European Language Grid – META-FORUM 2020
Filip Ginter (University of Turku, Finland)

ELG Pilot Project:

Textual paraphrase dataset for deep language modelling

Textual paraphrase dataset for deep language modelling

TURKUNLP .ORG

UNIVERSITY OF TURKU

Paraphrase

Paraphrase: “Two statements having the same contextual meaning, using a different wording”

“Calls from one EU country to another will have a price limit of 19 cents”

“EU-internal calls will cost no more than 19 cents”

- Deep language models should be able to model meaning and **encode mutual paraphrases into similar representations**
- **Paraphrase is a natural task for NLU model training**
- Human judgements needed to identify the truly **interesting cases which have little lexical overlap**

home / JennaS / rew-batch-01428-part-1.json / 0

JennaS

full context

Orig

Carter... Mitä tapahtuisi jos vain irrottaisimme töpselin?

Orig

Carter mitä tapahtuisi, jos vain... vedämme johdon irti?

As identified in the source document

Label 4>

Class

- 4 : Paraphrase
- 4< : Upper is more general
- 4> : Lower is more general
- 3 : Paraphrase here but not in general
- 2 : Related but not paraphrase
- 1 : Unrelated
- x : Skip

- s : Style (tone or register)
- i : Diff in number, person, etc

Copy to rewrite

Wipe

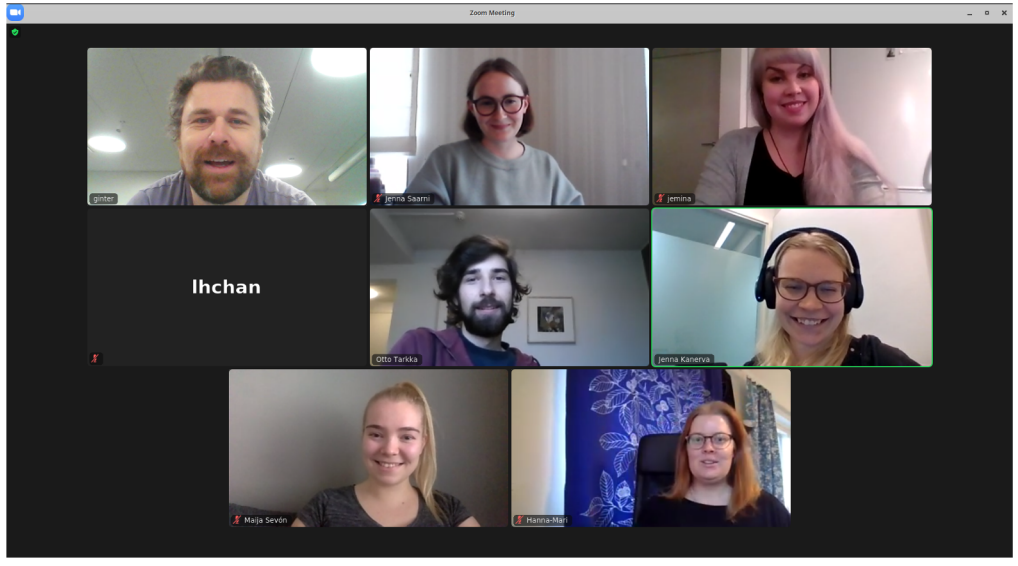
Carter... Mitä tapahtuisi jos vain irrottaisimme töpselin?

Carter mitä tapahtuisi, jos vain... vedämme töpselin irti?

Rewritten to a perfect paraphrase (where possible)

Primary outcome

- **The primary outcome is a large Finnish paraphrase dataset**
  - Pairs identified in text
  - Their context
  - Class
  - Rewritten version where applicable
- **100,000+ pairs** is the target, model training - finetuning - testing
- **Open license**, data release deadline **31.5.2021**



TURKUNLP .ORG

UNIVERSITY OF TURKU