whose language has been identified as English remain in the pipeline.

Sentence Detection

A common practice for writing image captions is to not produce grammatically correct sentences but rather sentence fragments. We opt for producing high quality passage-level paraphrases, and thus we try to exclude sentence fragments from the pipeline. Since fragment detection is a complex research problem, there are no simple heuristics that can be applied to distinguish sentences from fragments. Moreover, performance is a concern when we process multiple terabytes of data. Hence, we aim at rule-based filter heuristics based on parts-of-speech to decide wether a text is a fragment or a proper sentence. To find such rules, we split multi-sentence captions into sentence candidates and randomly sample 500 candidates for training and 100 for test from the Wikimedia dump. These train and test datasets of sentence candidates are

[9] `https://github.com/pemistahl/lingua`
[10] `https://tika.apache.org/`
[11] `https://opennlp.apache.org/`

Rule	Premise	Pattern
1	.* MD .*	.* MD RB? VB .*
2	.* (WDT|WP|WRB) .*	[¬(WDT|WP|WRB)]* (VBP|VBZ|VBD) .*
3	.* IN .*	[¬IN]* (VBP|VBZ|VBD) .*
4	⊥	.* (VBP|VBZ|VBD) .*

Table 4.3: POS patterns for sentence classification.

manually annotated with labels for sentences and sentence fragments. From the annotated training set, an expert linguist extracted POS tag patterns that separates grammatical sentences from fragments. Table 4.3 presents the constructed POS tag sequences with tags from the Penn Treebank [Taylor et al., 2003]. The rule numbers indicate the order of application. If the premise of a pattern is true and the pattern can be found in the POS sequence of a text, this text will be considered a sentence and none of the following rules have to be applied.

Rule 1 handles cases whenever a modal verb is contained in a sentence candidate. If a modal verb exists it needs to be directly succeeded by an optional particle and a verb in base form. An example sentence that would be accepted by this rule is "*The ultimate distribution can't be shown in this diagram*". Rule 2 and Rule 3 deal with sentences with subordinate clauses at the end and require that an inflected verb precedes a wh-word or a subordinate conjunction which separate the main clause. "*The responsibility is with whoever is taking care of the children*" is a sentence that would be classified as such by Rule 2. Rule 4 is taking responsibility for all other cases and models that a sentence has to comprise an inflected verb. For instance, "*Eventually the harbour became silted up, and the city lost its natural resources*" neither contains a subordinate clause nor a modal verb and thus is classified as a sentence due to its inflected verb.

4.6 Image Equivalence

Our proposed paraphrase acquisition approach links equal images in the web and considers their captions as paraphrases. Equivalence between images can be determined in multiple ways, which leads to different candidate sets.

4.6.1 Resource Identity

The most basic equality class clusters image references that refer to the exact same resource in the web. Each web resource is identified by a unique identi-

fier, the so called Uniform Resource Identifier (URI). Because of its uniqueness, every image inclusion that refers to the same URI embeds the physically identical resource and consequently are visually equal. A benefit of identifying equal images through their URI is that no actual pixel data is required to determine equivalence. The method is especially useful when working with web crawls that do not include (enough) images. Moreover, the probability of false-positive equivalences is close to zero. However, we may miss visually identical images which represent different resources by having divergent URIs.

4.6.2 Pixel Equality

With pixel equality, we cluster images that have the exact same color at every pixel. In this regard, it is convenient to use a hash function such as MD5 to generate an image fingerprint. MD5 is a 128-bit hash function that is frequently used as checksum for data integrity and thus perfectly suitable for this purpose. However, in image plagiarism, exact copies of an image are rare, and near duplicates, which are transformed versions of the original (e.g., compressed, scaled, watermarked) are much more common [Meuschke et al., 2018]. MD5 hashes of near duplicates will drastically change with respect to the original hash due to MD5's avalanche effect. This effect describes the property that small changes in the input result in significant changes in the output.

4.6.3 Perceptual Hashing

In order to be able to classify near duplicates as equivalent, we resort to transformational invariant image hash functions. Perceptual image hashes use visual features to compute image fingerprints that are robust against specific affine transformations. The following perceptual hash functions have the potential to detect near duplicate images; they will be investigated in Section 6.2.

1. **Average Luminosity Hash**
   An image is converted to the YCbCr color model and the average luminosity is computed on the Y values. This average luminosity is used as a threshold such that a hash bit is set to one if a corresponding pixel is above the threshold and to zero if it is below or equal.

2. **Perceptive Hash**
   A two-dimensional discrete cosine transformation (DCT) is applied to an image and the hash is computed analogously to the average luminosity hash by using the mean of the DCT coefficients as a threshold. The

discrete cosine transformation transforms an image in the frequency domain with lower frequencies in the upper and higher frequencies in the lower triangular matrix. The first element of a DCT matrix represents the hue of a whole pixel block and hence is an outlier with respect to the other coefficients. Moreover, a change in high frequencies in the DCT coefficients is not noticeable to the human eye. For these reasons, the computation of the threshold is done on a subset of the DCT matrix excluding very high and low frequencies.

3. **Rotational Invariant Perceptive Hash**
   The rotational invariant perceptive hash partitions the image into rings by rotating pixels around the center and grouping the luminance values of these in so called buckets. Each bucket is sorted to eliminate the influence of the order of rotation. A one-dimensional discrete cosine transformation is applied to each bucket and the hash is created analogously to the perceptive hash by comparing the DCT coefficients to the mean of each bucket.

4. **Rotational Invariant Average Hash**
   Similar to the rotational invariant perceptive hash, buckets of luminance values of ring partitions are created for the computation of the rotational invariant average hash. However, no cosine transformation is computed. Instead, average luminosities will be calculated and compared between each bucket to decide whether a one or a zero will be set in the resulting hash.

The average hash has been chosen due to its efficiency. The motivation for choosing the perceptive hash functions is its discrete cosine transformation. Since JPEG compression utilizes DCT coefficients, we expect that these hashes have the potential to be robust against this transformation. Based on their definition, we don't expect that the average hash or the perceptive hash yield the same hash for rotated images. This is why we include the rotational invariant hash functions as well.

The choice of a reasonable hash length has a significant influence on the properties of the resulting hash. If the number of bits for a perceptual hash is too low, many false-positive equivalences will be the result. If the number of bits is chosen too high, many equivalent images will be missed. These perceptual hash functions presumably handle varying bit-resolutions differently well. Nevertheless, we base our decision for a hash length on the number of present images in the web crawls. We assume that in the Web Archive are more than 1 billion images. In a worst-case scenario all images are different and this fact needs to be expressible through the hashes. Thus, we compute the hash length

l as follows.

$$l = \text{ceil}(\log_2(1,000,000,000)) = 29\text{bit} \tag{4.1}$$

To add extra headroom to the number of distinguishable hashes, we choose $l = 30$ which allows distinction of twice as much than the computed 29 bit.

Perceptual hashes have the benefit of not following the avalanche effect. Similar input images lead to close output hashes which allow us to reason about image similarity. One way of approximating image similarity from hash proximity is to compute the Hamming distance. The Hamming distance counts the number of differing characters in equally long strings.

4.6.4 Semantic Similarity

The challenge of finding images which show semantically similar sceneries is a step beyond visual equivalence. For example, all images that show cats would be assessed as semantically similar and thus can be considered an equivalence class. Semantic similarity assessment is a hard problem to tackle and a solution is not within the scope of this work. However, interesting research questions appear in this regard: If a pair of semantically similar images both have an explanatory caption, can this caption pair still be a paraphrase, and what are reasonable boundaries for the similarity of two images to obtain meaningful paraphrases?

4.7 Paraphrase Construction

As a result of equivalence linking, we obtain pairs of equivalent images which are identified by a unique image URI. Due to the indexing step (see Section 4.2), we are able to retrieve two sets of image references along with their extracted captions based on these URIs of those two images. The Cartesian product between the two sets of captions from all image references of the equivalent image pair yields the set of paraphrase candidates.

4.8 Paraphrase Filter

Main purpose of the paraphrase filtering step is to eliminate equivalent text pairs and duplicate paraphrases. Dependent on the source of images and the strategy of equivalence determination, it is certain that there is a significant amount of equivalent captions. Text pairs which only differ in punctuation or case are also unwanted since they do not offer any structural diversity. Thus, we discard these identical or near equivalent paraphrase candidates.

Due to the fact that all paraphrase candidates originate from captions of equivalent images, a semantic relation between an extracted text pair should exist to some extend. To train effective paraphrasing models, negative examples that are hard to distinguish from proper paraphrases are crucial [Zhang et al., 2019]. Thus, we do not filter paraphrase candidates based on semantic relationships but rather keep them in the pipeline and assess their textual similarities.

4.9 Paraphrase Similarity Assessment

Many different metrics have been proposed to determine textual similarity of paraphrases. Although semantic similarity is desirable to assess the quality of a paraphrase, it is a challenge to measure it. Thus, the majority of metrics use syntactical features (e.g., string overlap) to reason about semantics. The following metrics are commonly used in textual similarity measures by the paraphrasing community.

4.9.1 Levenshtein Distance

The Levenshtein distance [Levenshtein, 1966] is a metric that accumulates the minimum costs to transform a string into another by means of insertions, deletions, and substitutions of characters. The lower these costs are, the more similar are the compared strings. Originally, each operation has an assigned cost of 1. This method can easily be adapted to compute the cost upon word level and has already been used to measure paraphrase similarity [Burrows et al., 2013, Cordeiro et al., 2007a, Dolan et al., 2004]. A small Levenshtein distance does not necessarily indicate a paraphrase. Two sentences can have mutually contradicting interpretations while maintaining a reasonably low dissimilarity. "*This statement is true*" and "*This statement is false*" have a normalized word-based Levenshtein distance of 0.25 and are obviously not paraphrases. A pair of texts with low distance typically is of bad paraphrase quality since it maintains high text reuse and offers less structural diversity.

4.9.2 Word n-Gram Overlap

An n-gram is a consecutive sequence of n words within a text. To measure textual similarity between two texts we can compute the average number of shared n-grams between them. Let a_n and b_n be sequences of all possible n-grams of two texts and let N be an upper bound for the length of the considered n-grams. Then we can calculate the textual similarity as follows.

$$sim_o(a_n, b_n, N) = \frac{1}{N} \sum_{n=1}^{N} \frac{|a_n \cap b_n|}{\min(|a_n|, |b_n|)} \tag{4.2}$$

N is typically chosen to be 4 without justification [Barzilay and Lee, 2003, Cordeiro et al., 2007b, El-Alfy et al., 2015]. Most-likely $N = 4$ has been chosen to allow computation of this metric for paraphrases whose shorter text is at least 4 words long. Thus, N is a lower bound for the number of words in the shorter sentence of a paraphrase. As $N = 4$ is used in the majority of related literature, we adopt this choice to produce comparable results.

4.9.3 LCP n-Gram Overlap

In many NLP applications, longer matches between two strings are more significant than shorter ones and thus should have a greater impact on the textual similarity. With this in mind, Cordeiro et al. [2007a] proposed the n-gram overlap metric that is based on the Longest Common Prefix (LCP) paradigm [Yamamoto and Church, 2001]. To compute this metric, the longest common substrings have to be determined, and only the suffixes of these are taken into account. For example, if the only longest substring that is shared between two texts is a 4-gram, then its trailing 3-gram, 2-gram, and 1-gram are considered the only matching n-grams. With this strategy, redundant counts of n-gram overlaps are not considered, which is different compared to the simple n-gram overlap metric. Hence, it is also called exclusive n-gram overlap.

The normalization of this metric causes a specific difficulty as the maximum number of matching n-grams depends on the count of (n+1)-gram matches. This is why Cordeiro et al. [2007a] proposed to compute the LCP n-gram overlap in the following way.

$$sim_{exo}(a_n, b_n, N) = \max_{n \in 1,\ldots,N} \frac{|a_n \cap^{LCP} b_n|}{\min(|a_n|, |b_n|)} \tag{4.3}$$

The operator $\cap^{LCP}$ represents the n-gram matching strategy in the LCP sense.

4.9.4 The BLEU Metric

The original purpose of the BLEU metric was to automatically evaluate machine translation systems [Papineni et al., 2002]. Meanwhile it has been adapted for evaluation of many different text-to-text generation applications such as summarization [Graham, 2015], text simplification [Narayan and Gardent, 2014, Štajner et al., 2015, Xu et al., 2016], and grammatical error correction [Park and Levy, 2011]. Cordeiro et al. [2007a] were among the first

who proposed to adapt this metric for paraphrase assessment. Like the above described metrics, BLEU is also based on n-gram overlap and is defined in the following way.

$$bleu(a_n, b_n, N) = \exp\left(\sum_{n=1}^{N} w_n \cdot \log \frac{|a_n \cap b_n|}{\min\left(|a_n|, |b_n|\right)}\right) \tag{4.4}$$

The n-gram matching strategy in this metric can either be chosen to be exclusive (as in the LCP overlap) or non-exclusive (as in simple n-gram overlap). The parameter w_n allows to assign weights to different n-gram lengths. As El-Alfy et al. [2015] pointed out, if every n-gram length is equally weighted with $w_n = \frac{1}{N}$, the BLEU score equation can be re-written as follows:

$$bleu(a_n, b_n, N) = \sqrt[N]{\prod_{n=1}^{N} \frac{|a_n \cap b_n|}{\min(|a_n|, |b_n|)}} \tag{4.5}$$

Considering equally weighted n-gram lengths, Equation 4.5 computes the geometric mean of the n-gram match precision.

4.9.5 The Sumo Metric

All of the above metrics will favor text pairs that are exact or quasi-exact matches. However, these kind of paraphrases are not desirable. Rather, we aim at paraphrases that have a high degree of lexical reordering or different syntactic structures. The Sumo metric was introduced to automatically detect such texts pairs and is based on 1-gram exclusive overlap [Cordeiro et al., 2007a]. This means that if a 1-gram co-occurs in two sentences, a link between them will be established, and each word can only linked once. The number of links between two texts will be denoted as λ. First, the following function needs to be evaluated.

$$S(a_1, b_1, \lambda) = \alpha \log_2\left(\frac{\max\left(|a_1|, |b_1|\right)}{\lambda}\right) + \beta \log_2\left(\frac{\min\left(|a_1|, |b_1|\right)}{\lambda}\right) \tag{4.6}$$

α and β are weights that can be chosen according to the properties $\alpha, \beta \in [0, 1]$ and $\alpha + \beta = 1$. The Sumo metric computes as follows:

$$sumo(a_1, b_1, \lambda) = \begin{cases} S(a_1, b_1, \lambda) & S(a_1, b_1, \lambda) < 1 \\ e^{-k \cdot S(a_1, b_1, \lambda)} & \text{otherwise} \end{cases} \tag{4.7}$$

In the first branch of this equation, the penalizing factor for exact matches is the use of the binary logarithm. The purpose of the second branch is to ensure

an upper bounded of 1. In this branch the penalizing component is the use of a negative exponent of the exponential function. The penalizing potential can be scaled by the constant tuning parameter k. Cordeiro et al. [2007a] chose $k = 3$ for their experiments, and hence we adopt this value for the paraphrase similarity assessment.

4.10 Implementation

A challenge of implementing our paraphrase acquisition pipeline is to handle dozens of terabytes of data with regard to memory and runtime. Processing this amount of data in a sequential fashion would either exceed memory limits or acceptable processing times. Our implementation is based on the MapReduce programming model [Dean and Ghemawat, 2008] which allows parallel and distributed execution on a cluster. We use the open-source framework Apache Hadoop[12] which provides a MapReduce implementation and a distributed file system (HDFS), entirely written in Java. Although Apache Hadoop offers APIs for multiple programming languages, we also use Java for implementing the pipeline.

[12] `https://hadoop.apache.org/`

Chapter 5

Prototype Datasets

This chapter presents the creation and analysis of prototypal paraphrase datasets that were created with the proposed paraphrase acquisition approach on the Conceptual Captions dataset and the Wikimedia dumps.

5.1 Conceptual Captions Dataset

Extracting image captions from HTML is, as stated in Section 4.4, not a trivial tasks and can result in varying levels of textual quality. To skip this step in the paraphrase acquisition pipeline, we used an existing dataset of image caption pairs to analyze properties of the resulting paraphrase candidates. For this purpose, we use the Conceptual Captions dataset [Sharma et al., 2018], which comprises more than 3 million image-caption pairs. Because of its size and its similar caption acquisition methodology, we preferred this dataset over MS-COCO [Lin et al., 2014] among others. The captions from the Conceptual Captions dataset are extracted alt-texts of image tags in HTML web pages. These alt-texts are filtered, cleaned and hypernymed (i.e., named entities substitutions with hypernyms).

5.1.1 Image Acquisition

A drawback of the Conceptual Captions dataset is that the images are only present in form of their URIs. The image URIs in this dataset are unique, and thus we need access to the physical image data in order to compute pixel-based image equivalences to determine if this dataset contains duplicates. Hence, we downloaded images can be identified by their corresponding URI. Since this dataset was created in 2018, many of the images were not accessible anymore. 1.8 million images from the 3.3 million image URIs are still available and were downloaded for our experiment. As we plan to release these images as addition

to the Conceptual Captions dataset in the future, we will try to find the images are not available anymore in the Web Archive to extend our collection. This is not in the scope of this thesis though.

Among the downloaded images, there are artifacts such as corrupted or placeholder images introduced when the original is not available anymore. We reduced crawl artifacts by applying the following filter heuristics that were also used by the creators of the Conceptual Captions dataset. An image is required to be in the JPEG file format. This information is obtained from the content type of the HTTP response header during the download. Furthermore, corrupted images are excluded. We use Python's Pillow library[1], which is a fork of the well-known Python Imaging Library, to open all image files and to check their data integrity. With this library, we are also able to extract meta-information such as image width and height. The dimensions of an image should be at least 400x400 pixels. From the 1.8 million downloaded images, close to 70 thousand were deleted due to the application of the above mentioned filter criteria.

5.1.2 Paraphrase Acquisition

Since the majority of images from the Conceptual Captions dataset are stock photographs and since all images are in the same file format, there is no need to use perceptual hashing to find image duplicates. Since there are numerous offers for free stock photos, it is also not necessary to plagiarize them or to obfuscate them by applying transformations. It is satisfactory to determine image equivalence by comparing MD5 hashes. We found that there are in fact image duplicates within the Conceptual Captions dataset; 4500 images had at least one duplicate. From these, 1358 sets of equivalent images can be distinguished. Of the corresponding captions within an equivalence set, all possible combinations yield the set of paraphrase candidates. Identical text pairs within a paraphrase candidate or already seen candidates were discarded. Despite these basic filter heuristics, no other caption preprocessing or filtering steps have been applied because the respective steps were already done by the creators of the Conceptual Captions dataset.

Without applying any additional filter heuristics we constructed 415,710 paraphrase candidates from the Conceptual Captions dataset. At first glance, this sounds like a big number. However, when we looked into these paraphrase candidates, the majority of examples did not possess any semantic similarity whatsoever. Further, some sets of paraphrases (i.e., all text pairs that were generated from captions of a single set of equivalent images) were incomprehensible large. The largest paraphrase set comprised almost 350 thousand text

[1] `https://pillow.readthedocs.io/en/stable/`

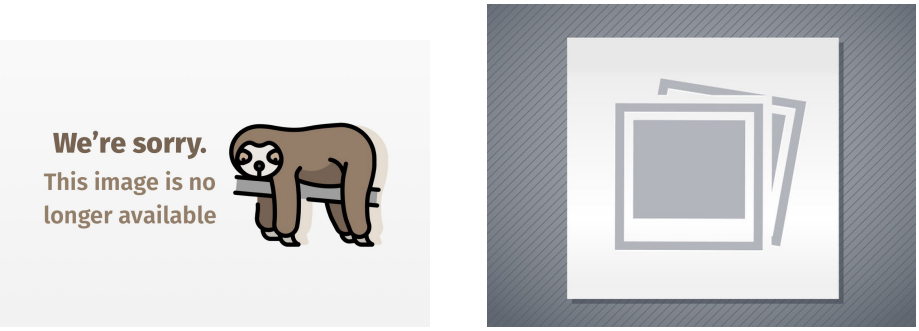


Figure 5.1: Exemplarily placeholder images from the Conceptual Captions dataset.

pairs. The corresponding images to that large set of paraphrases are equivalent placeholder images (i.e., replacement images if the original is not accessible anymore). Figure 5.1 shows examples of placeholder images that were downloaded with URIs from the Conceptual Captions dataset. For instance, 846 copies of a set of equivalent placeholder images were downloaded, which leads to the large paraphrase set of 350 thousand text pairs. This circumstance is also the cause for completely unrelated image captions being linked to paraphrase candidates. To separate actual paraphrases from semantically unrelated text pairs, implementing additional filter heuristics is required. Based on what we have seen, the quantity of images within an equivalence set is a promising indicator to eliminate placeholder images from the paraphrase acquisition pipeline. A reasonable upper bound for the accepted size of paraphrase sets was derived from z-scores. This metric computes the distance of a value x to the mean μ of a distribution in terms of the standard deviation σ.

$$z = \frac{x - \mu}{\sigma} \tag{5.1}$$

With $\mu = 3.13$ and a considerably large standard deviation $\sigma = 25.87$ of the distribution of paraphrase set sizes, we decided to choose a very narrow threshold of $z < 0.05$. If a paraphrase set size is 5% of the standard deviation above the mean, we discard the whole set. Thus, the upper bound for paraphrase set sizes would be 4.6, but since these number are discrete, we round the upper bound to 4. With this new threshold we obtained 119 paraphrase candidates.

5.1.3 Analysis

All of the 119 candidates were manually annotated regarding their quality. 68 of these are paraphrases while 51 of them are not. These are promising numbers in terms of suitability of image captions as paraphrases. Moreover,

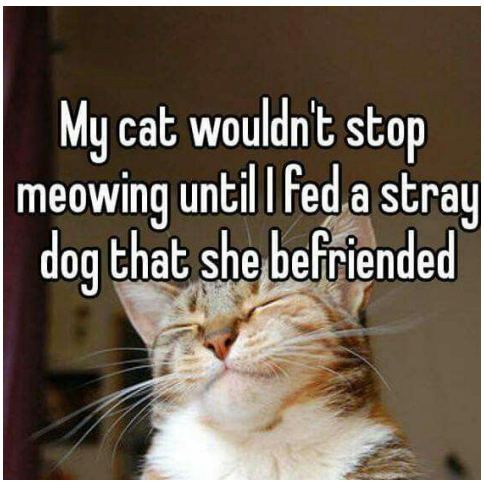


Figure 5.2: The caption that is associated with this image[2] is "*Not mine, but we could all learn a little something from this cat!*".

24 candidates, which do not classify as paraphrases are caption pairs of placeholder images, and this issue does not exist when we acquire paraphrases from web crawls. The remaining 26 text pairs that did not qualify as paraphrases expose an important property of image captions: Image captions from the web do not always reflect a description of the content of an image but rather give additional context. Figure 5.2 demonstrates an example where the caption does not give an explanation of the displayed content. Those image-caption pairs will most-likely not result in paraphrases when linked with duplicates. However, we can examine properties of image captions that depend on several factors such as origin and purpose of the image and the displayed content in the future. This will help us to tune our image filter heuristics to increase quality and quantity of acquired paraphrases.

The downside of using the Conceptual Captions dataset as source for the paraphrase acquisition is its high lexical similarity of the obtained paraphrases. Figure 5.3 presents an overview of the computed textual similarity measures on paraphrases and non-paraphrases that were acquired from the Conceptual Captions dataset. The difference in Levenshtein distance between paraphrases and non-paraphrases is extreme. This is not surprising since many of the negative examples originate from image captions of non-equivalent images. The low Levenshtein distance of paraphrases indicate that these paraphrases are very similar in their syntactic structure and therefore of rather low quality. The n-gram overlap metrics substantiate the high textual similarity. The Sumo metric favors paraphrases with high textual dissimilarity, and hence its score

[2] `http://whisper.sh/w/MTcwMjg5NDcx`

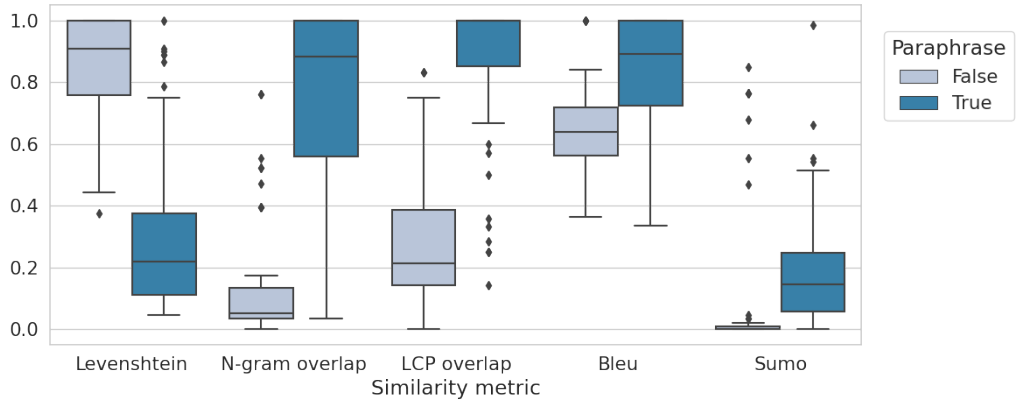


Figure 5.3: Textual similarity of paraphrases from the Conceptual Captions dataset.

is quite low. However, the outliers of the Sumo score indicate that there are a few promising paraphrases. For example, "*students standing on a stage in a line with their arms around each others*" and "*students forming a chain on a stage*" is a pair of texts that represent an outstanding example of a paraphrase with structural and lexical dissimilarities. This points out the importance of choice of the source of image captions. Another conclusion that we can draw is, that the Sumo metric is an important measure to determine the quality of a paraphrase. Nevertheless, this metric alone would not be a sufficient decision criteria to distinguish between paraphrases and non-paraphrases. To find such criteria, we have to analyze paraphrases from human-written image captions that were not artificially cleaned.

5.2 Wikipedia Paraphrase Dataset

As stated in Section 4.1, Wikipedia is an excellent source of human-written image captions with high textual quality and reliable caption extraction strategies. Therefore, we apply the paraphrase acquisition pipeline to the Wikimedia dump.

5.2.1 Paraphrase Acquisition

Wikimedia Commons is a free collection of media files which can be embedded in Wikipedia articles with the Wikitext markup language. The reuse of images from Wikimedia Commons is highly encouraged and they even detect if an image is used in an article and show information about the embedding articles

Pipeline step	Filter criteria	Effect	Image URIs
Input			5,745,684
Image filter	Number of occurrences	-1,516,715	4,228,969
Caption filter	No caption	-1,287,892	2,941,077
	Empty caption	-205	2,940,872
	Sentence heuristic	-2,710,127	230,745

Table 5.1: Effects of the filter critieria in the acquisition pipeline.

on the corresponding image page. The identifier for these images consists of a file or image prefix followed by the image name. Thus, we can detect embeddings of the same image with their associated Wikimedia Commons identifier, which we denote as resource identity (see Section 4.6.1).

As for the captions, we extract the displayed captions as well as the alt-texts from the file embedding syntax of Wikitext markup. Due to the additional alt-text extraction, empty captions can occur and they hence are discarded. Moreover, captions with less than 10 words are removed. Because of the clear separation of article languages in the Wikimedia dumps, no language detection is required to filter for English captions. Since Wikipedia is made for educational purposes, intentionally profane or sexual content is rarely found. Therefore, we do not have to apply profanity filters. Icons and other symbol images are also present in Wikipedia which encouraged us to limit the maximum number of occurrences of an image to 10.

In Table 5.1 we can see how many image references are discarded due to the application of the implemented filter criteria. A large number of references re dropped because these images occurred more than 10 times. Note that this number of image URIs referred to only 17 thousand different images. Inevitably, some images occurred many thousand times. It is also noticeable, that the majority of captions are discarded since they are not classified as sentences. This shows that using fragments for image captions and not authoring grammatically correct sentences is a common habit in Wikipedia.

Around 230 thousand image references remained for the paraphrase construction step. Thereof, close to 200 thousand unique URIs could be identified. By defining images with the same URI as equivalent, we are able to generate 323,807 paraphrase candidates. In Table 5.2, we can see the amount of paraphrase candidates that are discarded due to the corresponding filter heuristics. Since images in Wikipedia are centrally managed in Wikimedia Commons, it is not surprising that many repeating captions for an image can be found. Furthermore, we judge paraphrase candidates as “too similar” if the underlying

Pipeline step	Filter criteria	Effect	Paraphrases
Input			323,807
Paraphrase filter	Duplicates	-96,347	227,460
	Equal captions	-202,660	24,800
	Too similar captions	-806	23,994

Table 5.2: Effects of paraphrase filter on acquired candidates from Wikipedia

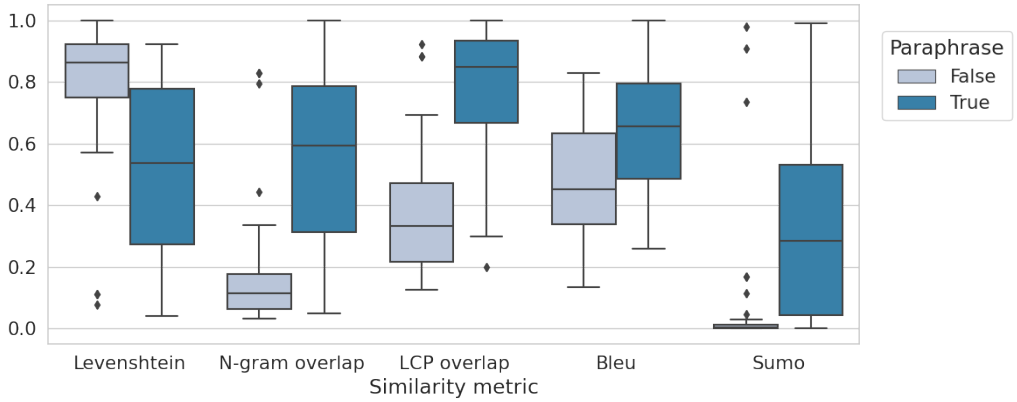


Figure 5.4: Textual similarity of paraphrases from Wikipedia

captions differ only in punctuation, segmentation or case. After application of these filters, we obtain a total of 23,944 paraphrase candidates, which is promising when we consider the size of the Wikimedia dumps.

5.2.2 Analysis

Approximately 24 thousand paraphrase candidates are too much for a manual annotation procedure within the scope of this thesis. Thus, we randomly sampled 100 from them to annotate paraphrase quality and assign entailment relations to it. The annotations are done by a single expert annotator but will be repeated in the future with multiple annotators to enable computation of inter-annotator agreement. This is important since the decision whether the information overlap is significant enough to consider a candidate as proper paraphrase is a subjective judgment.

Of 100 annotated candidates, 49 are judged to be proper paraphrases. Although this rate is lower compared to the Conceptual Captions paraphrase set, the textual quality and syntactic dissimilarity between a text pair of a paraphrase is way better. Figure 5.4 presents the textual similarity metrics in the

Caption A	Caption B	Sumo
"*The Fallen Astronaut memorial on the Moon includes the names of most of the known astronauts and cosmonauts who were killed before 1971*".	"*Commemorative plaque and the Fallen Astronaut sculpture left on the Moon in 1971 by the crew of Apollo 15 in memory of 14 deceased NASA astronauts and USSR cosmonauts.*"	0.9903
"*Marines demonstrate MCMAP in Times Square for Fleet Week 2010.*"	"*Marines demonstrate Marine Corps Martial Arts Program techniques at Times Square in 2010.*"	0.9797
"*Kyle Busch leads late in the Toyota Owners 400 at Richmond Raceway in April, a race he wins.*"	"*Eventual race winner Kyle Busch leads late in the race*"	0.9776

Table 5.3: Top 3 paraphrases by Sumo score from the annotated sample of the Wikipedia paraphrase corpus

Wikipedia paraphrase dataset, comparing paraphrases and non-paraphrases. The Levenshtein distance is quite high even for paraphrases which indicates structural dissimilarity. The n-gram overlap metrics are reasonably low for paraphrases and non-paraphrases which points towards lexical diversity. Since we learned that Sumo scores are especially important for rating syntactical and lexical dissimilarity, it is an achievement that the Wikipedia paraphrase corpus possesses paraphrases with very high Sumo scores. Since this metric is the most important indicator for high quality paraphrases, we want to demonstrate the top 3 paraphrases by Sumo score (see Table 5.3). 3 of the non-paraphrases also scored a comparably high Sumo score. This is due to high exclusive 1-gram overlap. For example, "*Worthy de Jong averaged the most steals in the 2015-16 season*" and "*Worthy de Jong won the inaugural award in 2011*" share many common 1-grams in a rearranged fashion and scored a Sumo score of 0.9797. However, we do not consider this example a paraphrase.

From the above example, we can learn the following. Although the subject of two captions are identical, they not necessarily paraphrase each other. Without the knowledge who Worthy de Jong is, we can not deduce the same information from both captions. However, if we replace "*inaugural award*" with "*DBL Most Improve Player award*", we could infer from both captions that Worthy de Jong is a basketball player. As explained in Section 3.2, we

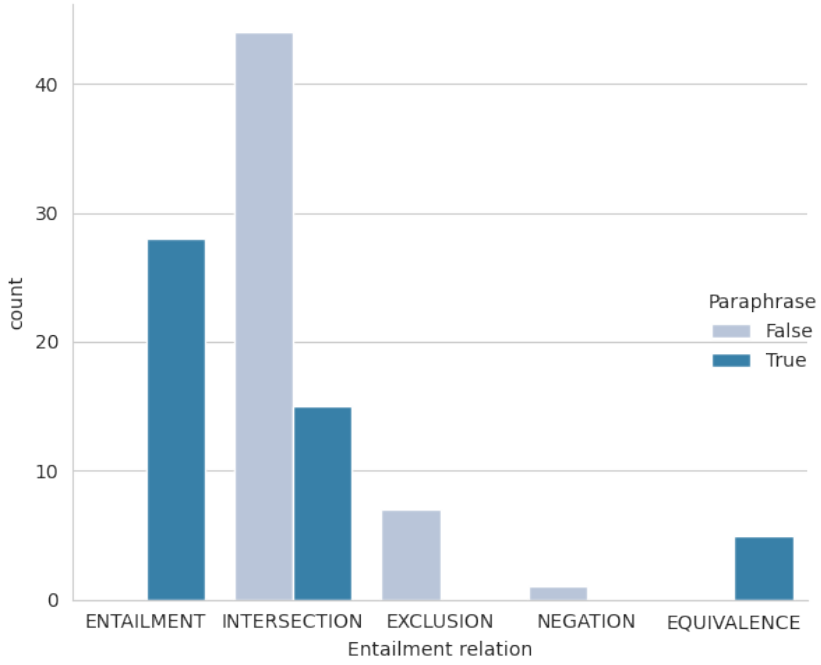


Figure 5.5: Entailment relation counts between paraphrases and non-paraphrases in the Wikipedia paraphrase corpus

consider this semantic relation an intersection.

Slight modifications of caption may result in a change of the entailment relation. Thus, assigning proper entailment relations is a challenging task even for a human. However, analyzing entailment relations gains important insights of the characteristics of image captions as paraphrases.

From Figure 5.5 we can deduce various important observations. The semantic equivalence relation is actually a rare occurrence rather than the norm for paraphrases. As already stated, semantic equivalence only occurs when two texts offers high lexical similarity. Consequently, a low number of semantically equivalent paraphrases is a benefit of this dataset. A paraphrase more frequently fulfills an entailment or intersection relation. Another important property is, that non-paraphrases that originate from our paraphrase acquisition pipeline most-frequently semantically intersect rather than exclude each other. This means that, being image captions of the same image establishes a strong semantic connection in most cases. This conforms our expectations.

It is hard to grasp how image captions of an identical image can semantically exclude or even contradict each other. Therefore, we had a look at those cases and analyzed what is displayed in those images. In most cases, single or multiple persons are shown in the image if its captions appear to be unrelated.

If there are multiple people in the image, a caption might refer to a different person as another one. If there is only a single person in the image, captions might describe individual aspects of a persons life or examines the image in an entirely different context. The only contradiction that we encountered in our sample are captions of an image of the professional stock car racing driver Elliot Sandler. The context of one of these captions is the NASCAR Race from 2011 and the other from the same race in 2012 where he scored a different number of points. Both captions state his score, which results in a negating caption pair. In the future it can be beneficial to find heuristics to exclude images which show single or multiple humans to reduce the amount of unrelated paraphrase candidates.

Chapter 6

Experiment and Evaluation

In the first half of this chapter, we are going to evaluate our rule-based sentence detection approach, which is an important step in the paraphrase acquisition pipeline. In the second half, we will analyze the suitability of perceptual image hashing to detect equivalences of slightly transformed duplicates.

6.1 Sentence Detection

The filter criteria for captions include the distinction between proper sentences and sentence fragments. As explained in Section 4.5.2, we randomly sample image captions from Wikipedia and manually annotate them to find POS tag sequences which distinguishes fragments from grammatically complete sentences. The result of this procedure are 4 rules (see Table 4.3) and a training and test set of 500 and 100 sentence candidates respectively. Around 30% of the training set examples were annotated as sentences and almost 20% of the test data. We use these data sets to conduct an experiment to evaluate the effectiveness of the classification that is based on the derived rule set.

Table 6.1 shows precision and recall of the rule-based sentence classifier on the labeled test and training sets. In order to maintain high textual quality and considering the amount of data available in the web crawls, maximizing precision has a much higher priority than achieving good recall. With this in mind, a precision of 94% on the unseen examples of the test set with a recall of almost 80% is a very good result. Precision-wise only one fragment in the

	Precision	Recall
Test set	0.94	0.79

Table 6.1: Effectiveness of the sentence classifier.

(a) Dices image[1] **(b)** Mars surface image[2]

Figure 6.1: Images used for perceptual hashing evaluation

test set was falsely classified as a sentence. The respective sentence fragment "*Two lively were-jaguar babies on the left side of La Venta Altar 5.*" is an example where the Stanford POS tagger misinterpreted a word (i.e., "*were*") as an inflected verb, which leads to a wrong classification. In fact, 8 fragments in the training set were misclassified as sentences due to wrong labeling of the tagger. Although these results are promising, the test set comprises only 100 examples and further evaluation is needed to validate these findings.

6.2 Case Study: Perceptual Hashing

To understand the potential of perceptual hashing to find visually identical images in the web, we conduct a case study based on a single example image. The goal of this experiment is to compare perceptual hash functions to evaluate which image transformations can be applied to an image without a change in the generated fingerprint.

To analyze properties of perceptual hash functions, we use two randomly chosen images: a PNG with transparent background from Wikipedia showing dices and a JPEG-compressed image of the surface of the Mars from NASA (see Figure 6.1). The dices image has a resolution of 800x600 and the Mars surface image is 4096x3879. We apply the following image transformations to the dices image to generate variants to compare hashes against.

[1] https://mars.nasa.gov/resources/25264/jezero-crater-as-seen-by-esas-mars-express-orbiter/
[2] https://commons.wikimedia.org/wiki/File:PNG_transparency_demonstration_1.png

1. **JPEG compression**
   The compression is applied with a quality of 80%. Since transparency can not be displayed in the JPEG format, all the transparent pixels in the original image are changed to white color.

2. **Flipping**
   The image is vertically flipped.

3. **Scaling**
   We uniformly scale the image down by 25% which result in a new resolution of 600x450 pixels.

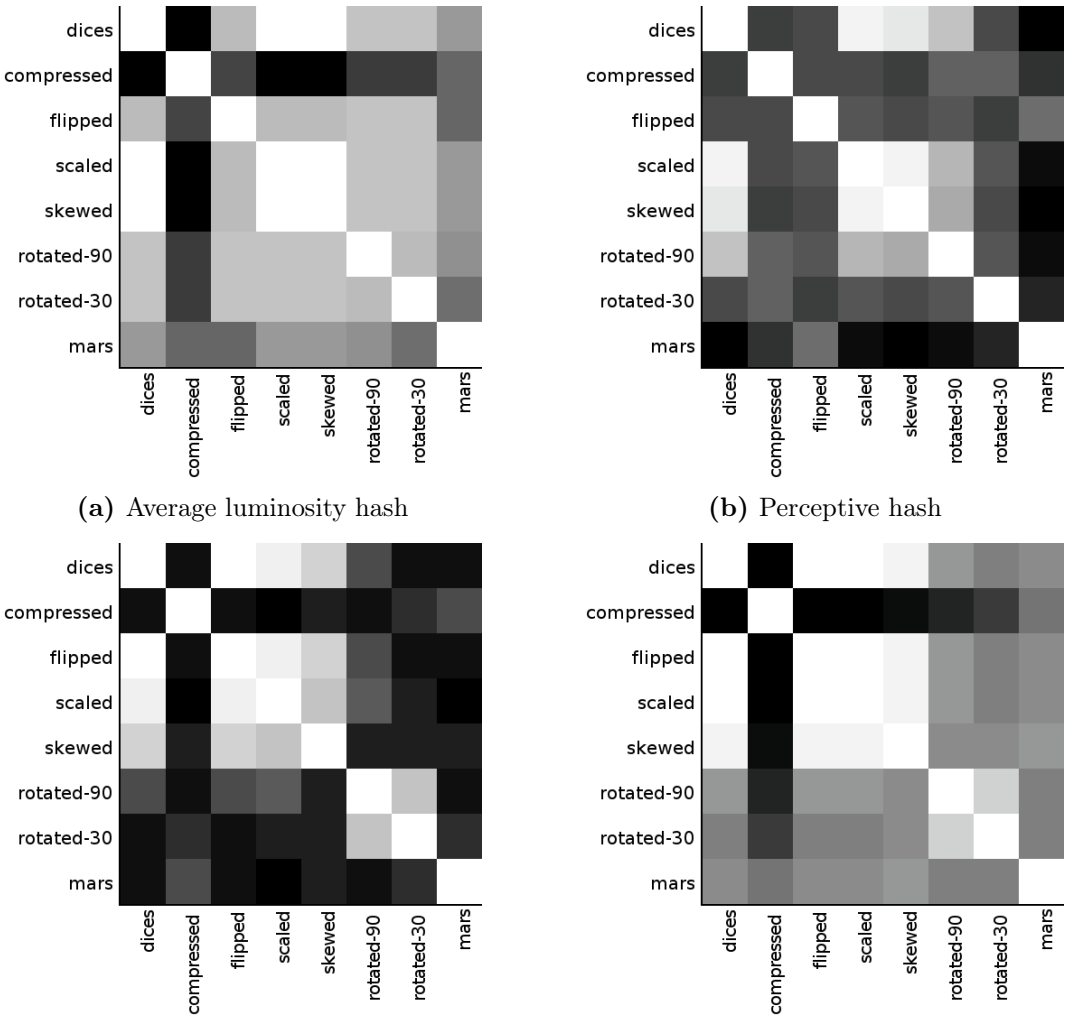
4. **Skewing**
   We apply non-uniform scaling in order to shrink the image to 300x100 pixels.

5. **Rotation**
   We generate two different rotated images from the dice image; one is rotated by 90°and one by 30°. All image formats require rectangular borders which means that the 30°-rotated image is filled with transparent pixels and has a new resolution of 992x919 pixels.

To compute perceptual hashes, we use JImageHash[3], a perceptual image fingerprinting library developed in Java. This library provides implementations for all hash functions described in Section 4.6.3 and provides methods for computing image similarity through normalized Hamming distance of the resulting hashes.

Figure 6.2 compares image similarities between different 30-bit perceptual hash functions in form of a heatmap. The average luminosity hash is able to detect equivalence between the original, the scaled, and the skewed image of the dices, but overall computed small distances among all transformed images but the compressed one. The rotational invariant version of the average hash is actually incapable of identifying equivalence between the rotated images but managed to identify the flipped image as equivalent to the original. Against our expectations, the perceptive hash does not classify the JPEG compressed image as equivalent. In fact, none of the algorithms were able to do so. A reason may be the lossy quantization step of the DCT coefficients during the JPEG compression that is not taken into account when computing the hash value. Unfortunately, the perceptive hash labels none of the transformed images as equivalent to the original. However, the rotational invariant perceptive hash performed best by grouping the flipped, the scaled, and the original image in

[3] `https://github.com/KilianB/JImageHash`



(a) Average luminosity hash

(b) Perceptive hash

(c) Rotational invariant average hash

(d) Rotational invariant perceptive hash

Figure 6.2: Image similarity between image transformations as measured by normalized Hamming distance. If a square is plain white, two images are identical (distance of 0) and if it is black they are completely unrelated (distance of 1).

one equivalence class while also maintaining low distance to the skewed image. Although this algorithm claim to be rotational invariant, it is also not able to detect the equivalence among the rotated images.

Chapter 7

Conclusion

A part of this thesis is to point out the difficulty of defining paraphrases. With our formalization we defined the concept of paraphrases in a logical sense. Further, we emphasized that it would be beneficial to distinguish semantic entailment relations to better understand different kinds of paraphrases.

As a main contribution of this thesis, we introduced a novel paraphrase acquisition approach that has the potential to create large paraphrase corpora, which overcome most of the shortcomings of existing datasets. We presented different variations of this method in terms of image equivalence determination, caption extraction strategies, and various filter criteria for images, captions, and paraphrases. Moreover, we presented several common metrics for assessment of textual similarity and outlined its properties.

With the creation of two different prototype datasets, we showed that this approach is capable of extracting many paraphrase candidates that have a good balance between paraphrases and non-paraphrases. Most of the candidates, which do not qualify as paraphrases, semantically intersect. Thus, these candidates are negative examples which are hard to distinguish from proper paraphrases. These kind of negative examples are important for training paraphrase models [Zhang et al., 2019]. Further, we found that many paraphrases that were acquired with our new approach have structural and lexical dissimilarities, which qualifies them as especially important for learning paraphrasing.

An important building block of our paraphrase acquisition pipeline is the caption extraction step along with the filtering step. A part of this is the detection of sentences, and with our evaluation we prove the effectiveness of our rule-based classification approach to eliminate sentence fragments. Equivalence determination between images is one of the most important steps of the proposed method of paraphrase acquisition. We evaluated the usefulness of perceptual image hashing to detect equivalences of transformed copies of an image. We found that compression is a challenging transformation for equiv-

alence detection but symmetric transformations can be detected sufficiently with perceptual hashes.

7.1 Future Work

An observation that we made is, due to "aggressive" filtering, large input datasets are required to obtain a considerable amount of paraphrase candidates. As our chosen input data for our prototype corpora are comparably small, we plan to create even larger paraphrase corpora by using the ClueWebs, Common Crawl, Web Archive crawls, or a combination of all. We showed that these datasets are magnitudes larger than the Wikimedia dumps. The use of these mentioned crawls raises new challenges since sophisticated caption extraction methods need to be implemented to maintain a similar textual quality as captions from Wikipedia. An analysis of images, which lead to the Wikipedia paraphrase dataset can help us develop new filter criteria to eliminate unpromising images from the other web crawls.

Since our paraphrase acquisition pipeline collects unlabeled paraphrase candidates, we will build automatic classifiers to label the resulting candidates. Useful indicators for a paraphrase are the presented textual similarity metrics. However, with these only, it will be presumably difficult to develop reliable paraphrase detection algorithms though. As we have access to meta-information such as source web pages of the captions, using meta-information as feature might also help to build such classifiers. Another classifier for assessment of semantic entailment relations might facilitate future research in textual entailment or summarization in the context of paraphrasing. We can base these classifiers on commonly used approaches that were specifically designed for paraphrases (e.g., Hickl et al. [2006]) rather than develop new ones from scratch.

Another open task is to do a sophisticated comparison between existing paraphrase corpora and paraphrases that were acquired with our approach. Therewith, we want to show and prove that our paraphrasing approach extracts paraphrases with distinctive textual properties and overcomes common issues of existing datasets. One important aspect of this comparison would be the vocabulary size and word frequencies to show textual diversity within our extracted corpora. With the analysis of our small sample we found promising textual similarity assessments. However, to verify these findings we need larger samples and compare textual similarity to other existing paraphrase corpora.

We evaluated our sentence detection method on a sample of 100 captions. To verify its effectiveness these experiments will be repeated in the future with larger caption samples.

Since analysis of the relation between text and images in the web offers many opportunities, we plan to dive deeper into this topic. Our presented approach can be easily adapted to be used in several applications such as acquisition of translations and multi-lingual paraphrases, plagiarism detection, or semantic similarity of entities in a knowledge base. The latter can be done since, with linking of equivalent images, we also link web pages that might have a semantic relation to some extend. This can be easily exploited in the future.

Bibliography

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263, 2015. 1

Colin J. Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 597–604, 2005. 2.1

Regina Barzilay and Lillian Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In Eduard Hovy, Marti Hearst, and Mari Ostendorf, editors, *Proceedings of the Third Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 16–23, Edmonton, Canada, May 2003. Association for Computational Linguistics. 4.9.2

Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France*, pages 50–57, 2001. 2.1

Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013. doi: 10.1162/COLI_a_00166. URL `https://doi.org/10.1162/COLI_a_00166`. 1

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *CoRR*, 2015. URL `http://arxiv.org/abs/1508.05326`. 2.3

Steven Burrows, Martin Potthast, and Benno Stein. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–21, 2013. 2.1, 3.1, 4.9.1

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. ClueWeb09 data set, 2009. 4.1

Jamie Callan et al. ClueWeb12 data set, 2013. 4.1

David Chen and William Dolan. Collecting Highly Parallel Data for Paraphrase Evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, June 2011. Association for Computational Linguistics. 2.2

Joao Cordeiro, Gael Dias, and Pavel Brazdil. A Metric for Paraphrase Detection. In Mihai Boicu, Jose Costa-Requena, Dominique Thiebaut, Ciprian Popoviciu, Bernard Tuy, and Gunter Van de Velde, editors, *Proceedings of the Second International Multi-Conference on Computing in the Global Information Technology*, pages 1–6, Gosier, Guadeloupe, March 2007a. IEEE. 4.9.1, 4.9.3, 4.9.4, 4.9.5, 4.9.5

Joao Cordeiro, Gael Dias, and Pavel Brazdil. New Functions for Unsupervised Asymmetrical Paraphrase Detection. *Journal of Software*, 2(4):12–23, October 2007b. 1, 2.3, 3.2.2, 4.9.2

Mathias Creutz. Open subtitles paraphrase corpus for six languages. *CoRR*, 2018. 2.1

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. 4.10

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. International Conference on Computational Linguistics, August 2004. 2.1, 4.9.1

William B. Dolan and Chris Brockett. Automatically Constructing a Corpus of Sentential Paraphrases. In Mark Dras and Kazuhide Yamamoto, editors, *Proceedings of the Third International Workshop on Paraphrasing*, pages 1–8, Jeju, South Korea, October 2005. Kazuhide Yamamoto. 2.1

El-Sayed M. El-Alfy, Radwan E. Abdel-Aal, Wasfi G. Al-Khatib, and Faisal Alvi. Boosting paraphrase detection through textual similarity metrics with abductive networks. *Applied Soft Computing*, 26:444–453, 2015. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2014.10.021. URL `https://www.sciencedirect.com/science/article/pii/S1568494614005316`. 4.9.2, 4.9.4

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, 2013. 1

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010. 2.2

Juri Ganitkevitch and Chris Callison-Burch. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer, 2014. 2.1

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 758–764, 2013. 2.1

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 1

Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1013. URL `https://www.aclweb.org/anthology/D15-1013`. 4.9.4

Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, and Ying Shi. Recognizing textual entailment with lcc's groundhog system. 02 2006. 7.1

Graeme Hirst. Paraphrasing paraphrased. In *Keynote address for The Second International Workshop on Paraphrasing: Paraphrase acquisition and Applications*, 2003. 1

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. First quora dataset release: Question pairs. *data. quora. com*, 2017. 2.1

Hamid Izadinia, Fereshteh Sadeghi, Santosh K. Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

Kevin Knight and Daniel Marcu. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence*, 139(1):91–107, July 2002. 2.1

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A continuously growing dataset of sentential paraphrases. *CoRR*, abs/1708.00391, 2017. URL `http://arxiv.org/abs/1708.00391`. 1, 2.1

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966. 4.9.1

Tao Li and Vivek Srikumar. Exploiting sentence similarities for better alignments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2193–2203, 2016. 1

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. 2.2, 5.1

Bill MacCartney and Christopher D Manning. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, 2009. 2.3, 3.2

Nitin Madnani and Bonnie J. Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36 (3):341–387, September 2010. 2.3

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`. 4.5.1

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, 2014. 1, 2.2, 2.3

Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel Keim, and Bela Gipp. An adaptive image-based plagiarism detection approach. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 131–140, 2018. 4.6.2

Shashi Narayan and Claire Gardent. Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics*, pages 435–445, 2014. 4.9.4

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, editor, *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, July 2002. Association for Computational Linguistics. 4.9.4

Y Albert Park and Roger Levy. Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 934–944, 2011. 4.9.4

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, 2015. 2.3

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek V. Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. *CoRR*, abs/1610.03098, 2016. URL `http://arxiv.org/abs/1610.03098`. 2.2

Chris Quirk, Chris Brockett, and William B. Dolan. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 142–149, 2004. 2.1

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010. 2.2

Yves Scherrer. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6868–6873, 2020. 1, 2.1

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390, 2015. 1

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://www.aclweb.org/anthology/P18-1238`. 4.4.1, 5.1

Sanja Štajner, Hannah Bechara, and Horacio Saggion. A deeper exploration of the standard pb-smt approach to text simplification and its evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 823–828, 2015. 4.9.4

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. *The Penn Treebank: An Overview*, pages 5–22. Springer Netherlands, Dordrecht, 2003. ISBN 978-94-010-0201-1. doi: 10.1007/978-94-010-0201-1_1. URL `https://doi.org/10.1007/978-94-010-0201-1_1`. 4.5.2

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2.2

Michael Völske, Janek Bevendorff, Johannes Kiesel, Benno Stein, Maik Fröbe, Matthias Hagen, and Martin Potthast. Web Archive Analytics: Infrastructure and Applications @ Webis (extended abstract). In *2nd International Symposium on Open Search Technology (OSSYM 2020)*, October 2020. 4.1

Luis von Ahn. List of bad words, 2019. 4.5.2

John Wieting and Kevin Gimpel. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *CoRR*, 2018. 2.1

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. Extracting lexically divergent paraphrases from twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448, 2014. 2.1

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016. 4.9.4

Mikio Yamamoto and Kenneth W Church. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30, 2001. 4.9.3

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. URL `https://doi.org/10.1162/tacl_a_00166`. 2.2

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: paraphrase adversaries from word scrambling. *CoRR*, 2019. URL `http://arxiv.org/abs/1904.01130`. 1, 2.1, 4.8, 7