LREC 2022
June 2022, Marseille, France

The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild

Taja Kuzman, Peter Rupnik and Nikola Ljubešić
Jožef Stefan Institute, Slovenia

macocu

Co-financed by the Connecting Europe Facility of the European Union

About us

- Department of Knowledge Technologies (Jožef Stefan Institute, Slovenia): language resources and technologies for South Slavic languages
- MaCoCu project: massive high-quality monolingual and parallel web corpora for 10+ under-resourced languages

Motivation

- Annotate MaCoCu corpora with genres to:

  a) analyse quality and content of the corpora

  b) allow further research using genre subcorpora

What is Automatic Genre Identification?

- Text classification task, focused on (web) genres
- Genres: text categories based on the author's purpose, the function and form of the document
- Examples: news article, recipe, legal text, etc.

Genre schema and annotation

Challenge: previous schemata created with the aim to reach high coverage did not result in reliable annotation

→ New genre schema (24 labels)

→ Improvements of the annotation procedure

GINCO genre schema

- High coverage: schema is based on previous high-coverage schemata, but further improved to achieve high reliability
- Labels of a reasonable granularity, recognizable to corpora users and annotators
- Genres are not fixed, static → category *Other*, possibility of adding new genre categories during the annotation process

GINCO genre schema

Purpose	Objectively inform/educate the reader	Subjectively report on an event	Convey opinion to persuade the reader	Sell/Promote	To discuss with other people	To give pleasure to the reader	Convey information in a typical form	Other/Unknown
Main annotation criteria	Objectivity	Subjective reporting	Subjective text – non-commercial	Subjective text - commercial	Interaction of multiple people	Literary text	The form of text	It is not possible to asign any other genre to the text
Category Group	**Objective Informative**	**Subjective Reporting**	**Opinion**	**Promotion**	**Dialogue**	**Literature**	**Formatted Text**	**Other category/Mixed**
Categories	News/Reporting	Opinionated News	Opinion/Argumentation	Promotion	Interview	Script/Drama	FAQ	Other
	Announcement		Review	Promotion of a Product	Forum	Lyrical	List of Summaries/Excerpts	
	Information/Explanation			Promotion of Services	Correspondence	Prose		
	Research article			Invitation				
	Instruction							
	Recipe							
	Call							
	Legal/Regulation							

Annotation procedure

- Expert annotators (linguists) instead of crowdsourcing
- Decision tree, guidelines with concrete characteristics of genres and examples
- Frequent meetings to discuss hard cases, updating guidelines to include new phenomena
- Hybrid texts: characteristics of multiple genres

→ multiple-label annotation (primary, secondary, tertiary label)

Annotation guidelines

Instruction

An objective text which instructs the readers on how to do something.

Common features:

- multiple steps/actions
- chronological order
- 1st person plural or 2nd person
- modality (must, have to, need to, can, etc.)

Note: If a text has features of an instruction, but it is subjective (contains subjective adjectives or adverbs, words that convey certainty), annotate it as Opinion/Argumentation, and use the secondary category Instruction.

Go to examples.

Go back to the Decision Tree or to the Table of Contents of Categories Explained.

Results of the annotation campaign

- Reliable annotation: inter-annotator agreement: 0.71 Krippendorff's alpha - above acceptable threshold
- First Slovene genre corpus: Genre Identification Corpus GINCO 1.0

GINCO 1.0 dataset

- Samples from two Slovenian web corpora to capture realistic conditions on the web:
  - slWaC 2.0 (2014)
  - MaCoCu-sl 1.0 (2021)

GINCO 1.0 dataset

- Challenge: uncurated web corpora (documents in other languages, machine translations, generated text ...) → manual annotation with Not Suitable categories
- The 2014 and 2021 part equally represented in the GINCO dataset, but 89% of “unsuitable” texts from the recent crawl

  → the quality of Slovene web deteriorated

GINCO size

subset	texts	pars	words
suitable	1,002	15,050	478,969
suitable (dedup.)	983	7,088	278,075
not suitable	123	3,402	173,778
both subsets	1,125	18,452	652,747

Machine Learning Experiments

1. Comparison of different technologies
2. Impact of using full texts of web documents instead of texts with near-duplicate paragraphs removed
3. Impact of using secondary labels as additional signal
4. Impact of training data size
5. Impact of downcasting the number of labels
6. Performance per category

Comparison of different technologies

- FastText not up to the task
- Transformer models achieve high results already on small dataset
- Monolingual SloBERTa and the multilingual XLM-RoBERTa comparable

classifier	micro F1	macro F1
stratified dummy	0.067	0.061
fastText	0.352 ± 0.038	0.217 ± 0.040
fastText + emb.	0.361 ± 0.007	0.219 ± 0.013
XLM-RoBERTa	0.624 ± 0.015	**0.579 ± 0.024**
SloBERTa	**0.629 ± 0.016**	0.575 ± 0.037

Interesting findings

- Monolingual model did not perform better than the multilingual one:
  - Pre-training did not play a large role?
  - AGI more generic linguistic task than other NLP tasks?
- Frequency does not correlate with F1-scores (Recipe, Research Article ...)

  → some classes more prototypical, easier to identify

Conclusions

Our contributions:

- Genre schema and annotation procedure for reliability and high coverage
- GINCO 1.0 - representative web-based sample of documents annotated for suitability and genre
- Promising results of the machine learning experiments with Transformer models

Future work

- Double the size of the Slovene dataset
- Annotation of Croatian and English corpora → cross-lingual experiments
- 10% of the data discarded manually - fully automated preparation of the corpora
- Further machine learning experiments

Thank you!

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).