Chapter 27
Textual Paraphrase Dataset for Deep Language Modelling

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and Otto Tarkka

Abstract The Turku Paraphrase Corpus is a dataset of over 100,000 Finnish paraphrase pairs. During the corpus creation, we strived to gather challenging paraphrase pairs, more suitable to test the capabilities of natural language understanding models. The paraphrases are both selected and classified manually, so as to minimise lexical overlap, and provide examples that are structurally and lexically different to the maximum extent. An important distinguishing feature of the corpus is that most of the paraphrase pairs are extracted and distributed in their native document context, rather than in isolation. The primary application for the dataset is the development and evaluation of deep language models, and representation learning in general.

1 Overview and Objectives of the Pilot Project

Natural language processing research focuses increasingly more at a deeper understanding of language meaning, which is the enabling factor for the next generation of language technology applications. Of especially recent interest are neural meaning representations that are robust to non-trivial re-phrasing of statements with equivalent or near-equivalent meaning. While deep learning methods have effectively solved many supervised learning tasks where large amounts of task-specific training data are available, their performance in representation learning tasks is much weaker (Glockner et al. 2018; Tsuchiya 2018; McCoy et al. 2019). In practical terms, we do not yet have well-proven general methods that, given arbitrary statements with the same contextual meaning but very different wording, would reliably produce highly similar representations for the statements. The fundamental limitation has been the lack of appropriate training data and learning procedures that are able to infer the projection from observable surface forms to faithful semantic representations.

In this ELG pilot project, we set out to address this limitation by building a fully manually annotated paraphrase corpus for Finnish, the Turku Paraphrase Corpus. In

Jenna Kanerva · Filip Ginter · Li-Hsin Chang · Valtteri Skantsi · Jemina Kilpeläinen · Hanna-Mari Kupari · Aurora Piirto · Jenna Saarni · Maija Sevón · Otto Tarkka
University of Turku, Finland, jmnybl@utu.fi, lhchan@utu.fi, figint@utu.fi

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_27

addition to building this resource, we also gathered experience and data regarding how such a resource can be built efficiently and what human resources are needed, built initial models based on the new resource, and produced baseline results.

2 Methodology

The primary distinguishing feature of our corpus compared to other related efforts is its fully manual annotation (as opposed to automatic candidate generation), resulting in paraphrase pairs that are non-trivial and challenging in not being highly lexically related. In other words, an important objective was to avoid bias due to automatic candidate selection so as to obtain a more realistic estimate of the performance of machine learning models on natural language understanding tasks. To this end, we gather source documents that are potentially rich in paraphrases for fully manual paraphrase candidate extraction. These documents include alternative translations of movie subtitles, news headings and articles reporting the same event, discussion forum messages with identical titles and topics, alternative student translations from translation course assignments, and student essays answering the same prompts.

Along with the manual extraction, all paraphrase candidates are manually classified into categories of paraphrases and non-paraphrases according to the developed annotation scheme. The design of the annotation scheme strives to capture varying levels of paraphrasability of candidate paraphrase pairs. We use a scale of four base labels, 1–4, similar to those used in some other paraphrase corpora (Creutz 2018). We define the four base labels as label *1* unrelated sentences, label *2* related but not paraphrases, label *3* paraphrases in the given context but not universally so, and label *4* universal paraphrases. In addition, label *4* paraphrases can be marked with optional flags $>$ or $<$ for subsumption, *s* for style, and *i* for minor deviations. These flags mark properties of the paraphrases that do not fulfill the strict universality criteria of the label *4* due to one of several defined reasons. The subsumption flag means that the paraphrasability is directional; one sentence can be universally substituted by the other, but not the other way around. The style flag means that the paraphrases convey the same meaning, but may have differing tones or registers, which make them not interchangeable in certain circumstances. The minor deviation flag marks minimal differences in meaning (for example, “this” vs. “that”), or grammatical number, person, tense, etc. that can be trivially identified automatically. These flags are independent of each other and thus one label *4* paraphrase pair can have multiple flags, disregarding the directional subsumption flags. More detailed description of the labels together with example annotations is given in the annotation guidelines (Kanerva et al. 2021a).

3 Implementation

The annotation work was carried out by six main annotators, each being a native Finnish speaker with a strong background in language studies by having completed or ongoing studies in a field related to languages or linguistics. Each annotator worked 5–9 months either full or part time in a strong collaboration with a broader project team including supportive roles in the annotation work.

An annotator starts the process by going through the automatically aligned source document pair presented side-by-side in a custom annotation tool[1] developed for the paraphrase extraction, and extracts all interesting paraphrase candidates by selecting the corresponding text passages from both documents. While saving the candidate, together with the text passage pair the tool also saves the actual position of the text passage in the original document, therefore supporting studying the paraphrase pairs in their original document context. To our knowledge, this is the first paraphrase corpus that includes the document context for the released paraphrase pairs. After extracting all interesting paraphrase candidates from the source document pair, the annotator marks the document finished and moves on to the next one.

The extracted paraphrase candidates are automatically transferred to a separate annotation tool[2] developed specifically for paraphrase labeling. In this tool, each pair of paraphrase candidates is shown separately, and the annotator can see the original contexts if necessary. The annotator labels the original paraphrase pair, and has the option to copy the original text and rewrite the texts into full paraphrases (label *4* without flags). In cases where the annotator decided to provide a rewritten pair, two or more pairs of paraphrases are obtained for the corpus: the original pair, and the rewritten pair(s). The annotators are instructed to rewrite the paraphrase candidates in cases where a simple edit, such as word deletion, insertion or synonym replacement, can be naturally constructed and does not require too much effort.

4 Evaluation

The paraphrase label annotation was guided using a shared annotation manual, daily meetings, and regularly assigned double annotation batches in order to ensure annotation consistency between the six annotators. The manual paraphrase extraction did not involve a similarly careful annotator training or consistency monitoring throughout the project. Instead of ensuring each annotator extracting the same segments if given the same text, the objective is to collect a diverse set of different paraphrase candidates, where minor deviations in the personal extraction habits only creates more diversity to the data. In order to study the extraction behaviour of the annotators, we measure the average number of paraphrase pairs extracted from one docu-

[1] https://github.com/TurkuNLP/pick-para-anno

[2] https://github.com/TurkuNLP/rew-para-anno

ment pair, indicating how eager the annotator was to include or exclude borderline uninteresting, extremely difficult or otherwise debatable pairs from the corpus.

While the data sources used in the paraphrase extraction step have distinct characteristics in terms of extraction ratios, we use the subset originating from the alternative subtitles (approx. 80% of the full corpus) for this study in order to account for differing source text proportions between the annotators. We measure the average number of paraphrases extracted from one subtitle document pair (about 15 minutes worth of the subtitled program's runtime), while taking into account all document pairs where the extraction and labeling was carried out by the same annotator, and the document pair resulted at least one extracted paraphrase. The statistics are shown in Table 1, the individual extraction rates falling between 13 and 50 pairs indicating some amount of diversity between the annotators. When measuring the mean lexical similarity of the extracted paraphrase pairs (together with standard deviation) as well as annotated paraphrase label distribution for each annotator, we do not notice any significant difference between annotators oriented towards higher or lower extraction rates. The label distributions are visualised in Figure 1. Finally, in Table 1 we measure the proportion of extracted paraphrase pairs each annotator chose to rewrite during the label annotation (row *Rewritten*), showing large differences among the annotators, between 1.4% and 29.5% of rewritten paraphrase pairs.

	Ann1	**Ann2**	**Ann3**	**Ann4**	**Ann5**	**Ann6**
Extracted pairs	28,685	18,908	9,553	7,713	6,359	1,897
Total extracted (%)	39.1	25.8	13.0	10.5	8.7	2.6
Extracted/doc	23.4	13.2	13.4	22.0	48.9	23.4
Rewritten (%)	6.8	23.4	1.3	29.5	14.9	1.4

Table 1 Comparison of the six annotators in terms of the average number of paraphrase pairs extracted from one 15-min subtitle pair (Extracted/doc), as well as the percentage of paraphrase pairs, where the annotator provided a rewrite (Rewritten); in addition to these two metrics, we also illustrate the total amount of the paraphrase pairs extracted by the annotator (both raw count and percentage); note that the number of extracted paraphrases does not sum up to the total corpus size as the comparison is done on the subtitle subset only (approx. 80% of the full corpus)

In order to ensure the consistency of the label annotation, approx. 2% of the paraphrase pairs are double annotated, where two different annotators annotate the labels independently from one another for the same paraphrase candidates. The two individual annotations are merged and conflicting labels resolved together with the annotation team, resulting in a consolidated subset of consensus annotation. The overall accuracy of the individual annotations against the consensus labels is around 70%, on the full set of labels permitted in the annotation scheme. The level of agreement is on par with similar numbers reported in other paraphrase studies (Dolan and Brockett 2005; Creutz 2018). The agreement measures when calculated separately for each annotator vary between 64% and 76%, the most common disagreements being between the semantically nearest labels (i. e., labels *3* and *4*</>, or labels *4*</> and *4*), or whether to include or not include the rare additional flags *s* or *i*.

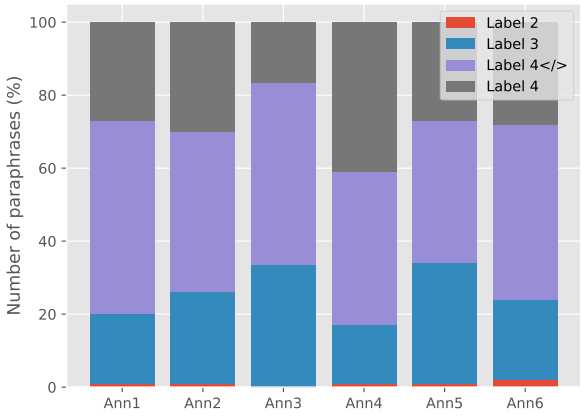

Fig. 1 Label frequencies illustrated separately for the six annotators using the same subtitle subset of the corpus as in Table 1

5 Conclusions and Results of the Pilot Project

The project resulted in a high quality corpus of Finnish paraphrases including a total of 104,645 manually classified pairs, 91,604 being naturally occurring pairs directly extracted from the source documents, while 13,041 are produced through manual rewriting. The manual extraction method presented in the article both skews the label distribution towards true paraphrases ensuring efficient use of human resources (98% being labeled positive) as well as preserves the original document context, making this the first released corpus of paraphrasing in context. The contextual information is used in Kanerva et al. (2021b), where we present a novel approach to paraphrase detection by framing the task as detecting the target paraphrase span from the given document, a similar setting as used in question answering. In addition to the actual corpus, the project also provided models trained for paraphrase classification and fine-tuned sentence representations.

All resources presented in this article are available through the European Language Grid[3] and also on the TurkuNLP website[4] under the CC-BY-SA license.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. In addition, this work was supported by the Academy of Finland and the Digicampus project. Computational resources were provided by CSC – IT Center for Science.

[3] https://live.european-language-grid.eu/catalogue/corpus/7754

[4] https://turkunlp.org/paraphrase.html

References

Creutz, Mathias (2018). "Open Subtitles Paraphrase Corpus for Six Languages". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 1364–1369.

Dolan, William B. and Chris Brockett (2005). "Automatically Constructing a Corpus of Sentential Paraphrases". In: *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, pp. 9–16.

Glockner, Max, Vered Shwartz, and Yoav Goldberg (2018). "Breaking NLI Systems with Sentences that Require Simple Lexical Inferences". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 650–655. DOI: 10.18653/v1/P18-2103. URL: https://aclanthology.org/P18-2103.

Kanerva, Jenna, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, et al. (2021a). "Annotation Guidelines for the Turku Paraphrase Corpus". In: *arXiv preprint arXiv:2108.07499*.

Kanerva, Jenna, Hanna Kitti, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter (2021b). "Semantic Search as Extractive Paraphrase Span Detection". In: *arXiv preprint arXiv:2112.04886*.

McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334.

Tsuchiya, Masatoshi (2018). "Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA, pp. 1506–1511. URL: https://aclanthology.org/L18-1239.