Automatic text simplification of Russian texts using control tokens

Anna Dmitrieva, University of Helsinki

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Motivation

- Simplified texts often need to be tailored to a specific group of readers;
- Therefore, we’d like to learn to control the linguistic properties of simplified texts.

Idea

- The source sentence gets augmented with special control tokens:

  *<CEFRgrade_0> <LevSim_0.4> <NbChars_1.15> Погода на завтра: преимущественно без осадков.**

  **Weather for tomorrow: mostly rainless.*

- The model learns the meanings of the tokens, thus allowing for direct output control.

Dataset

Dataset	Train	Dev	Test
Paraphraser.ru	338865	37652	7638
Opusparcus	103186	11465	2405
RuSimpleSentEval (dev)	2570	285	59
RuAdapt - literature	8530	948	169
RuAdapt - encyclopedic	2041	227	50
RuAdapt - fairytales	135	15	4
Total	**455327**	**50592**	**10325**
RuSimpleSentEval held out public test set			**3398**

Control tokens

- **NbChars**: the ratio between the lengths of target and source sentences in characters, represents the amount of compression
- **LevSim**: the Levenstein ratio between source and target sentences, represents the amount of paraphrasing
- **DepTreeDepth**: the ratio between the syntactic tree depths of target and source sentences, represents the syntactic complexity. Calculated using DeepPavlov.
- **CEFRgrade**: the CEFR (Common European Framework of Reference) grade level of the target sentence: from A1 to C2. Represent sentence complexity on multiple levels. Calculated using Textometr.

How are the tokens preprocessed?

The tokens are appended to the source sentence before preprocessing. So, for example, the sentence:

<CEFRgrade_1> <LevSim_0.4> <NbChars_1.25> Большинство людей так и живут.

Turns into this after SentencePiece tokenization:

_< CE FR grade _1 > _< Le v Sim _ 0.4 > _< N b Char s _1 . 25 > _Большинство _людей _так _и _живут .

All tokens except CEFRgrade are bucketed ratios and have 40 unique values from 0.05 to 2.0.

Other preprocessing

- If the source sentence's CEFR grade level happened to be lower than the target's, the sentences were swapped;
- Pairs with sentences shorter than 5 tokens were removed;
- ParaPhraser.ru and Opusparcus were additionally cleaned to avoid NER confusion. Pairs of sentences where the target had named entities not present in source were removed. NER recognition was done using the Natasha toolkit.

Models: choosing an architecture

- mBART cc25: mBART model with 12 encoder and decoder layers trained on 25 languages' monolingual corpus. Source: fairseq.
- T5: a version of the Google multilingual T5 with only Russian and some English embeddings left. Source: cointegrated/rut5-base @ Huggingface.

Models: choosing an architecture

Test set	mBART	T5
General	**44,3776**	40,781
RSSE	33,3876	**35,2519**

Table 1. Highest SARI scores for models with no control tokens.

Test set	mBART	T5
General test, true tags	**53,9269**	38,9376
General test, $NbChars_{0.95}$, $LevSim_{0.4}$	**43,1563**	40,0487
RSSE, $NbChars_{0.95}$, $LevSim_{0.4}$	**38,9894**	34,6402
RSSE, $NbChars_{1.0}$, $LevSim_{1.0}$	15,944	**35,1672**

Table 2. Highest SARI scores for models with NbChars and LevSim control tokens.

Do models actually understand control tokens?

Let's give the control tokens some arbitrary values and then look at the mean values of a given attribute in output texts.

Intuitively, if a model sees a control token <NbChars_0.95>, the average ratio between the lengths of source and output sentences should be close to 0.95.

	mBART		T5	
Token	**4 epochs**	**1 epoch**	**800k steps**	**500k**
$NbChars_{0.95}$	**0,9119**	0,8496	0,8976	0,7327
$LevSim_{0.4}$	**0,4812**	0,4980	0,5336	0,6666
$NbChars_{1.0}$	**0,9999**	0,9993	0,9914	0,8590
$LevSim_{1.0}$	**0,9990**	0,9987	0,8762	0,7085

Table 3. Mean attribute values calculated between the output and the source files (RSSE public test set).

New tokens: DepTreeDepth and CEFRgrade

- DepTreeDepth: the token was not learned properly, the model kept hallucinating throughout 11 epochs;
- CEFRgrade: the model understood the token, however, it could not be combined with other well-performing tokens (NbChars and LevSim).

CEFRgrade

- Scale: from 0 to 5 (A1 to C2)

Control token CEFR level	SARI
0 (A1)	**46,4875**
1 (A2)	44,8701
2 (B1)	42,2034
3 (B2)	38,0583
Actual target CEFR level (best model)	38,9731

Table 4. SARI scores on general test set for model with a CEFR grade level control token

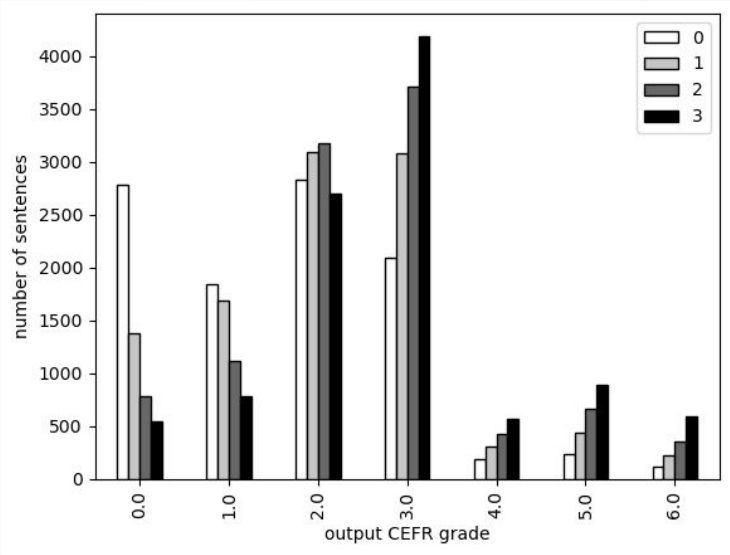


Fig. 1. Influence of the CEFR grade level control token on the output. General test set.

Conclusions

- mBART seems to be a better choice for models with control tokens, because of the training speed and the ability to comprehend the tokens' meaning;
- Some tokens and combinations of tokens damage the model's ability to learn anything;
- It was confirmed that the NbChars and LevSim tokens perform well on Russian texts;
- The CEFRgrade token also gives a small performance gain and is understood by the model, but only if used alone.

Examples

Partition	Text	Translation	Actual grade level
Source	Семья Березовского не дает согласия на закрытие уголовных дел против него	Berezovsky's family does not consent to the closure of criminal cases against him	B2
Target	Родственники Березовского не будут давать согласие на прекращение уголовных дел в отношении него	Berezovsky's relatives will not consent to the termination of criminal cases against [in relation to] him	B2
<CEFRgrade_0>	Семья Березовского не хочет закрывать дела	Berezovsky's family does not want to close cases	A1
<CEFRgrade_1>	Семья Березовского не хочет закрывать дела против него	Berezovsky's family does not want to close cases against him	A2
<CEFRgrade_2>	Семья Березовского не дает согласия на закрытие уголовных дел	Berezovsky's family does not consent to the closure of criminal cases	B1
<CEFRgrade_3>	Семья Березовского не согласна на закрытие уголовных дел против него	Berezovsky's family does not agree to the closure of criminal cases against him	B2

Source: ParaPhraser.ru

Examples

Partition	Text	Translation
Source	Андропов,военный атташе и водитель уцелели и пешком добрались до посольства.	Andropov, the military attache and the driver survived and reached the embassy on foot.
Target	Андропов вместе с военным атташе и водителем уцелели, но пешком два часа по ночному городу пробирались в посольство.	Andropov, along with the military attache and the driver, survived, but they made their way to the embassy on foot for two hours through the night city.
<NbChars_1.0> <LevSim_1.0>	Андропов,военный атташе и водитель уцелели и пешком добрались до посольства.	Andropov, the military attache and the driver survived and reached the embassy on foot.
<NbChars_0.95> <LevSim_0.4>	До посольства добрались Андропов, атташе и водитель.	Andropov, the attache and the driver reached the embassy.

Source: RuSimpleSentEval

Creating a parallel Finnish-Easy Finnish dataset from news articles

In collaboration with Aleksandra Konovalova, University of Turku

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Motivation

- Despite there being quite a few Easy Finnish resources, there seemingly has been no effort to create an aligned parallel corpus;
- This resource can help:
  - To study the simplification strategies used by simple language content providers;
  - Create and test automatic simplification tools for Finnish.

Raw data from Kielipankki

- Yle Finnish News Archive 2019-2020: http://urn.fi/urn:nbn:fi:lb-2021050401
  - Full articles
- Yle News Archive Easy-to-read Finnish 2019-2020: http://urn.fi/urn:nbn:fi:lb-2021050701
  - Short transcripts of the daily Easy Finnish radio broadcasts

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.

Copyright: Yle, URLs: https://yle.fi/uutiset/osasto/selkouutiset/tiistai_532019_radio/10674450, https://yle.fi/a/3-10673932

Article matching

- Only articles that came out on Yle and Easy Finnish Yle **on the same day** were 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) were matched
- Easy Finnish articles were matched to Standard Finnish articles
  - Sometimes Swedish Yle news get translated into Easy Finnish

Alignment

- How to align Standard and Easy Finnish articles?
- Two strategies were tried: Doc2Vec and Sentence Transformers vector matching.
- Doc2Vec:
  - GenSim implementation
  - Trained on the Yle News archives from years 2016-2018
  - Doc2Vec model trained on 202,656 articles, SentencePiece trained on 1 million randomly selected sentences.

Alignment

- Sentence Transformers:
  - SBERT implementation, model: distiluse-base-multilingual-cased-v2
  - Methodology: obtain vectors from the first 15 sentences of an article, then get an average from them.
- Vectors of Easy Finnish and Standard Finnish texts were compared pairwise. The pair that received the highest cosine similarity score was considered a true match.
- Based on a random sample of a couple of days' worth of articles, SBERT performed better and gave more representative scores.

Evaluation

- We ended up with 1920 pairs of articles with a cosine similarity score >= 0.6
- Aleksandra looked through all pairs and gave each a score:
  - "positive" - if the articles are definitely about the same topic,
  - "negative" - if the articles definitely talk about different topics,
  - "neutral" - if it cannot be definitively said whether or not the articles talk about the same topic.

Evaluation

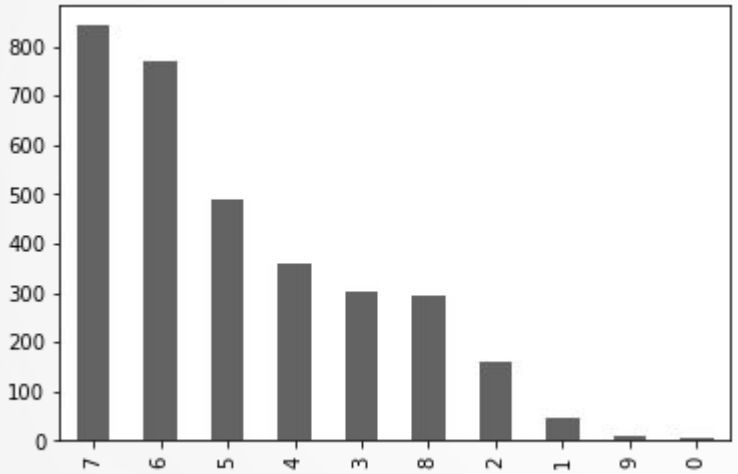


Figure 1. Distribution of cosine similarity scores across article pairs. X - approximated cosine similarity*10, y - number of pairs.

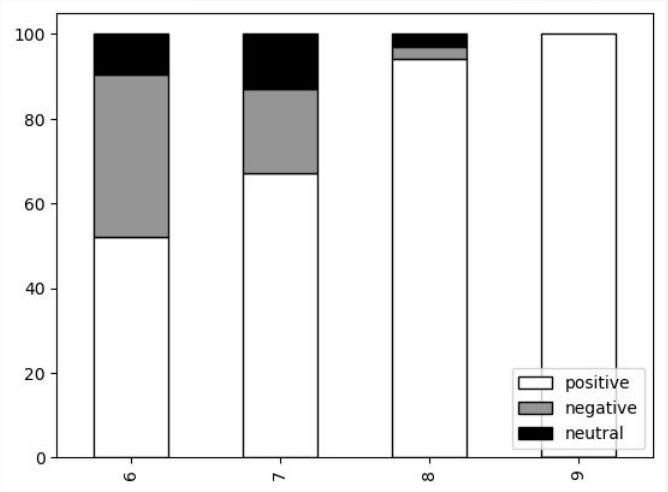


Figure 2. Percentages of labels given by the expert. X - approximated cosine similarity*10, y - percentage.

There are 1257 "positive", 470 "negative", and 192 "neutral" article pairs in the dataset.

Evaluation

- Difficult cases:
  - Hot topics spanning multiple days or months (Brexit, coronavirus, etc.), or repetitive topics (weather): because the number of articles on the same topic is high, it was difficult to establish if two similar articles are definitely talking about the same event.
  - Easy Finnish article covers a topic relevant to the entire country of Finland, but the Standard Finnish article is limited only to one particular region (e.g., Lapland).

Evaluation

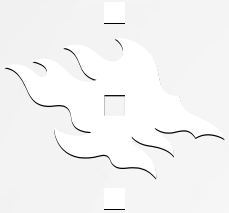
- In many ways, most Easy Finnish articles adhere to the guidelines found in Selkomittari 2.0:
  - The text mainly contains general vocabulary evaluated as familiar to the readers;
  - The text does not contain lots of long words;
  - Sentence structures are simple. For the most part, they only have one subordinate clause;
  - Clauses and sentences are mainly short;
  - …

Evaluation

- In rare cases, there are complex words and structures in Easy Finnish articles.
  - ***Amurinleopardikissapariskunta*** *on saanut pentuja Korkeasaaren eläintarhassa Helsingissä.*

    *[**A couple of Amur leopard cats** have had kittens at the Korkeasaari Zoo in Helsinki.]*
  - *Valtakunnansyyttäjä on määrännyt, {että poliisin pitää tutkia}, {onko Ano Turtiainen rikkonut lakia}, {kun hän on pilkannut mustaihoista miestä}, {joka kuoli USA:ssa}.*

    [The public prosecutor has ordered that {the police must investigate} {whether Ano Turtiainen broke the law} {when he mocked a black man} {who died in the USA}.]



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	

Check out the dataset: http://urn.fi/urn:nbn:fi:lb-2022111625
(ylenews-fi-2019-2020-selko-par-src)