Massively multilingual modeling of registers (or genres) in web-scale corpora

Veronika Laippala

TURKUNLP
.ORG

register* ~ genre

* typically used in text linguistic corpus studies, defined by Biber (1988) and Biber and Conrad (2009)

TURKUNLP
.ORG

- New registers (such as blogs)
- More variation (almost anybody can write online, no gatekeepers)
- More everything (makes some things easier)

WHAT DO THESE DOCS REPRESENT?

Article Talk

Borgio Verezzi

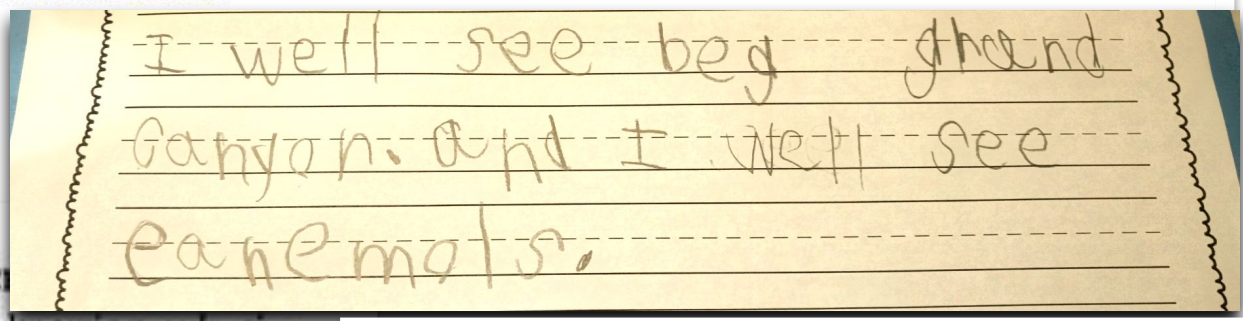
From Wikipedia, the free encyclopedia

Borgio Verezzi (Ligurian: *Bòrzi Veresso*) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about [illegible] southwest of Genoa and about 20 kilometres (12 mi) southwest of Savona.

Contents [hide]

1 Geography
2 Main sights
3 References

WIKIPEDIA
The Free Encyclopedia



"I am no longer surprised at your knowing *only* six accomplished women. I rather wonder now at your knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

WHAT DO THESE DOCS REPRESENT?

Article Talk

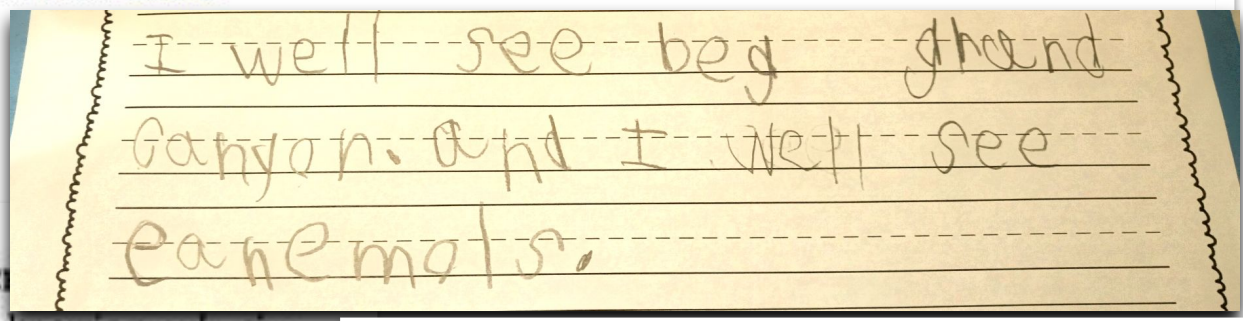
Borgio Verezzi

From Wi

Borgio
southwe

comune (municipality) in the Province of Savona in the Italian region Liguria, located about
(12 mi) southwest of Savona.

Conte
1 Geo
2 Main
3 Refe



WIKIPEDIA
The Free Encyclopedia

34 PRIDE AND PREJUDICE

"I am no longer surprised at your knowing *only* six accomplished women. I rather wonder now at your knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

WHAT DO THESE DOCS REPRESENT?

Article Talk

Borgio Verezzi

From Wi

Borgio

southwe

Conte

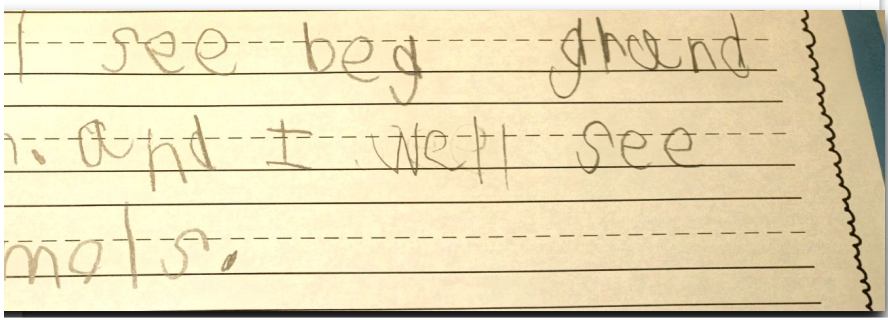
1 Geo

2 Main

3 Refe

WIKIPEDIA
The Free Encyclopedia

Savona in the Italian region Liguria, located about



34 PRIDE AND PREJU

"I am no longer surprised at yo
accomplished women. I rather w
knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

Riding the Rough Waves of Genre on the Web

Concepts and Research Questions

Marina Santini[1], Alexander Mehler[2], Serge Sharoff[3]

[1] HATII, University of Glasgow, UK
MarinaSantini.MS@gmail.com
[2] Faculty of Technology, Bielefeld University, Germany
Alexander.Mehler@uni-bielefeld.de
[3] Centre for Translation Studies, University of Leeds, UK
s.sharoff@leeds.ac.uk

2 Trying to Grasp the Ungraspable?

Although undeniably useful, the concept of genre is fraught with problems and difficulties. Social scientists, corpus linguists, computational linguists and all the computer scientists working on empirical and computational models for genre identification are well aware that one of the major stumbling blocks is the lack of a shared definition of genre, and above all, of a shared set of attributes that uncontroversially characterise genre.

The Web Library of Babel: evaluating genre collections

Serge Sharoff,[†] Zhili Wu,[†] Katja Markert[‡]

4. Conclusions

The results are relatively negative. The collections are not comparable to each other: even when categories in a collection are described in a very similar way, e.g., `FAQs` in SANTINIS and `help` in KI-04, their actual content is considerably different. When the similarity between genre collections is tested using cross-classification, the accuracy is also quite low. This shows the limits of the existing web-genre collections: if each of them is so different from any other, neither of them can be treated as a good representative for the entire web. The experiments also show that humans disagree on genre annotation of randomly selected webpages, throwing doubt on their reliability as well as on their representativeness.

The jury is still out on the best set of features useful for AGI. Character n-grams can capture many relevant generalisations not possible for other feature types, such as genre-specific prefixes and suffixes (unlike word forms), subcategories within general POS classes (unlike POS tags), but their efficiency is often related to the ability to identify *topics* exemplifying particular genres in available collections. This is the reason why the accuracy often drops when we go beyond the training set. In addition, as the datasets used

Corpus of online registers of English (CORE) (Egbert, Biber and Davies 2015)

- Unrestricted sample of the searchable web
- 48,571 documents, > 50 million words
- Manually annotated for registers
- Covers **the full range of registers and documents on the searchable Web!**
- Previous Web register corpora focused on pre-selected categories with manually chosen documents

TURKUNLP
.ORG

IN	INFORMATIONAL DESCRIPTION/EXPLANATION	NA	NARRATIVE	HI	HOW-TO/ INSTRUCTIONAL
cm	Course materials	ha	Historical article	fh	FAQ about how-to
dp	Description of a person	ma	Magazine article	ht	How-to
dt	Description of a thing	ne	News report/blog	oh	Other
en	Encyclopedia article	on	Other narrative	re	Recipe
fi	FAQ about information	pb	Personal blog	ts	Technical support
ib	Information blog	sr	Sports report	**LY**	**LYRICAL**
lt	Legal terms and conditions	ss	Short story	ol	Other
oi	Other information	tb	Travel blog	po	Poem
ra	Research article	**IP**	**INFORMATIONAL PERSUASION**	pr	Prayer
tr	Technical report	ds	Description with intent to sell	sl	Song lyrics
OP	**OPINION**	ed	Editorial	**SP**	**SPOKEN**
ad	Advertisement	oe	Other	fs	Formal speech
av	Advice	pa	Persuasive article or essay	it	Interview
le	Letter to editor	**ID**	**INTERACTIVE DISCUSSION**	os	Other
ob	Opinion blog	df	Discussion forum	ta	Transcript of video/audio
oo	Other opinion	of	Other forum	tv	TV/movie script
rs	Religious blogs/sermons	qa	Question/answer forum		
rv	Reviews	rr	Reader/viewer responses		

CORE annotation

- Each document coded by 4 annotators in MTurk
- At least 3-way agreement for 69.3% of documents for **main** register
- At least 3-way agreement for 51% of documents for **sub** register
- A part of disagreements in fact hybrid documents combining characteristics of several registers, such as personal blogs + opinion blogs

Text classification on CORE

- Question: how to assign register labels based on the four annotators' votes?

	Coder 1	Coder 2	Coder 3	Coder 4
Text 1	NARRATIVE News article	NARRATIVE News article	NARRATIVE Sports report	NARRATIVE Sports report
Text 2	INFORMATIONAL DESCRIPTION Encyclopedia article	INFORMATIONAL DESCRIPTION Research article	NARRATIVE Description of a person	NARRATIVE Historical article

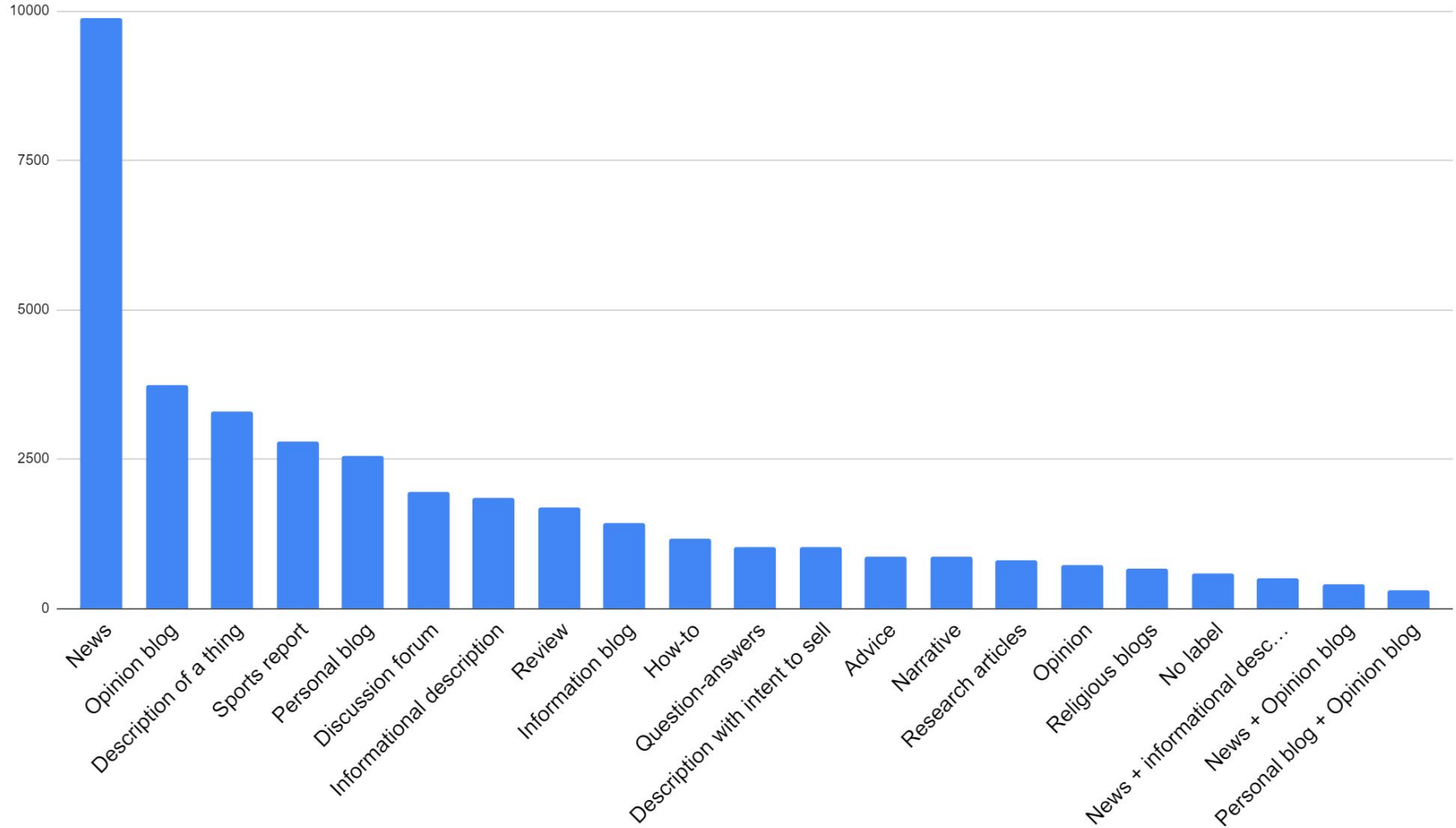
Text classification on CORE

- Solution: attribute a label always when at least two coders agreed on it

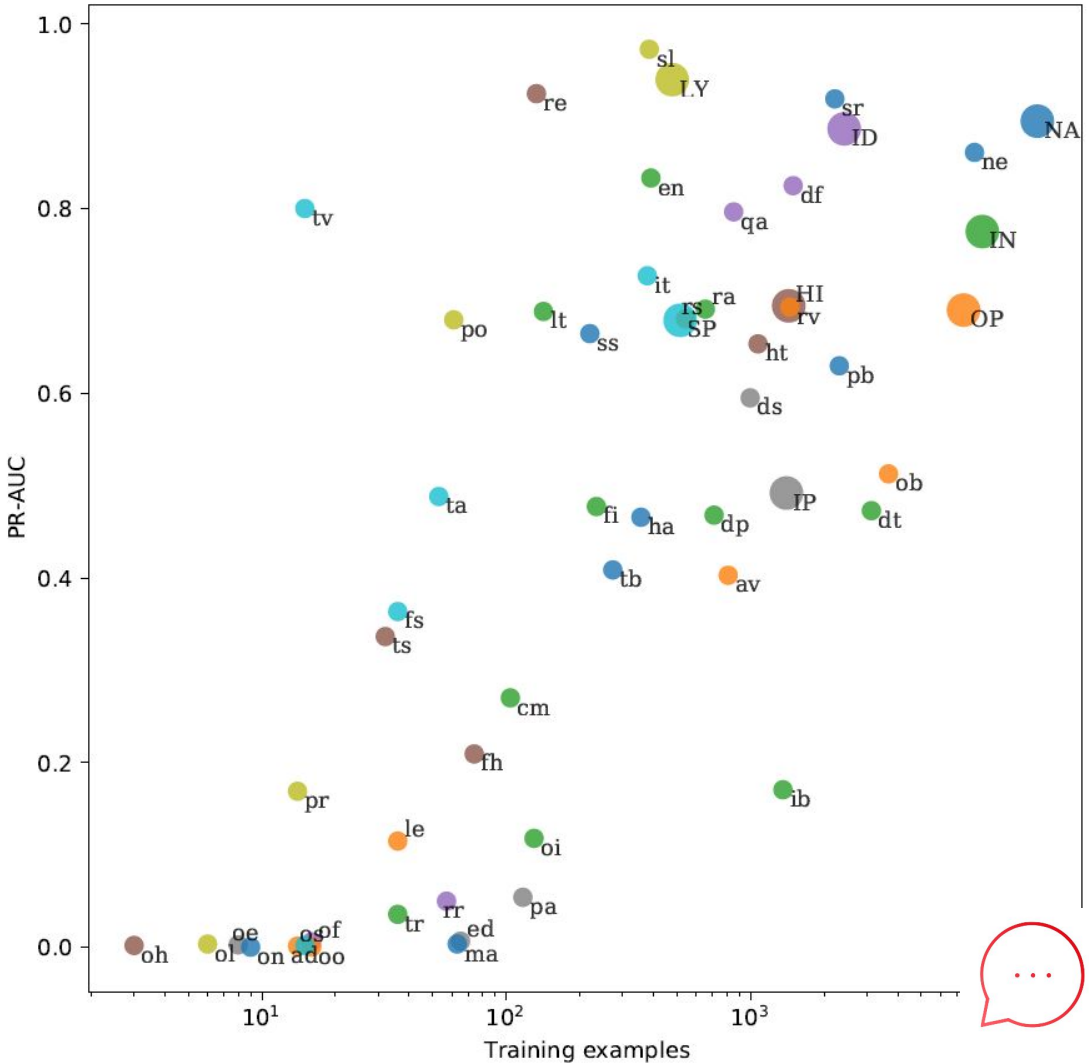
	Coder 1	Coder 2	Coder 3	Coder 4	Final labels
Text 1	NARRATIVE News article	NARRATIVE News article	NARRATIVE Sports report	NARRATIVE Sports report	NARRATIVE News article, Sports report
Text 2	INFORMATIONAL DESCRIPTION Encyclopedia article	INFORMATIONAL DESCRIPTION Research article	NARRATIVE Description of a person	NARRATIVE Historical article	INFORMATIONAL DESCRIPTION, NARRATIVE

- Experiments with
  - propagated / non-propagated labels (i.e. main register label repeated / not repeated together with the sub register)
  - multilabel/multilclass

TURKUNLP .ORG

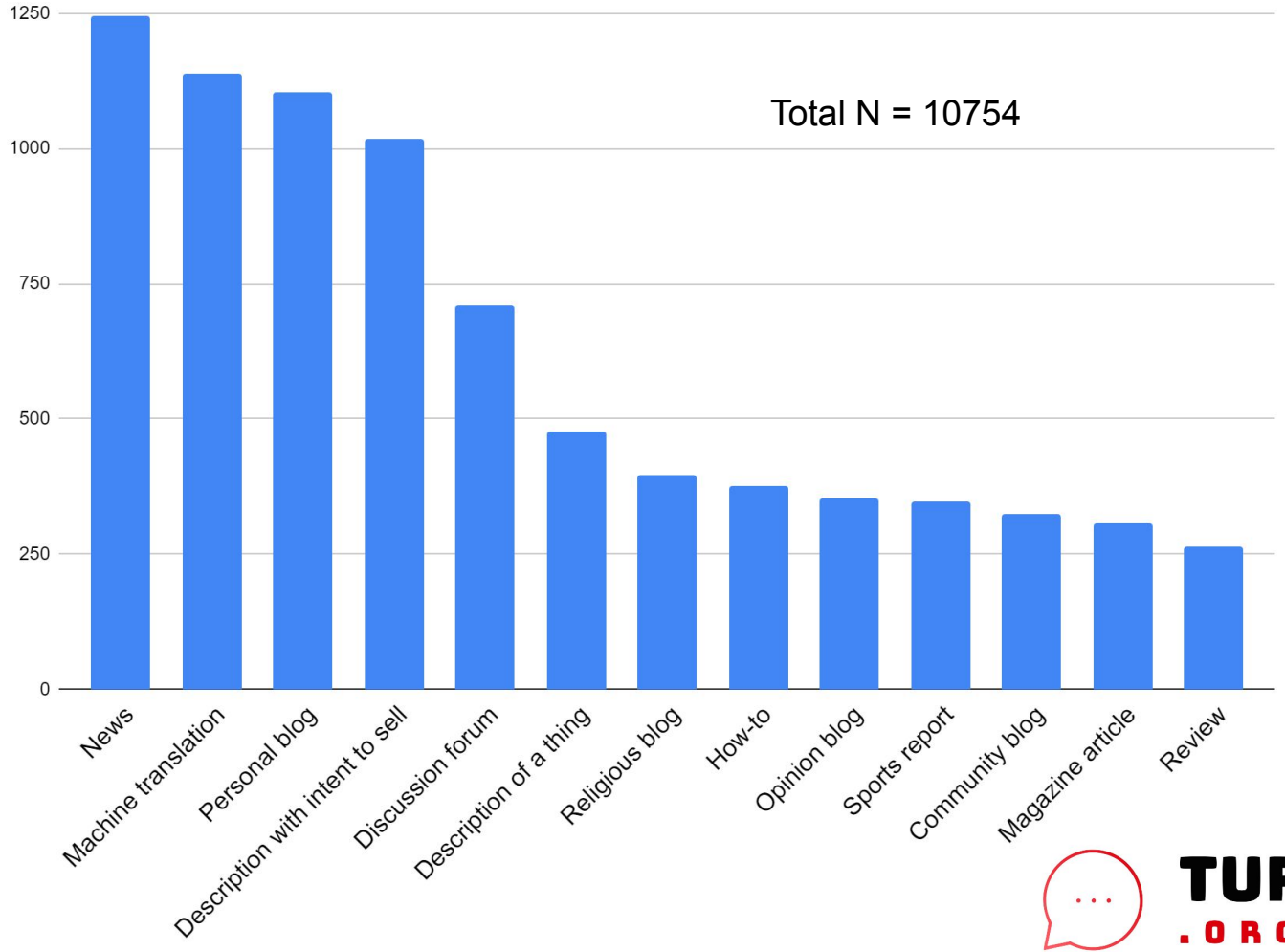

10000
7500
5000
2500
0
News
Opinion blog
Description of a thing
Sports report
Personal blog
Discussion forum
Informational description
Review
Information blog
How-to
Question-answers
Description with intent to sell
Advice
Narrative
Research articles
Opinion
Religious blogs
No label
News + informational desc…
News + Opinion blog
Personal blog + Opinion blog

Model	Dataset	Classes	F1% (SD)	Precision% (SD)	Recall% (SD)
BERT Large	Propagated	56	68 (0.33)	71 (0.66)	65 (0.07)
	Non-propagated	56	58 (0.21)	68 (1.23)	51 (0.37)
	Prop. multi-class	460	56 (0.12)	56 (0.12)	56 (0.12)
BERT Base	Propagated	56	67 (0.21)	69 (0.57)	66 (0.63)
	Non-propagated.	56	56 (0.02)	63 (0.07)	51 (0.03)
	Prop. multi-class	460	55 (0.21)	55 (0.21)	55 (0.21)
FastText	Propagated	56	62 (0.01)	56 (0.02)	69 (0.02)
	Non-propagated	56	52 (0.01)	53 (0.01)	52 (0.01)
	Prop. multi-class	460	53 (0.01)	48 (0.01)	58 (0.01)
CNN	Propagated	56	59 (0.26)	64 (0.77)	53 (0.31)
	Non-propagated	56	45 (0.39)	59 (1.18)	36 (0.86)
	Prop. multi-class	460	41 (0.08)	64 (0.69)	30 (0.74)



More languages: FinCORE

- 10,754 documents register-annotated manually, following the English CORE taxonomy
- Documents from Finnish Internet Parsebank
- Annotated by experts, IAA 79,66% across all the 41 classes (prior to discussions)
    - Sharoff et al. (2010) reported an IAA of 60%



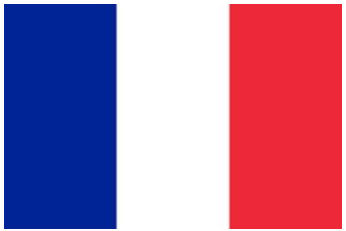
TURKUNLP
.ORG

Performance on FinCORE

classifier	(micro avg) f1-score
FinBERT	0.78
XLM-R	0.79
CNN	0.60

Multilingual modeling

- Annotated data also in French and Swedish
- 3900 documents in both languages
- Documents through a random sample of Common Crawl



TURKUNLP .ORG

Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers

Liina Repo[*†] Valtteri Skantsi[*°†] Samuel Rönnqvist[*] Saara Hellström[*]
Miika Oinonen[*] Anna Salmela[*] Douglas Biber[‡] Jesse Egbert[‡]
Sampo Pyysalo[*] Veronika Laippala[*]

Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification

Samuel Rönnqvist[*] Valtteri Skantsi[*°] Miika Oinonen[*] Veronika Laippala[*]

[*]TurkuNLP, University of Turku, Finland
[°]NSE, University of Oulu, Finland
{saanro,valtteri.skantsi,mhtoin,mavela}@utu.fi

Dataset sizes in multilingual experiments*

Lang.	Train	Dev.	Test	Total
En	33,915	4,845	9,692	48,452
Fi	1,559	222	445	2,226
Fr	909	363	546	1,818
Sv	1,093	435	654	2,182

*Experiments done on upper register labels only

TURKUNLP .ORG

	Zero-shot, from English (baseline)		Zero-shot, multilingual (ours)	
mBERT	Test		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	–	–	55.15	(2.58)
Fi	50.21	(0.74)	58.46	(0.76)
Fr	55.04	(0.66)	62.82	(1.86)
Sv	62.53	(0.78)	69.48	(0.72)
Average	–		61.48	
excl. En	55.93		63.59	
XLM-R				
Target	F1 (%)	Std.	F1 (%)	Std.
En	–	–	**63.32**	(0.25)
Fi	61.35	(1.26)	**69.60**	(0.55)
Fr	64.27	(1.58)	**72.85**	(1.74)
Sv	69.22	(1.66)	**79.49**	(0.95)
Average	–		**71.31**	
excl. En	64.95		**73.98**	

	Multilingual master model			
mBERT	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	71.32	(1.51)	66.27	(2.33)
Fi			65.27	(1.56)
Fr			69.76	(2.24)
Sv			77.92	(1.21)
Average			69.81	
excl. En			70.98	
XLM-R	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	78.20	(0.04)	72.37	(1.17)
Fi			75.05	(0.81)
Fr			78.81	(0.89)
Sv			82.36	(0.54)
Average			77.15	
excl. En			78.74	

71,31% Zero-shot avg

TURKUNLP .ORG

	Multilingual master model			
mBERT	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	71.32	(1.51)	66.27	(2.33)
Fi			65.27	(1.56)
Fr			69.76	(2.24)
Sv			77.92	(1.21)
Average			69.81	
excl. En			70.98	
XLM-R	Common dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	78.20	(0.04)	72.37	(1.17)
Fi			75.05	(0.81)
Fr			78.81	(0.89)
Sv			82.36	(0.54)
Average			77.15	
excl. En			78.74	

	Monolingual (baseline)			
mBERT	Dev.		Test	
Target	F1 (%)	Std.	F1 (%)	Std.
En	72.80	(0.21)	73.06	(0.09)
Fi	65.91	(0.85)	64.83	(1.16)
Fr	70.74	(1.67)	68.66	(0.63)
Sv	76.91	(0.45)	76.43	(0.46)
Average			70.75	
excl. En			69.97	
XLM-R				
Target	F1 (%)	Std.	F1 (%)	Std.
En	75.80	(0.12)	75.68	(0.05)
Fi	76.25	(0.45)	73.18	(1.35)
Fr	77.38	(0.51)	76.92	(0.24)
Sv	82.61	(0.37)	83.04	(0.62)
Average			**77.21**	
excl. En			**77.71**	

Then what?

- Train better models now with the full datasets and all the subregisters as well!
- While waiting for the final datasets:
  - Huggingface Big Science Project
  - Explaining black boxes to understand the predictions and registers

bigscience.huggingface.co

shar dead

BigScience

About Outcomes Events Ressources Join

A one-year long research workshop on large multilingual models and datasets

During one-year, from May 2021 to May 2022, 900 researchers from 60 countries and more than 250 institutions are creating together a very large multilingual neural network language model and a very large multilingual text dataset on the 28 petaflops Jean Zay (IDRIS) supercomputer located near Paris, France.

During the workshop, the participants plan to investigate the dataset and the model from all angles: bias, social impact, capabilities, limitations, ethics, potential improvements, specific domain performances, carbon impact, general AI/cognitive research landscape.

OSCAR

OSCAR or **O**pen **S**uper-large **C**rawled **A**ggregated co**R**pus is a huge multilingual corpus obtained by language classification and filtering of the Common Crawl corpus using the Ungoliant architecture.

Improving the quality of Oscar

1.

Rule-based filtering:
Parameters e.g.,

- Length of the document
- Average length of sentences
- Digit character ratio
- Foreign character ratio

Examples of texts filtered out

The DC War Memorial, honoring residents of Washington, DC, who fought in World War I (© Sean Pavone/Alamy)Poppies sit by names on a memorial wall during Remembrance Day in Kingston, Ontario (© Lars Hagberg/Alamy Stock Photo)ARCHIVES Select Month November 2018 (21) October 2018 (51) September 2018 (57) August 2018 (44) July 2018 (54) June 2018 (42) May 2018 (60) April 2018 (60) March 2018 (74) February 2018 (65) January 2018 (62) December 2017 (29) November 2017 (30) October 2017 (28) September 2017 (28) August 2017 (31) July 2017 (29) June 2017 (30) May 2017 (28) April 2017 (28) March 2017 (26) February 2017 (23) January 2017 (29) December 2016 (27) November 2016 (29) October 2016 (31) September 2016 (30) August 2016 (31) July 2016 (31) June 2016 (30) May 2016 (31) April 2016 (30) March 2016 (31) February 2016 (29) January 2016 (31) December 2015 (31) November 2015 (49) October 2015 (12) September 2015 (30) August 2015 (31) July 2015 (31) June 2015 (30) May 2015 (31) April 2015 (30) March 2015 (31) February 2015 (28) January 2015 (31) December 2014 (31) November 2014 (29) October 2014 (31) September 2014 (30) August 2014 (29) July 2014 (31) June 2014 (30) May 2014 (30) April 2014 (31) March 2014 (31) February 2014 (29) January 2014 (32) December 2013 (32) November 2013 (30) October 2013 (31) September 2013 (30) August 2013 (31) July 2013 (30) June 2013 (30) May 2013 (32) April 2013 (30) March 2013 (31) February 2013 (28) January 2013 (31) December 2012 (31) November 2012 (30) October 2012 (31) September 2012 (30) August 2012 (31) July 2012 (31) June 2012 (29) May 2012 (31) April 2012 (30) March 2012 (31) February 2012 (29) January 2012 (31) December 2011 (31) November 2011 (30) October 2011 (31) September 2011 (30) August 2011 (32) July 2011 (31) June 2011 (30) May 2011 (31) April 2011 (30) March 2011 (31) February 2011 (28) January 2011 (31) December 2010 (31) November 2010 (30) October 2010 (31) September 2010 (30) August 2010 (30) July 2010 (31) June 2010 (30) May 2010 (31) April 2010 (30) March 2010 (21)Feel free to join the discussion by leaving comments, and stay updated by subscribing to the RSS feed

3 person covered patio swing outdoor porch canopy bed furniture,patio swing canopy replacement sale porch swings lowes innovative technique orange county traditional,best outdoor swing with canopy ideas on hammock patio bed 3 person costco covered sets,3 person patio swing covered furniture canopy replacement sale swings strong zoom porch seat cushion repair,2 person canopy patio swing 3 with gazebo target set free home decor,porch swing with canopy gazebo outdoor covered patio deck 3 person bed replacement cushions cover for chair,3 person patio swing target canopy outdoor furniture porch replacement cushions covered bed 9 cool and cozy,patio swing canopy replacement hardware 3 person cushion outdoor garden covered double,2 person canopy patio swing 3 covered furniture wooden designs outdoor with target,3 person patio swing costco outdoors fabulous outdoor covered replacement canopy frame porch sets

Removing boilerplate remains

ok	I’ve written about the Chiricahua Mountains before, about how this mountain range that rises suddenly out of the southeast Arizona desert like an island and is, in fact, the largest of what are known as Sky Islands.
ok	I’ve talked about the birds, animals and plants that thrive nowhere else in the U.S. But with each visit to this unique area, I see, hear, feel something new that adds to my love of the place.
ok	It’s like hanging another jeweled charm from a favorite bracelet that holds cherished memories of special trips. “Phoenix Rising” can be printed on paper, canvas, note cards, t-shirts, mugs and more.
boilerplate	Be sure you’re on my subscriber list so you’ll get notified of when I’ve uploaded pieces.
boilerplate	Click here to go to the Gift Shop.Archives Select Month September 2018 (1) July 2018 (2) May 2018 (1) April 2018 (1) March 2018 (1) February 2018 (1) November 2017 (2) October 2017 (2) September 2017 (2) July 2017 (4) June 2017 (3)

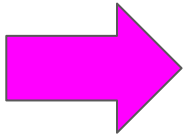
TURKUNLP .ORG

Removing boilerplate remains with junklines

	Texts	Lines total
French	1 982	84 103
Swedish	2 399	99 406
Finnish	137	4 277
English	191	6 172
Spanish	105	3 018
German	140	3 454
Total	4 954	200 430

Removing junklines

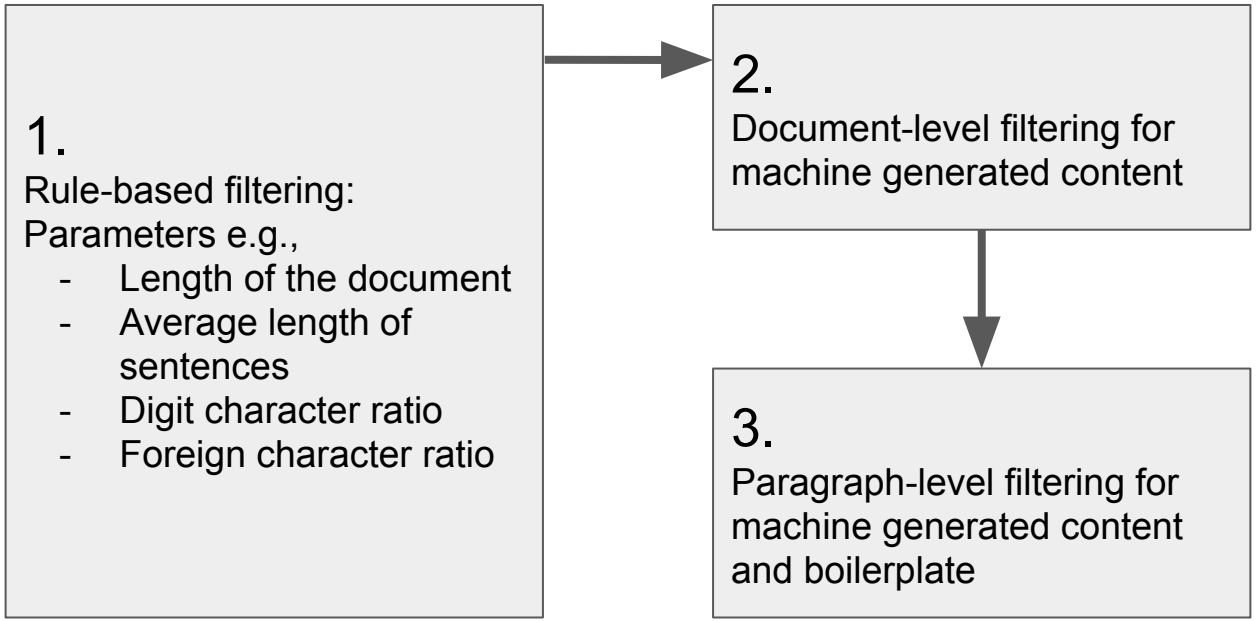
	Texts	Lines total
French	1 982	84 103
Swedish	2 399	99 406
Finnish	137	4 277
English	191	6 172
Spanish	105	3 018
German	140	3 454
Total	4 954	200 430

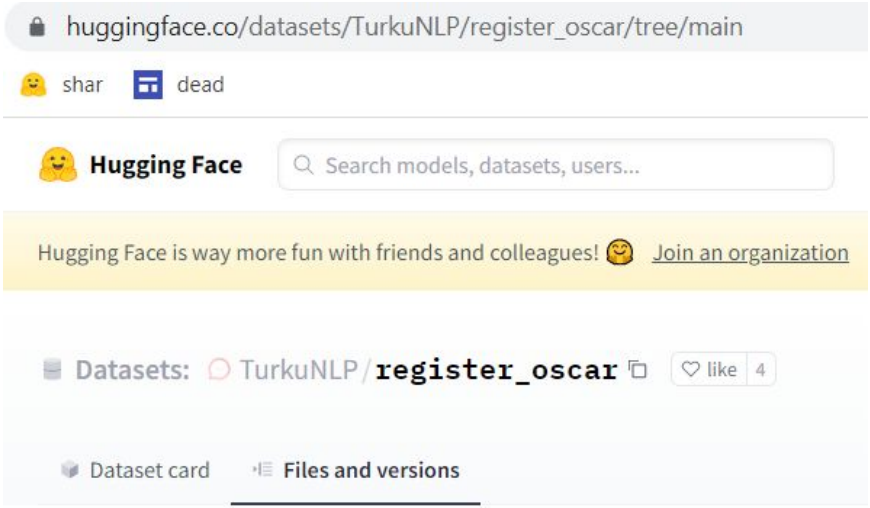


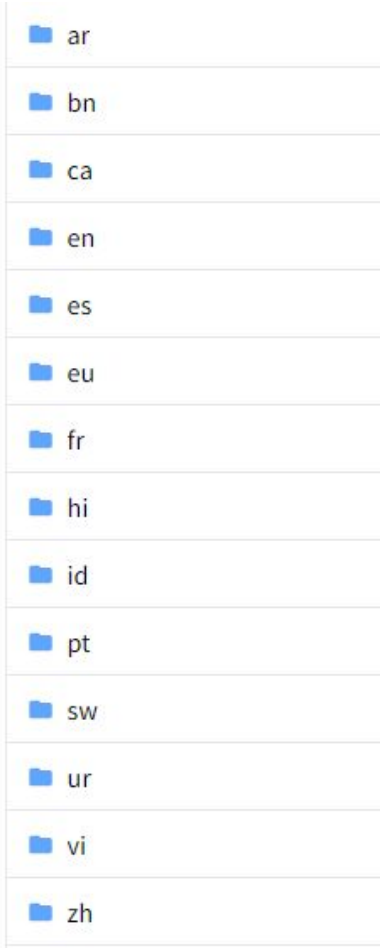
Accuracy 88%

(Multilingual LSTM based on XLM-R embeddings)

Improving the quality of Oscar







TURKUNLP .ORG

From somewhat multilingual to massively multilingual

- Collecting / collected evaluation datasets in further languages
  - **Done: Arabic, Catalan, Chinese, Indonesian, Portuguese, Turkish, Russian, Spanish, Japanese**
  - **Under annotation: Hindi, Urdu, Yoruba, Vietnamese**
- To what extent are Web registers culture-independent? How robust are the models?

Web register annotation guidelines

The annotation task consists of two steps: deciding whether to accept or reject a document, and giving a register label / labels to the accepted documents.

When to accept or reject a document

When to give a document several labels

Short list of register labels and their abbreviations

Video instructions to the annotation on Prodigy

Please note that

- You can have a look at how the document website looks like by following the document url on the annotator
- The annotation decision should, however, base on the text on the annotator

1. Is the web page **Machine translated or generated** from a template?
2. Is the web page **Lyrical**, such as songs or poems?
3. Is the web page originally spoken? (Texts composed of more than 50% spoken quotes classified as spoken)
    - If yes, is it an **Interview**?
    - If no, select **Other spoken** (e.g. formal speeches and TV/movie transcripts)
4. Is the web page **Interactive discussion** written by multiple participants in a discussion format (e.g. discussion or Q&A forum)? (Reader comments following e.g. an article or blog post are NOT included here)
5. Is the purpose of the document to narrate or report on EVENTS? If yes, select one of the following registers:
    - **News report**
    - **Sports report**
    - **Narrative blog** (e.g a travel blog or a personal blog)
    - **Other narrative** (e.g. fictional stories and magazine articles)

25 register classes

6. Is the purpose of the document to explain HOW-TO or INSTRUCTIONS?
    - If yes, is it a **Recipe**?
    - If no, select **Other how-to**. These are typically step-by-step, objective instructions on how to do something.
7. Is the purpose of the document to describe or explain INFORMATION? If yes, select one of the following registers:
    - **Encyclopedia article**
    - **Research article**
    - **Description of a thing or person**
    - **FAQ**

To conclude

- Since 2010, a lot of progress!
- Datasets
  - CORE ~50k documents
  - FinCORE ~10k
  - FreCORE + SweCORE ~4k + 4k
  - Further evaluation sets in 10+languages
- Decent register identification results
- Register-labeled Web-scale datasets

TURKUNLP
.ORG

TURKUNLP .ORG

UNIVERSITY OF TURKU