Genre-informed Multilingual Corpus for extractive Single-Document Text Summarization

Esther Sebastián Liso
University of Copenhagen
zxv330@alumni.ku.dk

1 Introduction

This paper presents a genre-informed single-document summarization corpus (GIC) for written Danish, English and Spanish. The corpus contains 450 documents divided in 5 genres and its respective summaries. It contains 30 documents in all three languages for each genre. We chose text documents that maximized comparability regarding to content and length document and summary between all three languages.

2 Text types

Even though NLP works are often domain-aware, we advocate for genre-aware summarization, understanding *genre* as the language and structure that is utilized in a specific domain, e.g. news articles or academic papers are likely to have a certain structure, regardless of language or geographical or historical context. These genres correspond approximately to different types of discourse or "superstructure", as defined by Van Dijk (1980). The types of discourse that have been chosen in this paper are the following: descriptive, narrative, expository, argumentative and mixed.

In Table 1, we can see how the average length varies in every genre from language to language.

Genre Document	DA	EN	ES
Descriptive	2433	5522	6315
Narrative	3108	4741	4328
Expository	631	715	828
Argumentative	337	115	141
Mixed	7415	7660	3120

Table 1: Average number of words for each document type and language. Although not reflected in this table, there is still a big variance among some documents withing the same language and genre.

2.1 Descriptive genre: Wikipedia articles

We have chosen five articles for six different topics: animals, capitals of Europe, diseases, elements of the periodic table, everyday objects and members of Europe's royal families. For each of these topics, we have selected five articles. The articles chosen are correlative articles in Danish, English and Spanish. We have chosen articles that maximize the size of their shortest variant (namely the Danish versions).

2.2 Narrative genre: Novels and tales

This genre is constituted by a collection of five chapters[1] of four classical novels and ten of Hans Christian Andersen's fairytales[2]. The novels chosen fall in two categories. Firstly, two originally published novels in English and their corresponding translations into Spanish; and, secondly, novels originally published in Spanish and their matching translations into English.

2.3 Expository genre: News articles

This genre is in this paper represented by a collection of 30 news articles in English (from the British metro.co.uk), Spanish (from the Spanish elpais.es) and in Danish (from jyllands-posten.dk) published from March through July 2015. The topics are many and various and include politics, migration, terrorism, food & ecology and weather among others.

2.4 Argumentative genre: Editorials

In this paper, this genre is constituted by gathering 30 letters to the editor in English (from the Canadian theglobeandmail.com), in Spanish (from el-

[1] Chapters are chosen in order to correspond to different parts of the book (the beginning, three chapters spread out in the middle of each novel and the ending of the book.

[2] It was not possible to find public summaries for the novels in Danish. In order to have 30 articles for this genre in Danish, we collected 20 additional Hans Christian Andersen's fairytales.

pais.es) and in Danish (from jyllands-posten.dk). The topics discussed in these letters are very different and refer to other articles published in the newspapers. The main topics of this genre are politics, the economy and culture education.

2.5 Mixed genre: Academic papers

Academic papers correspond to a mixture of text types, mainly descriptive, expository and argumentative, depending of the intention of their author. This corpus contains 30 different articles in each language. The main topics are in the social sciences (politics and sociology) and medicine.

3 Model summaries

We also provide a collection of human-generated model summaries. There is one human-generated summary per document either gathered from online resources or written by a human annotator when on-line resources were not available. As a result of this way of collecting model summaries, summaries are very different in terms of length, compression ratio, style, etc.

In Table 2, we can see that the length of the summaries is very different in some genres and languages. Individual summaries are inevitably very different in word length, writing style and content, especially in the narrative genre.

Genre Summary	DA	EN	ES	on-line
Descriptive	207	425	377	×
Narrative	379	255	198	×
Expository	66	94	58	∼
Argumentative	41	52	66	÷
Mixed	189	218	117	×

Table 2: Average number of words in model summaries for each genre and language and whether they were fetched on-line or not.

In Table 3, we show the average compression ratio for each genre as expressed in Ceylan et al. (2010). We can see that there is still discrepancy between genres and among languages inside every genre.

3.1 Descriptive genre: Wikipedia articles

We apply the introduction section of Wikipedia articles (which appear in the beginning of every Wikipedia article) as summaries for the entire Wikipedia article.

Compression Ratio	DA	EN	ES
Description	92%	92%	94%
Narration	94%	92%	94%
Exposition	91%	91%	89%
Argumentation	80%	64%	63%
Mixture	98%	98%	94%

Table 3: Average Compression Ratio for each genre of the corpus respectively.

3.2 Narrative genre: Novels and tales

Summaries of novels and tales were gathered from different on-line websites such as sparknotes.com (in English), monografias.com (in Spanish) and litteratursiden.dk (in Danish). When the required summaries were not available from these three sources, we searched for summaries in other websites from individuals, on-line reviews of the books or a plot or summary section in Wikipedia articles.

3.3 Expository genre: News articles

It was not always possible to find summaries of news articles in Danish, English and Spanish. However, sometimes we could find a short paragraph before the body of some news articles. When possible and long enough, those paragraphs were used as a model summary. When that information was not available, an annotator wrote a summary in English, Spanish and Danish. The annotators were instructed to write about 40-50 words (in at least two sentences) summaries and not to copy the exact same sentences that appeared in the original texts.

3.4 Argumentative genre: Editorials

Summaries for letters of the editor were not available, but it was the same annotators that wrote the Danish, English and Spanish summaries, respectively. The annotators were instructed to write about 40-50 words long summaries (in at least two sentences) and not to copy verbatim from the original documents.

3.5 Mixed genre: Academic articles

The abstracts of the academic articles were used as summaries in this genre. Because of author preferences when writing abstracts, the summaries of this genre are again very different in length and content of the summary, e.g. some abstracts seem more like an introduction to the paper while others focus on the results and conclusions of their articles.

4 Unsupervised generated summaries (extractive)

Finally, a collection of system-generated sentence rankings of the original documents is also included in this corpus. This section includes some baselines: document sentences in the reverse order and in a random order, and sentences ordered using the PageRank algorithm (Brin and Page, 1998) using an undirected graph and a directed graph (both forward and backwards).

5 Summarization Results

We use the ROUGE package (Lin, 2004)[3] to evaluate the system-generated summaries. Table 4, 5 and 6 reports ROUGE-1 recall scores for every genre in Danish, English and Spanish.

Although these results cannot be directly compared inter-language due to the differences in the original documents and their summaries, we can have an idea of which systems work better: generally a directed backward graph-based method.

Totally, there are six different sentence orders (including the normal order) for each document.

Although we use extractive summarization, summaries are new made and not only selecting the most important sentences.

References

Ceylan, Hakan, et al. ”Quantifying the limits and success of extractive summarization systems across domains.” Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

Lin, Chin-Yew. ”Rouge: A package for automatic evaluation of summaries.” Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.

Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). ”The PageRank citation ranking: Bringing order to the Web”. , published as a technical report on January 29, 1998

Van Dijk, Teun Adrianus. Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition. Lawrence Erlbaum Associates, 1980.

[3] We used version 1.5.5. with the following parameters: -n 2 -2 4 -U -a -x -c 95 -r 1000 -f A -p 0.5 -t 0 -l 400 -a.

ROUGE-1 (EN)	LEAD	LAST	RANDOM	D.FORWARD	D.BACKWARD	UNDIRECTED
Description	0.21	0.15	0.19	0.18	0.23	0.2
Narration	0.26	0.23	0.21	0.22	0.23	0.23
Exposition	0.58	0.2	0.26	0.22	0.59	0.35
Argumentation	0.46	0.43	0.43	0.43	0.44	0.42
Mixture	0.3	0.24	0.22	0.25	0.27	0.22

Table 4: Average ROUGE-1 50 word summary recall scores for every genre in English.

ROUGE-1 (ES)	LEAD	LAST	RANDOM	D.FORWARD	D.BACKWARD	UNDIRECTED
Description	0.29	0.2	0.25	0.25	0.29	0.25
Narration	0.3	0.26	0.25	0.27	0.29	0.27
Exposition	0.52	0.27	0.31	0.27	0.45	0.33
Argumentation	0.47	0.47	0.46	0.5	0.46	0.48
Mixture	0.38	0.28	0.27	0.3	0.34	0.26

Table 5: Average ROUGE-1 50 word summary recall scores for every genre in Spanish.

ROUGE-1 (EN)	LEAD	LAST	RANDOM	D.FORWARD	D.BACKWARD	UNDIRECTED
Description	0.22	0.18	0.22	0.2	0.23	0.22
Narration	0.23	0.2	0.18	0.19	0.2	0.2
Exposition	0.53	0.24	0.31	0.26	0.48	0.32
Argumentation	0.48	0.32	0.35	0.34	0.44	0.37
Mixture	0.3	0.21	0.19	0.22	0.26	0.22

Table 6: Average ROUGE-1 50 word summary recall scores for every genre in Danish.