

Description of 3 feature sets for automatic identification of genres in web pages

1 Three Feature Sets for Supervised Classification

1_set contains the following features:

- 50 most common words in English;
- 24 Part-of-Speech (POS) tags from Connexor (Tapanainen and Järvinen 1997);
- punctuation symbols;
- genre-specific facets for the 7-web-genre collection and 8 genre-specific facets for Meyer zu Eissen collection;
- 28 HTML tags (see Appendix)
- 1 nominal attribute representing the length of the web page (SHORT, MEDIUM and LONG).

2_set contains the following features:

- 100 POS trigrams for the 7-web-genre collection and 76 POS trigrams for Meyer zu Eissen collection;
- punctuation symbols;
- genre-specific facets for the 7-web-genre collection and 8 genre-specific facets for Meyer zu Eissen collection (as above);
- 28 HTML tags;
- 1 nominal attribute representing the length of the web page.

3_set contains the following features:

- 86 facets;
- genre-specific facets for the 7-web-genre web pages and 8 genre-specific facets for Meyer zu Eissen collection;
- 6 html facets;
- 1 nominal attribute representing the length of the web page.

2 Feature Description

2.1 Punctuation Symbols

Punctuation symbols have proved to be discriminating for genre. Kessler et al. (1997), Stamatatos et al. (2000), Lim et al. (2004) and others have successfully used punctuation marks, which are considered to be shallow features. The eight punctuation symbols used in this research are the following:

colon (:)	exclamation mark (!)
semi-colon (;)	question mark (?)
comma (,)	apostrophe (')
full stop (.)	double quotes (")

Description of 3 feature sets for automatic identification of genres in web pages

2.2 Frequencies of HTML Tags

The distribution of HTML tags is discriminating across different type of web pages. In particular, Lim et al. (2004) have specifically tested their discrimination power. The list of the 28 HTML tags used in this research are:

<a>	<h1>	<object>
<alt>	<h2>	
<applet>	<h3>	<p>
	<h4>	<script>
 	<hr>	
<div>	<mailto>	<table>
<dl>	<i>	<u>
		
	<input>	
<form>	<noscript>	

2.3 Frequencies of POS tags

The distribution of POS tags have proved to vary across different genres. Several studies have been conducted within corpus linguistics that prove this variation, for example Rayson et al. 2002 or Nakamura 90. POS tags have been included in several studies of genre identification, e.g. Karlgren and Cutting 1994 and Meyer zu Eissen 2004

The POS tags used in this thesis belong to Connexor tagset (Tapanainen & Järvinen, 1997) and are listed below¹:

¹ The complete tagset is available at: <http://www.connexor.com/demo/doc/enfdg3-tags.html>

Description of 3 feature sets for automatic identification of genres in web pages

ABBR=abbreviation	NEGPART=negative particle
ADJ=adjective	N=noun
ADV=adverb	NUM=numeral
CC=coordinating conjunction	PREP=preposition
CS=subordinating conjunction	PRON=pronoun
DET=determiner	PRON PERS=personal pronouns
EN=past participle	Rel=relative marker
Ex=existential 'there'	V AUX MOD=auxiliary
INFMARK=infinitive marker, i.e. 'to'	V IMP=imperative
ING=-ing marker for verb	V INF=infinitive
INTERJ=interjection	V PAST=past
Inter=interrogative marker	V PRES=present

2.4 Web Page Length

The length of a web page can be revealing for some web genres. For example, blogs tend to be long, because they have the form of a diary, where the narration of one day follows another, while search pages tend to be short, showing often only a few sentences and an input field for specifying keywords. Three different lengths are included in this facet, SHORT (<150 words per web page), MEDIUM (<500 words per web page) and LONG (>=500 words).

2.5 Frequencies of Function Words

The rationale behind the use of the frequencies of function words (a method used in stylometrics) is that they are expected to remain invariant for the same author. This assumption has been exported to genre by Stamatatos et al. (2000) who used the 50 most common word in the BNC, which mostly coincide with function words, over the Wall Street Journal corpus with good accuracy results. The 50 common words suggested by Stamatatos et al. (2000) and used in this research are shown in Figure 1.

1. the	11. with	21. are	31. or	41. her
2. of	12. he	22. not	32. an	42. n't
3. and	13. be	23. his	33. were	43. there
4. a	14. on	24. this	34. we	44. can
5. in	15. i	25. from	35. their	45. all
6. to	16. that	26. but	36. been	46. as
7. is	17. by	27. had	37. has	47. if
8. was	18. at	28. which	38. have	48. who
9. it	19. you	29. she	39. will	49. what
10. for	20. 's	30. they	40. would	50. said

Figure 1. The 50 most common words in the BNC (from Stamatatos et al., 2000)

2.6 Frequencies of POS Trigrams

The rationale behind the use of POS trigrams is that POS trigrams are large enough to encode useful syntactic information, and small enough to be computationally manageable. It is a feature borrowed from stylometrics and authorship attribution. Argamon et al. (1998) tried them for discrimination across different textual styles -- namely New York Times news, New York Times editorials, New York Daily news and Newsweek. In Chapter 6, we tried their discriminating power on 10 genres in the BNC with very good results. The selection of POS trigrams to be included in a feature set can be computed in different ways. In order to be revealing, POS trigrams should not be too rare or too frequent. While Argamon et al. (1998) selected POS trigrams that appeared in between 25 and 75% of the documents in their corpus for a total of 685 POS trigrams, in our experiment we selected POS trigrams by extracting all the POSs from the whole corpus and then computing POS trigrams using a freeware utility (kfNgram); only POS with a coverage from 40% to 70% were considered to be good candidates for genre discrimination.

2.7 Linguistic Facets

Linguistic facets are groups of features collected together because of the common communicative function they share². A linguistic facet can be seen as a macro-feature made up with computationally-extractable surface cues which can be interpreted functionally. The functional interpretation of co-occurring features was one of the original points of Biber's approach to the identification of text types (Biber 1988). The label 'linguistic facet' has been created to stress the fact that each of this groups of features highlights a facet, i.e. an aspect in the communicative context that is reflected in the use of language. The frequencies of occurrence, or sometimes the mere presence of a facet, help us understand the use and the variation of the language across different kinds of texts. 86 linguistic facets were employed: 43 are functional cues, 6 ambiguous subordinators, 29 syntactic patterns or subordinate clauses, and 8 syntactic patterns for simple sentences.

² Linguistic facets are fully described in: Marina Santini, *Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features*, Technical Report ITRI-05-02, 2004, ITRI, University of Brighton (UK), available at <http://www.nltg.brighton.ac.uk/home/Marina.Santini/>.

Description of 3 feature sets for automatic identification of genres in web pages

predicators	whereas
nominals	when
first_person	since
second_person	if
third_person	as
third_per_sing_inanimate	-----
present_tense_group	verb_to_clause
past_tense_group	concession_clause_init.
imperatives	concession_clause_final
active	concession_clause_spec.
passive	contrast_clause
time_markers	exception_clause
location	reason_clause_initial
instrument	reason_clause_final
advl_manner	space_clause_initial
negative_particles	space_clause_final
probability_markers	time_clause_initial
necessity_markers	time_clause_final
existential_there	time_clause_instruction.
expressiveness	time_clause_incidental
colon	conditional_clause_init.
question	conditional_clause_final
quotes	conditional_clause_spec.
activity_verbs	result_clause
communication_verbs	sim_manner_comp_clause
mental_verbs	complex_np
causative_verbs	verb_that_clause
occurrence_verbs	adjective_that_clause
existence_verbs	wh_clause
aspectual_verbs	adjective_to_clause
enumerative	verb_ing_clause
equative	purpose_clause
reinforcing	that_omission
summative	comparative_clause
appositive	relative_clause
resultative	-----
inferential	phen_registering
reformulatory	cont_action_recording
replacive	act_recording
antithetic	phen_identifying
concessive	phen_linking
discoursal	qua_attributing
temporal	phen_identifying_mod
-----	act_demanding_com
while	

2.8 Genre-Specific Words

Sets of genre-specific words were created both the 7-web-page collection and Meyer-zu-Eissen collection, one set per genre. Ideally, these groups of lexical items contain a set of words that are topic-independent, but are regularly used in a web page belonging to a specific genre.

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

blog_words (7-web-page collection)
eshop_words (7-web-page collection and Meyer-zu-Eissen web page collection)
faq_words (7-web-page collection)
frontpage_words (7-web-page collection)
listing_words (7-web-page collection)
php_words (7-web-page collection)
spage_words (7-web-page collection)

article_words (Meyer-zu-Eissen web page collection)
discussion_words (Meyer-zu-Eissen web page collection)
download_words (Meyer-zu-Eissen web page collection)
help_words (Meyer-zu-Eissen web page collection)
list_of_links_words (Meyer-zu-Eissen web page collection)
portrayal_non_priv (Meyer-zu-Eissen web page collection)
portrayal_priv (Meyer-zu-Eissen web page collection)

These words relate to genre-specific referential vocabulary, such as “Welcome to my personal home page”, “Thank you for stopping by my personal home page”, “This blog is my corner”, “Are you looking for something? ABACHO the powerful search engine will find it for you!”, etc. Additionally, they include a number of other words relating to the function of the web page. For example, eshop words contain terms such as *buy*, *order*, *purchase*, etc. connected to the function of “selling” and not to what is actually sold, which can vary from fruit to electronic equipment, to clothes to estates.

In order to select these sets of genre-specific words, we tried first with automatic extraction of the most common words per genre and then we measured the coverage of these words over the web pages belonging to a single genre. The original idea was to use words (not all words, but only lemmas from nouns, adjectives and verbs) that would occur in 80% of web pages per genre. The result of this automatic approach was quite disappointing. Here is an example of blog words with a coverage of at least 80% of blog web pages:

go	only
post	thing
comment	new
time	look
day	archive
good	way
find	work

While words like *post*, *day*, *comment*, can be good suggestions because blogs often show sentences like “posted by” or “comment #” or “number of comments” and so on. But words such as *good*, *find*, *new* cannot be considered blog words and presumably they don’t have much discriminating power. This list is too corpus-dependent (only 200 blogs are included in the web corpus) and even if the threshold of the coverage is decreased from 80% to 60% or 50%, the list of automatic extracted words still lacks generality.

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

In order to make a more comprehensive and potentially more general list, we worked out new lists based on a cursory qualitative analysis of the web genres. This qualitative list was conflated with some of the most interesting words coming from automatic extraction. Here are the lists of genre-specific words included in the final feature sets:

Blog words:

web log	sept
weblog	september
blog	oct
journal	october
diary	nov
posted by	november
comments	dec
archive	december
-----	-----
jan	mon
january	monday
feb	tues
february	tuesday
mar	wed
march	weds
apr	wednesday
april	thurs
may	thursday
jun	fri
june	friday
jul	sat
july	saturday
aug	sun
august	sunday

Eshop words:

£	pay
basket	price
buy	purchase
cart	rebate
catalogue	save
checkout	sell
cost	shipping
credit card	shop
debit card	store
delivery	story
offer	trolley
order	

FAQs words:

faq	enquir
frequently asked question	inquir
answer	

Front Page words:

frontpage	front page
-----------	------------

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

news
editor
column
story
stories

headline
opinion
report
newspaper

Listing words:

check list
checklist
hot list
hotlist
site map
sitemap

contents
toc
index
step
list
map

Personal home page words:

about me
curriculum vitae
cv
guest book
guestbook
home page
homepage
my page
interests
my site
my web page
my web site
my web-site

my webpage
personal page
personal web site
personal web-site
personal website
project
publications
research
resumé
vita
's web page
's webpage
's page

Search words:

advanced search
crawl
directories
engine

find
search
see

Article words

abstract
analys/analyz
approach
argument
article
book
chapter
case
challenge
claim
conclusion
design
development
document
explore
figure
function

introduction
investigat
issue
keyword
method
motiv
paper
press release
problem
proceeding
proces
project
publication
question
reference
report
research

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

result
review
section
solution

symposium
study
table

Discussion words

board
bulletin
discussion
fori
forum
mailing
message

moderator
post
problem
story
thread
topic
user

Download words

day
demo
download
freeware
groupware

shareware
software
trial
version

Help words

A:
answer
assistance
enquir
FAQ
Frequently Asked Question

help
inquir
q&a
Q:
question

List of links words

catalog
content
directory
index

link
list
map
toc

Portrayal (non-priv) words

about
brochure
bureau
business
club
compan
donat
flier

institution
logo
mission
organization/
portal
profile
public

Portrayal (priv) words

's page
's web page
's webpage
about me
curriculum vitae
cv

guest book
guestbook
home page
homepage
interest
my page

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

my site	personal web-site
my web page	personal website
my web site	project
my web-site	publication
my webpage	research
personal page	resum
personal web site	vita

2.9 HTML Facets

HTML facets contains are groups of HTML tags interpreted functionally. The importance of the textual interpretation of links become increasingly acknowledged. A number of qualitative analysis of links in relation to genre have been carried out hitherto, for example Haas and Grams (1998) Roberts (1998), Crowston and Williams (1999), Amitay (2000) and above all the very original interpretation of links of home pages in terms of text types by Askehave and Nielsen (2005).

HTML facets have been designed for an automatic approach to genre identification. The six HTML facets included in the final feature sets are the following:

layout	navigability_general
typography	navigability_external
functionality	navigability_internal

Layout. The layout facet contains HTML tags that relates the formatting of the text. Most of the web pages use lots of tags for layout. But some genres can be more regularly laid-out than others, for example FAQs or blogs are moderately laid-out, while front pages or eshops rely more on the visual impact of structured information. The impact of layout in different genres is not new (cf. Waller, 1989, but also the GEM project in Natural Language Generation at <http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/newframe.html>). The tags included in the layout facets are the following:

```
<br          # break the flow of a line,
<dl          # create a definition list
<dir         # create a directory list
<div        # division
<hr         # horizontal rule
<listing    # listing text (obsolete)
<menu       # crate a menu list
<ol         # ordered list
<p>         # paragrap
<pre        #preformatted text
<table      # table
<ul         # unordered list, i.e. bulleted list
```

Typography. Typographic devices are widely used in all web pages. Typography is another standard and traditional device of highlighting the importance of information

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

(cf. Waller, 1989). Again, some web genres rely more on typography than others (for example, front pages employ lots of headings while FAQs are plainer). The HTML tags included in this facets are:

```
<abbr           # abbreviavation
<acronym        # acronym
<address        # usually rendered in an italic font
<b>             # bold
<big>          # bit text
<blockquote    # a block quotation
<caption       # caption for a table
<center        # center the text
<cite          # identify a citation
<em>           # emphasis, usually italics
<font>        # change font characteristics
<h1>          # heading
<h2>          # heading
<h3>          # heading
<h4>          # heading
<h5>          # heading
<h6>          # heading
<i>            # italics
<img>         # a link to an external image
<q>           # quote
<s>           # strikeouts
<small>       # small text
<strike>      # strikeouts
<strong>      # usually bold
<style>       # style sheet definition
<sub>         # subscript
<sup>         # superscript
<tt>         # typewriter text
<u>           # underline
alt           # a short description of an image
```

Functionality. Functionality is mainly related to the possibility of user interaction with web pages. It is an important trait of many web pages, but it not present in all al them, for example academic paper usually do not show any interactivity, while search page are highly interactive. Another interesting set of functionality features for automatic genre classification is suggested by Shepherd et al. (2004). The HTML tags included in this facet are:

```
<applet        # calls a java applet
<bgsound       # background sound (obsolete)
<button        # create a button in a form
<embed         # embed multimedia (obsolete)
<fieldset      # groups together fields in a form
<form          # create a form for user input
<input         # an input element in a form
<legend        # provide a legend
<noscript     # for browsers not support scripts
```

Marina Santini (2005-2006)

Description of 3 feature sets for automatic identification of genres in web pages

```
<object          # can subsume images, applets etc.
<option         # a selectable option in a form
<optgroup       # provide a hierarchy of choices
<param         # a parameter passed to an object
<script        # insert an inline script
<select        # option selector element in a form
<text area     # a freeform text entry in a form
<var           # a program variable
mailto         # link to an email address
```

General Navigability. The general navigability facet measures the general level of hyperlinking of a web page. It includes both external and internal hyperlink tags.

```
<a
```

External Navigability. The external navigation facet measures the level of external navigability of a web page. The HTML tags included in this facet are only two:

```
href="http"
href="ftp"
```

Internal Navigability. The internal navigation facet measures the level of internal navigability of a web page. The HTML tag included in this facet is:

```
"href="#"
```

The end