# Towards Automatic Finnish Text Simplification

Anna Dmitrieva, Jörg Tiedemann
University of Helsinki

# Finnish text simplification: current state

- A lot of Easy Finnish data exists, but not a lot of parallel corpora;
- No Finnish-specific simplification models have been made before this work.

**Goal:** using data from Yle, which is available in the Finnish Language Bank (Kielipankki), make document-aligned and sentence-aligned datasets that can be used to train simplification models. Train baseline models on the obtained data.

# The Yle News

## Tänä vuonna jaetaan 2 kirjallisuuden Nobelia

Viime vuonna jakamatta jäänyt kirjallisuuden Nobel-palkinto jaetaan tänä vuonna. Kuva: Pelle T Nilsson / AOP

Ruotsin akatemia myöntää tänä vuonna 2 kirjallisuuden Nobel-palkintoa.

Ruotsalainen media kertoo, että Ruotsin akatemia myöntää tänä vuonna myös viime vuonna jakamatta jääneen palkinnon. Palkinto jätettiin viime vuonna myöntämättä, koska Ruotsin akatemiaa häiritsi seksuaaliseen häirintään liittynyt skandaali.

Kirjallisuuden Nobel-palkinto jaettiin ensimmäisen kerran vuonna 1901. Kirjallisuuden Nobel-palkinto on suuruudeltaan noin miljoona euroa.

## Viime vuoden kirjallisuuden Nobel-palkinto myönnetään tänä vuonna

Palkinto jätettiin viime vuonna myöntämättä Ruotsin akatemian ajauduttua sekasortoon seksuaalisen häirinnän johdosta.

Kuva: Pelle T Nilsson / AOP

**PETRI BURTSOV**

5.3.2019 12:41 • Päivitetty 5.3.2019 13:29 Jaa

Ruotsin akatemia myöntää tänä vuonna kaksi kirjallisuuden Nobel-palkintoa.

Vuoden 2019 palkinnon lisäksi akatemia myöntää tänä vuonna myös viime vuonna jakamatta jääneen vuoden 2018 palkinnon.

Palkinto jätettiin viime vuonna myöntämättä Ruotsin akatemian ajauduttua sekasortoon seksuaalisen häirinnän johdosta.

Ruotsalaismediat raportoivat viime vuonna, että päätös vuoden 2018 palkinnon jakamatta jättämisestä johtui osaksi Nobel-säätiön painostuksesta.

Nobel-säätiön puheenjohtaja **Lars Heikensten** on vaatinut Ruotsin akatemialta toimia, jotta luottamus siihen palkintoja jakavana tahona palautuisi.

Nobel-säätiö kertoo asiasta tiedotteessaan.

# Base data

Yle archives:

- Standard Finnish news:
  - Yle Finnish News Archive 2019-2020: http://urn.fi/urn:nbn:fi:lb-2021050401
  - Yle Finnish News Archive 2011-2018: http://urn.fi/urn:nbn:fi:lb-2017070501
- Easy Finnish news:
  - Yle News Archive Easy-to-read Finnish 2019-2020: http://urn.fi/urn:nbn:fi:lb-2021050701
  - Yle News Archive Easy-to-read Finnish 2011-2018: http://urn.fi/urn:nbn:fi:lb-2019050901
- All Yle-related data can be found here (type “yle” in the search at kielipankki.fi): https://www.kielipankki.fi/corpora/ylenews/

# The first Finnish-Easy Finnish parallel dataset

Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive **2019-2020**

- Document-aligned;
- All document pairs evaluated by a human expert after automatic alignment;
- 1919 document pairs: 1257 “positive”, 470 “negative”, and 192 “neutral”;
- License: CLARIN ACA - NC: Academic - Non Commercial Use, Attribution, No Redistribution;
- PID: http://urn.fi/urn:nbn:fi:lb-2022111625

# Example entry

| index_in_selko | index_in_regular | selko_text | regular_text | cos_sim | status | comments |
|---|---|---|---|---|---|---|
| 3-10977882_4 | 3-10976402 | Keskustalainen ministeri Annika Saarikko on saanut vauvan. Saarikko synnytti pojan viime yönä. Saarikko kertoo, että synnytys oli rankka ja kesti | Äitiysvapaalla oleva keskustaministeri Annika Saarikko on saanut pojan. Syyspoika syntyi viime yönä, Annika Saarikko (kesk.) kertoo tviitissään. Äityisvapaalla olevan Saarikon mukaan synnytys oli pitkä ja rankka. | 0,85956 | Positive | |
| 3-10753311_2 | 3-10752334 | Uudella eduskunnalla on tänään ollut ensimmäinen täysistunto. Istunnossa valittiin uudet puhemiehet. Eduskunnan puhemies on toistaiseksi suurimman puolueen puheenjohtaja eli SDP:n Antti Rinne. Ensimmäinen varapuhemies on perussuomalaisten kansanedustaja Juho Eerola. Toinen varapuhemies on kokoomuksen Paula | Istunnossa valitaneen eduskunnan puhemieheksi SDP:n puheenjohtaja Antti Rinne. Uusi eduskunta kokoontuu ensimmäiseen täysistuntoonsa puoliltapäivin. Istunnossa valitaan eduskunnan puhemies ja varapuhemiehet. Puhemieheksi valittaneen suurimman puolueen SDP:n puheenjohtaja Antti Rinne . | 0,85948 | Positive | Easy Finnish article has phrases that are not mentioned in original text |
| 3-11577969_0 | 3-11577656 | USA:n presidentti Donald Trump on siirretty sairaalaan. Trumpilla on koronatartunta. Trump ei ole saanut vakavia oireita. Lääkärit kuitenkin haluavat, että Trump | Valkoisen talon lääkärin mukaan Trump siirrettiin sairaalaan varotoimenpiteenä ja hänen oireensa ovat lieviä. Yhdysvaltain presidentti Donald Trump [on siirretty sairaalaan | 0,85844 | Positive | |
| 3-11017128_1 | 3-11015591 | Nobelin rauhanpalkinnon saa Etiopian pääministeri Abiy Ahmed . Palkinnon myöntää Norjan Nobel-komitea. Nobel-komitea sanoo, että Abiy Ahmed on rakentanut rauhaa ja | Rauhansopimuksen lisäksi Nobel-komitea kiittelee Abiy Ahmedia monista hänen aloittamistaan uudistuksista Etiopiassa. Etiopian pääministeri Abiy Ahmed on vuoden 2019 Nobelin rauhanpalkinnon | 0,85819 | Positive | |

# Tasks

1. Document alignment of earlier articles (2014-2018);
2. Sentence alignment of the entire collection;
3. Creating simplification models.

# Document alignment

**Tänä vuonna jaetaan 2 kirjallisuuden Nobelia**

Viime vuonna jakamatta jäänyt kirjallisuuden Nobel-palkinto jaetaan tänä vuonna. Kuva: Pelle T Nilsson / AOP

Ruotsin akatemia myöntää tänä vuonna 2 kirjallisuuden Nobel-palkintoa.

Ruotsalainen media kertoo, että Ruotsin akatemia myöntää tänä vuonna myös viime vuonna jakamatta jääneen palkinnon. Palkinto jätettiin viime vuonna myöntämättä, koska Ruotsin akatemiaa häiritsi seksuaaliseen häirintään liittynyt skandaali.

Kirjallisuuden Nobel-palkinto jaettiin ensimmäisen kerran vuonna 1901. Kirjallisuuden Nobel-palkinto on suuruudeltaan noin miljoona euroa.

**Viime vuoden kirjallisuuden Nobel-palkinto myönnetään tänä vuonna**

Palkinto jätettiin viime vuonna myöntämättä Ruotsin akatemian ajauduttua sekasortoon seksuaalisen häirinnän johdosta.

Kuva: Pelle T Nilsson / AOP

**"Toive ja paluu myyttiseen aikaan" – naisviha näkyy netin keskustelupalstoilla, joissa haaveillaan menneestä maailmasta**

Vihamielinen suhtautuminen naisiin on noussut keskusteluun Valkeakosken surman myötä.

PETRI [illegible]TSOV


15-vuotiaan tytön kuolema Valkeako[illegible] herättänyt kysymyksiä naisvihasta.

PETRA NYKÄNEN

**Yli 50 vuotta markkinoilla ollut lääke ei toimikaan – kysyimme miksi se on edelleen myynnissä**

Kyypakkauksen lääkkeen luultiin ennen tehoavan. Nykytietämyksen mukaan lääkkeellä ei ole vaikutusta käärmeen pureman aiheuttamiin vaurioihin.


Apteekeissa on myynnissä kaksi tuotetta, joita markkinoidaan käärmeen puremiin. Kummassakin vaikuttavana aineena on hydrokortisoni, joka ei tutkimusten mukaan auta kyyn puremaan. Kuva: Tero Kyllönen / Yle

# Document alignment: rules

- Only articles that came out on Yle and Easy Finnish Yle **on the same day** can be matched
  - Most articles, especially those deemed more important by the editors, come on air on Easy Finnish Yle within 24 hours after coming out on regular Yle;
  - Some articles can be translated after a couple days or (rarely) longer
- Only articles with **same subjects** (thematic tags) can be matched
- Easy Finnish articles are matched to Standard Finnish articles
  - Sometimes Swedish Yle news get translated into Easy Finnish
- Alignments are exclusive

# Document alignment

- Approaches:
  a. Document embedding is an average of its sentence embeddings;
  b. Vecalign document embeddings with or without candidate re-scoring.
- In both approaches, we used pre-trained embeddings to model sentences;
  a. LASER, LaBSE, MPNet, DistilUSE
- Both truncated and full embeddings were used;
- The results were compared to the human-evaluated data.

# Vecalign document embeddings

- Candidate generation

Find a fixed number, K (in our case, 5), of target documents as potential matches for each source document. Document embeddings are created by concatenating several sub-vectors, each emphasizing a different section of the document. Each sub-vector is the sum of the sentence embeddings for the entire document, after embeddings are weighted to emphasize a given region of the document and to de-emphasize boilerplate text.

- Candidate re-scoring

Align the sentences in the candidate document pairs and score the quality of the resulting sentence alignments in order to judge whether the proposed document pair appears to be a good translation pair.

Exploiting Sentence Order in Document Alignment (Thompson & Koehn, EMNLP 2020)

Vecalign document embeddings

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | **1,000** | **0,843** | **0,826** | **1,000** | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

Table 1: Document alignment with Vecalign document embeddings (Thompson and Koehn, 2020). "Sup-1" is support-1, the number of pairs deemed "positive" (true pairs) under the current threshold. "Sup-2" is support-2, the number of document pairs in the predicted sample that match the document pairs in the true dataset.

Sentences' embeddings average

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Cos. sim.↑** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| LaBSE | 0,68 | 0,717 | **1,000** | 0,835 | **0,812** | **1,000** | **0,896** | 1613 | 1613 |
| MPNet | 0,55 | 0,701 | **1,000** | 0,825 | 0,797 | **1,000** | 0,887 | 1628 | 1628 |
| DistilUSE | 0,47 | 0,689 | **1,000** | 0,816 | 0,783 | **1,000** | 0,878 | 1710 | 1710 |
| LASER | 0,80 | **0,719** | **1,000** | **0,836** | 0,810 | **1,000** | 0,895 | 1574 | 1575 |

Table 2: Document alignment by comparing averaged sentence embeddings.

# Document alignment: evaluation

Distance: cosine distance threshold between candidate sequences. Pairs with cosine distance below this threshold are considered good matches

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | **1,000** | **0,843** | **0,826** | **1,000** | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

# Document alignment: evaluation

"Positive" and "neutral" pairs in the reference dataset are considered "true"

Only "positive" pairs are "true"

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | **1,000** | **0,843** | **0,826** | **1,000** | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

# Document alignment: evaluation

Number of pairs deemed "positive" (true pairs) under the current threshold

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | **1,000** | **0,843** | **0,826** | **1,000** | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

# Document alignment: evaluation

Number of document pairs in the predicted sample that match the document pairs in the true dataset

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | **1,000** | **0,843** | **0,826** | **1,000** | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

# Document alignment: evaluation

Winning approach: Vecalign document embeddings with LASER pre-trained vectors and without candidate rescoring.

| | | Strict | | | Lax | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Embeddings** | **Dist↓** | **p** | **r** | **f1** | **p** | **r** | **f1** | **sup-1** | **sup-2** |
| **Truncated embeddings** | | | | | | | | | |
| LaBSE-128 | 0,9 | 0,723 | **1,000** | 0,840 | 0,820 | **1,000** | **0,901** | 1439 | 1439 |
| MPNet-128 | 0,9 | 0,718 | **1,000** | 0,836 | 0,814 | **1,000** | 0,898 | 1453 | 1453 |
| DistilUSE-128 | 0,9 | 0,712 | **1,000** | 0,832 | 0,808 | **1,000** | 0,894 | 1473 | 1473 |
| LASER-128 | 0,9 | **0,730** | 0,993 | **0,841** | **0,823** | 0,993 | 0,900 | 1319 | 1329 |
| **Full-size embeddings** | | | | | | | | | |
| LaBSE | 0,9 | 0,728 | **1,000** | 0,842 | 0,824 | **1,000** | 0,903 | 1424 | 1424 |
| MPNet | 0,9 | 0,717 | **1,000** | 0,835 | 0,814 | **1,000** | 0,897 | 1473 | 1473 |
| DistilUSE | 0,9 | 0,711 | **1,000** | 0,831 | 0,807 | **1,000** | 0,893 | 1504 | 1504 |
| LASER | 0,9 | **0,729** | 1,000 | **0,843** | **0,826** | 1,000 | **0,905** | 1188 | 1188 |
| **After candidate rescoring** | | | | | | | | | |
| LaBSE rescored | n/a | 0,701 | **1,000** | 0,824 | **0,805** | **1,000** | **0,892** | 743 | 743 |
| LASER rescored | n/a | **0,706** | **1,000** | **0,828** | 0,803 | **1,000** | 0,891 | 595 | 595 |

# Sentence alignment

## Tänä vuonna jaetaan 2 kirjallisuuden Nobelia

Viime vuonna jakamatta jäänyt kirjallisuuden Nobel-palkinto jaetaan tänä vuonna. Kuva: Pelle T Nilsson / AOP

Ruotsin akatemia myöntää tänä vuonna 2 kirjallisuuden Nobel-palkintoa.

Ruotsalainen media kertoo, että Ruotsin akatemia myöntää tänä vuonna myös viime vuonna jakamatta jääneen palkinnon. Palkinto jätettiin viime vuonna myöntämättä, koska Ruotsin akatemiaa häiritsi seksuaaliseen häirintään liittynyt skandaali.

Kirjallisuuden Nobel-palkinto jaettiin ensimmäisen kerran vuonna 1901. Kirjallisuuden Nobel-palkinto on suuruudeltaan noin miljoona euroa.

## Viime vuoden kirjallisuuden Nobel-palkinto myönnetään tänä vuonna

Palkinto jätettiin viime vuonna myöntämättä Ruotsin akatemian ajauduttua sekasortoon seksuaalisen häirinnän johdosta.

Kuva: Pelle T Nilsson / AOP

**PETRI BURTSOV**

5.3.2019 12:41 • Päivitetty 5.3.2019 13:29 Jaa

Ruotsin akatemia myöntää tänä vuonna kaksi kirjallisuuden Nobel-palkintoa.

Vuoden 2019 palkinnon lisäksi akatemia myöntää tänä vuonna myös viime vuonna jakamatta jääneen vuoden 2018 palkinnon.

Palkinto jätettiin viime vuonna myöntämättä Ruotsin akatemian ajauduttua sekasortoon seksuaalisen häirinnän johdosta.

Ruotsalaismediat raportoivat viime vuonna, että päätös vuoden 2018 palkinnon jakamatta jättämisestä johtui osaksi Nobel-säätiön painostuksesta.

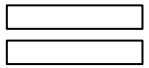
Nobel-säätiön puheenjohtaja **Lars Heikensten** on vaatinut Ruotsin akatemialta toimia, jotta luottamus siihen palkintoja jakavana tahona palautuisi.

Nobel-säätiö kertoo asiasta tiedotteessaan.

# Sentence alignment: criteria

- One-to-one, one-to-many, many-to-one, many-to-many alignments are possible;
- Crossing alignments/crossing links are allowed:

| Doc 1 | Doc 2 |
|---|---|
| A | a |
| B | b |
| C | c |
| D | d |

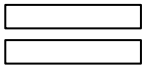


✅ BC -> a
✅ A -> d

# Sentence alignment: criteria

- Sentences within an alignment are consecutive;
- Alignments are exclusive;

| Doc 1 | Doc 2 |
|---|---|
| A | a |
| B | b |
| C | c |
| D | d |



❌ AC -> b
❌ A -> a, A -> b

- If the method uses embeddings, it should be possible to change the embedding model.

# Sentence alignment: strategies

- Aligners:
  - Vecalign: searching for similar sentence vectors with a scoring method based on cosine similarity;
  - Bertalign: searching for similar sentences based on BERT embeddings. First 1-to-1 alignments are found, then longer valid alignments;
  - Cosine similarity matrix: a simple method that satisfies all our criteria;
  - Baseline: MASSAlign - TF-IDF-based text comparison using a similarity matrix.
- Embedding models: LASER, LaBSE, MPNet, DistilUSE

# What is “cosine similarity matrix”?

Given document 1 with sentences ABC and document 2 with sentences abcd, take cosine similarities between each pair of individual sentences and each combination of <= n sentences (in our case, n = 3).

Strategy: find maximum similarity, eliminate this alignment, repeat.

| | a | ab | abc | b | bc | bcd | c | cd | d |
|---|---|---|---|---|---|---|---|---|---|
| A | 0,508 | 0,564 | 0,717 | 0,547 | 0,638 | 0,568 | 0,076 | 0,269 | 0,010 |
| AB | 0,245 | 0,167 | 0,638 | 0,606 | 0,270 | 0,322 | 0,298 | 0,225 | 0,120 |
| ABC | 0,139 | 0,897 | 0,809 | 0,369 | 0,302 | 0,295 | 0,168 | 0,218 | 0,815 |
| B | 0,815 | 0,434 | 0,896 | 0,508 | 0,924 | 0,449 | 0,482 | 0,212 | 0,975 |
| BC | 0,821 | 0,980 | 0,246 | 0,204 | 0,563 | 0,181 | **0,997** | 0,423 | 0,682 |
| C | 0,917 | 0,642 | 0,182 | 0,910 | 0,029 | 0,596 | 0,510 | 0,804 | 0,951 |

BC = c

# What is “cosine similarity matrix”?

Given document 1 with sentences ABC and document 2 with sentences abcd, take cosine similarities between each pair of individual sentences and each combination of <= n sentences (in our case, n = 3).

Strategy: find maximum similarity, eliminate this alignment, repeat.

| | a | ab | abc | b | bc | bcd | c | cd | d |
|---|---|---|---|---|---|---|---|---|---|
| A | 0,508 | **0,564** | | 0,547 | | | | | 0,010 |
| AB | | | | | | | | | |
| ABC | | | | | | | | | |
| B | | | | | | | | | |
| BC | | | | | | | | | |
| C | | | | | | | | | |

BC = c
A = ab

# Sentence alignment: evaluation

- Evaluated on 50 manually aligned documents;
- Best alignment results were obtained with Vecalign using LASER embeddings;
- The resulting dataset only contains sentence pairs with cosine distance <= 0.65.

| | Strict | | | Lax | | |
|---|---|---|---|---|---|---|
| **Embeddings** | **p** | **r** | **f1** | **p** | **r** | **f1** |
| **Vecalign** | | | | | | |
| LaBSE | 0,786 | 0,305 | 0,439 | 0,847 | 0,7 | 0,766 |
| MPNet | 0,788 | 0,3 | 0,435 | **0,852** | 0,704 | **0,771** |
| DistilUSE | 0,789 | 0,314 | 0,449 | 0,841 | 0,65 | 0,733 |
| LASER | **0,801** | **0,426** | **0,556** | 0,839 | 0,668 | 0,744 |
| **Bertalign** | | | | | | |
| LaBSE | 0,745 | 0,179 | 0,289 | 0,813 | 0,596 | 0,688 |
| MPNet | 0,77 | 0,269 | 0,399 | 0,822 | 0,601 | 0,694 |
| DistilUSE | 0,738 | 0,166 | 0,271 | 0,802 | 0,561 | 0,66 |
| LASER | 0,694 | 0,081 | 0,145 | 0,749 | 0,408 | 0,528 |
| **Cos. sim. matrix** | | | | | | |
| LaBSE | 0,34 | 0,368 | 0,353 | 0,585 | **0,726** | 0,648 |
| MPNet | 0,304 | 0,305 | 0,304 | 0,607 | 0,691 | 0,646 |
| DistilUSE | 0,301 | 0,336 | 0,318 | 0,514 | 0,632 | 0,567 |
| LASER | 0,311 | 0,269 | 0,288 | 0,601 | 0,614 | 0,608 |
| **MASSAlign** | | | | | | |
| n\a | 0,57 | 0,238 | 0,335 | 0,774 | 0,318 | 0,451 |

# New dataset 1

Parallel Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive **2014-2018**

- Document-aligned with Vecalign using LASER embeddings;
- 7004 document pairs;
- License: CLARIN ACA - NC: Academic - Non Commercial Use, Attribution, No Redistribution;
- PID: http://urn.fi/urn:nbn:fi:lb-2024011701

# New dataset 2

Parallel Sentence Aligned Corpus of Finnish and Easy-to-read Finnish from the Yle News Archive **2014-2020**

- Sentence-aligned with Vecalign using LASER embeddings;
- 11944 sentence pairs from 8261 document pairs;
- Includes 50 manually aligned documents;
- License: CLARIN ACA - NC: Academic - Non Commercial Use, Attribution, No Redistribution;
- PID: http://urn.fi/urn:nbn:fi:lb-2024011703

# Modeling

Goal: create baseline models for Finnish text simplification using newly obtained data.

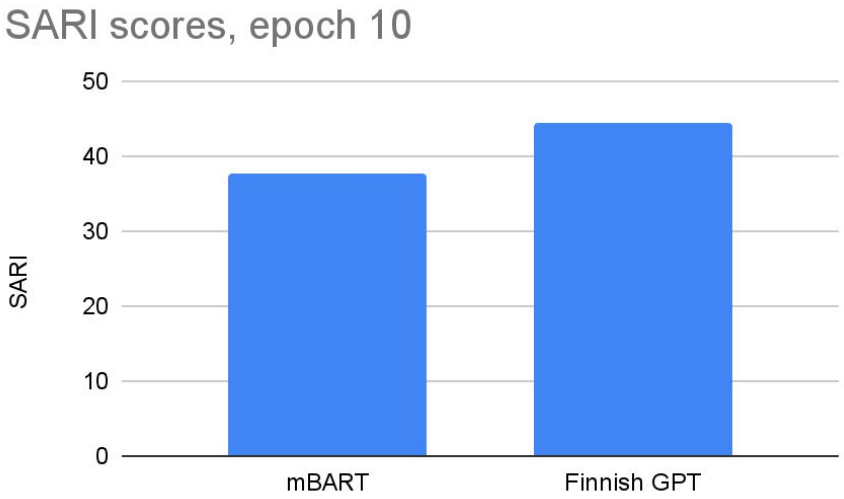
Used models:

- mBART (fairseq)
  - https://huggingface.co/Helsinki-NLP/fin-simple-mBART
- Finnish GPT XL (transformers)
  - https://huggingface.co/Helsinki-NLP/simple-finnish-gpt3-xl

# Models: evaluation

Metric used: **SARI** - an arithmetic average of n-gram precisions and recalls of editing operations: addition, keeping, and deletions between the source, output, and references.

SARI is widely used for evaluating text simplification. It is unable to consider grammaticality or coherence, but has a good correlation with human judgments of simplicity.



# Models: evaluation

Linguistic features of models' outputs (evaluated with EASSE):

1. How compressed is the output in comparison to the original;
2. Average number of sentence splitting performed by the system;
3. Levenshtein similarity between the original and the output;
4. Proportion of original sentences left untouched;
5. Proportions of added and deleted words.

The reference sentence from the dataset

| Feature | mBART | FinnGPT | Target |
|---|---|---|---|
| Compression | 0.710 | 0.680 | 0.743 |
| Sentence splits | 0.828 | 0.831 | 0.875 |
| Levenshtein | 0.782 | 0.610 | 0.559 |
| Exact copies | 0.181 | 0.036 | 0.020 |
| Additions | 0.057 | 0.297 | 0.403 |
| Deletions | 0.339 | 0.559 | 0.618 |

| | |
|---|---|
| Original | Uusi stadion korvaa Tokion vanhan olympiastadionin, joka purettiin vuonna 2015. Tuoreen urheilupyhätön rakennustyöt maksoivat noin 1,4 miljardia dollaria eli noin 1,27 miljardia euroa.<br><br>The new stadium replaces Tokyo's old Olympic stadium, which was demolished in 2015. The construction of the new sports sanctuary cost about 1.4 billion dollars, or about 1.27 billion euros. |
| Reference | Stadionin rakentaminen maksoi noin 1,27 miljardia euroa.<br><br>The construction of the stadium cost about 1.27 billion euros. |
| mBART | Uusi stadion korvaa Tokion vanhan olympiastadionin, joka purettiin vuonna 2015.<br><br>The new stadium replaces Tokyo's old Olympic stadium, which was demolished in 2015. |
| Finnish GPT | Stadion maksoi noin 1,3 miljardia euroa.<br><br>The stadium cost about 1.3 billion euros. |

Here and in the other examples the Finnish texts are from Yle.

| | |
|---|---|
| Original | Se matkaa Aurinkoa ympäröivän polttavan kuuman kaasukehän eli koronan lähituntumaan.<br><br>It travels to the close proximity of the scorching hot gas ring that surrounds the Sun, or the corona. |
| Reference | Laitteet tutkivat Auringon kaasukehää, koronaa.<br><br>The devices study the Sun's gas ring, the corona. |
| mBART | Se matkaa Aurinkoa ympäröivän kuuman kaasukehän eli koronan lähituntumaan.<br><br>It travels to the close proximity of the hot gas ring that surrounds the Sun, or the corona. |
| Finnish GPT | Se matkustaa koronaa suojaavan kaasun lähituntuma.<br><br>It travels in close contact with the gas that protects the corona. |

| | |
|---|---|
| Original | Maan ulkoministeriön perjantaina julkistaman ehdotuksen mukaan viisumia hakevan tulisi listata kaikki käyttämänsä sosiaalisen median palvelut sekä tilit, joita on käyttänyt kuluneen viiden vuoden aikana.<br><br>According to the proposal announced by the country's Ministry of Foreign Affairs on Friday, visa applicants should list all the social media services they use and the accounts they have used in the past five years. |
| Reference | Yhdysvaltojen ulkoministeriö ehdottaa, että viisuminhakijalta vaaditaan lista sosiaalisen median palveluista, joita hän käyttää.<br><br>The US Department of State is proposing that visa applicants be required to list the social media services they use. |
| mBART | Maan ulkoministeriö ehdottaa, että viisumia hakevan täytyy listata kaikki käyttämänsä sosiaalisen median palvelut ja tilit.<br><br>The country's Ministry of Foreign Affairs suggests that visa applicants must list all social media services and accounts they use. |
| Finnish GPT | Suomen ulkoministeriö sanoo, että viisumin hakijan pitää tietää kaikki sosiaalisen median palvelut sekä tilit.<br><br>The Ministry of Foreign Affairs of Finland says that visa applicants must know all social media services and accounts. |

# Thank you for your attention!

Paper: https://aclanthology.org/2024.determit-1.4.pdf