UNIVERSITY OF TURKU

Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora

TURKUNLP .ORG

Amanda Myntti, Liina Repo, Elian Freyermuth[1], Antti Kanner, Veronika Laippala, Erik Henriksson

University of Turku, [1]National Graduate School of Engineering of Caen

MOTIVATION

Lack of metadata often limits the use of large web corpora. We make use of advances in text classification by using two machine learning models, one focused on linguistic registers and another on literary genre, to classify documents form a large web corpus, Oscar [1], and evaluate the new metadata we create. This metadata supports new ways of studying digital cultural heritage by facilitating data selection and categorization.

Register	F1-score	Support
Lyrical (LY)	0.8949	135
Spoken (SP)	0.7032	146
Interactive Discussion (ID)	0.8475	686
Narrative (NA)	0.8405	4264
How-to Instructions (HI)	0.6788	411
Informational Description (IN)	0.7176	2596
Opinion (OP)	0.6854	2129
Informational Persuation (IP)	0.5591	402
μ* (micro)	0.74	18276

REGISTER

To train the **register classifier**, we finetune XLM-RoBERTa-Large [2] with the **Corpus of Online Registers (CORE**) [3,4]. This corpus has a hierarchical multilabel scheme covering the full range of web registers, e.g. *Opinion*, *News report*, and *Interactive discussion*. The resulting classifier was able to reach an F1-score of 0.74 which is in line with previous results [3]. See Table on the left.

The variablity in classification performance can be attributed to the features of the registers, which vary in the level of linguistic definition [5].

*Sublabels, such as Interview (under SP) or News report (under NA) omitted for simplicity.

GENRE

The **genre classifier** is similarly finetuned from XLM-RoBERTa-Large [2]. We use **Genre-6 corpus** [6], based on Kindle UK&US books. This corpus features a multilabel scheme with over 20 literary genre labels, such as *Politics & Social Sciences* and *Childrens' Books*. Using all of these labels in classification resulted in poor performance; thus we selected categories by evaluating candidate subsets while keeping in mind our target to cover the contents of a web corpus. This resulted in the selection of genre labels in the table below.

The classifier was able to reach 0.70 F1-score with variability in labelwise perfomance similarly to the registers. The Genre-6 is a small corpus and contains some noise in the labels, which could be mitigated using a label cleaning tools. Lastly, we acknowledge that the *Literature & Fiction* class is too broad.

EXAMPLE

Engineering & Transportation

Information Description, Research Article

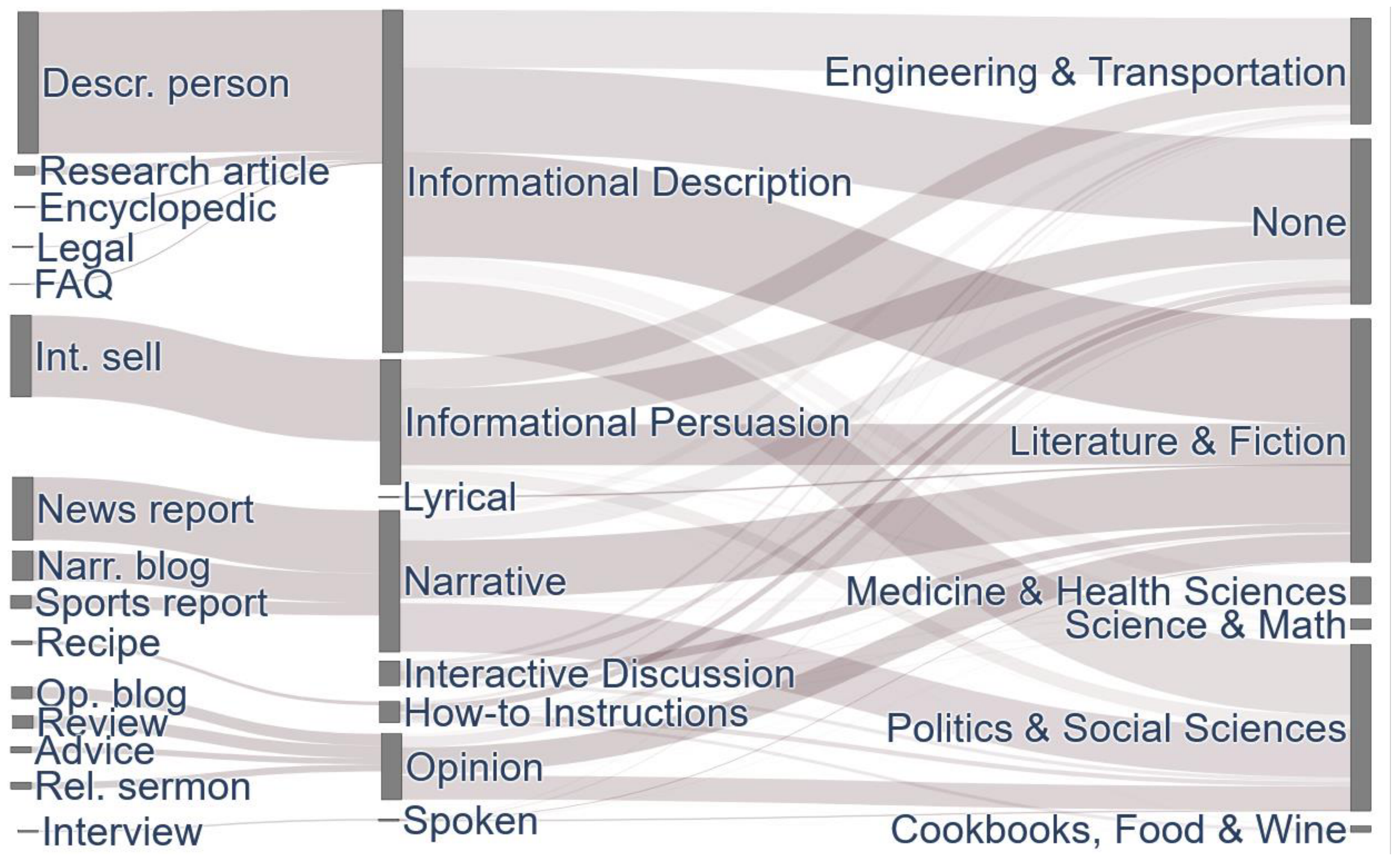
The management of existing road infrastructures is a multidisciplinary activity that involves structural engineering, material science, management, economics and ecology.
The objective is to achieve maximum availability of road links at minimum societal costs.

Genre	F1-score	Support
Cookbooks, Food & Wine (Cook)	0.59	35
Engineering & Transportation (Engn)	0.65	172
Literature & Fiction (Lit)	0.81	535
Medicine & Health Sciences (Med)	0.61	72
Politics & Social Sciences (Pol)	0.53	194
Science & Math (Sci)	0.45	144
μ (micro)	0.70	1152

INTERSECTION

We illustrate the intersection of the two labelling schemes with the figure below. The figure confirms that no register and genre categories fully overlap, demonstrating that cross-labelling with our setup achieves the intended outcome: it refines the classification and enriches the information for each document.

The results show expected combinations between certain registers and genres, such as the *Lyrical* register often aligning with the *Literature & Fiction genre*. However, most registers, such as *Interactive Discussion*, are divided across multiple genres, like *Engineering & Transportation and Politics & Social Sciences*, depending on the discussion topic.



TOPIC EVALUATION

To investigate whether the intersections between the labeling schemes are meaningful, we extracted topic keywords from all register-genre intersection classes, like *Informational Persuasion + Engineering & Transportation*. We use the Latent Dirichlet Allocation algorithm [7] to extract topic words. Our analysis shows the results match the expected subject matters and linguistic features contained in the intersection classes.

Opinion

White chocolate isn't really chocolate at all. While it contains the cocoa butter of true chocolate, it lacks cocoa solids, the element responsible for milk and dark chocolate's characteristic brown color and nutty roasted flavor.

Cookbooks, Food & Wine

EXAMPLE

QUANTITATIVE EVALUATION

Mutual information	Entropy (register)	Entropy (genre)	Entropy (register|genre)
0.109	3.370	2.229	5.443

We measured the mutual information between the register and genre labels, and the increase in entropy. These measures display that the register and genre labels are not redundant but complement each other.

References

[1] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache

[2] Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

[3] Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. Journal of the Association for Information Science and Technology, 66(9):1817–1831.

[4] Veronika Laippala et al. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. Language Resources and Evaluation, 57(3):1045–1079.

[5] Veronika Laippala, Jesse Egbert, Douglas Biber, et al. 2021. Exploring the role of lexis and grammar for the stable Identification of register in an unrestricted corpus of web documents. Language Resources and Evaluation, 55:757–788.

[6] Available https://huggingface.co/datasets/TurkuNLP/genre-6

[7] Available https://radimrehurek.com/gensim/models/ldamodel.html

Funded by the European Union
NextGenerationEU