Paraphrase corpora

Opusparcus

The paraphrases are extracted from the OpenSubtitles2016 corpus, which contains subtitles from movies and TV shows

The development and test sets consist of sentence pairs that have been checked manually; each set contains approximately 1000 sentence pairs that have been verified to be acceptable paraphrases by two annotators

Meidän kaikkien olisi pitänyt kuolla . Meidän kaikkien piti kuolla .

Koska laivat ovat valmiina ? Milloin laivat ovat valmiina ?

The Paraphrase Database (PPDB)

English portion (PPDB:Eng) contains over 220 million paraphrase pairs
73 million phrasal, 8 million lexical paraphrases, and 140 million paraphrase pattern
Extracted from bilingual parallel corpora totaling over 100 million sentence pairs and over 2 billion English words

Finnish: an automatically extracted database containing millions paraphrases in 16 different languages

Quality mostly (very) shitty, some good phrase pairs e.g. kolmesti ||| kolme kertaa, vuosittain ||| joka vuosi

PPDB: An example one-to-many pair

työtä kohtaan ||| työtä ||| Abstract=0 Adjacent=0 Alignment=0-0 CharCountDiff=-8 CharLogCR=-0.95551 ContainsX=1 GlueRule=0 Identity=0 Lex(e|f)=10.77032 Lex(f|e)=5.17257 Lexical=1 LogCount=0 Monotonic=1 PhrasePenalty=1 RarityPenalty=0.36788 SourceTerminalsButNoTarget=0 SourceWords=2 TargetTerminalsButNoSource=0 TargetWords=1 UnalignedSource=1 UnalignedTarget=0 WordCountDiff=-1 WordLenDiff=-1.00000 WordLogCR=-0.69315 p(LHS|e)=5.57678 p(LHS|f)=0 p(e|LHS)=12.13654 p(e|f)=1.26318 p(e|f,LHS)=2.56495 p(f|LHS)=12.96321 p(f|e)=7.66664 p(f|e,LHS)=3.39163

DIRT Paraphrase Collection

Discovery of Inference Rules from Text

Paraphrasing methods based on monolingual text corpora, like DIRT (Lin and Pantel, 2001), measure the similarity of phrases based on distributional similarity

Can’t find the corpus

Corpus	|Train|	|Dev|	|Test|	Task	Metric	Domain
				Single-Sentence Tasks		
CoLA	10k	1k	1.1k	acceptability	Matthews	linguistics literature
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
				Similarity and Paraphrase Tasks		
MRPC	4k	N/A	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman	misc.
QQP	400k	N/A	391k	paraphrase	acc./F1	social QA Questions
				Inference Tasks		
MNLI	393k	20k	20k	NLI	acc. (match/mismatch)	misc.
QNLI	108k	11k	11k	QA/NLI	acc.	Wikipedia
RTE	2.7k	N/A	3k	NLI	acc.	misc.
WNLI	706	N/A	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-Benchmark, which is a regression task. MNLI has three classes while all other classification tasks are binary.

Similarity and Paraphrase Tasks

MRPC: The Microsoft Research Paraphrase Corpus

- a corpus of sentence pairs automatically extracted from online news sources, with human annotations of whether the sentences in the pair are semantically equivalent
- class imbalance (68% positive, 32% negative)

QQP: The Quora Question Pairs

- a collection of question pairs from the community question-answering website Quora
- given two questions, the task is to determine whether they are semantically equivalent
- class imbalance (37% positive, 63% negative)

STS-B: The Semantic Textual Similarity Benchmark

QQP

question1	What is the step by step guide to invest in share market in india?
question2	What is the step by step guide to invest in share market?
is_duplicate	0

MRPC

They had published an advertisement on the Internet on June 10 , offering the cargo for sale , he added .
On June 10 , the ship 's owners had published an advertisement on the Internet , offering the explosives for sale .

Label: 1

Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

Label: 1

Paraphrase Grouped Corpora

How Well Sentence Embeddings Capture Meaning (White et al., 2015)

These corpora have been prepared by taking existing real-word based corpora: The Microsoft Research Paraphrase Corpus, and Opinosis, and partitioning them into groups of sentences which all share the same meaning.

MSRP

A subset of the Microsoft Research Paraphrase corpus was automatically grouped according to its original manually annotated meaning. This was done by taking the symetric and transitive closure over the original set of paraphrase pairs.

858 sentences divided into 273 groups

21,7 Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the rice crop as harvesting had just finished.
22,7 Agriculture Secretary Luis Lorenzo said there was no damage to the vital rice crop as the harvest had ended.
23,7 Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the vital rice crop as harvesting had just finished.

Paraphrase Grouped Corpora

Opinosis corpus: sentences extracted from reviews on 51 topics.

A subset of the Opinosis corpus was manually grouped according to its meaning.

668 sentences, divided into 162 groups

0,0 The keyboard, though slightly smaller than standard, is a pleasure to use .
1,0 keyboard is actually a pleasure to use inspite of small size
2,1 Speaking of which, the keyboard's relatively large size , at 92% of the size of a normal one , writing longer texts on it is no problem .
3,1 The keyboard, more than 90% standard size, is just large enough .
4,1 The keyboard is large enough to accommodate touch typing with ease .
5,1 Speaking of typing, the keyboard is great, almost full size .
6,1 It's 90% size keyboard doesn't take too long to get used to .
7,1 The keyboard is only slightly smaller than a regular one, so it's very comfortable .
8,1 The keyboard is nearly full size and very comfortable to type on for hours .

420,114 The food is excellent and the service great .
421,114 The food was good and the service was very good .
422,114 Food quality was good, service prompt, everything prepared as we had ordered .
423,114 Stunning food, amazing service .
424,114 The food is very, very good and the staff were very friendly .
425,114 Super dinner and attentive service .

566,137 Location was great , close to tube stop .
567,137 The location is EXCELLENT, just round the corner from the subway .
568,137 Speaking of the tube, the location really is great 1 block away !
569,137 The location is great and the Tube station is very close .
570,137 Perfect location, right around the corner from the tube .
571,137 Location was great come out of tube station, take first turn right and walk 150 yards you're there !
572,137 Location is good if you plan on taking the tube .
573,137 THe location was great, just aroung the corner from the tube .

Maintained by CLiC at the Universitat de Barcelona

Corpora to download

Name of corpus	Language	Sentences	Pairs/groups	Paraphrase phenomena
MSRP-A	en	7489	3900	22105
P4P	en	1712	856	11420
WRPA-person	en		3	
WRPA-person-2	en		12	
WRPA-authorship	es		81101	
WRPA-authorship-A	es	411	1000	1329

Paraphrase for Plagiarism (P4P)

A partition of the plagiarism cases in the PAN-PC-10 corpus manually annotated with the paraphrase phenomena they contain.

Composed of 847 source-plagiarism pairs in English.

MSRP-A

WRPA

ETPC

Paraphrase Adversaries from Word Scrambling (PAWS)

The dataset has two subsets, one based on Wikipedia and the other one based on the Quora Question Pairs (QQP) dataset.

The User-Language Paraphrase Corpus

Does not seem to be publicly available

Collected from the intelligent tutoring system iSTART

Sometimes blood does not transport enough oxygen, resulting in a condition called anemia.
Anemia is a condition that is happens when the blood doesn't have enough oxygen to be transported

Creation of a multi-paraphrase corpus based on various elementary operations

Does not seem to be publicly available

In this research, we construct a paraphrase corpus based on various elementary operations (reordering, substitution, deletion, insertion) in a crowdsourcing platform to generate multi- paraphrase sentences from a source sentence. These elementary paraphrase operations can be utilized for various applications (i.e., deletion for summarization and reordering for machine translation). Our evaluations show the richness and effectiveness of our created corpus.