Oversettelsesminner fra Målfrid

Ressursen inneholder oversettelsesminner fra 132 statlige domener for engelsk-bokmål, engelsk-nynorsk og bokmål-nynorsk.

Datagrunnlaget er hentet fra det såkalte Målfrid-prosjektet, der Nasjonalbiblioteket i samarbeid med Språkrådet høster statlige nettsider i forbindelse med språktilsyn. Se: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-69/

Datasettet er laget ved hjelp av det EU-finansierte verktøyet Bitextor (https://github.com/bitextor/bitextor), som er utviklet i forbindelse med ParaCrawl-prosjektet (https://paracrawl.eu/). Dataene ble høstet i slutten av 2021. Vi brukte Bitextor i versjon 8.2.

Domenene ble valgt ut på bakgrunn av tekstmengde totalt og andelen av språk på hvert domene. For at et domene skulle være med, måtte det ha mer enn 100.000 ord totalt og minst 2% av disse måtte være på ett av de tre kandidatspråkene bokmål, nynorsk eller engelsk.

Datasettet er vasket med Bitextor-verktøyet bicleaner. Alle segmenter som hadde en bicleaner-score mindre enn 0,5 og en hunalign-score mindre enn 1, ble fjernet. De anbefalte verdiene for bicleaner ligger mellom 0,5 og 0,7. Tabellen under viser antallet kandidater fra bitextor og antallet som ble igjen etter vasking med bicleaner:

språkpar	**antall kandidater**	**par etter vasking**
en-nb	2509662	145267
en-nn	255580	12268
nb-nn	6417830	50957

Vi ser en betydelig reduksjon pga. mye støy i materialet. Totalt inneholder datasettet 208 492 segmentpar. Selv etter vasking må man anta at det er en del støy i dette materialet. Datasettet tilbys derfor "som det er", uten noen form for garantier.

Format

Datasettet inneholder komprimerte TMX-filer for hvert domene/språkpar. TMX er standardformatet for oversettelsesminner i XML. Filene inneholder informasjon om språk, verdier for alignering/vasking, kildedokumenter og selve segmentene.

Veien videre

Bitextor bruker et eldre verktøy, hunalign, for å alignere tekster og setninger. Det siste året har prosjektet jobbet med å implementere støtte for store språkmodeller for å alignere tekster, som Googles Language-agnostic BERT Sentence Embedding (LABSE). Resultatet ser svært lovende ut og vi håper derfor at vi i neste utgave av dette datasettet vil kunne presentere et betydelig større antall segmentpar.

Lisens

Kontakt

Hvis du har spørsmål angående materialet, kontakt oss gjerne på sprakbanken@nb.no

Translation memories from Målfrid

This resource contains translation units from 132 public sector domains for English-Bokmål, English-Nynorsk and Bokmål-Nynorsk.

The data were taken from the so-called Målfrid project, where the National Library of Norway on behalf of the Ministry of Culture and in collaboration with the The Language Council of Norway collects and aggregates data for mapping the usage of Norwegian Bokmål and Norwegian Nynorsk in Norwegian state institutions. See: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-69/

The dataset was created using the EU-funded tool Bitextor (https://github.com/bitextor/bitextor), which has been developed within the ParaCrawl project (https://paracrawl.eu/). The data were crawled in the end of 2021. We used Bitextor version 8.2.

The domains were selected on the basis of the total amount of text and the proportion of languages within each domain. For a domain to be included, it had to comprise more than 100,000 words in total and at least 2% of these had to be in one of the three candidate languages Bokmål, Nynorsk or English.

The dataset was cleaned with the Bitextor tool bicleaner. All segments having a bicleaner score less than 0.5 and a hunalign score less than 1 were removed. The recommended values for bicleaner are between 0.5 and 0.7 (cf. Ram´ırez-S´anchez et al. 2020: 293). The table below shows the number of candidates from bitextor and the number that remained after using bicleaner:

pair	number of candidate pairs	pairs after cleaning
en-nb	2509662	145267
en-nn	255580	12268
nb-nn	6417830	50957

We see a significant reduction due to a lot of noise in the material. In total, the dataset contains 208,492 translation units. Even after cleaning, one must assume that there is some noise in this material. The dataset is therefore provided "as is", without any guarantees.

Format

The dataset is offered as compressed TMX files for each domain/language pair. TMX is the standard XML format for translation memories. The files contain information about language, scores for alignment/cleaning, URLs of the source documents and the segments themselves.

Outlook

Bitextor uses an older tool, hunalign, to align texts and sentences. In the last year the project has been working on implementing support for large language models for aligning texts, such as Google's Language-agnostic BERT Sentence Embedding (LABSE). The result looks very promising and we therefore hope that in the next edition of this data set we will be able to present a significantly larger number of segment pairs.

License

Contact

If you have questions regarding the material, please contact us at sprakbanken@nb.no