The Web Library of Babel:
evaluating genre collections

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

Centre of Translation Studies,[†] School of Computing[‡]
University of Leeds, Leeds, LS2 9JT, UK
{S.Sharoff,Z.Wu,K.Markert}@leeds.ac.uk

Abstract

We present experiments in automatic genre classification on web corpora, comparing a wide variety of features on several different genre-annotated datasets (HGC, I-EN, KI-04, KRYS-I, MGC and SANTINIS). We investigate the performance of several types of features (POS n-grams, character n-grams and word n-grams) and show that simple character n-grams perform best on current collections because of their ability to generalise both lexical and syntactic phenomena related to genres. However, we also show that these impressive results might not be transferrable to the wider web due to the lack of comparability between different annotation labels (many webpages cannot be described in terms of the genre labels in individual collections), lack of representativeness of existing collections (many genres are represented by webpages coming from a small number of sources) as well as problems in the reliability of genre annotation (many pages from the web are difficult to interpret in terms of the labels available). This suggests that more research is needed to understand genres on the Web.

1. Introduction

The possibility of dealing with a document depending on its genre rather than on its content is important in many NLP areas. In POS tagging, machine translation or discourse annotation knowing the genre of a document can help in selecting more appropriate language models. For example, the accuracy of POS tagging reaching 96.9% on newspaper texts drops down to 85.7% on forums (Giesbrecht and Evert, 2009), i.e., every seventh word in forums is tagged incorrectly. (Webber, 2009) showed that genres such as letters to the editor vs. newspaper articles differ in the distribution of particular discourse relations.

In information retrieval the user might be specifically interested in research papers or tutorials on a particular topic, but might find it hard to identify these given the large number of pages on any given topic (Vidulin et al., 2007). Due to the size of the Web as well as the multitude of traditional and new genres it contains, interest in automatic genre identification (AGI) has recently focused on automatically classifying web pages, using genre-annotated corpora of webpages, e.g., KI-04 (Meyer zu Eissen and Stein, 2004) or SANTINIS (Santini, 2010). These two collections have become the benchmarks for evaluation of genre identification algorithms, with the best reported results reaching 96.5% accuracy (with 10-fold cross-validation) on SANTINIS (Kanaris and Stamatatos, 2007).

In this paper, we evaluate the progress made by adressing both the question of reliability, representativeness and similarity of existing genre collections available for our research. We also investigate what features are useful for automatic genre identification. We report several experiments on:

1. the accuracy of genre identification on individual collections using different feature sets;
2. the similarity between identical or nearly identical categories in individual collections;
3. the accuracy across collections after mapping their subsets to a shared set of categories;
4. the agreement of human judgement on individual collections and their subsets.

2. Data preparation

The genre collections used in this experiment are summarised in Table 1.[1] In addition to web collections proper we extended some experiments to two classic English corpora, i.e., the Brown Corpus and the BNC. The following is a brief summary of each collection of webgenres:

HGC (Hierarchical Genre Collection) is based on a two-level hierarchy of genres, e.g., `Journalism/Reportage` or `Literature/Poem`. For each genre category, 40 example pages were collected in 2005/2006.

I-EN-Sample is a sample of 250 webpages randomly selected from I-EN, a corpus of 71,636 pages collected by random queries to Google in February 2005 (Sharoff, 2006). The pages in the sample have been annotated using the Functional Genre Classification (FGC) scheme, which reflects the main aim of text production, e.g., `information` (which includes homepages and encyclopedic entries) or `recreation` (fiction and popular lore).

KI-04 is based on eight genres, e.g., `article` or `portrayal (private)`, which were suggested in a study of genre usefulness. The pages were collected from the bookmarks used by five people and were downloaded in January 2004.

---

[1]The list of classes in each collection is described in the Webgenre Wiki `http://purl.org/webgenres`. The number of texts and genres corresponds to what came from actual data provided by the authors of each resource. In the case of HGC, MGC and KRYS-I there are minor discrepancies with the number listed in respective publications. Also, we rejected pages with no textual content or no genre labels.

Table 1: Genre collections used in the experiment

Source	# texts	# genres	Format
HGC (Stubbe and Ringlstetter, 2007)	1412	34	HTML only
I-EN-Sample (Sharoff, 2010)	250	7	TXT from HTML
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8	HTML only
KRYS I (Berninger et al., 2008)	6200	70	PDF
MGC (Vidulin et al., 2007)	1536	20	HTML with images
SANTINIS (Santini, 2010)	1400	7	HTML only
Combined (Santini and Sharoff, 2009)	9849	8	TXT from HTML
Brown Corpus (Kučera and Francis, 1967)	500	10	TXT
BNC (Lee, 2001)	4053	70	TXT

KRYS I is based on 70 genres combined into ten genre groups, e.g., `Poetry book` or `Magazine article`. The PDF files for each category were collected from the Web as well as from offline sources between 2005 and 2008.

MGC (Multi-labelled Genre Collection) is based on a list of webgenres adapted from (Lim et al., 2005), e.g., `Children's` or `Journalistic`. It contains three subsets: targeted (prototypical example pages for each category), Zeitgeist (pages collected from popular queries to Google), and webpages chosen using a random link generator (in a way similar to I-EN).

SANTINIS is based on seven genres which are unique to the Internet, e.g., `blogs` or `FAQs`. Examples were collected in 2005 from pages that claimed themselves as belonging to these genres (the principle of 'objective sources').

Each collection is relatively small and no collection on its own can be treated as a representative sample of genres on the Web. However, the collections vary in the way they represent webgenres, so that any operation of combining them to get a better picture of the Web is difficult. First, the collections use their own sets of genre labels, which in many cases require a many-to-many mapping, e.g., the category `Informative` from MGC contains encyclopedic entries, recipes, user manuals, lecture notes, each of which corresponds to different categories in other collections, or no label is available to cover such texts at all in, e.g., SANTINIS. Second, the genre classes in some collections are organised into a hierarchy (HGC, KRYS I, and to a lesser extent I-EN-Sample), while in others flat lists are used. Genre labels in some collections are quite specific, e.g., `Online Newspaper Front Page` in SANTINIS, while others use fairly broad clases, which cover a large number of diverse subtypes, e.g., `Informative` in MGC.

Third, the collections differ in their approach to the process of page selection. Some collections (like I-EN-Sample, KI-04 and a part of MGC) assume genre annotation of a diverse set of pages randomly selected from the web, while others define their genre palettes first and then target only suitable examples for these genres. In some collections such examples were selected from a small number of sources. This affects the variety of pages within each collection.

Finally, the collections also differ in the format used for the preservation of their pages: I-EN-Sample only contains running text extracted from HTML files, KRYS I consists of PDF pages, while MGC is the only collection that stores images along with HTML files. The absence of HTML tags in I-EN-Sample and KRYS I makes it impossible to use any structural HTML features (since the features need to be valid across all collections), while the use of visual features (how the page looks like) is possible only with MGC.

As a way of solving these problems we have reduced the richness of content in each collection to the least common denominator: plain text files with a flat list of genre labels. The experiments reported in Sections 3.1 and 3.2 are based on the original labels, while those in Section 3.3. involve mapping between **some** genre categories to make the collections more comparable.

Given the variety of genre labels in the collections, it is impossible to map all documents in each collection to a unified set of labels. In our earlier work we established the Functional Genre Classification scheme, containing eight generic (macro-) genre labels (this set was also used in annotating I-EN-Sample), and created a new collection listed as **Combined** in Table 1 by mapping all genre collections to this set (Santini and Sharoff, 2009). The only two labels that are consistent across most collections are `FAQ` and `Shop` (however, even in the latter case, HGC does not have this category at all, while MGC distinguishes between `Shopping` and `Promotional`). Other labels permit a degree of mapping into more general classes, even though this resulted in loss of information. For example, `Home pages` used in SANTINIS, `Portrayal (private)` in KI-04, `Personal` from MGC from HGC have been mapped into `information`. However, a single label in a source collection sometimes covers webpages of several different classes, so that its target label is not unique, e.g., `adult` in MGC covers lists of links, advertising, forms for accessing websites, legal disclaimers, instructions, etc. Manual remapping of each individual document with "ambiguious" labels was not feasible, so we had to discard such documents.

3. Experiments

3.1. Comparing features

KI-04 and SANTINIS are two more popular collections, which were used in several studies investigating differ-

Table 2: Existing studies on KI-04 and SANTINIS

Features in existing studies	KI-04	SAN
BOW, punctuation, HTML (Meyer zu Eissen and Stein, 2004)	70.0	-
BOW, punctuation, HTML (Boese and Howe, 2005)	74.8	-
POS, punctuation, HTML(Santini, 2007)	68.9	90.6
char n-grams (Kanaris and Stamatatos, 2007)	82.8	96.2
char n-grams, HTML (Kanaris and Stamatatos, 2007)	84.1	96.5
char n-grams (Mason et al., 2009)	-	94.6

Table 3: Our experiments on accuracy for features.

Features	HGC	I-EN-S	KI-04	KRYS-I	MGC	SAN	Comb	BNC	Brown
POS1	32.79	49.20	52.03	18.84	26.89	70.29	35.52	51.27	57.20
POS2	47.66	49.20	59.25	33.82	36.78	82.71	55.28	64.89	57.80
POS3	49.50	50.00	63.40	34.58	41.60	85.79	55.82	65.66	59.40
POS4	47.10	44.00	63.32	31.11	39.84	85.07	55.71	63.39	54.20
Char1	33.85	43.60	57.10	18.44	23.83	75.36	33.21	49.30	56.80
Char1-bin	17.28	35.60	38.17	13.76	16.41	67.07		17.02	43.00
Char2	54.39	49.60	76.10	44.63	42.77	90.93	59.76	69.95	56.60
Char2-bin	53.33	41.20	73.03	38.58	43.29	93.07	49.80	55.81	54.00
Char3	59.99	54.80	80.00	51.35	49.54	93.93	63.58	72.49	65.40
Char3-bin	63.31	54.40	81.91	57.77	53.26	96.21	62.36	71.80	62.60
Char4	59.91	52.40	79.25	50.90	50.91	94.43	65.22	73.62	64.80
Char4-bin	65.51	55.20	**85.81**	61.87	55.14	**97.14**	66.89	74.54	65.80
Char5	57.65	52.00	78.42	49.40	49.87	94.21	66.54	72.59	64.20
Char5-bin	65.72	56.80	85.48	61.85	56.45	97.14	68.90	75.33	65.40
Char6	56.09	50.40	77.34	47.24	49.93	93.86	66.39	72.05	60.60
Char6-bin	64.80	56.40	85.06	62.02	55.92	96.93	69.82	76.04	64.40
Char7	54.32	48.40	76.68	46.58	49.61	93.71	64.93	71.35	60.00
Char7-bin	63.10	56.40	83.73	60.13	53.71	96.14	70.28	76.17	62.40
Char8	53.90	44.00	77.59	47.10	50.33	93.79	62.98	70.54	57.60
Char8-bin	59.56	54.00	82.74	57.66	51.50	96.21	69.99	75.99	60.80
Char9	54.67	42.80	78.09	46.68	47.59	94.00	61.40	69.63	54.40
Char9-bin	56.94	49.60	81.24	54.42	48.63	95.21	69.62	75.35	60.00
Char10	54.11	42.00	78.01	46.63	47.40	93.21	60.06	68.27	52.00
Char10-bin	53.26	48.00	80.17	51.35	45.90	94.79		75.28	56.00
Word1sr	59.77	52.40	80.50	51.16	49.61	94.29	65.92	69.55	59.40
Word1sr-bin	59.49	56.80	82.49	59.08	50.85	95.36	62.67	73.38	63.80
Word2sr	47.03	39.20	71.95	44.50	42.58	86.79	61.76	59.81	51.40
Word2sr-bin	44.41	33.60	70.04	47.37	36.20	84.14	60.05	64.50	51.40
Word1	61.69	54.80	81.83	54.02	51.63	94.79	67.41	71.50	61.60
Word1-bin	59.06	60.40	84.15	59.05	51.63	95.86	63.31	75.03	64.00
Word2	54.11	44.00	79.17	49.97	48.89	91.29	64.18	67.33	50.00
Word2-bin	57.15	49.60	79.17	53.55	47.59	92.86	65.52	73.60	55.40

ent approaches to genre identification (Table 2). The approaches tested include the use of POS n-grams (e.g., RB VVZ DT), words and word n-grams (*also* or *also affects the*), character n-grams, either fixed (*als, lso, o a*) or of variable length (*tion, If the, alongside*). In addition to this, HTML features have been used, such as the frequency of individual tags (`<img>`, `<table>`) or the structure of the originating URL (`/cgi-bin/`, its length). However, to our knowledge there has been no thorough investigation of the performance of a wide range of features on a wide range of

genre collections (going beyond KI-04 and SANTINIS). Some corpora in our study did not have HTML tags at all, while the use of HTML markup has been shown to improve the accuracy only marginally, see (Kanaris and Stamatatos, 2007) and Table 2. Therefore, we decided to test the performance of textual features only, using word-based, POS-based and character-based features. Each webpage in each collection was converted to plain text using `lynx` or `pdftotext` (for KRYS-I). POS tags for POS n-grams were produced by TreeTagger (Schmid, 1994), character

n-grams were converted to lower case. For word-based features we made two experiments: with the full set of words and with stop words removed, using the list from the Rainbow package (McCallum, 1996). For testing each feature type in a corpus we collected up to the 1,000 most frequent features from each document and combined them together (thus, features were specific to each collection). This resulted in large feature vectors, reaching almost one million features. Two feature representations have been used for word and character features: normalised frequencies and binary features (a boolean value indicating whether a feature is present or absent in the list of features selected for a given document).

Similar to previous work on AGI we are aware of, we use a supervised classification framework. We used a speedup variant of linear SVM called Liblinear (Fan et al., 2008) and ten-fold cross-validation. Each individual collection has been tested on its original set of genre labels (with the exception of the Brown Corpus, for which 10 genre labels have been used by combining its genres of fiction, similar to prior work in (Karlgren and Cutting, 1994)).

Results. The results in Table 3 show that the bag of single words (with stop words included), and the binary versions of character trigrams and tetragrams are the best performing features. In this table *-bin* refers to the use of binary features, while *sr* to cases when the stop words have been removed.

The best results (highlighted in Table 3) for KI04 and SANTINIS outperform the best results previously reported in (Kanaris and Stamatatos, 2007), while our results are based on a simpler procedure (fixed- instead of variable-length n-grams used by Kanaris and Stamatatos). We also have not used any genre-based feature selection: it can improve accuracy, but if selection is based on the entire corpus, the cross-validation experiment is not theoretically sound.[2] We also achieve results of up to 66% on the Brown corpus, which compares favourably with the results by (Karlgren and Cutting, 1994), who use the same data for training and testing instead of cross-validation.

Discussion, Character n-grams, as well as individual words are fairly lexical features, which capture what is explicitly said in an individual document. The use of POS n-grams is aimed at capturing the syntactic complexity without the use of syntactic parsers, which are slow, unreliable for most genres and not available for many languages. The performance of POS trigrams was shown to be much less accurate than that of character n-grams. Character n-grams can capture some morphosyntactic properties (like the endings of verbs and adverbs), the use of punctuation marks (exclamations, parentheses, etc), as well as distinctions not captured by traditional POS tags, like the use of a particular class of modal constructions (*must, necessary*) or conjunctions (*however, actually*). All these properties can be potentially relevant for genres, but cannot be captured by POS n-grams.

Even so, the use of lexical features can lead to problems, as higher accuracy on a collection is sometimes achieved by detecting topics rather than genres of individual texts, so that the results are not applicable to another collection. Therefore, the high accuracy of lexical features on a single collection can be misleading as to their power of discriminating web genres in general.

To illustrate this point, we extracted the character tetragrams and words that are most specific to genre classes in individual collections. For each feature, its specificity to a class is proportional to the discrepancy between the weight of the class in the SVM classifier and the mean value of the weights of all classes. This shows why the SVM classifier selects this class. We present examples in Table 4.

We can see that for some categories the SVM classifier uses quite specific tetragrams or words to identify documents belonging to a particular class in a collection, e.g., *urri, cycl, tax_* or *hurricane, cyclone, tax* for FAQs in SANTINIS or *DNS, ISP, Palladium* for FAQs in KI04. This is an artefact as both KI04 and SANTINIS are highly targeted, for example all FAQs in SANTINIS come from two sources, a website with FAQs on hurricanes and another one with tax advice. In the end, an SVM classifier built for FAQs on these datasets relies on occasional properties of these two collections and will fail to spot any other FAQs. On the other hand, more generic corpora tend to have a larger number of more generic lexical items which cause greater confusion for the classifier, e.g., *make, people, build, locate* for `instruction` in I-EN-Sample (character n-grams were sometimes able to capture constructions specific to this genre, e.g., *?_wh* or *on't*).

3.2. Comparing labels

We combined two genre collections, KI-04 and SANTINIS, as they have a small number of somewhat compatible labels and both are commonly used in AGI experiments. In this experiment, we still used the original labels of each collection. Then we attempted to estimate, which classes are similar between them. The compatibility of two genres was measured by the pairwise distance of their SVM weight vectors for the pool of genres. Specifically when combining the two corpora, we have 15 genres in total, with 8 from KI-04 and 7 from SANTINIS. By building an SVM model on it, we obtain a weight vector per genre, 15 in total. If two vectors are very close in terms of a distance measure, that means the features are multiplied with similar weights and the two classes are likely hard to separate, partly because they are compatible genres. We would, for example, expect that KI04-shop and SAN-eShop are hard to distinguish. The results are displayed in a dendrogram in Figure 1.

Sometimes these expectations are met (such as in the shop example just mentioned). However, as a counterexample `Portrayal (priv)` in KI-04 corresponds to `PHP` (Personal home page) in SANTINIS, but they are considerably different in terms of ngrams (Figure 1). The same applies to `help` in KI-04 and `FAQ` in SANTINIS. This suggests that genre classes in each of the two collections are considerably different from the viewpoint of the SVM classifier.

---

[2]This is the reason for the higher accuracy of POS trigrams reported in (Sharoff, 2007).

Table 4: Tetragrams and words characteristic for some genre classes

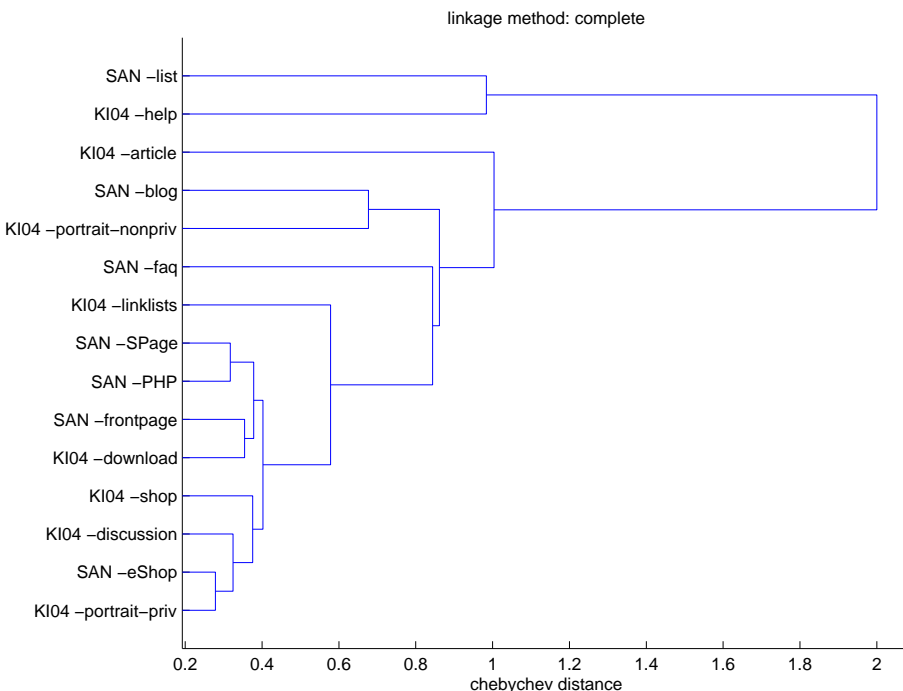| Corpus, genre	Features
I-EN-S, discussion, Char4:	,_yo _,_y liti is_s g_an weve ole_ as_w e_do owev itic rses _alt e_'s y_de than olit ds_o ded_ king
I-EN-S, discussion, Word:	including beginning trend part applications runs good facing things reach build launch perceptions system current made apply restaurant final transfer
KI-04, article, Char4:	n't_ e_ht _you f_a_ ee_h ach_ uch_ you_ _one _on_ s._t eali ml_s tic_ f_yo betw _giv ncep _and and_
KI-04, article, Word:	based <N> home information messenger values release institute mail html find scope page order case represented elsevier download classifications
MGC, shopping, Char4:	zon. trex garm rex_ armi _?10 e_et _etr wayp _waa waas peic _alk _pei y_aa bci6 lkal eice psu_ _bci
MGC, shopping, Word:	discount mobile dunning jamison unknowingly spotlights directory send micro crumb pleasingly tla
SAN, eshop, Char4:	offe _ord ffer rder tome orde ent_ news only _the _int pric mer_ ine_ bask te_m poun ment &pou und;
SAN, eshop, Word:	order basket pound catalogue offers conditions click dvd <N> prices customer orders phone customers find price news delivery web goods
KI-04, shop, Char4:	lsen usab book osau oddl _ord todd pric aur_ fist droi rice nosa _you dino _boo hop_ shop d_bo s_pf
KI-04, shop, Word:	home orders shipping cart shop order price copyright gift payment featured inconvenience page products selling amiga higher click prices web
HGC Help, Char4:	acit epai _fus q:_w cito _ac_ faq_ volt adap otok okan kata dojo _doj ob_r redm shod dapt kara a:_t
HGC Help, Word:	faq grow orders collection placing site uploading burlington utilizes grading owl frequently message invention migrating oratory inappropriateness coincidental underrepresented
I-EN-S, instruction, Char4:	_,_y ,_yo or_d _rem g_it _liv _?_w ?_wh _,_e er_w anot ly_a et_t ften eing houl ll_t _oft bein iden
I-EN-S, instruction, Word:	years make people build locate plain define tasks matters note kinds expected sort days makes public gave include disadvantage january
KI-04, FAQs, Char4:	n't_ ._wh on't stio ?_th _doe _of_ does _do_ _tc_ houl _que frit tc_a tc_w llad pall _tr tru ete-
KI-04, FAQs, Word:	answers isn tips asked <N> home online dns questions investing frequently source web erase billing autos sap adding isp
MGC, FAQs, Char4:	12;i _usi ivex vuln lner _tcp _sue -wri _tc_ tc_w tc_a e_tc t_tc _tr tru f_tc frit ritz e_" e"_
MGC, FAQs, Word:	micro site simpsons directory capitalized certifications incompatible <N> retype authenticate web spelled investigation beneficiaries maximum deducted prompted ssn
SAN FAQ, Char4:	ces: :_pu ing_ do_i opic lica ._ho orm_ _tro _cyc urri must rric pica clon trop lone cycl tax_ yclo
SAN FAQ, Word:	forecast observed early hurricanes office publication circular references notices introductory tropical forecasting copyright tax refer typhoons doc year atl

Figure 1: Similarity between classes across collections

Table 5: Cross-testing with character trigrams

test on →	HGC	I-EN-S	KI04	KRYS-I	MGC	SAN	BNC	BROWN
# mapped	1329	250	1205	4360	1305	1400	755	436
HGC	**63.31**	34.00	33.86	38.10	41.99	40.71	39.34	38.30
I-EN-S	35.59	**54.40**	25.31	25.64	27.82	28.07	47.15	29.13
KI04	35.14	33.60	**81.91**	33.07	32.26	56.36	34.83	24.08
KRYS-I	38.68	29.60	27.80	**57.77**	32.03	21.79	47.81	55.50
MGC	46.95	37.60	34.52	36.33	**53.26**	38.00	55.23	44.04
SAN	31.98	22.80	40.41	13.03	23.60	**96.21**	20.53	4.13
BNC	37.77	42.00	23.82	29.13	34.02	19.36	**71.80**	58.95
BROWN	28.07	21.60	20.83	34.01	25.44	11.29	45.17	**62.60**

3.3. Cross-testing the collections

Even if two collections differ from the viewpoint of their content and genre labels used, a successful AGI application based on one of them needs to be able to process any webpage. More specifically, this means that we need to test how well an SVM classifier trained on one collection predicts the genres in another one. This is difficult to achieve using the diverse set of original labels. As we mentioned above, only two categories, namely shop and FAQ, are present in some form in all collections (instruction used in I-EN-Sample also includes other types of advice, such as tutorials), while many others need to be mapped to a common representation. For mapping we used the macro-genres of FGC and applied the same procedure to the Brown Corpus and BNC to make them comparable to other webcollections.

The results in Table 5 show that the accuracy of cross classification using character trigrams as the features drops dramatically. The reason we used character trigrams and not tetragrams (which produced the best performance in Table 3) is that they had greater overlap between individual corpora, so they are more generalisable.

3.4. Human agreement

One problem with existing collections is that they are mostly annotated by a single person and therefore not tested for reliability. This issue affects SANTINIS and KI-O4.[3] HGC does report a very small-scale agreement study on 70 texts (approximately two of each genre) and two annotators with a percentage agreement of 76%. However, we are really looking for an agreement measure such as Cohen's kappa or Krippendorff's alpha that corrects for chance agreement as well as studies of principled errors or confusions (see (Arstein and Poesio, 2008) for a survey on agreement measures).

MGC as well as KRYS have been double-annotated (MGC fully and KRYS partially). MGC gives access to the decisions of individual annotators, so we were able to conduct an agreement study. As MGC allows for annotation of web pages with multiple categories we used both standard kappa as well as variant of alpha, the latter giving a portion of agreement to partial overlaps. Kappa was 0.57

but when partial overlap was taken into account alpha was 0.71, which at first glance looks encouraging. [4] However, the agreement is mainly achieved on the targeted pages the MGC corpus includes, i.e. pages which were specifically selected to represent the a priori agreed categories. Agreement on random web pages and on Zeitgeist was substantially lower with alpha of 0.55 and 0.56, respectively. In addition, when computing single-category reliability it emerges that only very few categories of 20 categories are reliable (for example, 5 out of 20 categories when looking at the subset of random web pages).

KRYS reports its own agreement study with only about 50-60% percentage agreement – as chance corrected agreement will be even lower, we can conclude that this annotation is also not reliable. Note, however, that KRYS makes very fine-grained distinctions of 70 categories.

I-EN-Sample was annotated originally by one annotator alone (the first author of this paper). To allow for a reliability study, the third author annotated all corpus entries separately. Alpha was used and stood at 0.55, similar to agreement on random pages of MGC. Only one of 8 categories (the category of regulation) could be assigned reliably. Especially high were the confusion rates between the categories of reporting, discussion and propaganda.

To conclude, current genre annotation schemes are either not evaluated or not evaluated on a reasonable scale for reliability or fall far short of the reliability normally expected in other annotation tasks. This is especially the case for portions of the web randomly extracted, which entails that the annotation schemes are not representative for the whole web, either.

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 webgenre collections: if each of them is so different from any

---

[3]As SANTINIS targets particular clear-cut examples of web pages only, this is less likely to be a problem.

[4]Normally an agreement above 0.67 is considered at least marginally reliable, although recent work such as (Reidsma and Carletta, 2008) reminds of the fact that apart from a single reliability figure, the annotator bias should also be taken into account.

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 might not be fully reliably annotated, some of the very impressive results reported for some collections in Table 3 might not be really applicable to the real web.

However, in addition to the negative results, this study suggests a potential for further AGI research. To achieve progress in this field we need a large reference corpus which is collected from a diverse range of sources, so that it can overcome the limitations of each individual collection. A simple concatenation of the available collections is not enough, as each collection is based on its own principles of genre selection and annotation. The reference corpus needs to be accompanied with a set of genre labels allowing consistently reliable annotation. The field also needs more research into detecting features which perform across a large number of texts and do not depend on accidental properties of an individual collection. Ideally, we need a range of corpora for several languages, so that we can learn features which can perform across a range of languages and cultures.

5. Acknowledgements

References

Arstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics (survey article). *Computational Limnguistics*, 34(4):555–596.

Berninger, V., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *Proceedings of the Corpus Profiling Workshop*, London.

Boese, E. and Howe, A. (2005). Effects of web document evolution on genre classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 639–646, Bremen.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Giesbrecht, E. and Evert, S. (2009). Part-of-Speech (POS) Tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35, Donostia-San Sebastián.

Kanaris, I. and Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. In *Proceedings of ICTAI*.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, pages 1071 – 1075, Kyoto, Japan.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Lim, C., Lee, K., and Kim, G. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management*, 41(5):1263–1276.

Mason, J., Shepherd, M., and Duffy, J. (2009). An N-gram based approach to automatically identifying web page genre. In *Proc. 42nd Hawaii International Conference on System Sciences*.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, Ulm, Germany.

Reidsma, D. and Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton.

Santini, M. (2010). Cross-testing a genre classification model for the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Santini, M. and Sharoff, S. (2009). Web genre benchmark under construction. *Journal for Language Technology and Computational Linguistics*, 25(1):125–141.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve.

Sharoff, S. (2010). In the garden and in the jungle: Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. In Santini, M. and Sharoff, S., editors, *Proc. Towards a Reference Corpus of Web Genres*.

Vidulin, V., Luštrek, M., and Gams, M. (2007). Using genres to improve search engines. In *Proc. Towards Genre-Enabled Search Engines: The Impact of NLP. RANLP-07*.

Webber, B. (2009). Genre distinctions for discourse in the Penn TreeBank. In *Proc the 47th Annual Meeting of the ACL*, pages 674–682.