EACL-2006

11th Conference
of the European Chapter of the
Association for Computational
Linguistics

Proceedings of the workshop on

NEW TEXT
Wikis and blogs and other dynamic text sources

April, 4, 2006
Trento, Italy

EACL-2006

11th Conference
of the European Chapter of the
Association for Computational
Linguistics

Proceedings of the workshop on

NEW TEXT
Wikis and blogs and other dynamic text sources

April, 4, 2006
Trento, Italy

Preface

New types of text sources, multi-lingual, with numerous cooperating or even adversarial authors and little or no editorial control are one effect of the recently dramatically lowered publication threshold. Many contain linguistic items or features classically associated with spoken language — combining the high interactivity of dialogue with the low bandwidth of written text and with the multicasting capabilities of digital communication.

New material published today most noticeably includes *blogs* — a genre that has evolved from diaries, logbooks, commentaries, columns, and editorials into a multi-faceted and networked churn of text with widely ranging viewpoints and perspectives and varying application and ambition on the part of the creator. One of the most noticeable charateristics of the blog genre is its opinionated nature and its timeliness. Blog texts are often ill-edited and hastily cobbled together in a language reminiscent of brief notes, spoken asides, or short letters, rather than of essays or newsprint. This, at any rate, is the public perception.

Another emergent genre is that of the *wiki*, a shared workspace for many more or less equal participants: wiki texts are written and edited by open teams of authors. The best known application of the wiki is the *wikipedia* — closely patterned on a classic text genre, that of the encyclopedia; other applications include e.g. project management or creative text authoring. In contrast to blogs, wikis (especially the wikipedia applications) tend to have high ambitions as regards factual correctness, persistence, editorial quality, and trustworthiness.

Bridging the two are genres such as discussion boards, web fora, and mailing lists.

Let us call these various new types of text (or indeed other modes of linguistic communication) collectively New Text. This workshop is intended to discuss the analysis and application of new text, formulate research measures that are crying out to be taken, discuss which methodological steps are obsoleted, and which babies can be saved from the bath water.

Challenge questions

NEW TEXT provides a number of research issues, immediately obvious questions, and tentative applications for our research fields:

1. New possibilities for the philologically inclined: How does new text cast new light on human communicative behaviour? This includes question on style and genre: the characteristics of new text and relations to traditional media. Do blogs in fact resemble spoken language in any important way? Do wikis hold up their promise of qualitative information dissemination? How has research in textuality, discourse and linguistic behaviour been hindered by reliance on well-edited and well-groomed data sets? Or, in more positive words: what advances can we expect, either in terms of application or in terms of understanding human behaviour, by the new data sources available to us now?

2. New challenges for building text analysis tools – how are the today's algorithms portable to new text? This includes questions on multilinguality,

code-switching, register variation, and formality melange apparent in new text.

3. New challenges for evaluation methodologies for information access systems:
   - Can new text, with dynamic information sources and streams of variable quality and impact be plugged into relevance-oriented evaluation frameworks without revising the target notion of text relevance?
   - Some new texts have high social impact; some sink without a trace; some have high import in tightly knit circles and communities. Traditional media have sales figures, citation indices, and distribution analyses. How can the impact of new texts be analyzed?
   - New texts have variable perceived intellectual status and quality – how can it be measured and predicted?
4. New opportunities for new services – e.g. linking different types of text in dynamic and interactive sessions of information refinement and elaboration.
5. How new is ”new”? Didn’t we use to have new text before? What is the difference between ”new” and ”old”, really?

Welcome!

Welcome to the workshop! Please join in the discussion!

Organizers

- Jussi Karlgren, SICS, Stockholm

Reviewers

- Shlomo Argamon, Illinois Institute of Technology, Chicago, IL
- Paul Clough, University of Sheffield
- Bjrn Gambck, SICS, Stockholm
- Michael Gamon, Microsoft, Redmond
- Julio Gonzalo, UNED, Madrid
- Gilad Mishne, University of Amsterdam
- Fredrik Olsson, SICS, Stockholm
- Martin Svensson, SICS, Stockholm
- zlem Uzuner, MIT, Cambridge

Workshop Program

Tuesday, April 4

9:00 - 10:30	**Session: Usage and the character of the net**
9:00 - 9:10	*Welcome*
	Jussi Karlgren
9:10 - 9:40	*Text Linkage in the Wiki Medium - A Comparative Study*
	Alexander Mehler
9:40 - 9:50	*Errors in wikis*
	Ann Copestake
9:50 - 10:30	*Discussion on quality, trust and authority*
10:30 - 11:00	**Italian Coffee**
11:00 - 12:30	**Session: Data**
11:00 - 11:20	*Linguistic features of Italian blogs: literary language*
	Mirko Tavosanis
11:20 - 11:40	*An analysis of Wikipedia digital writing*
	Antonella Elia
11:40 - 12:00	*Learning to Recognize Blogs: A Preliminary Exploration*
	Erik Elgersma, Maarten de Rijke
12:00 - 12:30	*Discussion on style*
12:30 - 14:30	**Italian lunch**
14:30 - 16:00	**Session: Experiments**
14:30 - 14:45	*Interpreting Genre Evolution on the Web*
	Marina Santini
14:45 - 15:00	*Novelle, a collaborative open source writing tool software*
	Federico Gobbo, Michele Chinosi, Massimiliano Pepe
15:00 - 15:15	*Anomaly Detecting within Dynamic Chinese Chat Text*
	Yunqing Xia, Kam-Fai Wong
15:15 - 15:30	*A proposal to automatically build and maintain gazetteers for Named Entity Recognition*
	Antonio Toral, Rafael Muoz
15:30 - 16:00	*Finding Similar Sentences across Multiple Languages in Wikipedia*
	Sisay Fissaha Adafre, Maarten de Rijke
16:00 - 16:30	**Italian Coffee**
16:30 - 18:00	**Winding Up**
16:30 - 16:40	*Multilingual interactive experiments with Flickr*
	Paul D Clough, Julio Gonzales, Jussi Karlgren
16:40 - 17:00	*Discussion on Common task*
17:00 - 17:30	*Discussion on Resources*
17:30 - 18:00	*Planning ahead: Continuing the discussion: How, Where, In what form?*

Contents

Text Linkage in the Wiki Medium – A Comparative Study

Alexander Mehler
Department of Computational Linguistics & Text Technology
Bielefeld University
Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Abstract

We analyze four different types of document networks with respect to their small world characteristics. These characteristics allow distinguishing wiki-based systems from citation and more traditional text-based networks augmented by hyperlinks. The study provides evidence that a more appropriate network model is needed which better reflects the specifics of wiki systems. It puts emphasize on their topological differences as a result of wiki-related linking compared to other text-based networks.

1 Introduction

With the advent of web-based communication, more and more corpora are accessible which manifest complex networks based on intertextual relations. This includes the area of *scientific communication* (e.g. digital libraries as CiteSeer), *press communication* (e.g. the New York Times which links topically related articles), *technical communication* (e.g. the Apache Software Foundation's documentations of open source projects) and *electronic encyclopedia* (e.g. Wikipedia and its releases in a multitude of languages). These are sources of *large* corpora of web documents which are connected by *citation links* (digital libraries), *content-based add-ons* (online press communication) or *hyperlinks* to related lexicon articles (electronic encyclopedias).

Obviously, a corpus of such documents is more than a set of textual units. There is structure formation above the level of single documents which can be described by means of graph theory and network analysis (Newman, 2003). But what is new about this kind of structure formation? Or do we just have to face the kind of structuring which is already known from other linguistic networks?

This paper focuses on the specifics of networking in wiki-based systems. It tackles the following questions: *What structure do wiki-based text networks have? Can we expect a wiki-specific topology compared to more traditional (e.g. citation) networks? Or can we expect comparable results when applying network analysis to these emerging networks?* In the following sections, these questions are approached by example of a language specific release of the Wikipedia as well as by wikis for technical documentation. That is, we contribute to answering the question why wiki can be seen as something new compared to other text types *from the point of view of networking*.

In order to support this argumentation, section (2) introduces those network coefficients which are analyzed within the present comparative study. As a preprocessing step, section (3) outlines a webgenre model which in sections (4.1) and (4.2) is used to represent and extract instances of four types of document networks. This allows applying the coefficients of section (2) to these instances (section 4.3) and narrowing down wiki-based networks (section 5). The final section concludes and prospects future work.

2 Network Analysis

For the time being, the overall structure of complex networks is investigated in terms of *Small Worlds* (SW) (Newman, 2003). Since its invention by Milgram (1967), this notion awaited formalization as a measurable property of large complex networks which allows distinguishing small worlds from random graphs. Such a formalization was introduced by Watts & Strogatz (1998) who characterize small worlds by two properties: First, other than in regular graphs, any randomly chosen pair of nodes in a small world has, on average, a considerably shorter *geodesic distance*.[1] Second, compared to random graphs, small worlds show a considerably higher level of *cluster formation*.

In this framework, cluster formation is measured by means of the average fraction of the number $\nabla(v_i)$ of triangles connected to vertex v_i and the number $\veebar(v_i)$ of triples centered on v_i (Watts and Strogatz, 1998):[2]

$$C_2 = \frac{1}{n}\sum_i \frac{\nabla(v_i)}{\veebar(v_i)} \tag{1}$$

Alternatively, the cluster coefficient C_1 computes the fraction of the number of triangles in the whole network and the number of its connected vertex triples. Further, the *mean geodesic distance* l of a network is the arithmetic mean of all shortest paths of all pairs of vertices in the network. Watts and Strogatz observe high cluster values and short average geodesic distances in small worlds which apparently combine cluster formation with shortcuts as prerequisites of efficient information flow. In the area of information networks, this property has been demonstrated for the WWW (Adamic, 1999), but also for co-occurrence networks (Ferrer i Cancho and Solé, 2001) and semantic networks (Steyvers and Tenenbaum, 2005).

In addition to the SW model of Watts & Strogatz, link distributions were also examined in order to characterize complex networks: Barabási & Albert (1999) argue that the vertex connectivity of social networks is distributed according to a scale-free power-law. They recur to the observation – confirmed by many social-semiotic networks, but not by instances of the random graph model of Erdős & Rényi (Bollobás, 1985) – that the number of links per vertex can be reliably predicted by a power-law. Thus, the probability $P(k)$ that a randomly chosen vertex interacts with k other vertices of the same network is approximately

$$P(k) \sim k^{-\gamma} \tag{2}$$

Successfully fitting a power law to the distribution of out degrees of vertices in complex networks indicates "that most nodes will be relatively poorly connected, while a select minority of *hubs* will be very highly connected." (Watts, 2003, p.107). Thus, for a fixed number of links, the smaller the γ value, the shallower the slope of the curve in a log-log plot, the higher the number of edges to which the most connected hub is incident.

A limit of this model is that it views the probability of linking a source node to a target node to depend solely on the connectivity of the latter. In contrast to this, Newman (2003) proposes a model in which this probability also depends on the connectivity of the former. This is done in order to account for social networks in which vertices tend to be linked if they share certain properties (Newman and Park, 2003), a tendency which is called *assortative mixing*. According to Newman & Park (2003) it allows distinguishing social networks from non-social (e.g. artificial and biological) ones even if they are uniformly attributed as small worlds according to the model of Watts & Strogatz (1998). Newman & Park (2003) analyze assortative mixing of vertex degrees, that is, the correlation of the degrees of linked vertices. They confirm that this correlation is *positive* in the case of social, but *negative* in the case of technical networks (e.g. the Internet) which thus prove disassortative mixing (of degrees).

Although these SW models were applied to citation networks, WWW graphs, semantic networks and co-occurrence graphs, *and thus to a variety of linguistic networks*, a comparative study which focuses on wiki-based structure formation in comparison to other networks *of textual units* is missing so far. In this paper, we present such a study. That is, we examine SW coefficients which allow distinguishing wiki-based systems from more "traditional" networks. In order to do that, a generalized web document model is needed to uniformly represent the document networks to be compared. In the following section, a webgenre model is outlined for this purpose.

3 A Webgenre Structure Model

Linguistic structures vary with the functions of the discourses in which they are manifested (Biber, 1995; Karlgren and Cutting, 1994). In analogy to the *weak contextual hypothesis* (Miller and Charles, 1991) one might state that structural differences reflect functional ones as far as they are confirmed by a significantly high number of textual units and thus are identifiable as recurrent pat-

[1] The geodesic distance of two vertices in a graph is the length of the shortest path in-between.

[2] A triangle is a subgraph of three nodes linked to each other. Note that all coefficients presented in the following sections relate by default to undirected graphs.

terns. In this sense, we expect web documents to be distinguishable by the functional structures they manifest. More specifically, we agree with the notion of *webgenre* (Yoshioka and Herman, 2000) according to which the functional structure of web documents is determined by their membership in *genres* (e.g. of *conference websites*, *personal home pages* or *electronic encyclopedias*).

Our hypothesis is that what is common to instances of different webgenres is the existence of an implicit *logical document structure* (LDS) – in analogy to textual units whose LDS is described in terms of section, paragraph and sentence categories (Power et al., 2003). In the case of web documents we hypothesize that their LDS comprises four levels:

- *Document networks* consist of documents which serve possibly heterogenous functions if necessary independently of each other. A web document network is given, for example, by the system of websites of a university.

- *Web documents* manifest – typically in the form of websites – pragmatically closed acts of web-based communication (e.g. *conference organization* or *online presentation*). Each web document is seen to organize a system of dependent subfunctions which in turn are manifested by modules.

- *Document modules* are, ideally, functionally homogeneous subunits of web documents which manifest single, but dependent subfunctions in the sense that their realization is bound to the realization of other subfunctions manifested by the same encompassing document. Examples of such subfunctions are *call for papers*, *program presentation* or *conference venue organization* as subfunctions of the function of *web-based conference organization*.

- Finally, elementary *building blocks* (e.g. *lists*, *tables*, *sections*) only occur as dependent parts of document modules.

This enumeration does not imply a one-to-one mapping between functionally demarcated *manifested* units (e.g. modules) and *manifesting* (layout) units (e.g. web pages). Obviously, the same functional variety (e.g. of a personal academic home page) which is mapped by a website of dozens of interlinked pages may also be manifested by a single page. The many-to-many relation induced by this and related examples is described in more detail in Mehler & Gleim (2005).

The central hypothesis of this paper is that genre specific structure formation also concerns document networks. That is, we expect them to vary with respect to structural characteristics according to the varying functions they meet. Thus, we do *not* expect that different types of document networks (e.g. systems of genre specific websites vs. wiki-based networks vs. online citation networks) manifest homogeneous characteristics, but significant variations thereof. As we concentrate on coefficients which were originally introduced in the context of small world analyses, we expect, more concretely, that different network types vary according to their fitting *to* or deviation *from* the small world model. As we analyze only a couple of networks, this observation is bound to the corpus of networks considered in this study. It nevertheless hints at how to rethink network analysis in the context of newly emerging network types as, for example, Wikipedia.

In order to support this argumentation, the following section presents a model for representing and extracting document networks. After that, the SW characteristics of these networks are computed and discussed.

4 Network Modeling and Analysis

4.1 Graph Modeling

In order to analyse the characteristics of document networks, a format for uniformly representing their structure is needed. In this section, we present *generalized trees* for this task. Generalized trees are graphs with a kernel tree-like structure – henceforth called *kernel hierarchy* – superimposed by graph-forming edges as models of hyperlinks. Figure (1) illustrates this graph model. It distinguishes three levels of structure formation:

1. According to the webgenre model of section (3), *L1-graphs* map document networks and thus corpora of interlinked (web) documents.

In section (4.3), four sources of such networks are explored: *wiki document networks*, *citation networks*, *webgenre corpora* and, for comparison with a more traditional medium, *networks of newspaper articles*.

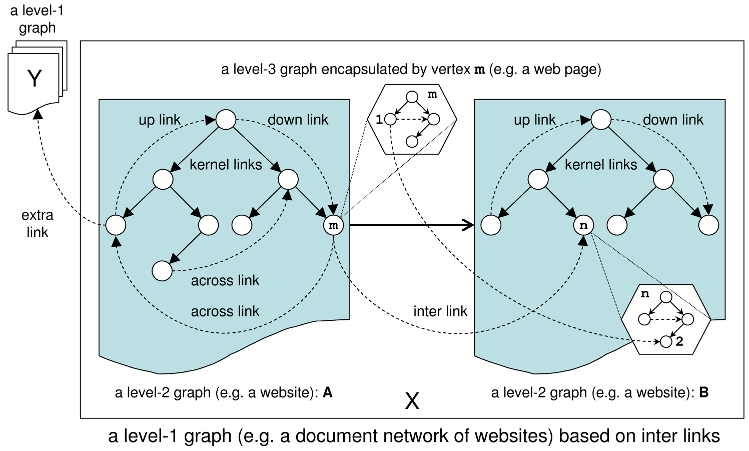

a level-1 graph
Y
a level-3 graph encapsulated by vertex m (e.g. a web page)
m
1
up link
down link
kernel links
extra link
m
across link
across link
inter link
up link
down link
kernel links
n
n
2
a level-2 graph (e.g. a website): A
X
a level-2 graph (e.g. a website): B
a level-1 graph (e.g. a document network of websites) based on inter links

Figure 1: The stratified model of network representation with kernel hierarchies of L2-graphs.

2. *L2-graphs* model the structure of web documents as constituents of a given network. This structure is seen to be based on kernel hierarchies superimposed, amongst others, by *up*, *down* and *across* links (see fig. 1).

In the case of webgenre corpora, L2-graphs model websites. In the case of citation networks, they map documents which consist of a scientific article and add-ons in the form of citation links. Likewise, in the case of online newspapers, L2-graphs model articles together with content-based hyperlinks. Finally, in the case of wikis, L2-graphs represent *wiki documents* each of which consists of a wiki article together with a corresponding discussion and editing page. According to the webgenre model of section (3), L2-graphs model web documents which consist of nodes whose structuring is finally described by L3-graphs:

3. *L3-graphs* model the structure of document modules.

In the case of webgenre corpora, *L3-graphs* map the DOM[3]-based structure of the *web pages* of the websites involved. In the case of all other networks distinguished above they represent the logical structure of single text units (e.g. the section and paragraph structuring of a lexicon, newspaper or scientific article). Note that the tree-like structure of a document module may be superimposed by hyperlinks, too, as illustrated in figure (1) by the vertices m and n.

The kernel hierarchy of an L2-graph is constituted by *kernel links* which are distinguished from *across*, *up*, *down* and *outside* links (Amitay et al., 2003; Eiron and McCurley, 2003; Mehler and Gleim, 2005). These types can be distinguished as follows:

- *Kernel links* associate dominating nodes with their immediately dominated successor nodes in terms of the kernel hierarchy.

- *Down links* associate nodes with one of their (mediately) dominated successor nodes in terms of the kernel hierarchy.

- *Up links* analogously associate nodes of the kernel hierarchy with one of their (mediately dominating) predecessor nodes.

- *Across links* associate nodes of the kernel hierarchy none of which is an (im-)mediate predecessor of the other in terms of the kernel hierarchy.

- *Extra* (or outside) *links* associate nodes of the kernel hierarchy with nodes of other documents.

Kernel hierarchies are exemplified by a *conference website* headed by a title and menu page referring to, for example, the corresponding *call for papers* which in turn leads to pages on the different conference sessions etc. so that finally a hierarchical structure evolves. In this example the kernel hierarchy evidently reflects navigational constraints. That is, the position of a page in the tree reflects

[3] I.e. D*ocument* O*bject* M*odel*.

the probability to be navigated by a reader starting from the root page and following kernel links only.

The kernel hierarchy of a wiki document is spanned by an *article page* in conjunction with the corresponding *discussion* (or *talk*), *history* and *edit this* or *view source* pages which altogether form a flatly structured tree. Likewise in the case of citation networks as the CiteSeer system (Lawrence et al., 1999), a document consists of the various (e.g. PDF or PS) versions of the focal article as well as of one or more web pages manifesting its citations by means of hyperlinks.

From the point of view of document network analysis, L2-graphs and inter links (see fig. 1) are most relevant as they span the corresponding network mediated by documents (e.g. websites) and modules (e.g. web pages). This allows specifying which links of which type in which network are examined in the present study:

- In the case of citation networks, citation links are modeled as interlinks as they relate (scientific) articles encapsulated by documents of this network type. Citation networks are explored by example of the CiteSeer system: We analyze a sample of more than 550,000 articles (see table 1) – the basic population covers up to 800,000 documents.

- In the case of newspaper article networks, content-based links are explored as resources of networking. This is done by example of the 1997 volume of the German newspaper Süddeutsche Zeitung (see table 1). That is, firstly, nodes are given by articles where two nodes are interlinked if the corresponding articles contain *see also* links to each other. In the online and ePaper issue of this newspaper these links are manifested as hyperlinks. Secondly, articles are linked if they appear on the same page of the same issue so that they belong to the same thematic field. By means of these criteria, a bipartite network (Watts, 2003) is built in which the top-mode is spanned by topic and page units, whereas the bottom-mode consists of text units. In such a network, two texts are interlinked whenever they relate to at least one common topic or appear on the same page of the same issue.

- In the case of webgenres we explore a corpus of 1,096 conference websites (see table 1 and 2) henceforth called `indogram` corpus.[4] We analyze the out degrees of all web pages of these websites and thus explore kernel, up, down, across, inter and outside links on the level of L2-graphs. This is done in order to get a *base line* for our comparative study, since WWW-based networks are well known for their small world behavior. More specifically, this relates to estimations of the exponent γ of power laws fitted to their degree distributions (Newman, 2003).

- These three networks are explored in order to comparatively study networking in Wikipedia which is analyzed by example of its German release `de.wikipedia.org` (see table 1). Because of the rich system of its node and link types (see section 4.2) we explore three variants thereof. Further, in order to get a more reliable picture of wiki-based structure formation, we also analyze wikis in the area of technical documentation. This is done by example of three wikis on open source projects of the Apache Software Foundation (cf. `wiki.apache.org`).

variable	value
number of web sites	1,096
number of web pages	50,943
number of hyperlinks	303,278
maximum depth	23
maximum width	1,035
average size	46
average width	38
average height	2

Table 2: A corpus of conference and workshop websites (counting unit: web pages).

In the following section, the extraction of Wikipedia-based networks is explained in more detail.

4.2 Graph Extraction – the Case of Wiki-based Document Networks

In the following section we analyze the network spanned by document modules of the German Wikipedia and their inter links.[5] This cannot simply be done by extracting all its article pages. The reason is that Wikipedia documents consist

[4] See `http://ariadne.coli.uni-bielefeld.de/indogram/resources.html` for the list of URLs of the documents involved.

[5] We downloaded and extracted the XML release of this wiki – cf. `http://download.wikimedia.org/wikipedia/de/pages_current.xml.bz2`.

network	network genre	node	$\lvert V\rvert$	$\lvert E\rvert$
`de.wikipedia.org`	electronic encyclopedia	wiki unit		
`variant I`		(e.g. article or talk)	303,999	5,895,615
`variant II`			406,074	6,449,906
`variant III`			796,454	9,161,706
`wiki.apache.org/jakarta`	online technical documentation	wiki unit	916	21,835
`wiki.apache.org/struts`	online technical documentation	wiki unit	1,358	40,650
`wiki.apache.org/ws`	online technical documentation	wiki unit	1,042	23,871
`citeseer.ist.psu.edu`	digital library	open archive record	575,326	5,366,832
`indogram`	conference websites genre	web page	50,943	303,278
`Süddeutsche Zeitung 1997`	press communication	newspaper article	87,944	2,179,544

Table 1: The document networks analyzed and the sizes $|V|$ and $|E|$ of their vertex and edge sets.

of modules (manifested by pages) of various types which are likewise connected by links of different types. Consequently, the choice of instances of these types has to be carefully considered.

Table (3) lists the node types (and their frequencies) as found in the wiki or additionally introduced into the study in order to organize the type system into a hierarchy. One heuristic for extracting instances of node types relates to the URL of the corresponding page. Category, portal and media wiki pages, for example, contain the prefix `Kategorie`, `Portal` and `MediaWiki`, respectively, separated by a colon from its page name suffix (as in `http://de.wikipedia.org/wiki/Kategorie:Musik`).

Analogously, table (4) lists the edge types either found within the wiki or additionally introduced into the study. Of special interest are *redirect* nodes and links which manifest transitive and, thus, mediate links of content-based units. An article node v may be linked, for example, with a redirect node r which in turn redirects to an article w. In this case, the document network contains two edges $(v, r), (r, w)$ which have to be resolved to a single edge (v, w) if redirects are to be excluded in accordance with what the MediaWiki system does when processing them.

Based on these considerations, we compute network characteristics of three extractions of the German Wikipedia (see table 1): *Variant I* consists of a graph whose vertex set contains all *Article* nodes and whose edge set is based on *Interlink*s and appropriately resolved *Redirect* links. *Variant II* enlarges variant I by including other content-related wiki units, i.e. *ArticleTalk*, *Portal*, *PortalTalk*, and *Disambiguation* pages (multiply typed nodes were excluded). *Variant III* consists of a graph whose vertex set covers all vertices and edges found in the extraction.

Type	Frequency
Documents total	796,454
Article	303,999
RedirectNode	190,193
Talk	115,314
ArticleTalk	78,224
UserTalk	30,924
ImageTalk	2,379
WikipediaTalk	1,380
CategoryTalk	1,272
TemplateTalk	705
PortalTalk	339
MediaWikiTalk	64
HelpTalk	27
Image	97,402
User	32,150
Disambiguation	22,768
Category	21,999
Template	6,794
Wikipedia	3,435
MediaWiki	1,575
Portal	791
Help	34

Table 3: The system of node types and their frequencies within the German Wikipedia.

4.3 Network Analysis

Based on the input networks described in the previous section we compute the SW coefficients described in section (2). Average geodesic distances are computed by means of the Dijkstra algorithm based on samples of 1,000 vertices of the input networks (or the whole vertex set if it is of minor cardinality). Power law fittings were computed based on the model $P(x) = ax^{-\gamma} + b$. Note that table (1) does not list the cardinalities of multi sets of edges and, thus, does not count multiple edges connecting the same pair of vertices within the corresponding input network – therefore, the numbers in table (1) do not necessarily conform to the counts of link types in table (4). Note further that we compute, as usually done in SW analyses, characteristics of *un*directed graphs. In the case of wiki-based networks, this is justified by the possibility to process *back links* in `Media Wiki` systems. In the case of the CiteSeer system this is justified by the fact that it always displays *citation*

Type	Frequency
Links total	17,814,539
Interlink	12,818,378
CategoryLink	1,415,295
Categorizes	704,092
CategorizedBy	704,092
CategoryAssociatesWith	7,111
TopicOfTalk	103,253
TalkOfTopic	88,095
HyponymOf	26,704
HyperonymOf	26,704
InterPortalAssociation	1,796
Broken	2,361,902
Outside	1,276,818
InterWiki	789,065
External	487,753
Intra	1,175,290
Kernel	1,153,928
Across	6,331
Up	6,121
Reflexive	5,433
Down	3,477
Redirect	182,151

Table 4: The system of link types and their frequencies within the German Wikipedia.

and *cited by* links. Finally, in the case of the newspaper article network, this is due to the fact that it is based on a bipartite graph (see above). Note that the `indogram` corpus consists of predominantly unrelated websites and thus does not allow computing cluster and distance coefficients.

5 Discussion

The numerical results in table (5) are remarkable as they allow identifying three types of networks:

- On the one hand, we observe the extreme case of the `Süddeutsche Zeitung`, that is, of the newspaper article network. It is the only network which, at the same time, has very high cluster values, short geodesic distances *and* a high degree of assortative mixing. Thus, its values support the assertion that it behaves as a small world in the sense of the model of Watts & Strogatz. The only exception is the remarkably low γ value, where, according to the model of Barabási & Albert (1999), a higher value was expected.

- On the other hand, the CiteSeer sample is the reverse case: It has very low values of C_1 *and* C_2, tends to show neither assortative, nor dis-assortative mixing, and at the same time has a low γ value. The small cluster values can be explained by the low probability with which two authors cited by a focal article are related by a citation relation on their own.[6]

- The third group is given by the wiki-based networks: They tend to have higher C_1 and C_2 values than the citation network does, but also tend to show stochastic mixing and short geodesic distances. The cluster values are confirmed by the wikis of technical documentation (also w.r.t their numerical order). Thus, these wikis tend to be small worlds according to the model of Watts & Strogatz, but also prove disassortative mixing – comparable to technical networks *but in departure from social networks*. Consequently, they are ranked in-between the citation and the newspaper article network.

All these networks show rather short geodesic distances. Thus, l seems to be inappropriate with respect to distinguishing them in terms of SW characteristics. Further, all these examples show remarkably low values of the γ coefficient. In contrast to this, power laws as fitted in the analyses reported by Newman (2003) tend to have much higher exponents – Newman reports on values which range between 1.4 and 3.0. This result is only realized by the `indogram` corpus of conference websites, thus, by a sample of WWW documents whose out degree distribution is fitted by a power law with exponent $\gamma = 2.562$.

These findings support the view that compared to WWW-based networks wiki systems behave more like "traditional" networks of textual units, *but are new in the sense that their topology neither approximates the one of citation networks nor of content-based networks of newspaper articles*. In other words: As intertextual relations are genre sensitive (e.g. citations in scientific communication vs. content-based relations in press communication vs. hyperlinks in online encyclopedias), networks based on such relations seem to inherit this genre sensitivity. That is, for varying genres (e.g. of scientific, technical or press communication) differences in topological characteristics of their instance networks are expected. The study presents results in support of this view of the genre sensitivity of text-based networks.

6 Conclusion

We presented a comparative study of document networks based on small world characteristics.

[6] Although articles can be expected which cite, for example, de Saussure and Chomsky, there certainly exist much less citations of de Saussure in articles of Chomsky.

instance	**type**	$\langle d \rangle$	l	γ	C_1	C_2	r
Wikipedia variant I	undirected	19.39	3.247	0.4222	0.009840	0.223171	-0.10
Wikipedia variant II	undirected	15.88	3.554	0.5273	0.009555	0.186392	-0.09
Wikipedia variant III	undirected	11.50	4.004	0.7405	0.007169	0.138602	-0.05
wiki.apache.org/jakarta	undirected	23.84	4.488	0.2949	0.193325	0.539429	-0.50
wiki.apache.org/struts	undirected	29.93	4.530	0.2023	0.162044	0.402418	-0.45
wiki.apache.org/ws	undirected	22.91	4.541	0.1989	0.174974	0.485342	-0.48
citeseer.ist.psu.edu	undirected	9.33	4.607	0.9801	0.027743	0.067786	-0.04
indogram	directed	5.95	× × ×	2.562	× × ×	× × ×	× × ×
Süddeutsche Zeitung	undirected	24.78	4.245	0.1146	0.663973	0.683839	0.699

Table 5: Numerical values of SW-related coefficients of structure formation in complex networks: the average number $\langle d \rangle$ of edges per node, the mean geodesic distance l, the exponent γ of successfully fitted power laws, the cluster values C_1, C_2 and the coefficient r of assortative mixing.

According to our findings, three classes of networks were distinguished. This classification separates wiki-based systems from more traditional text networks but also from WWW-based webgenres. Thus, the study provides evidence that there exist genre specific characteristics of text-based networks. This raises the question for models of network growth which better account for these findings. Future work aims at elaborating such a model.

References

Lada A. Adamic. 1999. The small world of web. In Serge Abiteboul and Anne-Marie Vercoustre, editors, *Research and Advanced Technology for Digital Libraries*, pages 443–452. Springer, Berlin.

Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, and Aya Soffer. 2003. The connectivity sonar: detecting site functionality by structural patterns. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia*, pages 38–47.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–512.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Béla Bollobás. 1985. *Random Graphs*. Academic Press, London.

Nadav Eiron and Kevin S. McCurley. 2003. Untangling compound documents on the web. In *Proceedings of the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK*, pages 85–94.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The small-world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, November.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of COLING '94*, volume II, pages 1071–1075, Kyoto, Japan.

Steve Lawrence, C. Lee Giles, and Kurt Bollacker. 1999. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71.

Alexander Mehler and Rüdiger Gleim. 2005. The net for the graphs — towards webgenre representation for corpus linguistic studies. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as corpus*. Gedit, Bologna, Italy.

Stanley Milgram. 1967. The small-world problem. *Psychology Today*, 2:60–67.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mark E. J. Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E*, 68:036122.

Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review*, 45:167–256.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Mark Steyvers and Josh Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Duncan J. Watts. 2003. *Six Degrees. The Science of a Connected Age*. Norton & Company, New York.

Takeshi Yoshioka and George Herman. 2000. Coordinating information using genres. Technical report, Massachusetts Institute of Technology, August.

Errors in wikis: new challenges and new opportunities — a discussion document

Ann Copestake
Computer Laboratory
University of Cambridge
aac@cl.cam.ac.uk

Abstract

This discussion document concerns the challenges to assessments of reliability posed by wikis and the potential for language processing techniques for aiding readers to decide whether to trust particular text.

1 Wikis and the trust problem

Wikis, especially open wikis, pose new challenges for readers in deciding whether information is trustworthy. An article in a wikipedia may be generally well-written and appear authoritative, so that the reader is inclined to trust it, but have some additions by other authors which are incorrect. Corrections may eventually get made, but there will be a time lag. In particular, many people are now using Wikipedia (www.wikipedia.org) as a major reference source, so the potential for misinformation to be spread is increasing. It has already become apparent that articles about politicians are being edited by their staff to make them more favourable and no doubt various interest groups are manipulating information in more subtle ways. In fact, as wikis develop, problems with reliability may get worse: authors who wrote an article several years ago won't care so much about its content and may not bother to check edits. When obscure topics are covered by a wiki, the community which is capable of checking facts may be small.

Of course errors arise in old text too, but a generally authoritative conventional article is unlikely to contain a really major error about a central topic. Different old text publications have different perspectives, political or otherwise, but the overall slant is usually generally known and hence not problematic. Non-wiki web pages may have unknown authors, but the domain offers some guide to reliability and to likely skew and the pages can be assessed as a whole. The issue here is not the overall number of errors in wikis versus published text or web pages, but how a reader can decide to trust a particular piece of information when they cannot use the article as a whole as a guide.

There is a need for automatic tools which could provide an aid for the reader who needs to assess trustworthiness and also for authors and moderators scanning changes. Similarly, moderators need tools for identification of vandalism, libel, advertising and so on.

Questions:

1. Is wiki reliability really a problem for readers, as I hypothesise? Perhaps readers who are not expert in a topic can detect problematic material in a wiki article, despite the multiple authorship.

2. Can we use language processing tools to help readers identify errors and misinformation in wiki pages?

2 Learning trustworthiness

The availability of change histories on wikis is a resource which could be exploited for training purposes by language processing systems designed to evaluate trustworthiness. If it is possible to categorise users as trustworthy or non-trustworthy/unknown by independent criteria (such as overall contribution level), then we can use changes made by trustworthy users that delete additions made by the unknown users as a means of categorising some text as bad. (Possibly the comments made by the editors could lead to subcategorization of the badness as error vs vandalism etc.) A tool for highlighting possible problem edits in wikis might thus be developed on the basis of a large amount of training data. Techniques derived from areas such as language-based spam detection, subjectivity measurement and so on could be relevant. However, one of the relatively novel aspects of the wiki problem is that we are looking at categorisation of small text snippets rather than larger quantities of text. Thus techniques that rely on stylistic cues probably won't work. Ideally, we need to be able to identify the actual information provided by individual contributors and classify this as reliable or unreliable. One way of looking at this is by dividing text into factoids (in the summarisation sense). Factoid identification is a really hard problem, but maybe the wiki edits themselves could help here.

Questions:

1. Can we automatically classify wiki contributors as reliable/unreliable?

2. Do trustworthy users' edits provide good training data?

3. Are there any features of text snippets that allow classification of reliability? (My guess: identification of vandalism will be possible but more subtle effects won't be detectable.)

4. What tools could be adapted from other areas of language processing to address these issues?

3 An ontology of errors?

As an extension of the ideas in the previous section, perhaps wiki histories could be mined as a repository of commonly believed false information. For instance, the EN wikipedia entry for University of Cambridge currently (Jan 5th, 2006) states:

> Undergraduate admission to Cambridge colleges used to depend on knowledge of Latin and Ancient Greek, subjects taught principally in the United Kingdom at fee-paying schools, called public schools.
> ('public schools' was linked)

One way in which this is wrong is that British 'public schools' (in this sense) are only a small proportion of the fee-paying schools, but equating public schools with all fee-paying schools is a common error. Suppose a trustworthy editor corrects this particular error in this article (and perhaps similar errors in the same or other articles). If we can automatically analyse and store the correction, we could use it to check for the same error in other text. As wikis get larger, this might become a useful resource for error detection/evaluation of many text types. Thus errors in wikis are an opportunity as well as a challenge.

Linguistic features of Italian blogs: literary language

Mirko Tavosanis
Dipartimento di Studi italianistici
Via del Collegio Ricci 10
I-56126 Pisa PI Italy
tavosanis@ital.unipi.it

Abstract

Preliminary surveys show that the language of blogs is not restricted to the more informal levels of expression. Instead blogs may include many kinds of written language: from simple personal notes to literary prose or poetry. The paper presents a sample of Italian blogs and comments on the results of the search of literary forms in two Web corpora using search engine queries.

1 Introduction

Close scrutiny of e-mails has revealed the presence of many different kinds of style in this medium (Baron 2000: 250-2; Pistolesi 2003: 178-184 for Italian). The same appears to be true for blogs. It is therefore difficult to determine specific linguistic features of blogs. Even occasional surveys however show that blogs are not limited to "a language reminiscent of brief notes, spoken asides, or short letters, rather than of essays or newsprint". Such language plays an important role in blogs, but accounts for only a small part of them. Many individual blogs aim instead at a true "literary" status and have a correspondingly high standard for word selection. Therefore, the linguistic equilibrium of this medium could be higher than expected. The paper will try to describe the general linguistic features of Italian blogs by contrasting them mainly with the language of newspapers, giving appropriate quantitative data.

2 Preliminary qualitative analysis: a textual sample

As a reference sample of blogs we can take ten blogs hosted by the Italian blog publishing site Splinder.com (arguably the most popular site of its kind). The sample was chosen by selecting the most recent blog appearing in the site list of the "Ultimi blog aggiornati" ('Most recently updated blogs'), and by selecting the first page of the postings published by the blog itself in November 2005 (if it had at least two postings in November). The selection was made at different times on one given day (29 December 2005) Some features of the selected blogs are described in Table 1.

A post taken from one of the less formal blogs in the sample (*di ritorno da...*) shows many of the linguistic features commonly ascribed to this kind of writing:

> giornata più tranquilla..sarà che sono a casa mia..a fare la mia vita..parlo appunto di mia perchè la vita parallela che sto facendo a milano non mi appartiene..quindi non posso dire che sia mia..che discorso complesso però ci stava dentro bene nel contesto..
>
> ho programmato il mio capodanno..dopo due anni consecutivi in una baita in montagna quest'anno lascio l'italia..pronta per 4 gg in scozia con tre amiche..
>
> non vedo l'ora..partirò il 30 di dicembre..aspetto quel giorno e intanto mi preparo per quattro esami all'università..
>
> ah...il viaggio al prezzo di 45 euro di volo e 90 di ostello..ultra risparmio...

The text describes the planning of a holday trip in Scotland for New Year's Eve and personal feeliings.The post shows also many unprofessional graphic choices (no capital letters for proper nouns or at the beginning of a sentence, no spacing after punctuation marks, *perchè* instead of *perché*) and the frequent use of three (or, wrongly, two) suspension points. The latter feature, also illustrated in Table 1, is considered one of the most common features of blogs and in this case it is surely used to give a feeling of “spoken language” to the text.

The language of the other blogs in the sample is, instead, very different. Suspension points are used in the blog *SoleLuna* in order to create high-pitched literary texts. This is a kind of lyrical description of a problematic relationship:

> Cancelli i tuoi passi nell'ombra di te stesso e scompari e compari quando e come più ti aggrada... Ed io mi lascio prendere dai pensieri e mi lascio intorpidire dai ricordi... Mi rivesto di te.. di noi.. Ho freddo.. Cerco di scaldarmi con il ricordo di un amore... Non ti amo. Amo il ricordo di quello che eravamo... E mi sfuggono via dalla mente le sensazioni e scivolano via gli odori.. si sbiadiscono i sapori... Tutto diventa la sfumatura del proprio colore.. la parodia, la beffa... Cerco di palpare le immagini e faccio attenzione a non sgualcirle.. più di quanto non lo sia io...

This kind of lyric language is heavily based upon the use of literary forms (“ti aggrada”), complex rhetorical constructions, “-*d* eufonica” and so on.

Midway between these two extremes we can find blogs like *Incontrista*. The posts of this blog are written in a language echoing newspaper editorials and brilliant prose. Significantly, they make no use of suspension points, as in this sectione (where the author contrasts the average psychological differences between female bloggers and female subscribers of dating sites):

> E' quindi un fatto ancora che le splinderine sono mediamente delle ragazze migliori delle meetiche proprio per questo motivo. Hanno capacità, caratteristiche e aspetti che a me piacciono, come la voglia di esprimersi, di raccontare, di scrivere, di comunicare, di leggere, di scegliere, di assumere posizioni critiche. Le splinderine fanno parte della fascia più esperta e innovativa degli utenti internet, quelli che ne fanno un utilizzo più consapevole, e che hanno una cultura più elevata della media. Le splinderine sono donne che hanno qualcosa da dire e vogliono compagni di alto livello con cui cercano il confronto serrato e accettano anche lo scontro.

It seems that none of those kinds of language could today claim to be the main model for blog writing. Free expression, literary prose and newspaper writing seem to co-exist without a clear dominant position. Of course, however, those impressions can be given substance only through extension of the field of search to a large set of blogs. This is partly made possible by the use of modern search engines.

3 Quantitative analysis of large corpora using search engines

Some linguistic features of blogs can now be probed and measured using search engines. For instance, preliminary searches show that from the orthographical point of view Italian blogs are much more correct than the average of the Italian web and that they are edited at least as well as online newspapers (Tavosanis in print b). Other indicators related to the use of “neo-standard” Italian forms (Berruto 1987) in the field of personal pronouns and demonstratives suggest a kinship between blogs and newspapers (Tavosanis in print a).

According to those searches, the main differences between blog posts and newspaper articles are not linked to writing accuracy or to different morphological choices. We can therefore assume as a working hypothesis that the main differences between blogs and newspapers in fact relate to lexicon and syntax.

The syntactic status of many blogs is probably well represented by the textual samples chosen above (widespread use of suspension points being the most conspicuous feature). However, close survey of this level can probably be obtained only through the encoding of a wide corpus with syntactic tagging.

The lexical features of blogs can instead be studied through simple search engine analysis (see again Tavosanis in print a and b for details of this method). Newspaper editing in Italy, enforced by a strong tradition and dedicated staff, excludes words considered too expressive (apart from those acknowledged by the same tradition: Bonomi 2002). Blogs, on the other hand, can include forms taken from every level of linguistic use. We can therefore expect that both literary and low forms are more used in blogs than in newspapers.

Two Web corpora were then selected: the web site of the newspaper *La Repubblica*, indexed and queried through the Google interface (= R), and the whole of the blogs indexed in the beta version of Blogsearch.google.com (= B). Of course, no exact data are available on the consistency of the two collections and the number of tokens indexed. The two corpora seem however roughly equal in size: the search of a common word like *questo* gives 427,000 occurrences in R and 467,000 in B; the search of *quello* 209,000 (R) and 257,608 occurrences (B); the search of *lui* 118,000 (R) and 159,970 occurrences (B); and so on. Of course, since word frequency is strongly correlated with the style and topic of the texts (for the Italian situation see Bortolini 1971: XIV-XV; Voghera 1993), this assessment cannot be considered an exact estimate. It does however give a preliminary quantitative estimate.

The highest frequency of vulgar words in the B corpus is of course undisputed, since newspaper editing is a strong barrier against this kind of language, and it needs no particular demonstration, e.g., we can find 30,310 occurrences of the word *cazzo* in B against 278 in R, and so on.

It is more difficult to demonstrate the highest frequency of literary language, which in the Italian tradition has a wide and varied lexicon. The abundance of synonyms and dispersion of forms lead one to focus searches on large groups of "weak" words instead of a limited set of "strong" words.

Next the list of "literary" verbs beginning with the letters *b*, *e* and *v* in the De Mauro (2000) dictionary was selected for analysis. The chosen verbs were 31 (*b*-), 47 (*e*-) and 49 (*v*-). Many of them also had non-literary uses and/or coincided with other Italian words: therefore only the words without homographs were used for the search, where every meaning recorded in the dictionary was marked at least as "obsolete" (code OB), "literary" (LE) or "bureaucratic" (BU). This left 23 (*b*-), 28 (*e*-) and 21 (*v*-) verbs. The two corpora were then searched for the infinitive forms of the verbs. Many of them did not appear at all: *baiare*, *balbuzzire*, *ballonzare* (1 occurrence in a text written in the dialect of Naples), *basciare*, *benedicere* (2 occurrences in two texts written in the dialect of Naples), *biancicare*, *biastemiare*, *blasmare*, *bombire*, *botare*, *botarsi*, *bravare*, *buccinare*, *bulicare*, *ebere*, *ecclissare*, *educere*, *enfiare*, *enfiarsi*, *escomunicare*, *escuotere*, *escusare*, *esinanire*, *espedire*, *esseguire*, *esterminare*, *estollere*, *estollersi*, *estorre*, *estruere*, *estruare*, *esturbare*, *esurire*, *evellere*, *evenire*, *vagheggiarsi*, *vanare*, *vengiare*, *vengiarsi*, *verberare*, *verdicare*, *verdire*, *vernare*, *verzicare*, *vilificare* and *vincire*.

The search also revealed that a verb marked in the dictionary as "literary" was instead widely used in both corpora: *vigilare*. While other forms occurred at most 94 times, in the corpora there are 644 occurrences of *vigilare*, evenly balanced (332 in B, 312 in R). It therefore appears more correct to consider this verb as a "common" word, without literary connotations, and to exclude it from further analysis.

In a second phase, many forms were excluded from counts since they resulted simple typos or broken forms of different words (e.g., many occurrences of *ventare* are in fact occurrences of widely used verbs like *diventare* or *inventare*, with incorrect spacing). Only words where the possibilities of misspellings seemed low were therefore included In the counts.

After this sifting, the forms represented in the corpus occurred as described in Table 2:

Form	**Occurrences in Blogs**	**Occurrences in *La Repubblica***
basire	1	0
bastarsi	12	1
beare	9	2
biasmare	0	1
biondeggiare	0	2
biscazzare	1	0
bruire	1	0
bruttare	0	1
bugiare	1	0
elicere	0	1
ergere	24	9
esondare	6	4
esperire	56	21
esplicare	79	15
estimare	2	0
evoluire	2	0
vacare	4	0
vagolare	7	1
vanire	1	0
vaticinare	17	6
ventare	2	0
vigoreggiare	2	0
villaneggiare	1	0
volvere	2	0
volversi	0	1
Total	**230**	**65**

Table 2: occurrences of literary forms

It seems therefore that some Italian blogs have in fact a higher proportion of a random selection of literary words than Italian newspapers. Further searches should be able to confirm or refute this finding.

4 Conclusion

Preliminary analyses of Italian blogs seem to confute the simple equivalence “blogs = informal text”. Clearly, both statistical tools and special monitoring software are needed to give this kind of search more focus and more depth. Future searches must also achieve a better understanding of the coverage of search engines and should be based upon different search engines. It would be useful, moreover, to identify and exploit other searchable indicators of the linguistic quality of a text. Anyway in future researches adequate space should be allowed for the assessment of the presence of literary features in many blogs.

Reference

[Baron 2000] Baron, N. *Alphabet to Email. How written English evolved and where it’s heading*. London-New York, Routledge.

[Berruto 1987] Berruto, G. *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia, Firenze

[Bonomi 2002] Bonomi, I. *L’italiano giornalistico. Dall’inizio del ’900 ai quotidiani on line*. Franco Cesati Editore, Firenze.

[Bortolini 1971] Bortolini, U., Tagliavini, C. and Zampolli, A. *Lessico di frequenza della lingua italiana contemporanea*. IBM Italia, Milano.

[De Mauro 2000] De Mauro, T. *Il dizionario della lingua italiana*. Paravia, Milano.

[Pistolesi 2004] Pistolesi, E. *Il parlar spedito.* Esedra, Padova.

[Tavosanis in print a] Tavosanis, M. *Traditional corpora and the Web as corpus: the Italian newspapers case study*. Presented at CLIN 2005, Amsterdam, 16 December 2005.

[Tavosanis in print b] Tavosanis, M. *Are blogs edited? A linguistic survey of Italian blogs using search engines*. To be presented at AAAI-CAAW 2006, Stanford, 27-29 March 2006.

[Voghera 1993] Voghera, M. “Le variabili testuali e pragmatiche”. In T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*. Etaslibri, Milano: 32-38.

Title	Description	Suspension dots	Other graphical highlightings	Images	Quoting	Smileys	Literary language, poetry	Alleged sex of main poster
cavi & gu	Personal exchanges by a couple of fiancées	x	x			x		M+F
CominciareDall'Inizio	Notes on her work by a high-school Physics teacher		x	x	x	x		F
di ritorno da_	A female student in Milan gives vent to her feelings about the big city	x		x	x			F
Incontrista	Newspaper-quality reports by a dedicated user of dating sites in Italy			x	x	x (character sequences)		M
NuVoLaBiAnCa In MeZzO a NuVoLe gRiGie	Personal rantings, poetry, reading notes	x	x	x	x	x	x	F
Pillola Rossa	Political rantings from a leftist position	x		x	x	x		M
53r3n4	Personal posts: feelings, song lyrics	x		x	x	x		F
se_lo_dici_è_vero	Collection of quotations	x	x	x	x	x		F
Servo del Signore	Prayers, thoughts about religion		x	x	x	x		M
SoleLuna	Poetry, recipes	x	x	x	x		x	F
VolObliquo	Poetry, both quoted and original		x	x	x		x	F

An analysis of Wikipedia digital writing

dott. Antonella Elia
Dipartimento di Scienze Statistiche - Sezione Linguistica
Facoltà di Scienze Politiche - Università degli Studi di Napoli Federico II
Napoli, Italy
aelia@unina.it

Abstract

This paper is a presentation of a doctoral research in progress focused on a new genre: online encyclopaedias. The introduction to Wikipedia and Encyclopaedia Britannica Online will be followed by a presentation of wiki as a new textual genre. Wikipedia analysis will focus firstly on the investigation of the "WikiLanguage", the language used in official encyclopaedic articles. Secondly, the "WikiSpeak", the spoken-written language used by Wikipedians in their backstage and informal community, will be taken into account. The initial findings of this research seem to suggest that, the language of the Wikipedia's co-authored articles is formal and standardized in a way similar to that found in Encyclopaedia Britannica Online. By contrast, the WikiSpeak, as a new variety of NetSpeak Jargon, can be considered as a creative domain, an independent and individual expression of linguistic freedom of self-representation, characterizing the wiki Computer Mediated Discourse Community.

1. Introduction

The encyclopaedia's structure, either hierarchical or alphabetically ordered, with its evolving nature is particularly adaptable to a disk-based or online format. All major printed encyclopaedias have moved to this method of delivery. Online E-ncyclopedias can include multimedia (such as video, sound clips and animated illustrations) unavailable in the printed format. They can make use of hypertext cross-references between conceptually related items and, furthermore, they offer the additional advantage of being dynamic: new and frequently updated information can be presented almost immediately, rather than waiting for the next release of a static format (as with a paper or disk publication).

This research is based particularly on a contrastive linguistic analysis of Wikipedia and Encyclopaedia Britannica Online. The latter is considered one of the greatest examples of general encyclopaedias in the English speaking world. It contains 120,000 articles which are commonly considered accurate, reliable and well-written. Brief article summaries can be viewed for free on the net, while the full text is available only for individuals with monthly or yearly subscription.

On the other hand, Wikipedia is a collaborative authoring project on the web, a repository of encyclopaedic knowledge, an example of a collaborative hypermedium focused on a common project. It is one of the most popular reference websites receiving around 50 million hits per day. It is a social e-democracy environment, designed with the goal of creating a free encyclopedia containing information on all subjects written collaboratively by volunteers. At the time of writing this paper the project has produced over two and half million articles and has been officially recognized as the largest international online community. It consists of 200 independent language editions and the English version is the biggest one with more than 962,995 articles (up to January 2006).

2. Wiki as new textual genre

With reference to the extensive empirical studies of Susan Herring on CMC, wikis and blogs considered as spaces belonging to the second web generation, can be regarded as adding new peculiarities to the existing synchronous and asynchronous tools of the first CMC generation (such as e-mail, mailing list, forum and chat). It is well known in media studies that "the medium is the message" as McLuhan (1964) pointed out in the sixties, and in fact the medium adds unique properties to the web genre in terms of production, function, and reception which cannot be ignored. Wikis are co-authoring tools which allow collective collaboration. They can be, simultaneously, a repository of information and an asynchronous tool of communication and discussion across the web (see Wikipedia). All wikis have integrated search engines for locating content and are open to anyone since they are considered a public space, even though they can be protected against unauthentic users.

Their main aim is to create documents. Wikis, unlike traditionally designed web sites, encourage "topical writing" by using wiki links and creating a wide network of interconnected pages. The interlinking process becomes simpler to type by just putting the word(s) in square brackets. It simultaneously creates a new topic title (a WikiWord), a new writing space for that topic and a link to that space. Once created, a topic will be available anywhere on the wiki as whenever the WikiWord is typed, it will link to the writing space of that topic (Morgan, 2006).

The writer, the supreme authority in print, is considered the one who transmits content through paper pages, to passive readers, whose role is merely to decode and interpret their message. The electronic writing space, being hypertextual and extremely flexible, changes the landscape. Writers can create multiple structures from the same topics (hierarchy, web, spiral, etc.) and readers can enter, browse and leave text at many points. In the hypertext, the author creates different paths for the reader, although there is neither a canonical path nor a defined page order to follow. The new active readers making their choices, become co-authors of the hypertext (Bolter, 1991). This idea is more pronounced on a wiki than elsewhere, because in an open wiki the reader can (if allowed) really interrupt the process, re-writing, changing, erasing and modifying the original text or creating new topics.

Traditional writing creates a gap between writer and reader. Wiki technology mediates the gap because the two actors assume interchangeable roles in this new open e-environment. To conclude, wiki text is never static as it is considered revisable, a-temporal as nodes continually change through the collaborative writing process, creating a never ending evolving network of topics. Thus, knowledge becomes webbed, contextualized though it remains temporary as it can always be changed or vandalized. Luckily, the original version can always, and easily, be recovered by SysOps[1], through page histories[2] (Morgan, 2006).

Wikis offer two different writing modes. The first one is known as "document mode". When it is used, contributors create documents collaboratively and can leave their additions to articles. Multiple authors can edit and update the content of documents which gradually become representations of contributors' shared knowledge (Leuf and Cunningham, 2001). Wikis have two states, "Read" and "Edit".

"Read state" is by default. In this case, wiki pages look just like normal webpages. When the user wants to edit a page, he/she must only access the "edit state".

"Document mode" is expository, extensive, monological, formal, refined and less creative than "thread mode". It is in third person and unsigned. "Document mode" demonstrates that knowledge is collective and that the ideas, not the writers, are the main focus. Writers contribute to "document mode" *refactoring*, reorganizing, incorporating and synthesizing "thread mode" comments in encyclopaedic articles and changing the first to third person (Morgan, 2006).

The second wiki writing mode is "thread mode". Contributors carry out discussions by posting signed messages in the discussion page connected to the main article. Others reply to the original message and so a group of threaded messages evolves (Morgan, 2006).

"Thread mode" is dialogical, open, collective, dynamic and informal. It develops organically, without a predictive structure. It expresses public thinking, presents multiple positions and is exploratory. Entries are phrased in first person and are signed. Rather then replying to a discussion entry, the writer can *refactor* the page to incorporate suggestions made, then delete the comment.

"Thread mode" demonstrates that knowledge is the result of constructivist collaboration and not a lonely production.

[1] SysOp is the abbreviation for "systems operator", and is a commonly used term for the administrator of a special-interest area of an online service.

[2] The page history of all versions of previous pages is available on Wikipedia. It consists of text, date , time and editing authors.

3. Research objectives and methodolology

3.1. Wikipedia vs Britannica

The first objective of this research has been directed towards the investigation of Wikipidia articles and on what has been defined, in this paper as "WikiLanguage", the formal, neutral and impersonal language used in the official encyclopedic articles. In this phase, an analysis of randomly selected sample articles has been carried out. The data for this research in progress has been based on two corpora. Up to now, they include a collection of txt files made up of one hundred articles representing topics taken from the Wiki Folksonomy's [3] eight categories (culture, geography, history, life, mathematics, science, society, technology) and on a contrastive analysis of the same articles found in Encyclopaedia Britannica Online.

The purpose of the quantitative research has been the empirical measurement of some linguistic features in order to define the degree of formality in the WikiLanguage. The sample articles have been analyzed through the ConcApp Concordancer Program. Different factors have been taken into consideration in order to define the formality of Britannica vs Wikipedia. The first aspect has been articles' length (total words) as conciseness was found to be a feature of formal written discourse (Chafe, 1982). The second, average word length (in letters) as short words have been considered a characteristic of informal genres (Biber, 1988). A high level of lexical density (Halliday, 1985) has been found in formal academic writing. It has been considered the main stylistic difference between speech and writing (Biber, 1988).

Subsequently, the number of unique lexical items in the two corpora has been measured. With reference to the findings of Heylighen and Dewaele (1999), frequency of word suffixes typical in formal genres (such as *-age, -ment, -ance/ence, -ion, -ity, -ism*) and impersonal pronouns (it/they) have been calculated. A contrastive frequency of meaningful keywords has also been investigated. The informality of the language has been measured through the frequency of abbreviations, acronyms, contractions (I'm, don't, he's, etc.) and personal pronouns (I, we, you, he/she, they) which have been found to be typical of informal genres, such as face-to-face and phone conversations (Biber, 1988). As shown in Appendix A (Fig.1), the first results of this research conducted on one hundred articles have highlighted a number of differences and similarities between Wikipedia and Britannica.

Articles in Britannica have proven to be shorter than those in Wikipedia (average length: 1728 vs 3510 words) and they have shown a higher lexical density (44.9% vs 31.4%). Although the level of total formality is clearly higher in Britannica (50.2% vs 36.6%), the frequency of formal nouns and impersonal pronouns typical of the formal discourse (5.3 vs 5.2) and the average word length (in letters 5.4 vs 5.2) has proven to be very similar. The divergent value is related to lexical density, but if text length varies widely (as happens in the two e-ncyclopedias) the different lexical items will appear to be much higher in the shortest text as their relationship is not linear. Each additional one hundred words of text adds fewer and fewer additional unique words (Biber, 1988). Thus, an interpretation of the collected data seems to suggest that thanks to the collective editorial control, the WikiLanguage of the co-authored articles shows a formal and standardized style similar to that found in Britannica. A table representing a part of the collected data, and their graphical representation, has been provided in Appendix A (Fig. 2,3,4).

3.2 Web analysis

Particular attention has been devoted to Wikipedia digital style due to the importance of the interplay between genre and medium when dealing with web-mediated texts. The layout of sample articles has been investigated (table of content, sections and sub-sections extension) as well as multimodality (tables, graphs, images, audio recordings and videos) and hypertextuality [explicative (internal bookmarks), associative (wikilinks) and explorative links (external weblinks)]. At present Wikipedia does not seem to fully exploit the potential offered by multimodality (and Britannica even less), showing few audio

[3] Folksonomy is a neologism which indicates a practice of collaborative categorization which makes use of freely chosen keywords. Taxonomy derives from Greek "taxis" and "nomos". "Taxis" means classification, "nomos" (or nomia) management and "folk" people; so folksonomy means people's classification management.

recordings and videos. This is probably due to the feature of Open Source software, keeping with hackers' simple and essential style (i.e. Slashdot and Everything2), to the contributors' average technical skills and to the philosophical choice which grants a privilege to information and content over appearance. One of the prominent properties of Wikipedia is its highly dense hypertextuality when compared to Britannica. The analysis of the articles clearly reveal the abundance of Wikipedia's nodes interlinking and dynamism, made possible by wiki software and, by contrast, the isolation, linearity (page structure) and static nature of corresponding Britannica articles. In this case using Finnemann's (1999) concept of "modal shifts" with reference "to reading mode" and "navigating mode", it is evident that Wikipedia articles actively stimulates the latter allowing the reader to construct his/her own personal pathway, browsing inside and outside the website.

4. WikiSpeak

The second phase of this research will focus on Wikipedia as "Computer Mediated Discourse Community" and on the language, defined in this paper as "WikiSpeak", the language spoken-written by Wikipedians in their informal backstage community. The medium has developed its own wired style and specific glossary, which resembles in some aspects the hackers' Jargon File. The main WikiSpeak distinctiveness lies in the lexicon used.

WikiSpeak is an unofficial and high-context language which can be considered as a new variety of the Netspeak, one of the most creative domains of contemporary English. Its peculiarity is immediately evident in the "wikilogisms" found in the Community Portal homepage (i.e. *stub, NPV, wikify, backlogs, FAQ, village pump, etc.)* which can be considered, for its lexical density, a supreme synthesis of WikiSpeak, as well as a political manifesto as the wiki philosophical essence and its informal community style are clearly disclosed here[4].

The present investigation has started from its analysis in order to measure the impact of the community front door (content, form, functionality) on the reader, and it will go on analysing the WikiSpeak used in discussion pages connected to the selected articles.

A large number of new words have emerged. WikiSpeak is an informal and colloquial language rich, for example, in acronyms [i.e. NPOV (Neutral Point Of View), COTW (Collaboration Of The Week), IFD (Image For Deletion), etc]. Plenty of abbreviations are also found. They are individual words reduced to two or three letters, [i.e. pls (please), bb ppls (bye bye peoples), etc]. Some abbreviations are like rebuses, as the sound value of the letter, or numeral, acts as a syllable of a word [i.e. B4N (bye for now), CYL (see you later), etc]. Wiki acronyms used in wiki CMC (discussion pages, mailing lists, IRC channels, instant messaging and personal user pages) are not restricted to words or short phrases, but can be sentence-length [i.e. WDYS (what did you say?), CIO (check it out), etc].

Many word processes take place in WikiSpeak, including several ludic innovations. A popular method of creating *wikilogisms* is to combine two separate words to make new *compound words*. Some elements turn up repeatedly, i.e. Wiki (WikiPage, WikiBooks, WikiLink, WikiStress, etc.)[5]. In addition, WikiSpeak makes large use of *blends* (namespace, infobox, quickpoll, etc.) and *semantic shifts* [i.e. orphan, mirror, stub, etc] shown in the wiki glossary available for the newbies.

Distinctive *graphology* is also an important feature of WikiSpeak. All orthographic features have been affected. For example, the status of capitalization varies greatly. There is a strong tendency to use lowercase everywhere on the net. The *lower-case default mentality* means that any use of capitalization is a marked form of communication. Messages wholly in capitals are considered to be shouting and usually avoided. A distinctive feature of Wiki graphology is the way two capitals are used: one initial, one medial.

This phenomenon is called *BiCapitalization* (BiCaps or CamelCase[6]) and is widespread in

[4] In the Community Portal homepage, of 1604 words used, 809 are unique words. The lexical density is very high 50.4%. The keywords are: *help* (19), *you* (16), *article* (16), collaboration (8), *free* (7).

[5] In Wikipedia veterans avoid their use as it is considered cliché. However it is tolerated when it refers to technical terms (i.e. *wikilinks*).

[6] CamelCase is the practice of writing compound words or phrases where the words are joined without spaces, and each word is capitalized within the compound. The

Wiki community (i.e. MediaWiki, WikiProject, etc.). It is a very interesting example of how a programming language influences the wired style, as BiCaps were used in hackers' communities as a word joiner alternative to the underscore based style and, in the original wiki convention to create links before the invention of [[_]] square brackets. Now it has become fashionable in marketing for names of products and companies. Outside these contexts, however, BiCaps are rarely used in formal written English, and most style guides recommend against it.

Spelling practice is also a WikiSpeak distinctive character. New spelling conventions have emerged, such as the replacement of plural –s by –z. Emotional expressions make use of a varying number of vowels and consonants (yayyyyyyy) and repeated punctuation (WHAT????), but punctuation sometimes tends to be minimalist or completely absent, a great deal depends on the user's personality: some Wikipedians are scrupulous about maintaining a traditional punctuation while some do not use it at all. On the other hand, there is an increased use of symbols not normally part of the traditional punctuation system, such as # , or repeated dots (…), hyphens (---), repeated use of commas („,) or asterisks (***). WikiSpeak, as a new variety of the NetSpeak Jargon, can be considered as a creative domain, an independent and individual expression of the linguistic freedom of self-representation in the wiki community of practice.

This research will make use of textual linguistics and corpus linguistics for the investigation of the interactions expressed in the unofficial and informal Wiki CMC.

5. Conclusions

In conclusion, this research project has two main focuses: defining the Wikipedia language variations within a dual context of use: official encyclopaedic entries (WikiLanguage) vs backstage community Speak (WikiSpeak).

Wikipedia, as a new expression for the encyclopedic genre, appears very similar to traditional printed encyclopedias due to its stylistic homogeneity, expressed Neutral Point of View[7] and formal style. The first findings of this research in progress seem to demonstrate how Wikipedia succeeds in reproducing an extant traditional genre even if applied to a collaborative and constructivist scenario.

According to Shephered and Watters (1998), extant subgenres are based on already existing genres in other media forms which have been converted into digital form (i.e. newspaper into electronic news); on the contrary, novel subgenres are entirely dependent on the new medium (i.e. homepages, search engines, webgames, etc.). They stated that when an extant genre migrates to a digital environment, it will initially be faithfully replicated: content and form will be preserved and the capabilities of the new medium will not be fully exploited (see Britannica). At a later stage in the evolution, variant genres are created. This process is driven by the technical capabilities of the new medium. It is the point of view of this study that Wikipedia can be taken as an example of the evolution of an extant traditional genre (encyclopedias) which has been officially preserved in the articles' superficial form, but not in the writing and reading processes (social editing, intertextuality, high informativeness and browsing mechanisms). The articles' textual form seems to suggest that when collaborative users have to respect stylistic established norms (see Wiki Manual of Style[8]) and shared social working ethics (see Wikiquette[9]), diversity and controversy are erased and the official requested style is respected within the open editing system. Nevertheless, technological advantages offered by collaborative software, reinforce the variety, the quick updating and interconnection of the information provided by the contributors' multitude. Their voices, even if individually, originally and democratically expressed in the CMC wiki community, are merged and homogenized in the articles' neutrality and formality.

name comes from the uppercase "bumps" in the middle of the compound word, suggesting the humps of a camel.

[7] A Neutral Point Of View (NPOV) is writing free from bias. It is generally considered desirable for journalistic and encyclopedic writings. According to the Wikipedia's founder, Jimbo Wales, NPOV is an "absolute and non-negotiable" principle in Wiki Manual of Style.

[8] Manual of Style is a style guide for Wikipedia's contributors. It has the purpose of making the editing easier by following a consistent format.

[9] Principles of Wikiquette are the guidelines on how to work with others on Wikipedia.

Linguistic analysis cannot be separated from the investigation of the main philosophical and political goals of Wikipedia whose main aim is to pursue freedom of content and information expressed through the Wikipedian "Collective" (Lèvy, 1994) and "Connective" Intelligence (de Kerckove, 1997) in this new acentric rhizomatic environment[10] (Deleuze-Guattari, 1980). Encyclopaedia Britannica is a knowledge compendium without any political meaning hosted by a commercial website (*.com*). In the 18th century, the original French "Encyclopédie" from Diderot and D'Alambert was mainly a political project designed to propagate the ideas of Enlightenment and to establish the reign of reason in Europe (Soufron, 2004). Similarly, Wikipedia in the current I.C.T. age, can be considered as a post-modern Encyclopaedia, a copyleft reference work with a non-profit cultural goal (.org) affording a political project rather than merely a scientific one. It is aimed at changing the society of the 21st century by giving control over content to everyone and thus enhancing freedom of expression and recovering the original aim of the World Wide Web inventor: Sir Tim Berners Lee wanted the web to be a boundless library of Babel and not a global supermarket as it has become in the *dot.com* era.

References

Biber Douglas. 1988. *Variation across speech and writing*. Cambridge University Press. Cambridge, UK.

Bolter Jay David. 1991, *Writing Space: Computers, Hypertext, and the Remediation of Print,* Lawrence Erlbaum Associates, N.Y., USA

Chafe Wallace L. 1982. "Integration and involvement in speaking, writing, and oral literature" in D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy* (pp. 35-53). Norwood, NJ: Ablex.

Crystal David. 2001. *Language and the Internet*, Cambridge University Press, Cambridge, UK.

Crystal David 2004. *The Cambridge Encyclopedia of English Language,* Cambridge University Press, Cambridge, UK.

De Kerkhove Derrick. 1997. *Connected Intelligence, the Arrival of the Web Society,* Somerville House Toronto, Canada.

Deleuze Gilles and Guattari Felix. 1980. tr. Eng., 1987, *A Thousand Plateaus: Capitalism and Schizophrenia,* University of Minnesota Press Minneapolis, USA.

Emigh William, Herring Susan. 2005. *Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias.* Proceedings of the Thirty-Eighth Hawai'i., International Conference on System Sciences (HICSS-38), IEEE Press, Los Alamitos, USA.

Encyclopaedia Britannica Online
http://www.britannica.com

Finnemann Niels Ole. 1999 *Hypertext and theRepresentational Capacities of the Binary Alphabet* http://www.hum.au.dk/ckulturf/pages/publications/nof/hypertext.htm

Herring Susan. 1996. *Computer Mediated Communication, linguistic, social and cross-cultural perspectives,* John Benjamins Publishing Company. Amsterdam, Philadelphia.

Heylighen Francis and Dewaele JM. 1999. *Formality of language: Definition, measurement and behavioral determinants.* Internal Report, Center "Leo Apostel", Free University of Brussels, Belgium .

Leuf Bo and Cunningham Ward. 2001. *The Wiki Way: Quick Collaboration on the Web,* Addison-Wesley, New York, USA.

Lèvy Pierre. 1994. *L'intelligence Collective. Pour une antropologie du cyberspace*, La Découverte, Paris, France.

McLuhan Marshall. 1964. "The Medium is the Message" in *Understanding Media: The Extensions of Man*, Signet, New York, USA.

Morgan M.C. *BlogsandWikis*
http://biro.bemidjistate.edu/~morgan/wiki/wiki.ph

Shepherd Michael, Watters Carolyn. 1998. *The evolution of cybergenres* in International Conference on System Sciences (HICSS-31). Hawai'i, vol. II, p. 97-109 cit. in "Literature Genre & Cybergenre".

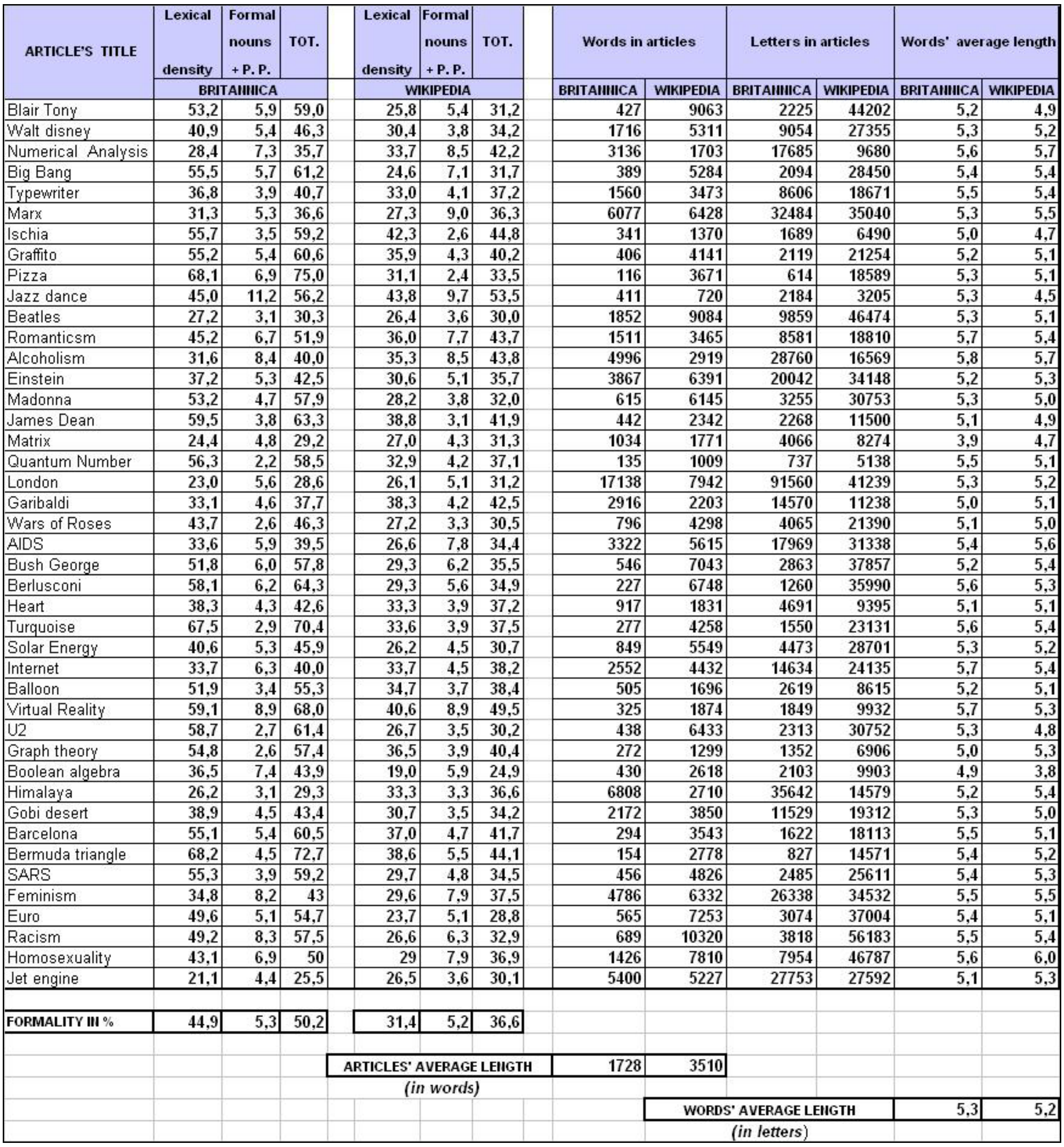
Soufron Jean Baptiste. 2004. *The political importance of the Wikipedia project: Wikipedia toward a new electronic enlightment era?* http://soufron.free.fr

Swales John M. 1990. *Genre Analysis. English in Academic and Research Setting*, Cambridge University Press, Cambridge, UK.

Wikipedia
http://www.wikipedia.org

[10] In *A Thousand Plateaus: Capitalism and Schizophrenia,* Deleuze and Guattari state that a rhizome is any structure in which each point is necessarily connected to each other point, where no location may become a beginning or an end, so the whole is heterogeneous. Deleuze labels the rhizome as a "multiplicity" resistant to structures of domination.

APPENDIX A



ARTICLE'S TITLE	Lexical density	Formal nouns + P. P.	TOT.	Lexical density	Formal nouns + P. P.	TOT.	Words in articles		Letters in articles		Words' average length	
	BRITANNICA			WIKIPEDIA			BRITANNICA	WIKIPEDIA	BRITANNICA	WIKIPEDIA	BRITANNICA	WIKIPEDIA
Blair Tony	53,2	5,9	59,0	25,8	5,4	31,2	427	9063	2225	44202	5,2	4,9
Walt disney	40,9	5,4	46,3	30,4	3,8	34,2	1716	5311	9054	27355	5,3	5,2
Numerical Analysis	28,4	7,3	35,7	33,7	8,5	42,2	3136	1703	17685	9680	5,6	5,7
Big Bang	55,5	5,7	61,2	24,6	7,1	31,7	389	5284	2094	28450	5,4	5,4
Typewriter	36,8	3,9	40,7	33,0	4,1	37,2	1560	3473	8606	18671	5,5	5,4
Marx	31,3	5,3	36,6	27,3	9,0	36,3	6077	6428	32484	35040	5,3	5,5
Ischia	55,7	3,5	59,2	42,3	2,6	44,8	341	1370	1689	6490	5,0	4,7
Graffito	55,2	5,4	60,6	35,9	4,3	40,2	406	4141	2119	21254	5,2	5,1
Pizza	68,1	6,9	75,0	31,1	2,4	33,5	116	3671	614	18589	5,3	5,1
Jazz dance	45,0	11,2	56,2	43,8	9,7	53,5	411	720	2184	3205	5,3	4,5
Beatles	27,2	3,1	30,3	26,4	3,6	30,0	1852	9084	9859	46474	5,3	5,1
Romanticsm	45,2	6,7	51,9	36,0	7,7	43,7	1511	3465	8581	18810	5,7	5,4
Alcoholism	31,6	8,4	40,0	35,3	8,5	43,8	4996	2919	28760	16569	5,8	5,7
Einstein	37,2	5,3	42,5	30,6	5,1	35,7	3867	6391	20042	34148	5,2	5,3
Madonna	53,2	4,7	57,9	28,2	3,8	32,0	615	6145	3255	30753	5,3	5,0
James Dean	59,5	3,8	63,3	38,8	3,1	41,9	442	2342	2268	11500	5,1	4,9
Matrix	24,4	4,8	29,2	27,0	4,3	31,3	1034	1771	4066	8274	3,9	4,7
Quantum Number	56,3	2,2	58,5	32,9	4,2	37,1	135	1009	737	5138	5,5	5,1
London	23,0	5,6	28,6	26,1	5,1	31,2	17138	7942	91560	41239	5,3	5,2
Garibaldi	33,1	4,6	37,7	38,3	4,2	42,5	2916	2203	14570	11238	5,0	5,1
Wars of Roses	43,7	2,6	46,3	27,2	3,3	30,5	796	4298	4065	21390	5,1	5,0
AIDS	33,6	5,9	39,5	26,6	7,8	34,4	3322	5615	17969	31338	5,4	5,6
Bush George	51,8	6,0	57,8	29,3	6,2	35,5	546	7043	2863	37857	5,2	5,4
Berlusconi	58,1	6,2	64,3	29,3	5,6	34,9	227	6748	1260	35990	5,6	5,3
Heart	38,3	4,3	42,6	33,3	3,9	37,2	917	1831	4691	9395	5,1	5,1
Turquoise	67,5	2,9	70,4	33,6	3,9	37,5	277	4258	1550	23131	5,6	5,4
Solar Energy	40,6	5,3	45,9	26,2	4,5	30,7	849	5549	4473	28701	5,3	5,2
Internet	33,7	6,3	40,0	33,7	4,5	38,2	2552	4432	14634	24135	5,7	5,4
Balloon	51,9	3,4	55,3	34,7	3,7	38,4	505	1696	2619	8615	5,2	5,1
Virtual Reality	59,1	8,9	68,0	40,6	8,9	49,5	325	1874	1849	9932	5,7	5,3
U2	58,7	2,7	61,4	26,7	3,5	30,2	438	6433	2313	30752	5,3	4,8
Graph theory	54,8	2,6	57,4	36,5	3,9	40,4	272	1299	1352	6906	5,0	5,3
Boolean algebra	36,5	7,4	43,9	19,0	5,9	24,9	430	2618	2103	9903	4,9	3,8
Himalaya	26,2	3,1	29,3	33,3	3,3	36,6	6808	2710	35642	14579	5,2	5,4
Gobi desert	38,9	4,5	43,4	30,7	3,5	34,2	2172	3850	11529	19312	5,3	5,0
Barcelona	55,1	5,4	60,5	37,0	4,7	41,7	294	3543	1622	18113	5,5	5,1
Bermuda triangle	68,2	4,5	72,7	38,6	5,5	44,1	154	2778	827	14571	5,4	5,2
SARS	55,3	3,9	59,2	29,7	4,8	34,5	456	4826	2485	25611	5,4	5,3
Feminism	34,8	8,2	43	29,6	7,9	37,5	4786	6332	26338	34532	5,5	5,5
Euro	49,6	5,1	54,7	23,7	5,1	28,8	565	7253	3074	37004	5,4	5,1
Racism	49,2	8,3	57,5	26,6	6,3	32,9	689	10320	3818	56183	5,5	5,4
Homosexuality	43,1	6,9	50	29	7,9	36,9	1426	7810	7954	46787	5,6	6,0
Jet engine	21,1	4,4	25,5	26,5	3,6	30,1	5400	5227	27753	27592	5,1	5,3
FORMALITY IN %	44,9	5,3	50,2	31,4	5,2	36,6						
				ARTICLES' AVERAGE LENGTH			1728	3510				
					(in words)							
								WORDS' AVERAGE LENGTH			5,3	5,2
									(in letters)			

Figure 1. Linguistic formality: Britannica vs Wikipedia

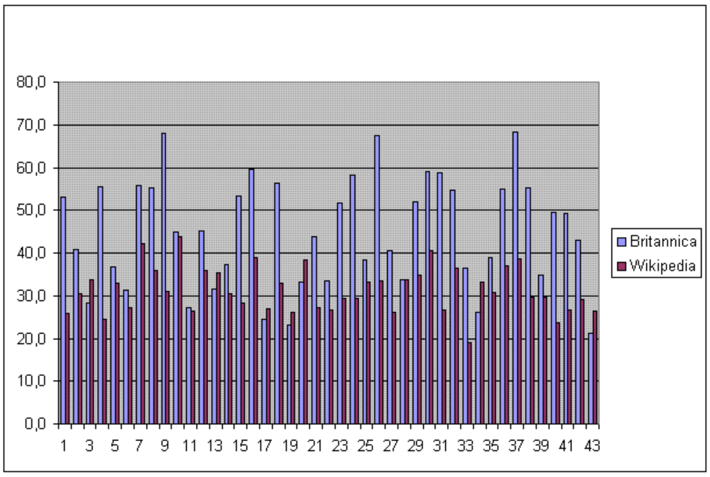


Figure 2. Lexical density

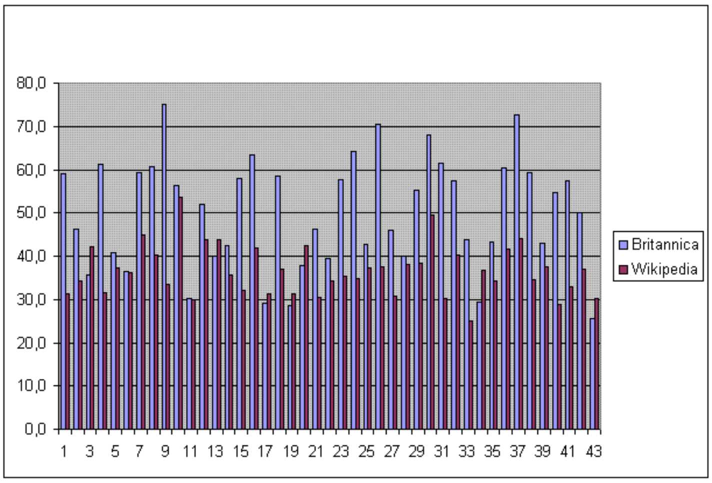


Figure 3. Total formality in percentage

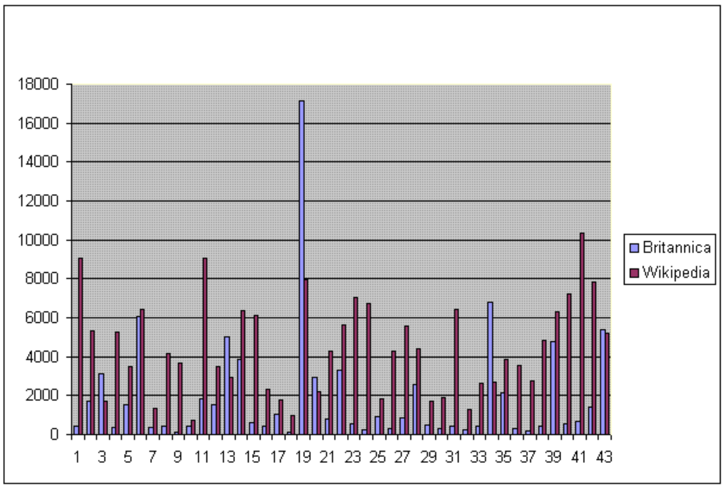


Figure 4. Articles' length in words

Learning to Recognize Blogs: A Preliminary Exploration

Erik Elgersma and **Maarten de Rijke**
ISLA, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
erik@elgersma.net, mdr@science.uva.nl

Abstract

We present results of our experiments with the application of machine learning on binary blog classification, i.e. determining whether a given web page is a blog page. We have gathered a corpus in excess of half a million blog or blog-like pages and pre-classified them using a simple baseline. We investigate which algorithms attain the best results for our classification problem and experiment with resampling techniques, with the aim of utilising our large dataset to improve upon our baseline. We show that the application of off-the-shelf machine learning technology to perform binary blog classification offers substantial improvement over our baseline. Further gains can sometimes be achieved using resampling techniques, but these improvements are relatively small compared to the initial gain.

1 Introduction

In recent years, weblogs (online journals in which the owner posts entries on a regular basis) have not only rapidly become popular as a new and easily accessible publishing tool for the masses, but its content is becoming ever more valuable as a “window to the world,” an extensive medium brimming with subjective content that can be mined and analysed to discover what people are talking about and why. In recent years the volume of blogs is estimated to have doubled approximately every six months. Technorati[1] report that about 11% of internet users are blog readers and that about 70 thousand new blogs are created daily. Popular blogosphere (the complete collection of all blogs) analysis tools estimate the blogosphere to contain anywhere between 20[1] and 24 million[2] blogs at time of writing. Given this growing popularity and size, research on blogs and the blogosphere is also increasing. A large amount of this research is being done on the content provided by the blogosphere and the nature of this content, like for example (Mishne and de Rijke, 2005), or the structure of the blogosphere (Adar et al., 2004).

In this paper, however, we address the task of binary blog classification: given a (web) document, is this a blog or not? Our aim is to base this classification mostly on blog characteristics rather than content. We will by no means ignore content but it should not become a crucial part of the classification process.

Reliable blog classification is an important task in the blogosphere as it allows researchers, ping feeds (used to broadcast blog updates), trend analysis tools and many others to separate real blog content from blog-like content such as bulletin boards, newsgroups or trade markets. It is a task that so far has proved difficult as can be witnessed by checking any of the major blog update feeds such as weblogs.com[3] or blo.gs.[4] Both will at any given time list content that clearly is not a blog. In this paper we will explore blog classification using machine learning to improve blog detection and experiment with several methods to try and further improve the percentage of instances classified correctly.

The main research question we address in this paper is exploratory in nature:

- How hard is binary blog classification?

Put more specifically,

[1] Intelliseek’s BlogPulse, http://www.blogpulse.com
[2] Technorati, http://www.technorati.com
[3] Weblogs.com, http://www.weblogs.com
[4] Blo.gs, http://blo.gs

- What is the performance of basic off-the-shelf machine learning algorithms on this task?

and

- Can the performance of these methods be improved using resampling methods such as bootstrapping and co-training?

An important complicating factor is the lack of labeled data. It is widely accepted that given a sufficient amount of training data, most machine learning algorithms will achieve similar performance levels. For our experiments, we will have a very limited amount of training material available. Therefore, we expect to see substantial differences between algorithms.

In this paper we will first discuss related work in the following section, before describing the experiments in detail and reporting on the results. Finally, we will draw conclusions based on the experiments and the results.

2 Related Work

Blog classification is still very much in its infancy and to date no directly related work has been published as far as we are aware. There is, however, work related to several aspects of our experiments.

Nanno et al. (2004) describe a system for gathering a large collection of weblogs, not only those published using one of the many well-known authoring tools but also the hand-written variety. A very much comparable system was developed and used for these experiments. Members of the BlogPulse team also describe blog crawling and corpus creation in some detail (Glance et al., 2004), but their system is aimed more at gathering updates and following active blogs rather than gathering as many blogs in their entirety, as our system is set up to do.

As to the resampling methods used in this paper—bootstrapping and co-training—, Jones et al. (1999) describe the application of bootstrapping to text learning tasks and report very good results applying this method to these tasks. Even though text learning is a very different genre, their results provide hope that the application of this method may also prove useful for our blog classification problem.

Blum and Mitchell (1998) describe the use of separate weak indicators to label unlabeled instances as "probably positive" to further train a learning algorithm and gathered results that suggested that their method has the potential for improving results on many practical learning problems. Indeed their example of web-page classification is in many ways very similar to our binary blog classification problem. In these experiments however we will use a different kind of indicators on the unlabeled data, namely the predictions of several different types of algorithms.

3 Binary blog classification

In our first experiment, we attempted binary blog classification ("is this a blog or not?") using a small manually annotated dataset and a large variety of algorithms. The aim of this experiment was to discover what the performance of readily available, off-the-shelf algorithms is given this task.

We used a broad spectrum of learners implemented in the well-known Weka machine learning toolkit (Witten and Frank, 2005).

3.1 Dataset

For our later resampling experiments, a large amount of data was gathered, as will be explained further on in this paper. To create a dataset for this experiment, 201 blog / blog-like pages were randomly selected from the collection, processed into Weka's arff format and manually annotated. These instances were then excluded from the rest of the collection. This yielded a small but reliable dataset, which we hoped would be sufficient for this task.

3.2 Attribute selection

All pages were processed into instances described by a variety of attributes. For binary blog classification to succeed, we had to find a large number of characteristics with which to accurately describe the data. This was done by manually browsing the HTML source code of several blogs as well as some simple intuition. These attributes range from "number of posts" and "post length" to checking for characteristic phrases such as "Comments" or "Archives" or checking for the use of style sheets. Interesting attributes are the "firstLine" / "lastLine" attributes, which calculate a score depending on the number of tokens found in those lines, which frequently occur in those lines in verified blog posts. The "contentType" attribute does something very similar, but based on the complete clean text of a page rather than particular lines in posts. It counts how many of the 100 most frequent tokens in clean text versions of actual blogs, are found in a page and returns a true value if more than 60% of these are found, in which case the page is probably a blog. The “frequent terms”-lists for these attributes were generated using a manually verified list gathered from a general purpose dataset used for earlier experiments. A “host”-attribute is also used, which we binarised into a large number of binary host name attributes as most machine learning algorithms cannot cope with string attributes. For this purpose we took the 30 most common hosts in our dataset, which included Livejournal,[5] Xanga,[6] 20six,[7] etc., but also a number of hosts that are obviously not blog sites (but host many pages that resemble blogs). Negative indicators on common hosts that don’t serve blogs are just as valuable to the machine learner as the positive indicators of common blog hosts. Last but not least a binary attribute was added that acts as a class label for the instance. This process left us with the following 46 attributes:

Attribute	Type
nrOfPosts	numeric
avgPostLength	numeric
minPostLength	numeric
maxPostLength	numeric
firstLine	numeric
lastLine	numeric
containsBlog	numeric
containsMetaTag	binary
contentType	binary
containsComment	binary
containsPostedBy	binary
containsRSS	binary
containsArchives	binary
containsPreviousPosts	binary
StyleSheetsUsed	binary
livejournal.com	binary
msn.com	binary
wretch.cc	binary
xanga.com	binary
diaryland.com	binary
abazy.com	binary
20six.fr	binary
research101-411.com	binary
search-now700.com	binary
search-now999.com	binary
search-now600.com	binary
20six.co.uk	binary
research-bot.com	binary
blogsearchonline.com	binary
googane.com	binary
typepad.com	binary
findbestnow.com	binary
myblog.de	binary
quick-blog.com	binary
findhererightnow.com	binary
findfreenow.com	binary
websearch010.com	binary
twoday.net	binary
websearch013.com	binary
tracetotal.info	binary
kotobabooks.com	binary
cocolog-nifty.com	binary
20six.de	binary
is-here-online.com	binary
4moreadvice.info	binary
blog	binary

Table 1: Attributes selected for our experiments.

3.3 Experimental setup

For this experiment, we trained a wide range of learners using the manually annotated data and tested using ten-fold cross-validation. We then compared the results to a baseline.

This baseline is based mostly on simple heuristics, and is an extended version of the WWW-Blog-Identify[8] perl module that is freely available online. First of all, a URL check is done which looks for a large number of the well-known blog hosts as an indicator. Should this fail, a search is done for metatags which indicate the use of well-known blog creation tools such as Nucleus,[9] Greymatter,[10] Movable Type[11] etc. Should this also fail, an actual content search is done for other indicators such as particular icons blog creation tools leave on pages (“created using… .gif” etc). Next, the module checks for an RSS feed, and as a very last resort checks the number of times the term “blog” is used on the page as an indicator.

In earlier research, our version of the module was manually tested by a small group of individuals and found to have an accuracy of roughly 80% which means it is very useful as a target to aim for with our machine learning algorithms and a good baseline.

[5] http://www.livejournal.com

[6] http://www.xanga.com

[7] http://www.20six.com

[8] http://search.cpan.org/~mceglows/WWW-Blog-Identify-0.06/Identify.pm

[9] http://nucleuscms.org

[10] http://www.noahgrey.com/greysoft/

[11] http://www.movabletype.org/

3.4 Results: single classifiers

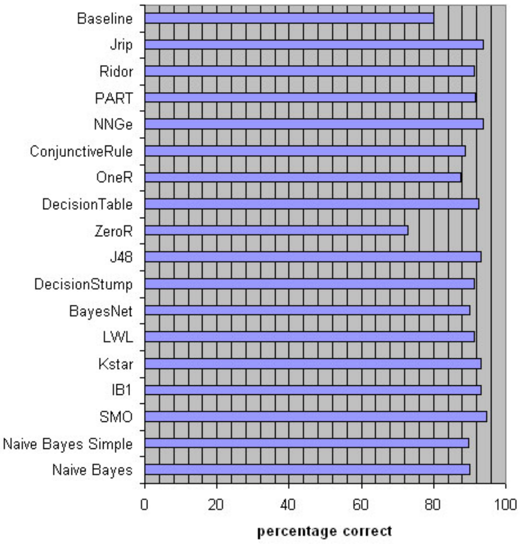


Figure 1: Chart showing the percentage correct predictions for each algorithm tested.

It is clear that all algorithms bar ZeroR perform well, most topping 90%. ZeroR achieves no more than 73%, and is the only algorithm that actually performs worse than our baseline. The best algorithm for this task, and on this dataset, is clearly the support vector-based algorithm SMO, which scores 94.75%. These scores can be considered excellent for a classification task, and the wide success across the range of algorithms shows that our attribute selection has been a success. The attributes clearly describe the data well.

Full results of this experiment can be found in Appendix A.

4 Resampling

Now we turn to the second of our research questions: to what extent can resampling methods help create better blog classifiers.

As reported earlier, the blogosphere today contains millions of blogs and therefore potentially plenty of data for our classifier. However, this data is all unlabeled. Furthermore, we have a distinct lack of reliably labeled data. Resampling may provide us with a solution to this problem and allow us to reliably label the data from our unlabeled data source and further improve upon the results gained using our very small manually annotated dataset.

For these experiments we selected two resampling methods. The first is ordinary bootstrapping, which we chose because it is the simplest way of relabeling unlabeled data on the basis of a machine learning model. Additionally, we chose a modified form of co-training, as co-training is also a well-known resampling method, which was easily adaptable to our problem and seemingly offered a good approach.

4.1 Data set

To gather a large data set containing both blogs and non-blogs, a crawler was developed that included a blog detection module based on the heuristics in our baseline module mentioned earlier. After downloading a page judged likely to be a blog by the module on the basis of its URL, several additional checks were done by the blog detection module based on several other characteristics, most importantly the presence of date-entry combinations. Pages judged to be a blog and those judged not to be even though the URL looked promising, were consequently stored separately. Blogs were stored in html, clean text and single entry (text) formats. For non-blogs only the html was stored to conserve space while still allowing the documents to be fully analysed post-crawling.

Using this system, 227.380 blog- and 285.337 non-blog pages (often several pages were gathered from the same blog, so the actual number of blogs gathered is significantly lower) were gathered in the period from July 7 until November 3, 2005. This amounts to roughly 30Gb of HTML and text, and includes blogs from all the well-known blog sites as well as personal handwritten blogs and in many different languages.

The blog detection module in the crawler was used purely for the purpose of filtering out URLs and webpages that bear no resemblence to a blog. By performing this pre-classification, we were able to gather a dataset containing only blogs and pages that in appearance closely resemble blogs so that our dataset contained both positive examples and useful negative examples. This approach should force the machine learner to make a clear distinction between blogs and non-blogs. However, even though this data was pre-classified by our baseline, we treat it as unlabeled data in our experiments and make no further use of this pre-classification whatsoever.

For our resampling experiments, we randomly divided the large dataset into small subsets containing 1000 instances, one for each iteration. This figure ensures that the training set grows at a reasonable rate at every iteration while preventing the training set from becoming too large too quickly which would mean a lot of unlabeled instances being labeled on the basis of very few labeled instances and the model building process would take too long after only a few iterations.

For training and test data we turned back to our manually annotated dataset used previously. Of this set, 100 instances were used for the initial training and the remaining 101 for testing.

4.2 Experimental setup: bootstrapping

Generally, bootstrapping is an iterative process where at every iteration unlabeled data is labeled using predictions made by the learner model based on the previously available training set (Jones et al., 1999). These newly labeled instances are then added to the training set and the whole process repeats. Our expectation was that the increase in available training instances should improve the algorithm’s accuracy, especially as it proved quite accurate to begin with so the algorihm’s predictions should prove quite reliable. For this experiment we used the best performing algorithm from Section 3, the SMO support-vector based algorithm. The bootstrapping method is applied to this problem as follows:

- Initialisation: use the training set containing 100 manually annotated instances to predict the labels of the first subset of 1000 unlabeled instances.
- Iterations: Label the unlabeled instances according to the algorithm’s prediction and add these instances to the previous training set to form a new training set. Build a new model based on the new training set and use it to predict the labels of the next subset.

4.3 Results: bootstrapping

We now present the results of our experiment using normal bootstrapping. After every iteration, the model built by the learner was tested on our manually annotated test set.

Iteration	Nr. of training instances	Correctly / incorrectly classified (%)	Precision (yes/no)	Recall (yes/no)
init	100	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
1	1100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
2	2100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
3	3100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
4	4100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
5	5100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
6	6100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
7	7100	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
8	8100	93.07 / 6.93	0.952 / 0.925	0.769 / 0.987
9	9100	93.07 / 6.93	0.952 / 0.925	0.769 / 0.987
10	10100	93.07 / 6.93	0.952 / 0.925	0.769 / 0.987
11 - 42	11100 – 42100	92.08 / 7.92	0.95 / 0.914	0.731 / 0.987

Table 2: Overview of results using normal bootstrapping.

After 36 iterations, the experiment was halted as there was clearly no more gain to be expected from any further iterations. Clearly, ordinary bootstrapping does not offer any advantages for our binary blog classification problem. Also, the availability of larger amounts of training instances does nothing to improve results as the results are best using only the very small training set.

Generally, both precision and recall slowly decrease as the training set grows, showing that classifier accuracy as a whole declines. However, recall of instances with class label “no” (non-blogs) remains constant throughout. Clearly the classifier is able to easily detect non-blog pages on the basis of the attributes provided, and is thwarted only by a small number of outliers. This can be explained by the fact that the learner recognizes non-blogs mostly on the basis of the first few attributes having zero values (nrOfPosts, minPostLength, maxPostLength etc.). The outliers consistently missed by the classifier are probably blog-like pages in which date-entry combinations have been found but which nevertheless have been manually classified as non-blogs. Examples of this are calendar pages commonly associated with blogs (but which do not contain blog content), or MSN Space pages on which the user is using the photo album but hasn’t started a blog yet. In this case the page is recognized as a blog, but contains no blog content and is therefore manually labeled a non-blog.

4.4 Experimental setup: co-training

As mentioned in Section 2, we will use the predictions of several of the most successful learning algorithms from Section 3 as our indicators in this experiment. The goal of our co-training experiment is to take unanimous predictions from the three best performing algorithms from Section 3, and use those predictions, which we assume to have a very high degree of confidence, to bootstrap the training set. We will then test to see if it offers an improvement over the SMO algorithm by itself. By unanimous predictions we mean the predictions of those instances, on which all the algorithms agree unanimously after they have been allowed to predict labels using their respective models.

As instances for which the predictions are unanimous can be reasoned to have a very high level of confidence, the predictions for those instances are almost certainly correct. Therefore we expect this method to offer substantial improvements over any single algorithm as it potentially yields a very large number of correctly labeled instances for the learner to train on.

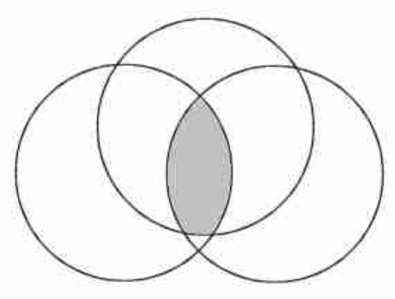


Figure 2: Visual representation of our implementation of the co-training method.

We chose to adapt the co-training idea in this fashion as we believe it to be a good way of radically reducing the fuzziness of potential predictions and a way to gain a very high degree of confidence in the labels attached to previously unlabeled data. Should the algorithms disagree on a large number of instances there would still not be a problem as we have a very large pool of unlabeled instances (133.000, we only used part of our corpus for our experiments as our dataset was so large that there was no need to use all the data available). The potential maximum of 133 iterations should prove quite sufficient even if the growth of the training set per iteration proves to be very small.

The algorithms we chose for this experiment were SMO (support vector), J48 (decision tree, a C4.5 implementation) and Jrip (rule based). We chose not to use nearest neighbour algorithms for this experiment even though they performed well individually as we feared it would prove a less successful approach given the large training set sizes. Indeed, an earlier experiment done during our blog classification research showed the performance of near neighbour algorithms bottomed out very quickly so no real improvement can be expected from those algorithms given larger training sets and given the unanimous nature of this method of co-training it may spoil any gain that might otherwise be achieved.

The process started with the manually annotated training set and used the predictions from the three algorithms, for unlabeled instances they agree unanimously on, to label those instances. Those instances were subsequently added to the trainingset and using this new trainingset, a number of the instances in another unlabeled set (1000 instances per set) were to be labeled (again, only those instances on which the algorithms agree unanimously). Once again, those instances are added to the training set and so on and so forth for as many iterations as possible.

4.5 Results: co-training

We now turn to the results of our experiment using our unanimous co-training method described above. The experiment was halted after 30 iterations, as Weka ran out of memory. The experiment was not re-run with altered memory settings as it was clear that no more gain was to be expected by doing so. Again, testing after each iteration was performed by building a model using the SMO support-vector learning algorithm and testing classifier accuracy on the manually annotated test set.

Iteration	Nr. of training instances	Correctly/ incorectly classified (%)	Precision (yes/no)	Recall (yes/no)
init	100	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
1	1000	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987
2	1903	93.07 / 6.93	0.952 / 0.925	0.769 / 0.987
3	2798	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
4	3696	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
5	4566	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
6	5458	96.04 / 3.96	0.958 / 0.961	0.885 / 0.987
7	6351	96.04 / 3.96	0.958 / 0.961	0.885 / 0.987
8	7235	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
9	8149	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987

10	9041	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
11	9929	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
12	10810	95.05 / 4.95	0.957 / 0.949	0.846 / 0.987
13 - 43	11684 - 38510	94.06 / 5.94	0.955 / 0.937	0.808 / 0.987

Table 3: Overview of results using our unanimous co-training method.

Even though the "steps" in test percentages shown represent only one more blog being classified correctly (or incorrectly), the classifier does perform better than it did using only the manually annotated training set at some stages of the experiment. This means that gains in classifier accuracy can be achieved by using this method of co-training on this problem. Also the classifier generally performs better than in our bootstrapping experiment, which shows that the instances unanimously agreed on by all three algorithms are certainly more reliable than the predictions of even the best algorithm by itself, as predicted.

Clearly this method offers potential for an improvement even though the SMO algorithm was already very accurate in our first binary blog classification experiment.

5 Discussion

As the title suggests, these experiments are of a preliminary and exploratory nature. The high accuracy achieved by almost all algorithms in our binary classification experiment show that our attribute set clearly defines the subject well. However, these results must be viewed with an air of caution as they were obtained using a small subset and as such the data may not represent the nature of the complete dataset well. Indeed, how stable are the results obtained?

Later experiments using a (disjoin, but) larger manually annotated dataset containing 700 instances show that the results obtained here are optimistic. The extremely diverse nature of the blogosphere means that describing an entire dataset using a relatively small subset is very difficult and as such both the performance and ranking of off-the-shelf machine learning algorithms will vary among different datasets. Off-the-shelf algorithms do however still perform far better than our baseline and the best performing algorithms still achieve accuracy rates in excess of 90%.

Two aspects of our attribute set that need to be worked on in future are date detection and content checks. Outliers are almost always caused by the date detection algorithm not detecting certain date formats, and pages containing date-entry combinations but no real blog content. Therefore, although it is possible to perform binary blog classification based purely on the particular characteristics of blog pages with high accuracy, content checks are invaluable. The rise of blogspam, which cannot be separated from real blogs on the basis of page characteristics at all, further emphasises this. We have already developed a document frequency profile and replaced the contentType attribute used in these experiments, to extend the content-based attributes in our dataset and hopefully improve blog recognition.

6 Conclusion

Our experiments have shown that binary blog classification can be performed successfully if the right attributes are chosen to describe the data, even if the classifier is forced to rely on a small number of training instances. Almost all basic off-the-shelf machine learning algorithms perform well given this task, but support vector based algorithms performed best in this experiment. Notable was that the best algorithms of each type achieved almost the same accuracy, all over 90% and the difference is never larger than a few percent even though they approach the problem in completely different manners.

The performance of these algorithms can be improved by using resampling methods, but not all resampling methods achieve gains and those that do gain very little. The extremely high success rates of the plain algorithms means that there is very little room for improvement, especially as the classification errors are almost always caused by outliers that none of the algorithms manage to classify correctly.

The results of later experiments with larger numbers of manually annotated instances show that a lot of work remains to be done and that although this paper shows that the application of machine learning to this problem offers substantial improvements over our baseline, this problem is still far from solved.

Future work will include further analysis of the results obtained using larger manually annotated subsets as well as a detailed analysis of the contributions of the different features in the feature set described in Section 3.

Acknowledgements

The authors wish to thank Gilad Mishne for his input and valuable comments during these experiments and the writing of this piece.

Maarten de Rijke was supported by grants from the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 640.001.501, and 640.002.501.

References

G. Mishne, M. de Rijke. 2006. Capturing Global Mood Levels using Blog Posts, In: *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW 2006)

E. Adar, L. Zhang, L. Adamic, R. Lukose. 2004. Implicit Structure and the Dynamics of Blogspace, In: *Workshop on the Weblogging Ecosystem, WWW Conference*, 2004

T. Nanno, Y. Suzuki, T. Fujiki, M. Okumura. 2004. Automatic Collection and Monitoring of Japanese Weblogs, In: *Proceeding of WWW2004: the 13th international World Wide Web conference*, New York, NY, USA, 2004. ACM Press.

N. Glance, M. Hurst, T. Tomokiyo. 2004. BlogPulse: Automated Trend Discovery for Weblogs, In: *Proceeding of WWW2004: the 13th international World Wide Web conference*, New York, NY, USA, 2004. ACM Press.

R. Jones, A. McCallum, K. Nigam, and E. Riloff. 1999. Bootstrapping for Text Learning Tasks, In: *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, p52-63.

A. Blum, T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training, In: *Proceedings of the 1998 Conference on Computational Learning Theory*.

I. Witten, E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Appendix A. Full results of our binary blog classification experiment

Algorithm	Type	Percentage correct predictions
Naïve Bayes	Bayes	90.07
Naïve Bayes Simple	Bayes	89.64
SMO	**Support Vector**	**94.75**
IB1	Instance based	93.00
KStar	**Instance based**	**93.30**
LWL	Instance based	91.25
BayesNet	Bayes	90.08
DecisionStump	Tree	91.25
J48	**Tree**	**93.29**
ZeroR	Rule-based	73.00
DecisionTable	Rule-based	92.55
OneR	Rule-based	87.60
ConjunctiveRule	Rule-based	88.75
NNGe	Rule-based	93.73
PART	Rule-based	91.67
Ridor	Rule-based	91.26
JRip	**Rule-based**	**93.73**

Interpreting Genre Evolution on the Web: Preliminary Results

Marina Santini
University of Brighton
Lewes Rd, Moulsecoomb Campus,
Brighton, BN2 4GJ (UK)
Marina.Santini@itri.brighton.ac.uk

Abstract

The study presented in this paper explores the current state of genre evolution on the web through web users' perception. More precisely, it explores the perception of genres when users are faced not only with prototypical genre exemplars but also with hybrid or individualized web pages, and interpret the subjects' perception in term of genre evolution Although this exploration is partial (23 labels to be assigned to 25 web pages), it offers an interesting section of the genre repertoire on the web. This study can be also seen as a confirmatory study, because it confirms that a number of recent web genres, unprecedented in the paper world (such as home page, FAQs, and blog) can be recognized by the subjects; others have not fully emerged and many web users are not familiar with their new genre labels; finally some web pages show a high level of ambiguity and web users largely disagree on assigning labels to them.

1 Introduction

The study presented in this paper has been designed to explore the current state of genre evolution on the web through web users' perception. The web can be interpreted, among other things, as a genre repertoire in evolution because there are still many genre labels which have not been consolidated and many web pages that cannot be sorted into a recognized and acknowledged genre. The interpretation of the web as a genre repertoire in evolution has been developed within a research project on automatic identification of genre in web pages (Santini, 2006b). This interpretation is an attempt to explain the high level of hybridism and individualization of many web pages, which result in classification intractability. The study reported in this paper shows that humans have the same problems as classification algorithms when it comes to less standardized and conventionalized web pages.

As it has often been pointed out (for example, cf. Kwasnik and Crowston, 2005), it is hard to pin down the concept of genre from a single perspective or to find an agreed definition of what genre is. This lack is also experienced in the more restricted world of non-literary or non-fictional genres, such as professional or instrumental genres, where the variation due to personal style is less pronounced than in literary genres. In particular, scholars working with practical genres focus upon a specific environment. For instance Swales (1990) develops his notion of genre in academic and research settings, Bathia (1993) and Trosborg (2000) in professional settings, Yates and Orlikowsky (1992) within organizational communication. Despite the lack of an agreed theoretical notion, genre is a well-established term (cf. Karlgren, 2004), intuitively understood in its vagueness. Classifying documents by genre is a common operation that humans perform with more or less effort.

Genres can be seen as "artifacts", i.e. cultural objects created to meet and streamline communicative needs. These cultural objects represent the role that a certain type of documents plays in an environment. Each genre shows a set of standardized or conventional characteristics that makes it recognizable among others, and this kind of identity raises specific expectations in the recipients, despite the fuzziness of genre labels (cf. Santini, 2005). Being cultural objects, showing common conventions and raising similar expectations are unifying traits. Together with these, there is a number of separating traits, such as hybridism, individualization, and evolution. In fact, genres are not mutually exclusive and different genres can be merged in a single document, generating hybrid forms. Genres are based on conventions, but allow a certain freedom of variation and consequently can be individualized.

Being *artifacts*, sharing *conventions* and *expectations*, showing *hybridism* and *individualization*, and undergoing *evolution* are important traits characterizing all sorts of genres. More precisely, genres can be defined as *cultural artifacts*, i.e. objects linked to a culture, a society or a community, bearing standardized features (conventions) but leaving space for creativity (individualization). On the one hand, standardized and recurrent features induce predictable expectations in the receivers. On the other hand, the freedom allowed by creativity allow genres to change, evolve, and be created to meet new needs (genre evolution), especially under the impulse of a new communication medium. While the change is still ongoing, i.e. before a modified genre is redefined, or a new genre is identified with a new name, documents show mixed forms and functions (genre hybridism).

This view of genre is flexible enough to encompass not only paper genres (both literary and practical genres), but also digital genres and, more specifically, web genres, such as the personal home page. The personal home page has no evident antecedent in the paper world (cf. Dillon and Gushrowski, 2000). It sprang up on the web as a new cultural object servicing the community of web users. When browsing a personal home page, web users expect a blend of standardized information (self-narration, personal interests, contact details, and often pictures related to one's life) and personal touch.

Another important thing to notice is that before genre conventions become fully standardized, genres do not have an official name. A genre name becomes acknowledged when the genre itself has a role and a communicative function in a community or society (Görlach, 2004: 9). Before this acknowledgement, a genre shows hybrid or individualized forms, and undefined functions. For example, before 1998 web logs (or blogs) were already present on the Web, but they were not identified as a genre. They were just "web pages", with similar characteristics and functions. In 1999, suddenly a community sprang up using this new genre (Blood, 2000). Only at this point, the genre label "web log" or "blog" started spreading and being recognized.

Genre hybridism and individualization are evident on the web, and play an important role in the change and the creation of new genres. In fact, web pages – which can be considered as a new kind of document, much more unpredictable and customized than paper documents (Santini, 2006a) – are often hybrid because of intra-genre and inter-genre variations. They are also highly individualized because of the creative freedom provided by HTML or XML tags (the building blocks of web pages) or programming languages such as Javascript. On the web, new genres are constantly added (blogs, clogs, eshops, wikis, etc.) and traditional genres are adapted or updated in order to include more or different functionalities (online front pages, ezines, net ads, etc.). Genres such as emails, newsletters, search pages, eshops, etc. were a futuristic prophecy only 10 or 15 years ago, while today they belong to the normal life of a web user. Presumably, other genres will soon be added to meet new communicative needs brought about by new technologies.

As any other evolutions, also genre evolution proceeds along the axis of time. It is a diachronic process. There must be a 'before' and an 'after'. What often hallmarks a 'before' and an 'after' is the introduction of a new communication medium within a culture, a society, a community. The added value of studying genres on the web (a new medium) is represented by the possibility of following the development of genres and genres repertoires *live*, i.e. while it is taking place, and not *a posteriori*. That is, on the web we can capture synchronically a diachronic process. From a synchronic point of view, the genre repertoire is a continuum, where there are three forces interacting: what we bring from the past (reproduced genres), what is new or adapted to the new environment (novel genres and adapted genres), what is going to emerge and is not fully formed yet (emerging genres).

This view of genre evolution complements previous studies on the same subject (cf. Crowston and Williams, 1997; Shepherd and Watters, 1998; Kwasnik and Crowston, 2005). The main contribution of the synchronic continuum is that an additional force has been acknowledged to take part in the evolution process, i.e. emerging genres. Emerging genres are those that are not fully standardized, that are still in formation and for which a genre label has not been created or have a label which is still opaque to the majority of users. Currently many web pages are in this phase of evolution, showing a high level of hybridism or individualization. We suggest that the subjects' perception of these web pages can be interpreted in term of genre evolution.

The study reported in this paper provides a snapshot of the current state of the genre repertoire of web pages seen through the perception of web users. Although this view is partial (23 labels to be assigned to 25 web pages), it offers an interesting section of the genre repertoire on the web. This study can be also seen as a confirmatory study, because it confirms that a number of recent web genres, unprecedented in the paper world (such as home page, FAQs, blog) can be recognized by the subjects; others have not fully emerged and many web users are not familiar with their new genre labels; finally some web pages show a high level of ambiguity and web users largerly disagree on assigning labels to them.

The article is organized as follows: Section 2 presents a short overview of previous work; Section 3 describes the web study and presents preliminary results; in Section 4 some conclusions are drawn.

2 Previous Work

No studies have been carried out so far on users' perception of a genre repertoire in transition.

Crowston and Williams (1997) were the first who reported on the genre repertoire on the web. They identified 48 reproduced and emergent genres in a sample of about 1,000 web pages.

A few user studies were carried out with the more pragmatic approach of exploring the usefulness of genre to improve web searches and defining a genre palette appropriate for this purpose. The most comprehensive study related to genre effectiveness for web searching is recent. Rosso (2005) carried out a series of four linked experiments, all based on human subjects. Quite surprisingly, the conclusion drawn by the author was that genre-annotated search results produced no significant improvement in participants' ability to make more consistent or faster assessment on the relevance of search results (Rosso, 2005: 133-179). In fact, only 17 of 32 participants reported noticing the genre label (Rosso, 2005: 176). Most probably, as pointed out by the author, this outcome was influenced by the difficulty and complexity of the task, together with the limitations of the setting (Rosso, 2005: 170-172).

Rosso's attempt to assess the relevance of search results including genre labels was almost unique. All other studies with web users, in contrast, did not provide any assessment of how well genres improved a web search. These studies are more like surveys on users' preferences in terms of useful non-topical categories that can help restrict web searches.

Along this line, Meyer zu Eissen and Stein (2004) built a genre palette for the web using two criteria: usability and feasibility. Their user study was based on a questionnaire where they asked about search engine use, usefulness of genre classification, and usefulness of genre classes. Interestingly, the authors note that one of the inherent problems of genre classification is that "even humans are not able to consistently specify the genre of a given page" because web pages have different functions, i.e. they might be hybrid forms, as in the case of product information sites that are combined with a shopping interface.

Roussinov et al. (2001) carried out a exploratory study of web users in order to identify what genres they most/least frequently come in contact with, and what genres most/least address their information needs. In their study, carried out in 2000, 116 different genres were identified, but not all web pages could be classified.

Karlgren (2000: 99 ff.), a pioneer in building a genre palette, tried to collect genres that were both consistent with what users expect as well as conveniently computable. He sent around a questionnaire where the core question was: "What genres do you feel you find on the WWW?". He ended up with a palette of 11 genres. One frequent comment by the respondents was that the genres in the palette were not mutually exclusive, in other words they showed some level of hybridism.

Very informative in many respects, these studies have in common the practical aim of improving web searches. This might explain why they overlook difficult issues such as the hybridism or the individualization of many web pages, which are nonetheless perceived by the subjects. The authors must necessarily focus on unambiguous exemplars, showing clear-cut conventions and expectations.

The present study, on the other hand, explores the perception of genres when users are faced not only with prototypical genre exemplars but also with hybrid or individualized web pages, and interpret the subjects' perception in term of genre evolution.

3 Web Study

The study described in this section was web-based. It was uploaded on to one of the servers at University of Brighton at the end of February 2005, and kept online for one month.

The study is based on participants who volunteered within the University of Brighton (UK), University of Sussex (UK), Dalhousie University (Canada), Syracuse University (USA), plus other academics (interested in genre-related issues) in other universities and research institutes in Europe. Potential participants were sent an email containing the URL of the study on the web.

3.1 Population and Sample: Academic Environment

Genre recognition and acknowledgement is based on elements like education, culture, community, and society. The academic population on which the study is built upon has three elements in common:

- it is a medium-high educated population (from administrative people to students and professors);
- it is very used to computer-meditated communication;
- it is familiar with the Web.

3.2 Web Pages and Web Genres

Web pages were chosen by the author of this paper from the *live* Web and from the SPIRIT collection of web pages (Joho and Sanderson, 2004). Three typologies of web genres and web pages were hypothesized for the selection and for the study (the web pages included in the study are available, together with their URLs, at http://www.nltg.brighton.ac.uk/home/Marina.Santini/:

1) Easy web genres:

1. eshop (web_page_01)
2. personal home page (web_page_02)
3. front page (web_page_04)
4. search page (web_page_05)
5. corporate home page (web_page_11)
6. FAQs (web_page_12, the word "FAQs" was deleted from the heading)
7. splash screen (web_page_24)
8. net ad (web_page_2)

2) Ambiguous web genres:

9. email (web_page_03, because of the format and the granularity: email vs. mailing list)
10. sitemap (web_page_06, the words "sitemap" and "hotlist" were deleted from the heading)
11. hotlist (web_page_15,the word "hotlist" was deleted from the heading),
12. academic personal home page (web_page_08)
13. about page (web_page_10)
14. organizational home page (web_page_14)
15. blog (web_page_07)
16. clog (web_page_16,blog and clog could be swapped in their interpretation)
17. search by multiple fields (web_page_17)
18. online form (web_page_10, online forms and search by multiple field can appear very similar)
19. newsletter (web_page_19, which was presented truncated),
20. howto page (web_page_20)
21. online tutorial (web_page_22, online tutorial is a super-genre of howto pages)

3) Difficult web pages:

22. ezine cover (web_page_13)
23. "Adirondack Orienteering Klub" (web_page_18, the author could not find a genre for it)
24. CitiDex (web_page_21, the author could not find a genre for it)
25. Collimating Lens Holder (web_page_23, the author could not find a genre for it)

The expectation was that easy web genres would collect the highest rate of agreement, ambiguous web genres would receive a lower agreement rate, while difficult web pages were expected to be the most controversial in users' perception.

The term "genre" was never mentioned in the whole study in order not to influence or confuse the participants. The goal of the study was not declared either because the idea was to ask for a genre classification of web pages implicitly and study the reactions. Participants were simply told to assign "labels" to web page "types".

3.3 Participants' Task and Sample Size

The task of participants was straightforward. They had to go through 25 screenshots of web pages and assign one of the 23 labels to each of them.

The total number of users who started the experiment was 198. 135 participants went through the whole study and provided valid responses for the experiment.

3.4 Results

Currently, there is no standard test largely agreed upon that can be used for experiments where subjects can make choices from a large number of categories (23 labels) for a large number of objects (25 web pages). In the following paragraphs some views and interpretations of the data are presented, namely raw counts and percentages, Fisher's exact test, and adjusted residuals.

Raw Counts and Percentages. A view on the data is offered in Table 2, which shows the number of subjects assigning a particular label to a particular web page and the percentage of the most voted label. For example, the label *eshop* (8th row) was assigned to WP1[1] (first column) by 119 subjects (highlighted cell), which corresponds to 88.15% (bottom row). Four subjects thought that WP1 was a *corporate home pages* (around 2.9%), seven selected *net ad* (around 5%), one subject chose *front page* (around 0.7%), one *hotlist*, one *did not know*, two added a new label for it (around 1.4%).

Three ranges of agreement can be identified out of this table. The top range includes web pages with a percentage of agreement above 80%; the middle range groups web pages with an agreement between 79% and 50%; finally the bottom range contains web pages with an agreement between 49 % and 20%. Table 1 lists the web pages by percentage of agreement.

From these ranges a first conclusion can be drawn. According to the ranges shown in Table 1, participants show the highest agreement on what we selected as "easy web genres", except in three cases: front pages, net ad and splash screen, which seem among the least agreed upon (see bottom range). The middle range includes most of the ambiguous web genres together with ezine, which was deemed to be difficult by the author. The bottom range includes the rest of the ambiguous genres, together with other difficult web pages and three web pages from the top range, webpage_type_04 (front page), webpage_type_24 (splash screen) and webpage_type_25 (net ad).

We have now a first picture of users' perception of some web pages in relation to some web genre labels. The hypothesized genre recognition pattern was mostly confirmed in the top range, but slightly reshuffled in the middle and bottom ranges. Figure 1 shows the charted percentages.

Fisher's Exact Test. The percentages at the bottom row in Table 2 can be interpreted in terms of conditional distribution on the most voted label (response variable) per web page type (explanatory variable). In other words, they refer to the sample distribution of most voted labels, *conditional* to the web page type. In terms of association, this means that the distribution of the response variable (the label) changes with the value of the explanatory variable (the web page type) if the two variables are related. Table 2 suggests the existence of an association or correlation between the label and the web page to which this label was assigned. But as Table 2 refers to the sample rather than the population, it provides evidence but not the final answer to whether labels and web page types are associated in the way suggested by the percentages. In order to see if it is plausible that labels and web page types are associated in the population, Fisher's exact test can be calculated. The value returned for this test by SPSS is 9292.275, which is large enough to reject the hypothesis that labels and web page are independent[2]. This statistically significant association shows that the web pages chosen by the author to represent some web genres mostly map the subjects' perception of these web pages. It also shows that many genre labels are acknowledged by the users and are consistently associated to web pages.

Adjusted Residuals: A test statistic, such as Fisher's exact test, and statistical significance summarize the strength of evidence against the null hypothesis of independence, but does not indicate how many and which cells deviate greatly from this hypothesis. Residuals, i.e. the differences between expected and observed cell frequencies can help in this task. In particular, adjusted residuals can indicate if the cell counts are significantly different from what independence predicts. A large adjusted residual provides evidence against independence of a cell. As Table 3 mostly maps Table 2, a significant association between genre labels and web page types on the cells containing the most voted labels is then confirmed.

[1] WP1, WP2, WP3, etc. are short form of webpage_01, webpage_02, webpage_03, etc.

[2] The larger the value, the greater the evidence against the null hypothesis of independence.

3.5 Discussion

The original impression that there were different degree of perception of genres of web pages was confirmed by these preliminary results. Also the rough distinction into three levels of genre awareness (easy, ambiguous and difficult) was confirmed. Three ranges of perception came out clearly from percentages, but the distribution of the web pages into these three ranges is slightly different from what was expected.

The general view of the results (Fisher's test) reveals that there is a significant association between the 25 web pages and the 23 labels. The analysis of adjusted residuals support this interpretation.

The agreement among subjects on the label to assign to a particular web pages can be divided into three levels.

At the first level, which can be interpreted as the highest perception of web genres, there are web pages labelled as personal home page (webpage_type_02), eshop (webpage_type_01), corporate home page (webpage_type_11), FAQs (webpage_type_12), and search pages (webpage_type_05). We can define these labels as stable web genres.

At a middle level of perception, there are web genres still emerging. Most of the labels are fairly novel (ezine, clog, blog, about, how to), sometimes not entirely transparent, and some of them are specialized (academic home page, organizational home page, online tutorial). Probably the textual conventions of these genres are not entirely standardized yet and can cause oscillation in users' perception. This level offers the most interesting view on a genre repertoire which is moving and evolving and it is not consolidated yet.

The bottom range shows a blurred level of perception for different reasons. For some genres such as email and newsletter, the presentation in form of screenshots was not ideal. Subjects could not navigate through the web page and they could not resolve the level of granularity. For instance, for webpage_type_03 (the web page selected by the author to represent an email), 66 subjects chose email, but 34 subjects preferred to add a new label for it and 20 thought it was an about page. Surprisingly, labels such as splash screen and front page for webpage_type_04 and webpage_type_24 were not favoured by the respondents who preferred to add their own labels in many cases. For webpage_type_06, subjects preferred the label search page instead of sitemap. Another interesting case is net ad (webpage_type_25), which was often assessed as eshop, probably because the concept of advertising and selling are closely related. The most opaque label seems to be hotlist (webpage_type_15) because most subject preferred to add their own label. Three of the four web pages that were classified by the author as "I don't know" belong to this level of perception. While webpage_type_21 fell into the middle range because most of the subjects perceive it as a search page, the genre perception or interpretation of webpage_type_17, webpage_type_18, and webpage_type_23 is not so straightforward. For instance, webpage_type_17 was assessed as online form (57 subjects), search page (26 subjects), an eshop (26 subjects) and probably it is has all these functions at the same.

4 Conclusions and Future Work

The study shows a composite picture of the perception of the genre repertoire on the Web. This picture focuses on recent genres only, overlooking those more based on paper genres because, in our opinion, this hot area can reveal more about the dynamics behind genre evolution.

Preliminary findings coming out from this study confirm the initial hypothesis and show that users' perception can be divided into three ranges. These three ranges can be interpreted in terms of genre evolution: high perception for the most stable and acknowledge genres; medium perception for emerging genres, not fully acknowledged by the majority or still unstable, and finally low perception for the highly ambiguous genres (for different reasons). Some of the new web genres can be unambiguously perceived (for example, personal home page, eshop, corporate home page, FAQs and search page).

Web users can also handle a certain degree of granularity, for example by distinguishing a personal home page from a corporate home page, but the boundary between academic home pages and organizational home pages is still too fuzzy for them.

The approach to the web as a genre repertoire in evolution and these preliminary findings can turn out to be useful when building web genre palettes or when designing new genre identification experiments.

Future work includes the computation of agreement coefficients. K statistic is largely used but still controversial and mostly used for measuring the agreement of two or three raters. Two new interesting measures to assess users' recognition of web page genres were used by Rosso (2005: 109 ff.), but their full interpretation is still under study. The challenging follow up of these preliminary results is to find an objective coefficient of agreement applicable for 135 raters that can choose among 23 categories to classify 25 objects.

References

Bathia Vijay (1993), *Analysing Genre. Language Use in Professional Settings*, Longman, London-NY.

Blood R. (2000),Weblogs: A History and Perspective, http://www.rebeccablood.net/essays/weblog_history.htm

Crowston K. and Williams M. (1997), Reproduced and Emergent Genres of Communication on the World-Wide Web, *Proc. 30th Hawaii International Conference on System Sciences*.

Dillon A. and Gushrowski B. (2000), Genres and the Web: is the personal home page the first uniquely digital genre?, *Journal of the American Society for Information Science*, 51, 2.

Görlach M. (2004), *Text Types and the History of English*, Mouton de Gruyter, Berlin-NY.

Joho H. and Sanderson M. (2004), The SPIRIT collection: an overview of a large web collection, *SIGIR Forum*, December 2004 (Vol. 38, N. 2)

Karlgren J. (2000), *Stylistic Experiments for Information Retrieval*, Thesis submitted for the degree of PhD, Stockholm University, Sweden.

Karlgren J. (2004), The Wheres and Whyfores for Studying Textual Genre Computationally, *Papers from the AAAI Fall Symposium*.

Kwasnik, B. and Crowston, K. (2005), Genres of digital documents: Introduction to the special issue, *Information, Technology & People*, 18(2), 76–88.

Meyer zu Eissen S. and Stein B. (2004), Genre Classification of Web Pages: User Study and Feasibility Analysis, in Biundo S., Fruhwirth T. and Palm G. (eds.), *KI 2004: Advances in Artificial Intelligence*, Springer, Berlin-Hedelberg-NY, p. 256-269.

Rosso M. (2005), *Using Genre to Improve Web Search*, Thesis submitted for the degree of PhD, University of North Carolina at Chapel Hill, USA.

Roussinov D., Crowston K., Nilan M., Kwasnik B., Cai J., Liu X. (2001), Genre Based Navigation on the Web, *Proc. 34th Hawaii International Conference on System Sciences*.

Santini M. (2005), Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis, *Proc. CLUK 05*.

Santini M. (2006a), Web pages, text types, and linguistic features: Some issues, *ICAME Journal,* Vol. 30.

Santini M. (2006b), Automatic Identification of Genres in Web Pages, *forthcoming*.

Shepherd M. and Watters C. (1998), The Evolution of Cybergenre, *Proc. 31st Hawaii International Conference on System Sciences*.

Swales J. (1990), *Genre Analysis. English in academic and research settings*, Cambridge University Press, Cambridge.

Trosborg Ann (ed.) (2000), *Analysing Professional Genres*, John Benjamins Publishing Company, Amsterdam-Philadelphia.

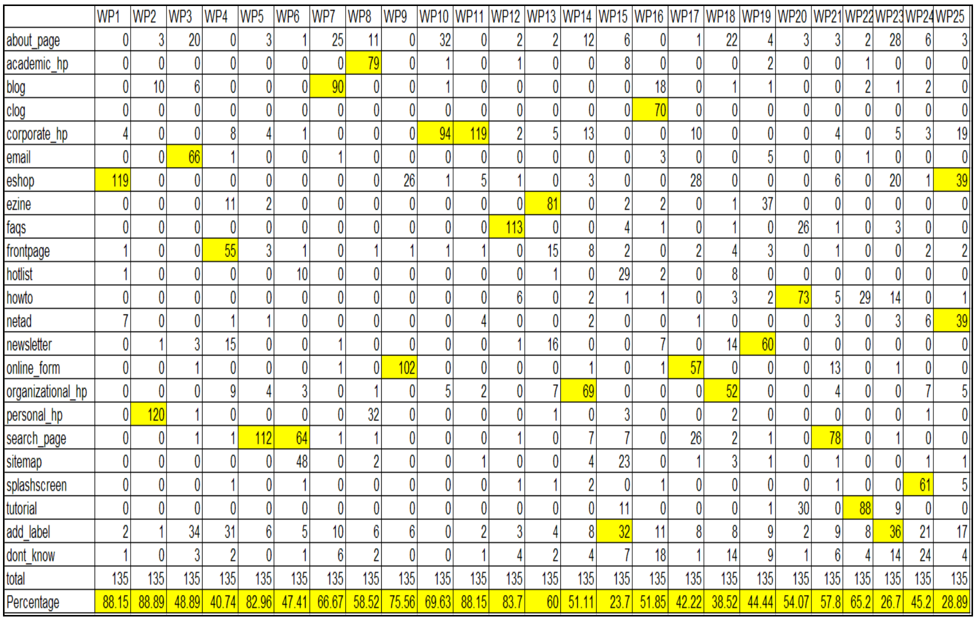
Yates J., and Orlikowski W. (1992), Genres of organizational communication: A structural approach to studying communications and media, *Academy of Management Review*, Vol. 17, 2, 229-326.

Appendix

Ranges		Web page type and related web genre suggested by the author
Top: Above 80%	5	webpage_type_01 (eshop), webpage_type_02 (personal_hp), webpage_type_05 (search_page), webpage_type_11 (corporate_hp), webpage_type_12 (faqs)
Middle: From 79% to 50%	10	webpage_type_07 (blog), webpage_type_08 (academic_hp), webpage_type_09 (online_form) webpage_type_10 (about_page) webpage_type_13 (ezine) webpage_type_14 (organiz_hp) webpage_type_16 (clog) webpage_type_20 (howto) webpage_type_21 (dontknow) webpage_type_22 (tutorial)
Bottom: From 49% to 20%	10	webpage_type_03 (email) webpage_type_04 (frontpage) webpage_type_06 (sitemap) webpage_type_15 (hotlist) webpage_type_17 (dontknow) webpage_type_18 (dontknow) webpage_type_19 (newsletter) webpage_type_23 (dontknow) webpage_type_24 (splashscreen) webpage_type_25 (netad)

Z

Table 1. Ranges of percentages



	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	WP10	WP11	WP12	WP13	WP14	WP15	WP16	WP17	WP18	WP19	WP20	WP21	WP22	WP23	WP24	WP25
about_page	0	3	20	0	3	1	25	11	0	32	0	2	2	12	6	0	1	22	4	3	3	2	28	6	3
academic_hp	0	0	0	0	0	0	0	79	0	1	0	1	0	0	8	0	0	0	2	0	0	1	0	0	0
blog	0	10	6	0	0	0	90	0	0	1	0	0	0	0	0	18	0	1	1	0	0	2	1	2	0
clog	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0	0	0	0	0	0	0
corporate_hp	4	0	0	8	4	1	0	0	0	94	119	2	5	13	0	0	10	0	0	0	4	0	5	3	19
email	0	0	66	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	5	0	0	1	0	0	0
eshop	119	0	0	0	0	0	0	0	26	1	5	1	0	3	0	0	28	0	0	0	6	0	20	1	39
ezine	0	0	0	11	2	0	0	0	0	0	0	0	81	0	2	2	0	1	37	0	0	0	0	0	0
faqs	0	0	0	0	0	0	0	0	0	0	0	113	0	0	4	1	0	1	0	26	1	0	3	0	0
frontpage	1	0	0	55	3	1	0	1	1	1	1	0	15	8	2	0	2	4	3	0	1	0	0	2	2
hotlist	1	0	0	0	0	10	0	0	0	0	0	0	1	0	29	2	0	8	0	0	0	0	0	0	0
howto	0	0	0	0	0	0	0	0	0	0	0	6	0	2	1	1	0	3	2	73	5	29	14	0	1
netad	7	0	0	1	1	0	0	0	0	0	4	0	0	2	0	0	1	0	0	0	3	0	3	6	39
newsletter	0	1	3	15	0	0	1	0	0	0	0	1	16	0	0	7	0	14	60	0	0	0	0	0	0
online_form	0	0	1	0	0	0	1	0	102	0	0	0	0	1	0	1	57	0	0	0	13	0	1	0	0
organizational_hp	0	0	0	9	4	3	0	1	0	5	2	0	7	69	0	0	0	52	0	0	4	0	0	7	5
personal_hp	0	120	1	0	0	0	0	32	0	0	0	0	1	0	3	0	0	2	0	0	0	0	0	1	0
search_page	0	0	1	1	112	64	1	1	0	0	0	1	0	7	7	0	26	2	1	0	78	0	1	0	0
sitemap	0	0	0	0	0	48	0	2	0	0	1	0	0	4	23	0	1	3	1	0	1	0	0	1	1
splashscreen	0	0	0	1	0	1	0	0	0	0	0	1	1	2	0	1	0	0	0	0	1	0	0	61	5
tutorial	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	1	30	0	88	9	0	0
add_label	2	1	34	31	6	5	10	6	6	0	2	3	4	8	32	11	8	8	9	2	9	8	36	21	17
dont_know	1	0	3	2	0	1	6	2	0	0	1	4	2	4	7	18	1	14	9	1	6	4	14	24	4
total	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135	135
Percentage	88.15	88.89	48.89	40.74	82.96	47.41	66.67	58.52	75.56	69.63	88.15	83.7	60	51.11	23.7	51.85	42.22	38.52	44.44	54.07	57.8	65.2	26.7	45.2	28.89

Table 2. Users' assignment and percentages

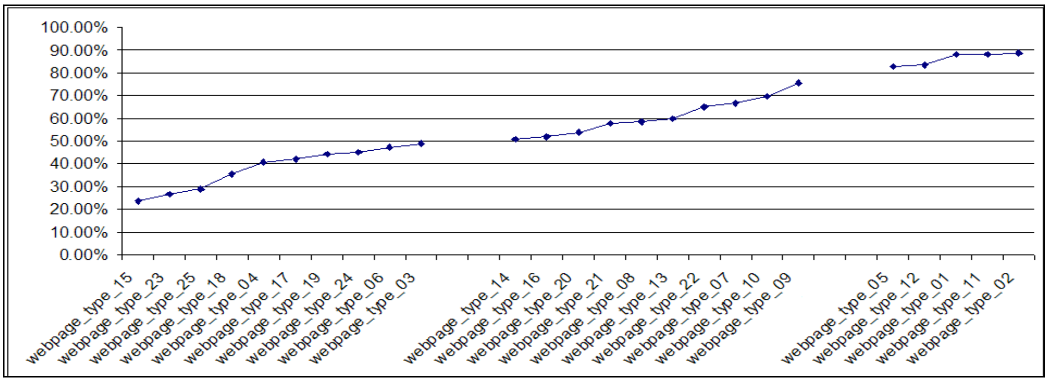
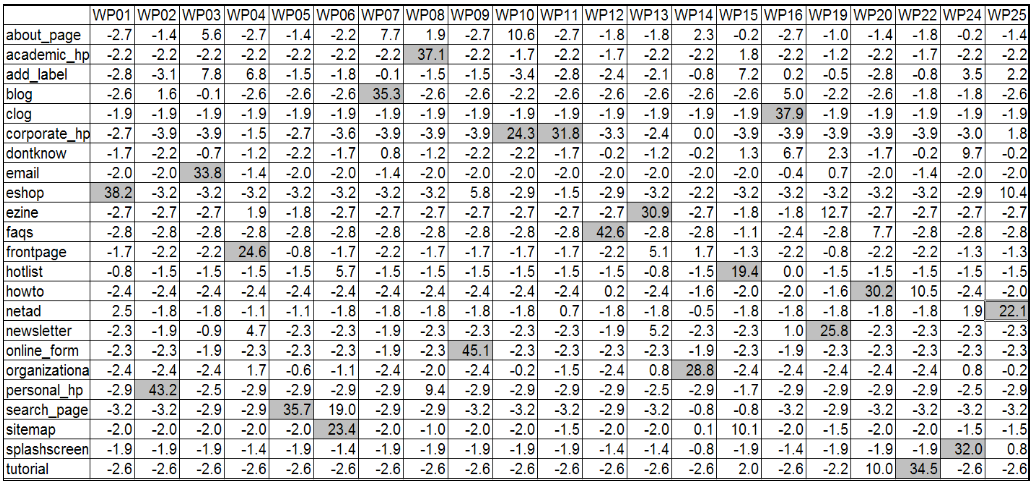


Figure 1. Charted percentages



	WP01	WP02	WP03	WP04	WP05	WP06	WP07	WP08	WP09	WP10	WP11	WP12	WP13	WP14	WP15	WP16	WP19	WP20	WP22	WP24	WP25
about_page	-2.7	-1.4	5.6	-2.7	-1.4	-2.2	7.7	1.9	-2.7	10.6	-2.7	-1.8	-1.8	2.3	-0.2	-2.7	-1.0	-1.4	-1.8	-0.2	-1.4
academic_hp	-2.2	-2.2	-2.2	-2.2	-2.2	-2.2	-2.2	37.1	-2.2	-1.7	-2.2	-1.7	-2.2	-2.2	1.8	-2.2	-1.2	-2.2	-1.7	-2.2	-2.2
add_label	-2.8	-3.1	7.8	6.8	-1.5	-1.8	-0.1	-1.5	-1.5	-3.4	-2.8	-2.4	-2.1	-0.8	7.2	0.2	-0.5	-2.8	-0.8	3.5	2.2
blog	-2.6	1.6	-0.1	-2.6	-2.6	-2.6	35.3	-2.6	-2.6	-2.2	-2.6	-2.6	-2.6	-2.6	-2.6	5.0	-2.2	-2.6	-1.8	-1.8	-2.6
clog	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	37.9	-1.9	-1.9	-1.9	-1.9	-1.9
corporate_hp	-2.7	-3.9	-3.9	-1.5	-2.7	-3.6	-3.9	-3.9	-3.9	24.3	31.8	-3.3	-2.4	0.0	-3.9	-3.9	-3.9	-3.9	-3.9	-3.0	1.8
dontknow	-1.7	-2.2	-0.7	-1.2	-2.2	-1.7	0.8	-1.2	-2.2	-2.2	-1.7	-0.2	-1.2	-0.2	1.3	6.7	2.3	-1.7	-0.2	9.7	-0.2
email	-2.0	-2.0	33.8	-1.4	-2.0	-2.0	-1.4	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-0.4	0.7	-2.0	-1.4	-2.0	-2.0
eshop	38.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	5.8	-2.9	-1.5	-2.9	-3.2	-2.2	-3.2	-3.2	-3.2	-3.2	-3.2	-2.9	10.4
ezine	-2.7	-2.7	-2.7	1.9	-1.8	-2.7	-2.7	-2.7	-2.7	-2.7	-2.7	-2.7	30.9	-2.7	-1.8	-1.8	12.7	-2.7	-2.7	-2.7	-2.7
faqs	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8	42.6	-2.8	-2.8	-1.1	-2.4	-2.8	7.7	-2.8	-2.8	-2.8
frontpage	-1.7	-2.2	-2.2	24.6	-0.8	-1.7	-2.2	-1.7	-1.7	-1.7	-1.7	-2.2	5.1	1.7	-1.3	-2.2	-0.8	-2.2	-2.2	-1.3	-1.3
hotlist	-0.8	-1.5	-1.5	-1.5	-1.5	5.7	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-0.8	-1.5	19.4	0.0	-1.5	-1.5	-1.5	-1.5	-1.5
howto	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	0.2	-2.4	-1.6	-2.0	-2.0	-1.6	30.2	10.5	-2.4	-2.0
netad	2.5	-1.8	-1.8	-1.1	-1.1	-1.8	-1.8	-1.8	-1.8	-1.8	0.7	-1.8	-1.8	-0.5	-1.8	-1.8	-1.8	-1.8	-1.8	1.9	22.1
newsletter	-2.3	-1.9	-0.9	4.7	-2.3	-2.3	-1.9	-2.3	-2.3	-2.3	-2.3	-1.9	5.2	-2.3	-2.3	1.0	25.8	-2.3	-2.3	-2.3	-2.3
online_form	-2.3	-2.3	-1.9	-2.3	-2.3	-2.3	-1.9	-2.3	45.1	-2.3	-2.3	-2.3	-2.3	-1.9	-2.3	-1.9	-2.3	-2.3	-2.3	-2.3	-2.3
organizationa	-2.4	-2.4	-2.4	1.7	-0.6	-1.1	-2.4	-2.0	-2.4	-0.2	-1.5	-2.4	0.8	28.8	-2.4	-2.4	-2.4	-2.4	-2.4	0.8	-0.2
personal_hp	-2.9	43.2	-2.5	-2.9	-2.9	-2.9	-2.9	9.4	-2.9	-2.9	-2.9	-2.9	-2.5	-2.9	-1.7	-2.9	-2.9	-2.9	-2.9	-2.5	-2.9
search_page	-3.2	-3.2	-2.9	-2.9	35.7	19.0	-2.9	-2.9	-3.2	-3.2	-3.2	-2.9	-3.2	-0.8	-0.8	-3.2	-2.9	-3.2	-3.2	-3.2	-3.2
sitemap	-2.0	-2.0	-2.0	-2.0	-2.0	23.4	-2.0	-1.0	-2.0	-2.0	-1.5	-2.0	-2.0	0.1	10.1	-2.0	-1.5	-2.0	-2.0	-1.5	-1.5
splashscreen	-1.9	-1.9	-1.9	-1.4	-1.9	-1.4	-1.9	-1.9	-1.9	-1.9	-1.9	-1.4	-1.4	-0.8	-1.9	-1.4	-1.9	-1.9	-1.9	32.0	0.8
tutorial	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	-2.6	2.0	-2.6	-2.2	10.0	34.5	-2.6	-2.6

Table 3. Adjusted residuals

Novelle, a collaborative open source writing tool software

Federico Gobbo
DICOM
21100, Insubria University
Varese, Italy
federico.gobbo@uninsubria.it

Michele Chinosi
DICOM
21100, Insubria University
Varese, Italy
michele.chinosi@gmail.com

Massimiliano Pepe
DICOM
21100, Insubria University
Varese, Italy
massimiliano.p@gmail.com

Abstract

In this paper we discuss the notions of hypertext, blog, wiki and cognitive mapping in order to find a solution to the main problems of processing text data stored in these forms. We propose the structure and architecture of Novelle as a new environment to compose texts. Its flexible model allows the collaboration for contents and a detailed description of ownership. Data are stored in a XML repository, so as to use the capabilities of this language. To develop quickly and efficiently we choose AJAX technology over the Ruby on Rails framework.

1 Introduction

Computational linguists are facing the explosion of new forms of writing as a mass phenomenon. Telling personal and collaborative stories throught web technologies is known under the etiquettes of 'blog' and 'wiki'. It therefore brings new challenges to the field of natural language processing. We are trying to address them by rendering explicitly the structure of these new forms of text in a way suitable for linguistic computation. In order to do so, we are building an open source writing tool software, called Novelle.

1.1 Hypertext as a New Writing Space

Bolter (1991) was the first scholar who stressed the impact of the digital revolution to the medium of writing. Terms as 'chapter', 'page' or 'footnote' simply become meaningless in the new texts, or they highly change their meaning. When Gutenberg invented the printing press and Aldo Manuzio invented the book as we know it, new forms of writings arose. For example, when books shouldn't be copied by hand any longer, authors took the advantage and start writing original books and evaluation – i.e. literary criticism – unlike in the previous times (Eisenstein, 1983). Nowadays the use of computers for writing has drammatically changed, expecially after their interconnection via the internet, since at least the foundation of the web (Berners-Lee, 1999). For example, a 'web page' is more similar to an infinite canvas than a written page (McCloud, 2001). Moreover, what seems to be lost is the relations, like the texture underpinning the text itself. From a positive point of view these new forms of writing may realize the postmodernist and decostructionist dreams of an 'opera aperta' (open work), as Eco would define it (1962). From a more pessimistic one, an author may feel to have lost power in this openness. Henceforth the collaborative traits of blogs and wikis (McNeill, 2005) emphasize annotation, comment, and strong editing. They give more power to readers, eventually filling the gap - the so-called active readers become authors as well. This situation could make new problems rise up: Who owns the text? Which role is suitable for authors? We have to analyse them before presenting the architecture of Novelle.

1.2 Known problems

It is certainly true that wikis and blogs are new forms of text. It is also true that we have already met these problems in the first form of purely digital texts – hypertexts. Now we are facing the same question during processing texts in blogs and wikis. We consider hypertexts as parents of blogs and wikis. Our aim is to use the analysis of hypertexts for interesting insights, useful for blogs and wikis too.

Following the example of Landow (1994), we will call the autonomous units of a hypertext *lexias* (from 'lexicon'), a word coined by Roland Barthes (1970). Consequently, a hypertext is a set of lexias. In hypertexts transitions from one lexia to another are not necessarily sequential, but navigational. The main problems of hypertexts, acknowledged since the beginning, have been traced as follows (Nelson, 1992):

- *The framing problem*, i.e. creating arbitrary closed contexts of very large document collections. When extracting sub-collections, some links may be cut off.
- *Comparing complex alternatives*, i.e. to get parallel or alternate versions of the same document in a simple and effective way, one of the main goal of Xanadu, the ultimate "global hypertext" dreamt by Nelson.
- *Typology of links*, i.e. when links become too many, we need a typology for links, avoiding confusion to the reader/author.
- *Version control*, as the system should keep track of the history of every document, saving differences and pointing out correspondencies.

We take from wikis the concept of document history and its consequences. We consider it as a good approximation of the 'version control' concept as shown above.

In wikis every document keeps track of its own history: *creating* a document means to start a history, *editing* a document to move ahead, *restoring* to move back onto the history timeline, *destroying* a document to stop the history itself. Moreover, a *sandbox* is a temporary view of a document itself - i.e. a sandbox can not cause a change in the history (Cunningham and Leuf, 2001). Figure 1 shows the model.

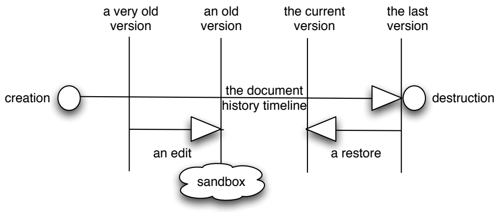


Figure 1: The document history model

History snapshots of the timeline may be considered as permanent views, i.e. views with a timestamp. Consequently, except in the case of sandboxes, every change in the document cannot be erased. This model will have a strong impact on the role of links and on the underpinning structure of Novelle itself.

2 The Structure of Novelle

Our aim is to create an open source hypertext modeling software, called Novelle. 'Novelle' is an Italian old-fashioned word meaning 'novels', and in German it means 'novel' too. It resembles the English word 'novel' and the French word 'nuovelle'. We believe that this name is clearly understable to every people educated in a European-based culture, and this is why we have chosen it.

The emphasis on narrativity takes into account the use of blogs as public diaries on the web, that is still the main current interpretation of this literary genre, or *metagenre* (McNeill, 2005). Furthermore we noticed that blogs and wikis are currently subjected to osmosis, because they have in common the underlying core technology. So blogs are a literary metagenre which started as authored personal diaries or journals. Now they try to collect themselves in so-called 'blogspheres'. On the other side, wikis started as collective works where each entry is not owned by a single author - e.g. Wikipedia (2005). Now personal wiki tools are arising for brainstorming and mind mapping. See Section 4 for further aspects.

2.1 The Problem of Ownership

The main difference between blogs and wikis is in the ownership of documents. Most blogs follow the *annotation model*, where a single lexia is central and the others are comments, sometimes in threads. Every lexia is authored and changes are minimal. People prefer commenting instead of editing. The paradigm is "write once, read many".

On the contrary, in wikis no lexia is authored and there is no hierarchy between lexias. In fact a document is still a set of lexias, but every document is only the set of historical versions of the document itself. Generally, people avoid commenting, preferring to edit each document. The paradigm is "write many, read many" (Cunningham and Leuf, 2001).

We believe that ownership has an important role and we do not want to force our users to take a non-attributive copyright licence to their work. We consider the Creative Commons model as the most suitable one to let each author choose the rights to reserve (Lessig, 2004). Narrative writings or essays are creative works and they generally treat ownership as authorship, even for the most enthusiastic fellows of free culture (Stallman, 2001).

2.2 The Representation of Context

In the structure of Novelle we are trying to retain authorship and the core concept of document history of wikis through a typology of links, taking what we consider the best of the two worlds of blogs and wikis.

In Novelle each user owns his own lexias, and the relations between them, i.e. links. Furthermore authors are free to read and to link other users' lexias. In other words, each user does permit everyone to link its own lexias for free, at the condition that the others do the same. Every user may recall the link list on each element (e.g. a single word) of his lexias at every time, but he can not destroy them. Lexias may be commented by every user, but the author may retain for himself the right to edit. This decision has to be taken when a lexia is created.

If a user lets others edit some lexias, he has the right to retain or refuse the attribution when other users have edited it. In the first instance, the edited version simply moves ahead the document history. In the second one, the last user, who has edited the lexia, may claim the attribution for himself. The lexia will be marked as a derivative work from the original one, and a new document history timeline will start (see Figure 2). Authors may choose this right with the No-Deriv option of the Creative Commons licences (Lessig, 2004).

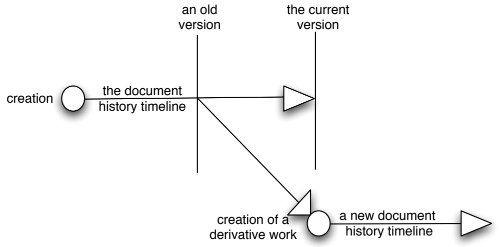

an old version
the current version
creation
the document history timeline
creation of a derivative work
a new document history timeline

Figure 2: How to create derivative works

If nobody claims the document for himself, it will fall in the public domain. The set of lexias in the public domain will form a special document, owned by a special user, called Public Domain. If the author refuses the permission to create derivative works, i.e. to edit his own lexias, users still have the right to comment the author's work. So as to come to terms with this idea, we need a concept invented by Nelson (1992), i.e. *transclusion*.

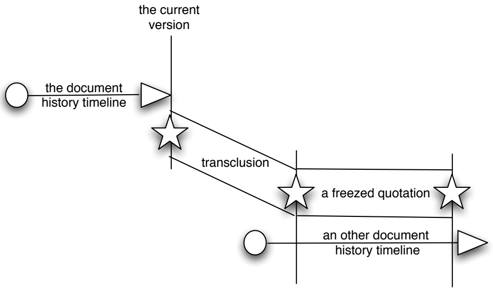

the current version
the document history timeline
transclusion
a freezed quotation
an other document history timeline

Figure 3: How transclusion works

Rather than copy-and-paste contents from a lexia, a user may recall a quotation of the author's lexia and write a comment in the surroundings. In doing so, the link list of the author's lexia will be updated with a special citation link marker, called *quotation link* (see later for details). Usually, the quotation will be 'frozen', as in the moment where it was transcluded (see Figure 3). Consequently the transclusion resembles a copied-and-pasted text chunk, but the link to the original document will always be consistent, i.e. neither it expires nor it returns an error. Otherwise the user who has transcluded the quotation may choose to keep updated the links to the original document. This choice has to be made when the transclusion is done.

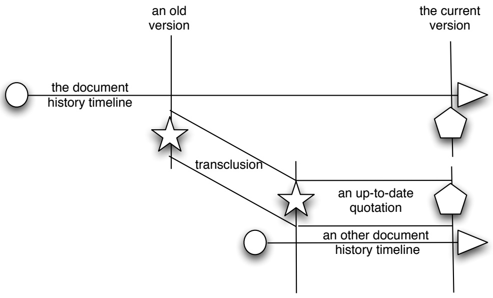


Figure 4: An up-to-date transclusion

If so, the transcluded quotation will update automatically, following the history timeline of the original document. For example, if the original document changes topic from stars to pentagons, the quotation transcluded will change topic too (see Figure 4).

2.3 Contents and the Typology of Links

Following our model of ownership, there are at least two categories of links: shallow links and deep links. By *shallow links* we mean visual links occurring in a single canvas, usually owned by the same author. These will represent iconically the relationship space of lexias, as explained by McCloud, talking about web comics (2001). They are particularly useful when comparing parallel versions of the same text, e.g. digital variants (see Conclusions).

We consider a web page, or better a web canvas, as a *view of lexias*, i.e. a group of lexias and their relations visually shown with shallow links. A set of lexias is a *document*. Every author has the right to decide the relation type of a set of lexias, i.e. to form a document. A document can also be considered as a collection of history timelines, i.e. the set of related lexias and their versions. The set of documents is the *docuverse*, a word coined by Nelson (1992). We use the word docuverse, unlikely the original sense, with the meaning of a set of documents owned by a single author.

Every document can be viewed within a web canvas, but users may click on a deep link and so change view. With *deep links* we mean links which let the user change view, i.e. rearrange elements in the web canvas for revealing shallow links between lexias, belonging to the same document or not. Therefore a web canvas may show relations between views owned by different authors. We consider quotation links, i.e. links created by transclusion, as a special kind of deep links. Authors may create specific views adding labels to links. The set of labels will form a typology of links, customized by every user and even shared, on demand of users' desires.

With our typology of links, we aim to solve the framing problem as defined in Section 1.2. We want to model views as dynamic objects - the creation of context will be still arbitrary, but changes are very easily. We would also provide a user facility for choosing the right licence for every lexia, following the model of Creative Commons licences (Lessig, 2004).

3 The Architecture of Novelle

We have considered many hypotheses in order to choose a first layer architecture to save a repository. We used a multi-tier model based on XML. Our idea is based on merging together some of the most common design techniques used in blogs and wikis. Recently previous implementation techniques have been studied from their new aspects to find innovative web technologies. A basic scheme of Novelle architecture is presented in Figure 5. The first layer is the most important. It is based on an infrastructure for storing effectively data repository in order to obtain the best performances. We have studied two alternatives for the repository. On one side we have different techniques to map XML trees onto a database management system. On the other side we may map XML trees directly on a filesystem – see below for details.

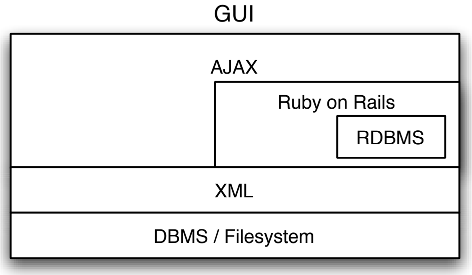


Figure 5: Novelle: multi-tier architecture

The second layer is represented by XML. Messages, data and metadata are exchanged between layers using the capability of this language. This allows to treat data and metadata on different level of abstraction.

The *Ruby on Rails* (2006) framework permits us to quickly develop web applications without rewriting common functions and classes.

We used the *Asyncronous Javascript And XML* (or AJAX) paradigm to create the graphical user interface. AJAX function lets the communication works asyncronously between a client and a server through a set of messages based on HTTP protocol and XML (Garrett, 2005).

3.1 XML repository

We chose to use XML trees to store together data, metadata, messages and their meanings because it has some benefits. The most important is storing

XML data. The other benefits of a native XML solution are: the storing without mapping your XML to some other data structure like objects, classes or tables; the neatness of the structure; the underlying technology from the abstract layer to the physical ones is based on a unique standard, widely accepted by the community. Data may be entered and retrieved as XML. Another advantage is flexibility, gained through the semi-structured nature of XML and the schema independent model used by most of native XML databases. This is especially valuable when you have very complex XML structures that would be difficult or impossible to map to a more structured database. At this time there are not XML databases so stable to be used into project of this kind.

Xindice (developed by Apache Group) proved better than others. Apache Xindice is a database designed from the ground up to store XML data or what is more commonly referred to as a native XML database. It stores short XML documents in collections with runtime generation of indexes. Unfortunately Xindice seems not to have been developed any more since April 2004.

Another native XML database, more usable and supported, is *eXist*. eXist is growing quickly and it implements some functionalities of Xindice. The standards support is not completed and some functions are currently being rewritten directly embedded into the software. After doing many tests on it, it reveals worse performances with respect to other platforms, even if it is more complete in comparison to the others.

Anyway our interest keeps focusing on them waiting for the first stable release effectively usable in Novelle. We have considered the possibility to map XML trees to relational or object-oriented database management systems that support XML. We can map directly an XML tree into a memory tree structure, made up with classes and objects, with object-oriented databases, as we can see in *Ozone* project (2006). The last stable version of Ozone was released in 2004. The main problem with Ozone - and with others OODBMS - is the overhead requested to the memory for storing a complex tree. On the other side, many RDBMS with XML support map directly an XML tree to an entity-relationship schema. In some instances XML trees are stored as Binary Large Object, or BLOB, into one big table. In other situations XML trees are parsed, splitted and finally stored in tables where attributes have the same names than XML nodes.

Ronald Bourret (2006) mantains and updates a very comprehensive list of native XML databases on his web site.

While we are waiting for a native XML database stable and useful for our project, we have decided to get inspiration from the common idea used in many blogs and wikis. Most of these architectures are used to store messages in a structure that is similar to a directory tree saved on filesystem. Often this idea is only developed to present to users messages organized in collection ordered by time (e.g. blogs), but all the platforms are based on RDBMS. We have found in our research only one other project in which messages are stored directly on filesystem: the *Gblog* project (Gblog, 2005). Nobody usually adopt this solution because the security of the web site is less strong. In order to represent messages archives the most common structure is the triple *../year/month/day/..*. In our assumption, a message is a history. Therefore a structure of this kind works very well with our idea. We are going to build a filesystem time-based structure in which we can directly map our messages i.e. our histories. This structure is also a tree. We can write also an XML document that mantains an architecture scheme with some indexes to speed up queries. Moreover, we store with a message another XML document representing all the past history (i.e. the paths) of the message.

So as to sum up, every time a user stores a message, he has to save the first XML document with the message, then saves or updates a second XML document representing its past history and finally saves or updates a third XML message with filesystem directory tree. The overhead on bandwith and net speed of this solution does not let users notice significant differences, even though it is necessary to grant writing permissions to everyone on the entire repository. Having a native XML database will give the advantage of saving XML documents in a rapid, neat and indicized way, in order to be able to execute efficient queries on the repository.

3.2 eXtensible Markup Language

We chose XML as language and meta-language because we needed to be able to save messages with their meanings. Every lexia is saved with some tags and attributes which describe its meaning. The possibility of storing separately data from their representations lets a system access more quickly to a data and extract the requested information. XML is a W3C standard and this makes our project ready to be changed and extended, as well as to be connected with other applications and services (XML, 2005). XML will be used to represent data, metadata, link typing, messages and paths map, and to exchange messages betweeen different layers.

3.3 Ruby on Rails

Ruby on Rails, or RoR, is a framework rich in extensions and libraries with licences suitable for our usage, in particular *XML Builder* and *gdiff/gpatch*. The first library offers a set of classes which allows to generate XML code in a simple way (Builder, 2006). Gdiff/gpatch library is an implementation of the *gdiff* protocol, that creates a patch from two files and then a new file from one of the previous files and the patch (Gdiff, 2005). Using this library we are going to be able to store the history and the last version in an easy way and saving space. Creating a document is therefore a sequence of patches. Storing works in the same way, that is executing a gdiff protocol and storing the new patch. Moving across the document history means retrieving a number of patch commands until you reach the desired version of the document.

Ruby on Rails does not support native XML databases at this time, therefore in our architecture there will be provisionally a relational DBMS dedicated to RoR, which had no problem with a filesystem repository.

3.4 Asyncronous Javascript And XML

AJAX is not a technology in itself but a term that refers to the use of a group of technologies together, in particular Javascript and XML. In other words AJAX is a web development technique for creating interactive web applications using a combination of XHTML and CSS, Document Object Model (or DOM), the *XMLHTTPRequest* object (Wikipedia, 2005).

AJAX paradigm has been recently defined, when someone has rediscovered a simple function originally developed by Microsoft as ActiveX control. This function, named *XMLHTTPRequest* lets clients ask servers for some particular data using asyncronous handshake. In this way users can continue using web application (typically filling web forms) while the client and the server exchange data and messages. Other developers have published a concurrent version of this function for other browsers than Internet Explorer, like Mozilla/Gecko, Opera and Safari. The web pages builded with this technology give the impression to have dynamic content. Important examples builded with AJAX paradigm are Gmail by Google, Writely, Kiko, Webnote, Meebo. Using AJAX to develop web applications and web services needs some attention. First of all Javascript must not be disabled in browsers. It is also necessary to pay attention to estimate correctly the time spent in exchanging messages between client and server so to exploit the good capabilities gained with AJAX, fully supported by and integrated in Ruby on Rails.

3.5 Access points

We are going to add to every view of Novelle a search engine that returns a list of meanings and a set of link between them. These links are represented in our project with images. Every image is itself a map that the user can surf and/or open to increase details level. When the user has found the message, he can access to it simply clicking on it. An user can comment or modify every lexia, if these actions are granted by the original author, as explained above. Users can create new links between lexias and they can describe what kind of link they intend to create through appropriate link type. These modifications are stored using the document history model of Novelle through following patch.

4 Related Works

The main source of Novelle are wikis and blogs. While wikis have spread from a detailed design (Cunningham and Leuf, 2001), unfortunately blogs have not been designed under a model. So we have tested and compared the most used tools available for blogging: Bloggers, WordPress, MovableType and LiveJournal.

Generally speaking, we find that the personal public diary metaphor behind blogs (McNeill, 2005) may bring to an unsatisfactory representation of the context. The only way to retrieve information is through a search engine or a calendar, i.e. the date of the 'post' – a lexia in the jargon of bloggers.

Moreover, we use some new web applications to take and share notes or to browser everyone's bookmarks, e.g. del.icio.us. Mostly, these web applications oriented to writing give a strong emphasis on collaboration and sharing. This led us to rethink ownership and to use the Creative Commons model to design the contents of Novelle.

Finally, we noticed that personal wikis are used for storing cognitive maps of individuals and brainstorming. This use was already thought by the founders of wikis (Cunningham and Leuf, 2001), but it has not been widely explored in practics, as far as the authors know. However, this direction of work is not actually new - concept and mind mapping, the two main paradigms for cognitive maps, have been used for several years.

Concept mapping has been used at least in education for over thirty years, in particular at the Cornell University, where Piaget's ideas gave the roots to the *assimilation theory* by David Ausubel. Very briefly, concept maps show the relationships between concepts labelling both nodes and arcs. Every arc always has a definite direction, i.e. arcs are arrows (Novak, 1998).

In contrast, mind maps spread from a centre, with branches radiating out. Furthermore, mind maps, as thought and copyrighted by Tony Buzan, can label only nodes, not arcs. The resulting shape of mind maps is sometimes similar to neurons' (Buzan, 2000).

We have tested both concept and mind mapping software tools, available for free or in a trial period. In particular, CmapTools software is currently used at the Cornell University and it is free as a client. It may run on CmapServers, and it is a very good way to share the knowledge stored in cognitive maps. Unfortunately, it does not collect data in a format suitable for the web, and it does not permit to view concepts across cognitive maps owned by different users (Tergan, 2005). More, concept maps require a learning curve very high when started to be used, at least in our experience. On the contrary, mind maps are by far more intuitive.

There are a lot of mind mapping tools, which are clones of MindJet MindManager, the official software for Buzan's mind mapping. The mind mapping tool we were looking for should have had an open source licence, used a format for data storage suitable for the web, and been cross-platform. In fact, Freemind, as the closest approximation of our needs (Mueller, 2000), succeeded in running on the three major operating systems available, without sensible differences.

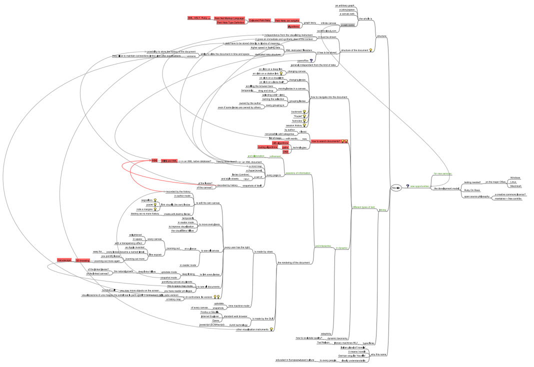


Figure 6: Our free mind map for Novelle

Even if we like the idea behind mind maps, we need to have a multiauthored environment, where arcs may be labeled. In other terms, the centre of the map should change according to the user's desire. That is why we thought about web canvas as document views. If we consider documents as free mind maps, the nodes will be lexias and the arcs will be links.

Apart from wikis, blogs, and cognitive mapping, we were also inspired by the experiences of early hypertext writing tools, in particular Intermedia and Storyspace. In fact, they were used expecially in academic writing with some success. Intermedia is no more developed and nobody of us had the opportunity to try it (Landow, 1994). Storyspace is currently distributed by Eastgate (2005), and we have used it for a time. However, in our opinion Storyspace is a product of its time and in fact it isn't a web application. Although it is possible to label links, it lacks a lot of features we need. Moreover, no hypertext writing tool available is released under an open source licence. We hope that Novelle will bridge this gap - we will choose the exact licence when our first public release is ready.

We are persuaded that there is no contradiction in collaborative mind mapping and academic writing. Maybe it is not by chance that Eastgate has also released a "personal content management assistant" (Eastgate, 2006). Our purpose is to bring back again collaborative writing and free brainstorming, as it should be.

5 Conclusions and Further Works

We are currently developing a prototype of Novelle. We argue that the model under Novelle would be an explicit representation of the context and a clear model for the contents. One of the main application of our software is natural language processing. We are going to test it expecially on digital variants of literary texts.

Acknoweledgements

We want to acknowledge Dr. Marco Benini, Dr. Alberto Trombetta for their deep insights and Gabriella Canciani for having reviewed the final draft of this paper.

References

Roland Barthes. 1970. *S/Z*. Editions du Seuil, Paris.

Tim Berners-Lee. 1999. *Weaving the Web*. Harper, San Francisco.

Jay David Bolter. 1991. *Writing Space: the Computer, Hypertext, and the History of Writing*. Erlbaum Associates, Hillsdale, N.J.

Ronald Bourret. 2006. *XML Database Products*. Url: http://www.rpbourret.com/xml/. Retrieved the 3^{rd} of January.

Builder library.for example 2006. *Project: Builder. Provide a simple way to create XML markup and data structures*. Url: http://builder.rubyforege.org/. Retrieved the 4^{th} of January.

Tony Buzan and Barry Buzan. 2000. *The Mind Map Book*. BBC Worldwide Limited, London.

Ward Cunningham and Bo Leuf. 2001. *The Wiki Way - Quick Collaboration on the Web*. Addison-Wesley, Boston.

Eastgate 2005. *Storyspace*. Url: http://www.eastgate.com/storyspace. Retrieved the 31^{st} of December.

Eastgate 2006. *Tinderbox*. Url: http://www.eastgate.com/tinderbox. Retrieved the 2^{nd} of January.

Umberto Eco. 1962. *Opera aperta*. Bompiani, Milan, Italy.

Elizabeth L. Eisenstein. 1983. *The Printing Revolution in Early Modern Europe*. Cambridge University Press, Cambridge, UK.

Jesse James Garrett. 2005. *Ajax: A New Approach to Web Applications*. Url: http://www.adaptivepath.com/publications/essays/ /archives/000385.php. Retrieved the 22^{nd} of December.

Gblog 2.0. 2005. *Gblog 2.0. Blog, reloaded*. Url: http://gblog.com/. Retrieved the 27^{th} of December.

Gdiff/Gpatch library. 2005. *Gdiff/Gpatch. An implementation of the W3C gdiff protocol*. Url: http://ruby.brian-schroeder.de/gdiff/. Retrieved the 28^{th} of December.

George P. Landow 1994. *Hypertext 2.0. The Convergence of Contemporary Critical Theory and Technology*. The Johns Hopkins University Press, Baltimore, Maryland.

Lawrence Lessig 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. Penguin Books.

Scott McCloud. 2001. *Understanding Comics*. Paradox Press, NY.

Laurie McNeill. 2005. Genre Under Construction: The Diary on the Internet. *Language@Internet*, 1.

Joerg Mueller. 2000. *FreeMind*. Url: http://freemind.sourceforge.net. Retrieved the 31^{st} of December 2005.

Theodor Holm Nelson. 1992. *Literary Machines 90.0*. Muzzio, Padua, Italy.

Joseph Donald Novak. 1998. *Learning, Creating, and Using Knowledge: Concept Maps As Facilitative Tools in Schools and Corporations*. Lawrence Erlbaum Associates.

Ozone Database Project. 2006. *Ozone Database Project open initative*. Url: http://ozone-db.org/. Retrieved the 03^{rd} of January.

Ruby on Rails. 2006. *Ruby on Rails. Web developement that doesn't hurt*. Url: http://www.rubyonrails.org/. Retrieved the 03^{rd} of January.

Richard M. Stallman. 2001. *Free Software, Free Society: Selected Essays of Richard M. Stallman*. GNU Press, Cambridge, Massachusetts.

Sigmar-Olaf Tergan and Tanja Kellers. 2005. *Knowledge And Information Visualization: Searching for Synergies*. Springer, Berlin.

Dave Thomas and David Heinemeier Hansson. 2005. *Agile Web Development with Rails - A pragmatic guide*. Pragmatic Bookshelf.

Wikipedia. 2005. *Wikipedia. From Wikipedia, the free encyclopedia*. Url: http://en.wikipedia.org/wiki/Wikipedia. Retrieved the 31^{st} of December.

XML, eXtensible Markup Language. 2005. *Extensible Markup Language (XML)*. Url: http://www.w3.org/XML/. Retrieved the 27^{th} of December.

Anomaly Detecting within Dynamic Chinese Chat Text

Yunqing Xia
Department of S.E.E.M.
The Chinese University of Hong Kong
Shatin, Hong Kong
yqxia@se.cuhk.edu.hk

Kam-Fai Wong
Department of S.E.E.M.
The Chinese University of Hong Kong
Shatin, Hong Kong
kfwong@se.cuhk.edu.hk

Abstract

The problem in processing Chinese chat text originates from the anomalous characteristics and dynamic nature of such a text genre. That is, it uses ill-edited terms and anomalous writing styles in chat text, and the anomaly is created and discarded very quickly. To handle this problem, one solution is to re-train the recognizer periodically. This costs a lot of manpower in producing the timely chat text corpus. The new approaches are proposed in this paper to detect the anomaly within dynamic Chinese chat text by incorporating standard Chinese corpora and chat corpus. We first model standard language text using standard Chinese corpora and apply these models to detect anomalous chat text. To improve detection quality, we construct anomalous chat language model using one static chat text corpus and incorporate this model into the standard language models. Our approaches calculate confidence and entropy for the input text and apply threshold values to help make the decisions. The experiments prove that performance equivalent to the best ones produced by the approaches in existence can be achieved stably with our approaches.

1 Introduction

Network Informal Language (NIL) refers to the special human language widely used in the community of network communication via platforms such as chat rooms/tools, mobile phone short message services (SMS), bulletin board systems (BBS), emails, blogs, etc. NIL is ubiquitous due in special to the rapid proliferation of Internet applications. As one important type of NIL text, chat text appears frequently within increasing volume of chat logs of online education (Heard-White, 2004) and customer relationship management (Gianforte, 2003) via chat rooms/tools. In wed-based chat rooms and BBS a large volume of NIL text is abused by (McCullagh, 2004). A survey by the Global System for Mobile Communication (GSM) showed that Germans send 200 million messages a year (German News, 2004). All the facts disclose the growing importance in processing NIL text.

Chat text holds anomalous characteristics in forming non-alphabetical characters, words, and phrases. It uses ill-edited terms and anomalous writing styles. Typical examples of anomalous Chinese chat terms can be found in (Xia et. al., 2005a). Besides the anomalous characteristics, our observations reveal remarkable dynamic nature of the chat text. The anomaly is created and discarded very quickly. Although there is no idea how tomorrow's chat text would look like, the changing will never stop. Instead, the changing gets faster and faster.

The challenging issues originates from the dynamic nature are two-fold. On the one hand, anomalous chat terms and writing styles are frequently found in chat text. Knowledge about chat text is urgently required to understand the anomaly. On the other hand, the dynamic nature of the chat text makes it nearly impossible to maintain a timely chat text knowledge base. This claim has been proved by (Xia et. al., 2005a) in which experiments are conducted with an SVM classifier. The classifier is trained on chat text created in an earlier period and tested on chat text created in a later period. In their experiments, performance of the SVM classifier becomes lower when the two periods are farther. This reveals that chat text is written in such a style that changes constantly along with time. A straightforward solution to this problem is to re-train the SVM classifier periodically with timely chat text collections. Unfortunately, this solution costs a lot of manpower in producing new chat text corpora. The supervised learning technique becomes ineffective in processing chat text.

This paper proposes approaches to detecting anomaly in dynamic Chinese chat text by incorporating standard Chinese corpora and a static chat corpus. The idea is basically error-driven. That is, we first create standard language models using trigram on standard Chinese corpora. These corpora provide negative training samples. We then construct anomalous chat language model using one static chat text corpus which provides positive training samples. We incorporate the chat language model with the standard language models and calculate confidence and entropy to help make decisions whether input text is anomalous chat text. We investigate two types of trigram, i.e. word trigram and part-of-speech (POS) tag trigram in this work.

The remaining sections of this paper are organized as follow. In Section 2, the works related to this paper are addressed. In Section 3, approaches of anomaly detection in dynamic Chinese chat text with standard Chinese corpora are presented. In Section 4, we incorporate the NIL corpus into our approaches. In section 5, experiments are described to estimate threshold values and to evaluate performance of the two approaches with various configurations. Comparisons and discussions are also reported. We conclude this paper and address future works in Section 6.

2 Related Works

Some works had been carried out in (Xia et. al., 2005a) in which an SVM classifier is implemented to recognize anomalous chat text terms. A within-domain open test is conducted on chat text posted in March 2005. The SVM classifier is trained on five training sets which contain chat text posted from December 2004 to February 2005. The experiments show that performance of the SVM classifier increases when the training period and test period are closer. This reveals that chat text is written in a style that changes quickly with time. Many anomalous popular chat terms in last year are forgotten today and new ones replace them. This makes SVM based pattern learning technique ineffective to reflect the changes.

The solution to this problem in (Xia et. al., 2005b) is to re-train the SVM classifier periodically. This costs a lot of manpower in producing the timely chat text corpora, in which each piece of anomalous chat text should be annotated with several attributes manually.

We argue that the anomalous chat text can be identified using negative training samples in static Chinese corpora. Our proposal is that we model the standard natural language using standard Chinese corpora. We incorporate a static chat text corpus to provide positive training samples to reflect fundamental characteristics of anomalous chat text. We then apply the models to detect the anomalous chat text by calculating confidence and entropy.

Regarding the approaches proposed in this paper, our arguments are, 1) the approaches can achieve performance equivalent to the best ones produced by the approaches in existence; and 2) the good performance can be achieved stably. We prove these arguments in the following sections.

3 Anomaly Detection with Standard Chinese Corpora

Chat text exhibits anomalous characteristics in using or forming words. We argue that the anomalous chat text, which is referred as anomaly in this article, can be identified with language models constructed on standard Chinese corpora with some statistical language modeling (SLM) techniques, e.g. trigram model.

The problem of anomaly detection can be addressed as follows. Given a piece of anomalous chat text, i.e. $W = \{w_1, w_2, ..., w_n\}$, and a language model $LM = \{p(x)\}$, we attempt to recognize W as anomaly by the language model. We propose two approaches to tackle this problem. We design a confidence-based approach to calculate how likely that W fits into the language model. Another approach is designed based on entropy calculation. Entropy method was originally proposed to estimate how good a language model is. In our work we apply this method to estimate how much the constructed language models are able to reflect the corpora properly based on the assumption that the corpora are sound and complete.

Although there exist numerous statistical methods to construct a natural language model, the objective of them is one: to construct a probabilistic distribution model $p(x)$ which fits to the most extent into the observed language data in the corpus. We implement the trigram model and create language models with three Chinese corpora, i.e. People’s Daily corpus, Chinese Gigaword and Chinese Pen Treebank. We investigate quality of the language models produced with these corpora.

3.1 The N-gram Language Models

N-gram model is the most widely used in statistical language modeling nowadays. Without loss of generality we express the probability of a word sequence $W = \{w_1, ..., w_n\}$ of *n* words, i.e. $p(W)$ as

$$p(W) = p(w_1, ..., w_n) = \prod_{i=1}^{n} p(w_i \mid w_0, w_1, ..., w_{i-1}) \quad (1)$$

where w_0 is chosen appropriately to handle the initial condition. The probability of the next word w_i depends on the history h_i of words that have been given so far. With this factorization the complexity of the model grows exponentially with the length of the history.

One of the most successful models of the past two decades is the trigram model (n=3) where only the most recent two words of the history are used to condition the probability of the next word.

Instead of using the actual words, one can use a set of word classes. Classes based on the POS tags, or the morphological analysis of words, or the semantic information have been tried. Also, automatically derived classes based on some statistical models of co-occurrence have been tried (Brown et. al., 1990). The class model can be generally described as

$$p(W) = \prod_{i=1}^{n} p(w_i \mid c_i) p(c_i \mid c_{i-2}, c_{i-1}) \quad (2)$$

if the classes are non-overlapping. These tri-class models have had higher perplexities than the corresponding trigram model. However, they have led to a reduction in perplexity when linearly combined with the trigram model.

3.2 The Confidence-based Approach

Given a piece of chat text $W = \{w_1, w_2, ..., w_n\}$ where each word w_i is obtained with a standard Chinese word segmentation tool, e.g. ICTCLAS. As ICTCLAS is a segmentation tool based on standard vocabulary, it means that some unknown chat terms (e.g., "介个") would be broken into several element Chinese words (i.e., "介" and "个" in the above case). This does not hurt the algorithm because we use trigram in this method. A chat term may produce some anomalous word trigrams which are evidences for anomaly detection.

We use non-zero probability for each trigram in this calculation. This is very simple but naïve. The calculation seeks to produce a so-called confidence, which reflects how much the given text fits into the training corpus in arranging its element Chinese words. This is enlightened by the observation that the chat terms use element words in anomalous manners which can not be simulated by the training corpus.

The confidence-based value is defined as

$$C(W) = \left(\prod_{i=1}^{K} C(T_i) \right)^{\frac{1}{K}} \quad (3)$$

where *K* denotes the number of trigrams in chat text *W* and T_i is the *i*-th order trigram. $C(T_i)$ is confidence of trigram T_i. Generally $C(T_i)$ is assigned probability of the trigram T_i in training corpus, i.e. $p(T_i)$. When a trigram is missing, linear interpolation is applied to estimate its probability.

We empirically setup a confidence threshold value to determine whether the input text contains chat terms, namely, it is a piece of chat text. The *input* is concluded to be *stand* text if its confidence is bigger than the confidence threshold value. Otherwise, the *input* is concluded to be *chat* text. The confidence threshold value can be estimated with a training chat text collection.

3.3 The Entropy-based Approach

The idea beneath this approach comes from entropy based language modeling. Given a language model, one can use the quantity of entropy to get an estimation of how good the language model (LM) might be. Denote by *p* the true distribution, which is unknown to us, of a segment of new text *x* of *k* words. Then the entropy on a per word basis is defined as

$$H = \lim_{k->\infty} -\frac{1}{k} \sum_x p(x) \ln p(x) \quad (4)$$

If every word in a vocabulary of size |*V*| is equally likely then the entropy would be $\log_2 |V|$; $H \le \ln |V|$ for other distributions of the words.

Enlightened by the estimation method, we compute the entropy-based value on a per trigram basis for the input chat text. Given a standard LM denoted by $\tilde{p}$ which is modeled by trigram, the entropy-value is calculate as

$$\tilde{H}_K = -\frac{1}{K}\sum_{i=1}^{K}\tilde{p}(T_i)\ln\tilde{p}(T_i) \qquad (5)$$

where K denotes number of trigrams the input text contains. Our goal is to find how much difference the input text is compared against the LM. Obviously, bigger entropy discloses a piece of more anomalous chat text. An empirical entropy threshold is again estimated on a training chat text collection. The *input* is concluded to be *stand* text if its entropy is smaller than the entropy threshold value. Otherwise, the *input* is concluded to be *chat* text.

4 Incorporating the Chat Text Corpus

We argue performance of the approaches can be improved when an initial static chat text corpus is incorporated. The chat text corpus provides some basic forms of the anomalous chat text. These forms we observe provide valuable heuristics in the trigram models. Within the chat text corpus, we only consider the word trigrams and POS tag trigrams in which anomalous chat text appears. We thus construct two trigram lists. Probabilities are produced for each trigram according to its occurrence. One chat text example EXP1 is given below.

EXP1: 介个故事听起来 8 错。
SEG1: 介 个 故事 听 起来 8 错 。

SEG1 presents the word segments produced by ICTCLAS. We generate chat text word trigrams based on SEG1 as follow.

TRIGRAM1: (1)/介 个 故事/
(2)/个 起来 8 /
(3)/起来 8 错/
(4)/ 8 错 。/

For each input trigram T_i, if it appears in the chat text corpus, we adjust the confidence and entropy values by incorporating its probability in chat text corpus.

4.1 The Refined Confidence

For each $C(T_i)$, we assign a weight ϖ_i, which is calculated as

$$\varpi_i = e^{p_n(T_i)-p_c(T_i)} \qquad (6)$$

where $p_n(T_i)$ is probability of the trigram T_i in standard corpus and $p_c(T_i)$ probability in chat text corpus. Equation (3) therefore is re-written as

$$\begin{aligned} C'(W) &= \left(\prod_{i=1}^{K}\varpi_i C(T_i)\right)^{\frac{1}{K}} \\ &= \left(\prod_{i=1}^{K} e^{p_n(T_i)-p_c(T_i)} p_n(T_i)\right)^{\frac{1}{K}} \end{aligned} \qquad (7)$$

The intention of inserting ϖ_i into confidence calculation is to decrease confidence of input chat text when chat text trigrams are found. Normally, when a trigram T_i is found in chat text trigram lists, $p_n(T_i)$ will be much lower than $p_c(T_i)$; therefore ϖ_i will be much lower than 1. By multiplying such a weight, confidence of input chat text can be decreased so that the text can be easily detected.

4.2 The Refined Entropy

Instead of assigning a weight, we introduce the entropy-based value of the input chat text on the chat text corpus, i.e. $\tilde{H}_K^c$, to produce a new equation. We denote $\tilde{H}_K^n$ the entropy calculated with equation (5). Similar to $\tilde{H}_K^n$, $\tilde{H}_K^c$ is calculated with equation (8).

$$\tilde{H}_K^c = -\frac{1}{K}\sum_{i=1}^{K}\tilde{p}_c(T_i)\ln\tilde{p}_c(T_i) \qquad (8)$$

We therefore re-write the entropy-based value calculation as follows.

$$\begin{aligned} \tilde{H}_K &= \tilde{H}_K^n + \tilde{H}_K^c \\ &= -\frac{1}{K}\sum_{i=1}^{K}\left(\tilde{p}_n(T_i)\ln\tilde{p}_n(T_i) + \tilde{p}_c(T_i)\ln\tilde{p}_c(T_i)\right) \end{aligned} \qquad (9)$$

The intention of introducing $\tilde{H}_K^c$ in entropy calculation is to increase the entropy of input chat text when chat text trigrams are found. It can be easily proved that $\tilde{H}_K$ is never smaller than H_K^n. As bigger entropy discloses a piece of more anomalous chat text, we believe more anomalous chat texts can be correctly detected with equation (9).

5 Evaluations

Three experiments are conducted in this work. The first experiment aims to estimate threshold values from a real text collection. The remaining experiments seek to evaluate performance of the approaches with various configurations.

5.1 Data Description

We use two types of text corpora to train our approaches in the experiments. The first type is standard Chinese corpus which is used to construct standard language models. We use People's Daily corpus, also know as Peking University Corpus (PKU), the Chinese Gigaword (CNGIGA) and the Chinese Penn Treebank (CNTB) in this work. Considering coverage, CNGIGA is the most excellent one. However, PKU and CPT provide more syntactic information in their annotations. Another type of training corpus is chat text corpus. We use NIL corpus described in (Xia et. al., 2005b). In NIL corpus each anomalous chat text is annotated with their attributes.

We create four test sets in our experiments. We use the test set #1 to estimate the threshold values of confidence and entropy for our approaches. The values are estimated on two types of trigrams in three corpora. Test set #1 contains 89 pieces of typical Chinese chat text selected from the NIL corpus and 49 pieces of standard Chinese sentences selected from online Chinese news by hand. There is no special consideration that we select different number of chat texts and standard sentences in this test set.

The remaining three test sets are used to compare performance of our approaches on test data created in different time periods. The test set #2 is the earliest one and #4 the latest one according to their time stamp. There are 10K sentences in total in test set #2, #3 and #4. In this collection, chat texts are selected from YESKY BBS system (http://bbs.yesky.com/bbs/) which cover BBS text in March and April 2005 (later than the chat text in the NIL corpus), and standard texts are extracted from online Chinese news randomly. We describe the four test sets in Table 1.

Test set	# of standard sentences	# of chat sentences
#1	49	89
#2	1013	2320
#3	1013	2320
#4	1014	2320

Table 1: Number of sentences in the four test sets.

5.2 Experiment I: Threshold Values Estimation

5.2.1 Experiment Description

This experiment seeks to estimate the threshold values of confidence and entropy for two types of trigrams in three Chinese corpora.

We first run the two approaches using only standard Chinese corpora on the 138 sentences in the first test set. We put the calculated values (confidence or entropy) into two arrays. Note that we already know type of each sentence in the first test set. So we are able to select in each array a value that produces the lowest error rate. In this way we obtain the first group of threshold values for our approaches.

We incorporate the NIL corpus to the two approaches and run them again. We then produce the second group of threshold values in the same way to produce the first group of values.

5.2.2 Results

The selected threshold values and corresponding error rates are presented in Table 2~5.

Trigram option	Threshold	Err rate
word of CNGIGA	1.58E-07	0.092
word of PKU	7.06E-07	0.098
word of CNTB	2.09E-06	0.085
POS tag of CNGIGA	0.0278	0.248
POS tag of PKU	0.0143	0.263
POS tag of CNTB	0.0235	0.255

Table 2: Selected threshold values of confidence for the approach using standard Chinese corpora and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	3.762E-056	0.099
word of PKU	5.683E-048	0.112
word of CNTB	2.167E-037	0.169
POS tag of CNGIGA	0.00295	0.234
POS tag of PKU	0.00150	0.253
POS tag of CNTB	0.00239	0.299

Table 3: Selected threshold values of entropy for the approach using standard Chinese corpora and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	4.26E-05	0.089
word of PKU	3.75E-05	0.102
word of CNTB	6.85E-05	0.092
POS tag of CNGIGA	0.0398	0.257
POS tag of PKU	0.0354	0.266
POS tag of CNTB	0.0451	0.249

Table 4: Selected threshold values of confidence for the approach incorporating the NIL corpus and error rates.

Trigram option	Threshold	Err rate
word of CNGIGA	8.368E-027	0.102
word of PKU	3.134E-019	0.096
word of CNTB	5.528E-021	0.172
POS tag of CNGIGA	0.00465	0.241
POS tag of PKU	0.00341	0.251
POS tag of CNTB	0.00532	0.282

Table 5: Selected thresholds values of entropy for the approach incorporating the NIL corpus and error rates.

We use the selected threshold values in experiment II and III to detect anomalous chat text within test set #2, #3 and #4.

5.3 Experiment II: Anomaly Detection with Three Standard Chinese Corpora

5.3.1 Experiment Description

In this experiment, we run the two approaches using the standard Chinese corpora on test set #2. The threshold values estimated in experiment I are applied to help make decisions.

Input text can be detected as either standard text or chat text. But we are only interested in how correctly the anomalous chat text is detected. Thus we calculate precision (p), recall (r) and F_1 measure (f) only for chat text.

$$p=\frac{a}{a+c} \quad r=\frac{a}{a+b} \quad f=\frac{2\times p\times r}{p+r} \qquad (10)$$

where *a* is the number of true positives, *b* the false negatives and *c* the false positives.

5.3.2 Results

The experiment results for the approaches using the standard Chinese corpora on test set #2 are presented in Table 6.

5.3.3 Discussions

Table 4 shows that, in most cases, the entropy-based approach outperforms the confidence-based approach slightly. It can thus be conclude that the entropy-based approach is more effective in anomaly detection.

It is also revealed that both approaches perform better with word trigrams than that with POS tag trigrams. This is natural for class based trigram model when number of class is small. Thirty-nine classes are used in ICTCLAS in POS tagging Chinese words.

When the three Chinese corpora are compared, the CNGIGA performs best in the confidence-based approach with word trigram model. However, it is not the case with POS tag trigram model. Results of two approaches on CNTB are best amongst the three corpora. Although we are able to draw the conclusion that bigger corpora yields better performance with word trigram, the same conclusion, however, does not work for POS tag trigram. This is very interesting. The reason we can address on this issue is that CNTB probably provides highest quality POS tag trigrams and other corpora contain more noisy POS tag trigrams, which eventually decreases the performance. An observation on word/POS tag lists for three Chinese corpora verifies such a claim. Text in CNTB is best-edited amongst the three.

5.4 Experiment III: Anomaly Detection with NIL Corpus Incorporated

5.4.1 Experiment Description

In this experiment, we incorporate one chat text corpus, i.e. NIL corpus, to the two approaches. We run them on test set #2, #3 and #4 with the estimated threshold values. We use precision, recall and F_1 measure again to evaluate performance of the two approaches.

5.4.2 Results

The experiment results are presented in Table 7~ Table 9 on test set #2, #3 and #4 respectively.

5.4.3 Discussions

We first compare the two approaches with different running configurations. All conclusions made in experiment II still work for experiment III. They are, i) the entropy-based approach outperforms the confidence-based approach slightly in most cases; ii) both approach perform better with word trigram than POS tag trigram; iii) both approaches perform best on CNGIGA with word trigram model. But with POS tag trigram model, CNTB produces the best results.

An interesting comparison is conducted on F_1 measure between the approaches in experiment II and experiment III on test set #2 in Figure 1 (the left two columns). Generally, F_1 measure of anomaly detection with both approaches with word trigram model is improved when the NIL corpus is incorporated. It is revealed in Table 7~9 that same observation is found with POS tag trigram model.

We compare F_1 measure of the approaches with word trigram model in experiment III on test set #2, #3 and #4 in Figure 1 (the right three columns). The graph in Figure 1 shows that F_1 measure on three test sets are very close to each other. This is also true the approaches with POS tag trigram model as showed in Table 7~9. This provides evidences for the argument that the approaches can produce stable performance with the NIL corpus. Differently, as reported in (Xia et. al., 2005a), performance achieved in SVM classifier is rather unstable. It performs poorly with training set C#1 which contains BBS text posted several months ago, but much better with training set C#5 which contains the latest chat text.

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.685	0.737	0.710	0.722	0.761	0.741	0.614	0.654	0.633	0.637	0.664	0.650
PKU	0.699	0.712	0.705	0.701	0.738	0.719	0.619	0.630	0.624	0.625	0.648	0.636
CNTB	0.653	0.661	0.657	0.692	0.703	0.697	0.651	0.673	0.662	0.684	0.679	0.681

Table 6: Results of anomaly detection using standard Chinese corpora on test set #2.

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.821	0.836	0.828	0.857	0.849	0.853	0.653	0.657	0.655	0.672	0.678	0.675
PKU	0.818	0.821	0.819	0.838	0.839	0.838	0.672	0.672	0.672	0.688	0.679	0.683
CNTB	0.791	0.787	0.789	0.821	0.811	0.816	0.691	0.679	0.685	0.712	0.688	0.700

Table 7: Results of anomaly detection incorporating NIL corpus on test set #2

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.819	0.841	0.830	0.849	0.848	0.848	0.657	0.659	0.658	0.671	0.677	0.674
PKU	0.812	0.822	0.817	0.835	0.835	0.835	0.663	0.671	0.667	0.687	0.681	0.684
CNTB	0.801	0.783	0.792	0.822	0.803	0.812	0.689	0.677	0.683	0.717	0.689	0.703

Table 8: Results of anomaly detection incorporating NIL corpus on test set #3

Corpus	Word trigram						POS tag trigram					
	confidence			entropy			confidence			entropy		
	p	r	f	p	r	f	p	r	f	p	r	f
CNGIGA	0.824	0.839	0.831	0.852	0.845	0.848	0.651	0.654	0.652	0.674	0.674	0.674
PKU	0.815	0.825	0.820	0.836	0.84	0.838	0.668	0.668	0.668	0.692	0.682	0.687
CNTB	0.796	0.785	0.790	0.817	0.807	0.812	0.694	0.681	0.687	0.713	0.686	0.699

Table 9: Results of anomaly detection incorporating NIL corpus on test set #4

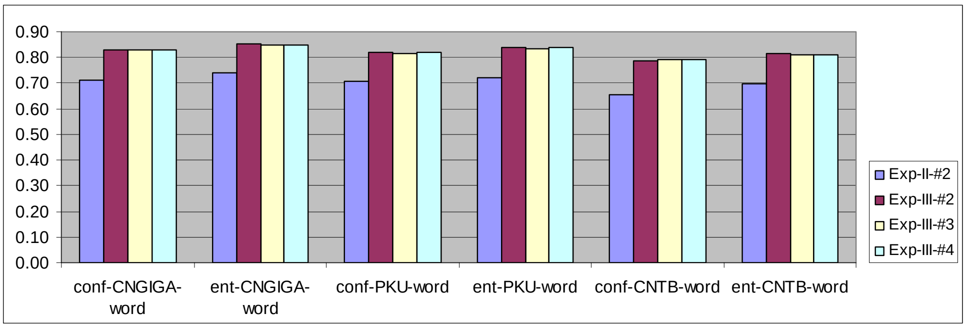


Figure 1: Comparisons on F_1 measure of the approaches with word trigram on test set #2, #3 and #4 in experiment II and experiment III.

We finally compare performance of our approaches against the one described in (Xia, et. al., 2005a). The best F_1 measure achieved in our work, i.e. 0. 853, is close to the best one in their work, i.e. 0.871 with training corpus C#5. This proves another argument that our approaches can produce equivalent performance to the best ones achieved by the approaches in existence.

6 Conclusions

The new approaches to detecting anomalous Chinese chat text are proposed in this paper. The approaches calculate confidence and entropy values with the language models constructed on negative training samples in three standard Chinese corpora. To improve detection quality, we incorporate positive training samples in NIL corpus in our approaches. Two conclusions can be made based on this work. Firstly, F_1 measure of anomaly detection can be improved by around 0.10 when NIL corpus is incorporated into the approaches. Secondly, performance equivalent to the best ones produced by the approaches in existence can be achieved stably by incorporating the standard Chinese corpora and the NIL corpus.

We believe some strong evidences for our claims can be obtained by training our approaches with more chat text corpora which contain chat text created in different time periods. We are conducting this experiment seeks to find out whether and how our approaches are independent of time. This work is still progressing. A report on this issue will be available shortly. We also plan to investigate how size of chat text corpus influences performance of our approaches. The goal is to find the optimal size of chat text corpus which can achieve the best performance. The readers should also be noted that evaluation in this work is a within-domain test. Due to shortage of chat text resources, no cross-domain test is conducted. In the future cross-domain test, we will investigate how our approaches are independent of domain.

Eventual goal of chat text processing is to normalize the anomalous chat text, namely, convert it to standard text holding the same meaning. So the work carried out in this paper is the first step leading to this goal. Approaches will be designed to locate the anomalous terms in chat text and map them to standard words.

Acknowledgement

Research described in this paper is partially supported by the Chinese University of Hong Kong under the Direct Grant Scheme (No: 2050330) and Strategic Grant Scheme project (No: 4410001) respectively.

Reference

Brown, P. F., V. J. Della Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. 1990. Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, Paris, France.

Finkelhor, D., K. J. Mitchell, and J. Wolak. 2000. Online Victimization: A Report on the Nation's Youth. Alexandria, Virginia: National Center for Missing & Ex-ploited Children, page ix.

German News. 2004. Germans are world SMS champions, 8 April 2004, http://www.expatica.com/source/site_article.asp?subchannel_id=52&story_id=6469.

Gianforte, G.. 2003. From Call Center to Contact Center: How to Successfully Blend Phone, Email, Web and Chat to Deliver Great Service and Slash Costs. RightNow Technologies.

Heard-White, M., Gunter Saunders and Anita Pincas. 2004. Report into the use of CHAT in education. Final report for project of Effective use of CHAT in Online Learning, Institute of Education, University of London.

McCullagh, D.. 2004. Security officials to spy on chat rooms. News provided by CNET Networks. November 24, 2004.

Xia, Y., K.-F. Wong and W. Gao. 2005a. NIL is not Nothing: Recognition of Chinese Network Informal Language Expressions, 4th SIGHAN Workshop on Chinese Language Processing at IJCNLP'05, pp95-102.

Xia, Y., K.-F. Wong and R. Luk. 2005b. A Two-Stage Incremental Annotation Approach to Constructing A Network Informal Language Corpus. In Proc. of NTCIR-5 Meeting, pp. 529-536.

Zhang, Z., H. Yu, D. Xiong and Q. Liu. 2003. HMM-based Chinese Lexical Analyzer ICTCLAS. SIGHAN’03 within ACL’03, pp. 184-187.

A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia

Antonio Toral
University of Alicante
Carretera San Vicente S/N
Alicante 03690, Spain
atoral@dlsi.ua.es

Rafael Muñoz
University of Alicante
Carretera San Vicente S/N
Alicante 03690, Spain
rafael@dlsi.ua.es

Abstract

This paper describes a method to automatically create and maintain gazetteers for Named Entity Recognition (NER). This method extracts the necessary information from linguistic resources. Our approach is based on the analysis of an on-line encyclopedia entries by using a noun hierarchy and optionally a PoS tagger. An important motivation is to reach a high level of language independence. This restricts the techniques that can be used but makes the method useful for languages with few resources. The evaluation carried out proves that this approach can be successfully used to build NER gazetteers for location (F 78%) and person (F 68%) categories.

1 Introduction

Named Entity Recognition (NER) was defined at the MUC conferences (Chinchor, 1998) as the task consisting of detecting and classifying strings of text which are considered to belong to different classes (e.g. person, location, organization, date, time). Named Entities are theoretically identified and classified by using evidence. Two kinds of evidence have been defined (McDonald, 1996). These are internal and external evidence. Internal evidence is the one provided from within the sequence of words that constitute the entity. In contrast, external evidence is the criteria that can be obtained by the context in which entities appear.

Since the time NER was introduced, mainly two approaches have been adopted to deal with this task. One is referred as knowledge-based and uses explicit resources like rules and gazetteers, which commonly are hand-crafted. The other follows the learning paradigm and usually uses as a resource a tagged corpus which is used to train a supervised learning algorithm.

In the knowledge-based approach two kind of gazetteers can be distinguished. On one hand there are trigger gazetteers, which contain key words that indicate the possible presence of an entity of a given type. These words usually are common nouns. E.g. ms. indicates that the entity after it is a person entity. On the other hand there are entity gazetteers which contain entities themselves, which usually are proper nouns. E.g. Portugal could be an instance in a location gazetteer.

Initially, and specially for the MUC conferences, most of the NER systems developed did belong to the knowledge-based approach. This approach proved to be able to obtain high scores. In fact, the highest score obtained by a knowledge-based system in MUC-7 reached F 93.39 % (Mikheev et al., 1998). However, this approach has an important problem: gazetteers and rules are difficult and tedious to develop and to maintain. If the system is to be used for an open domain, linguistic experts are needed to build the rules, and besides, it takes too much time to tune these resources in order to obtain satisfactory results. Because of this, lately most of the research falls into the learning-based paradigm.

Regarding the creation and maintenance of gazetteers, several problems have been identified, these are mainly:

- Creation and maintenance effort
- Overlaps between gazetteers

The first problem identified assumes that the gazetteers are manually created and maintained. However, this is not always the case. Gazetteers could be automatically created and maintained by extracting the necessary information from available linguistic resources, which we think is a promising line of future research.

Several research works have been carried out in this direction. An example of this is a NER system which uses trigger gazetteers automatically extracted from WordNet (Magnini et al., 2002) by using wordnet predicates. The advantage in this case is that the resource used is multilingual and thus, porting it to another language is almost straightforward (Negri and Magnini, 2004).

There is also a work that deals with automatically building location gazetteers from internet texts by applying text mining procedures (Ourioupina, 2002), (Uryupina, 2003). However, this work uses linguistic patterns, and thus is language dependent. The author claims that the approach may successfully be used to create gazetteers for NER.

We agree with (Magnini et al., 2002) that in order to automatically create and maintain trigger gazetteers, using a hierarchy of common nouns is a good approach. Therefore, we want to focus on the automatically creation and maintenance of entity gazetteers. Another reason for this is that the class of common nouns (the ones being triggers) is much more stable than the class of proper names (the ones in entity gazetteers). Because of this, the maintenance of the latter is important as new entities to be taken into account appear. For example, if we refer to presidents, the trigger word used might be 'president' and it is uncommon that the trigger used to refer to them changes over time. On the other hand, the entities being presidents change as new presidents appear and current presidents will disappear.

Our aim is to find a method which allow us to automatically create and maintain entity gazetteers by extracting the necessary information from linguistic resources. An important restriction though, is that we want our method to be as independent of language as possible.

The rest of this paper is structured as follows. In the next section we discuss about our proposal. Section three presents the results we have obtained and some comments about them. Finally, in section four we outline our conclusions and future work.

2 Approach

In this section we present our approach to automatically build and maintain dictionaries of proper nouns. In a nutshell, we analyse the entries of an encyclopedia with the aid of a noun hierarchy. Our motivation is that proper nouns that form entities can be obtained from the entries in an encyclopedia and that some features of their definitions in the encyclopedia can help to classify them into their correct entity category.

The encyclopedia used has been Wikipedia[1]. According to the English version of Wikipedia [2], Wikipedia is a multi-lingual web-based, free-content encyclopedia which is updated continuously in a collaborative way. The reasons why we have chosen this encyclopedia are the following:

- It is a big source of information. By December 2005, it has over 2,500,000 definitions. The English version alone has more than 850,000 entries.
- Its content has a free license, meaning that it will always be available for research without restrictions and without needing to acquire any license.
- It is a general knowledge resource. Thus, it can be used to extract information for open domain systems.
- Its data has some degree of formality and structure (e.g. categories) which helps to process it.
- It is a multilingual resource. Thus, if we are able to develop a language independent system, it can be used to create gazetteers for any language for which Wikipedia is available.
- It is continuously updated. This is a very important fact for the maintenance of the gazetteers.

The noun hierarchy used has been the noun hierarchy from WordNet (Miller, 1995). This is a widely used resource for NLP tasks. Although initially being a monolingual resource for the English language, a later project called EuroWordNet (Vossen, 1998), provided wordnet-like hierarchies

[1] http://www.wikipedia.org

[2] http://en.wikipedia.org/wiki/Main_Page

for a set of languages of the European Union. Besides, EuroWordNet defines a language independent index called Inter-Lingual-Index (ILI) which allows to establish relations between words in wordnets of different languages. The ILI facilitates also the development of wordnets for other languages.

From this noun hierarchy we consider the nodes (called synsets in WordNet) which in our opinion represent more accurately the different kind of entities we are working with (location, organization and person). For example, we consider the synset 6026 as the corresponding to the entity class Person. This is the information contained in synset number 6026:

```
person, individual, someone,
somebody, mortal,
human, soul -- (a human being;
"there was too much for one person
to do")
```

Given an entry from Wikipedia, a PoS-tagger (Carreras et al., 2004) is applied to the first sentence of its definition. As an example, the first sentence of the entry Portugal in the Simple English Wikipedia [3] is presented here:

```
Portugal portugal NN
is be VBZ
a a DT
country country NN
in in IN
the the DT
south-west south-west NN
of of IN
Europe Europe NP
. . Fp
```

For every noun in a definition we obtain the synset of WordNet that contains its first sense[4]. We follow the hyperonymy branch of this synset until we arrive to a synset we have considered belonging to an entity class or we arrive to the root of the hierarchy. If we arrive to a considered synset, then we consider that noun as belonging to the entity class of the considered synset. The following example may clarify this explanation:

```
portugal --> LOCATION
country --> LOCATION
south-west --> NONE
europe --> LOCATION
```

As it has been said in the abstract, the application of a PoS tagger is optional. The algorithm will perform considerably faster with it as with the PoS data we only need to process the nouns. If a PoS tagger is not available for a language, the algorithm can still be applied. The only drawback is that it will perform slower as it needs to process all the words. However, through our experimentation we can conclude that the results do not significantly change.

Finally, we apply a weighting algorithm which takes into account the amount of nouns in the definition identified as belonging to the different entity types considered and decides to which entity type the entry belongs. This algorithm has a constant Kappa which allows to increase or decrease the distance required within categories in order to assign an entry to a given class. The value of Kappa is the minimum difference of number of occurrences between the first and second most frequent categories in an entry in order to assign the entry to the first category. In our example, for any value of Kappa lower than 4, the algorithm would say that the entry Portugal belongs to the location entity type.

Once we have this basic approach we apply different heuristics which we think may improve the results obtained and which effect will be analysed in the section about results.

The first heuristic, called is_instance, tries to determine whether the entries from Wikipedia are instances (e.g. Portugal) or word classes (e.g. country). This is done because of the fact that named entities only consider instances. Therefore, we are not interested in word classes. We consider that an entry from Wikipedia is an instance when it has an associated entry in WordNet and it is an instance. The procedure to determine if an entry from WordNet is an instance or a word class is similar to the one used in (Magnini et al., 2002).

The second heuristic is called is_in_wordnet. It simply determines if the entries from Wikipedia have an associated entry in WordNet. If so, we may use the information from WordNet to determine its category.

[3] http://simple.wikipedia.org/wiki/Portugal

[4] We have also carried out experiments taking into account all the senses provided by WordNet. However, the performance obtained is not substantially better while the processing time increases notably.

3 Experiments and results

We have tested our approach by applying it to 3517 entries of the Simple English Wikipedia which were randomly selected. Thus, these entries have been manually tagged with the expected entity category[5]. The distribution by entity classes can be seen in table 1:

As it can be seen in table 1, the amount of entities of the categories Person and Location are balanced but this is not the case for the type Organization. There are very few instances of this type. This is understandable as in an encyclopedia locations and people are defined but this is not the usual case for organizations.

According to what was said in section 2, we considered the heuristics explained there by carrying out two experiments. In the first one we applied the is_instance heuristic. The second experiment considers the two heuristics explained in section 2 (is_instance and is_in_wordnet). We do not present results without the first heuristic as through our experimentation it proved to increase both recall and precision for every entity category.

For each experiment we considered two values of a constant Kappa which is used in our algorithm. The values are 0 and 2 as through experimentation we found these are the values which provide the highest recall and the highest precision, respectively. Results for the first experiment can be seen in table 2 and results for the second experiment in table 3.

As it can be seen in these tables, the best recall for all classes is obtained in experiment 2 with Kappa 0 (table 3) while the best precision is obtained in experiment 1 with Kappa 2 (table 2).

The results both for location and person categories are in our opinion good enough to the purpose of building and maintaining good quality gazetteers after a manual supervision. However, the results obtained for the organization class are very low. This is mainly due to the fact of the high interaction between this category and location combined with the practically absence of traditional entities of the organization type such as companies. This interaction can be seen in the in-depth results which presentation follows.

In order to clarify these results, we present more in-depth data in tables 4 and 5. These tables present an error analysis, showing the false positives, false negatives, true positives and true negatives among all the categories for the configuration that provides the highest recall (experiment 2 with Kappa 0) and for the one that provides the highest precision (experiment 1 with Kappa 2).

In tables 4 and 5 we can see that the interactions within classes (occurrences tagged as belonging to one class but NONE and guessed as belonging to other different class but NONE) is low. The only case in which it is significant is between location and organization. In table 5 we can see that 12 entities tagged as organization are classified as LOC while 20 tagged as organization are guessed with the correct type. Following with these, 5 entities tagged as location where classified as organization. This is due to the fact that countries and related entities such as ”European Union” can be considered both as organizations or locations depending on their role in a text.

4 Conclusions

We have presented a method to automatically create and maintain entity gazetteers using as resources an encyclopedia, a noun hierarchy and, optionally, a PoS tagger. The method proves to be helpful for these tasks as it facilitates the creation and maintenance of this kind of resources.

In our opinion, the principal drawback of our system is that it has a low precision for the configuration for which it obtains an acceptable value of recall. Therefore, the automatically created gazetteers need to pass a step of manual supervision in order to have a good quality.

On the positive side, we can conclude that our method is helpful as it takes less time to automatically create gazetteers with our method and after that to supervise them than to create that dictionaries from scratch. Moreover, the updating of the gazetteers is straightforward; just by executing the procedure, the new entries in Wikipedia (the entries that did not exist at the time the procedure was performed the last time) would be analysed and from these set, the ones detected as entities would be added to the corresponding gazetteers.

Another important fact is that the method has a high degree of language independence; in order to apply this approach to a new language, we need a version of Wikipedia and WordNet for that language, but the algorithm and the process does not change. Therefore, we think that our method can be useful for the creation of gazetteers for lan-

[5] This data is available for research at `http://www.dlsi.ua.es/˜atoral/index.html\#resources`

Entity type	**Number of instances**	**Percentage**
NONE	2822	
LOC	404	58
ORG	55	8
PER	236	34

Table 1: Distribution by entity classes

k	**LOC**			**ORG**			**PER**		
	prec	**rec**	**F**$_{\beta=1}$	**prec**	**rec**	**F**$_{\beta=1}$	**prec**	**rec**	**F**$_{\beta=1}$
0	66.90	94.55	78.35	28.57	18.18	22.22	61.07	77.11	68.16
2	86.74	56.68	68.56	66.66	3.63	6.89	86.74	30.50	45.14

Table 2: Experiment 1. Results applying is_instance heuristic

k	**LOC**			**ORG**			**PER**		
	prec	**rec**	**F**$_{\beta=1}$	**prec**	**rec**	**F**$_{\beta=1}$	**prec**	**rec**	**F**$_{\beta=1}$
0	62.88	96.03	76.00	16.17	20.00	17.88	43.19	84.74	57.22
2	77.68	89.60	83.21	13.95	10.90	12.24	46.10	62.71	53.14

Table 3: Experiment 2. Results applying is_instance and is_in_wordnet heuristics

Tagged	**Guessed**			
	NONE	**LOC**	**ORG**	**PER**
NONE	2777	33	1	11
LOC	175	229	0	0
ORG	52	1	2	0
PER	163	1	0	72

Table 4: Results fn-fp (results 1 k=2)

Tagged	**Guessed**			
	NONE	**LOC**	**ORG**	**PER**
NONE	2220	196	163	243
LOC	8	387	5	4
ORG	20	12	20	3
PER	30	9	2	195

Table 5: Results fn-fp (results 2 k=0)

guages in which NER gazetteers are not available but have Wikipedia and WordNet resources.

During the development of this research, several future works possibilities have appeared. Regarding the task we have developed, we consider to carry out new experiments incorporating features that Wikipedia provides such as links between pairs of entries. Following with this, we consider to test more complex weighting techniques for our algorithm.

Besides, we think that the resulting gazetteers for the configurations that provide high precision and low recall, although not being appropriate for building gazetteers for NER systems, can be interesting for other tasks. As an example, we consider to use them to extract verb frequencies for the entity categories considered which can be later used as features for a learning based Named Entity Recogniser.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV04B-268.

We also would like to specially thank Borja Navarro for his valuable help on WordNet.

References

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC Conference*.

N. Chinchor. 1998. Overview of MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

B. Magnini, M. Negri, R. Preete, and H. Tanev. 2002. A wordnet-based approach to named entities recognition. In *Proceedings of SemaNet '02: Building and Using Semantic Networks*, pages 38–44.

D. McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus Processing for Lexical Aquisition*, pages 21–39, chapter 2.

A. Mikheev, C. Grover, and M. Moens. 1998. Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, Virginia, 29 April-1 May*.

G. A. Miller. 1995. Wordnet: A lexical database for english. *Communications of ACM*, (11):39–41.

M. Negri and B. Magnini. 2004. Using wordnet predicates for multilingual named entity recognition. In *Proceedings of The Second Global Wordnet Conference*, pages 169–174.

O. Ourioupina. 2002. Extracting geographical knowledge from the internet. In *Proceedings of the ICDM-AM International Workshop on Active Mining*.

O. Uryupina. 2003. Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 18–25.

P. Vossen. 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32:73–89.

Finding Similar Sentences across Multiple Languages in Wikipedia

Sisay Fissaha Adafre **Maarten de Rijke**
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
sfissaha,mdr@science.uva.nl

Abstract

We investigate whether the Wikipedia corpus is amenable to multilingual analysis that aims at generating parallel corpora. We present the results of the application of two simple heuristics for the identification of similar text across multiple languages in Wikipedia. Despite the simplicity of the methods, evaluation carried out on a sample of Wikipedia pages shows encouraging results.

1 Introduction

Parallel corpora form the basis of much multilingual research in natural language processing, ranging from developing multilingual lexicons to statistical machine translation systems. As a consequence, collecting and aligning text corpora written in different languages constitutes an important prerequisite for these research activities.

Wikipedia is a multilingual free online encyclopedia. Currently, it has entries for more than 200 languages, the English Wikipedia being the largest one with 895,674 articles, and no fewer than eight language versions having upwards of 100,000 articles as of January 2006. As can be seen in Figure 1, Wikipedia pages for major European languages have reached a level where they can support multilingual research. Despite these developments in its content, research on Wikipedia has largely focused on monolingual aspects so far; see e.g., (Voss, 2005) for an overview.

In this paper, we focus on multilingual aspects of Wikipedia. Particularly, we investigate to what extent we can use properties of Wikipedia itself to generate similar sentences acrose different languages. As usual, we consider two sentences similar if they contain (some or a large amount of) overlapping information. This includes cases in which sentences may be exact translations of each other, one sentence may be contained within another, or both share some bits of information.

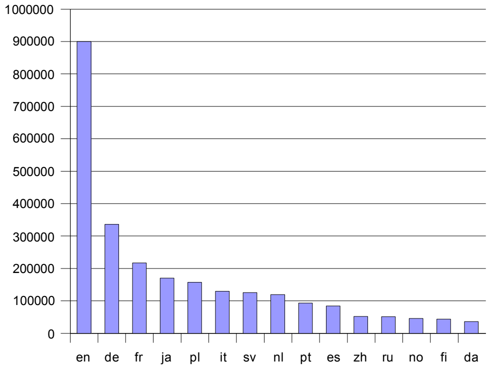


Figure 1: Wikipedia pages for the top 15 languages

The conceptually simple but fundamental task of identifying similar sentences across multiple languages has a number of motivations. For a start, and as mentioned earlier, sentence aligned corpora play an important role in corpus based language processing methods in general. Second, in the context of Wikipedia, being able to align similar sentences across multiple languages provides insight into Wikipedia as a knowledge source: to which extent does a given topic get different kinds of attention in different languages? And thirdly, the ability to find similar content in other languages while creating a page for a topic in one language constitutes a useful type of editing support. Furthermore, finding similar content acrose different languages can form the basis for multilingual summarization and question answering support for

Wikipedia; at present the latter task is being developed into a pilot for CLEF 2006 (WiQA, 2006).

There are different approaches for finding similar sentences across multiple languages in non-parallel but comparable corpora. Most methods for finding similar sentences assume the availability of a clean parallel corpus. In Wikipedia, two versions of a Wikipedia topic in two different languages are a good starting point for searching similar sentences. However, these pages may not always conform to the typical definitions of a bitext which current techniques assume. Bitext generally refers to two versions of a text in two different languages (Melamed, 1996). Though it is not known how information is shared among the different languages in Wikipedia, some pages tend to be translations of each other whereas the majority of the pages tend to be written independently of each other. Therefore, two versions of the same topic in two different languages can not simply be taken as parallel corpora. This in turn limits the application of some of the currently available techniques.

In this paper, we present two approaches for finding similar sentences across multiple languages in Wikipedia. The first approach uses freely available online machine translation resources for translating pages and then carries out monolingual sentence similarity. The approach needs a translation system, and these are not available for every pair of languages in Wikipedia.

This motivates a second approach to finding similar sentences across multiple languages, one which uses a bilingual title translation lexicon induced automatically using the link structure of Wikipedia. Briefly, two sentences are similar if they link to the same entities (or rather: to pages about the same entities), and we use Wikipedia itself to relate pages about a given entity across multiple languages. In Wikipedia, pages on the same topic in different languages are topically closely related. This means that even if one page is not a translation of another, they tend to share some common information. Our underlying assumption here is that there is a general agreement on the kind of information that needs to be included in the pages of different types of topics such as a biography of a person, and the definition and description of a concept etc., and that this agreement is to a consderable extent "materialized" in the hypertext links (and their anchor texts) in Wikipedia.

Our main research question in this paper is this: how do the two methods just outlined differ? A priori it seems that the translation based approach to finding similar sentences across multiple languages will have a higher recall than the link-based method, while the latter outperforms the former in terms of precision. Is this correct?

The remainder of the paper is organized as follows. In Section 2, we briefly discuss related work. Section 3 provides a detailed description of Wikipedia as a corpus. The two approaches to identifying similar sentences across multiple languages are presented in Section 4. An experimental evaluation is presented in Section 5. We conclude in Section 6.

2 Related Work

The main focus of this paper lies with multilingual text similarity and its application to information access in the context of Wikipedia. Current research work related to Wikipedia mostly describes its monolingual properties (Ciffolilli, 2003; Viégas et al., 2004; Lih, 2004; Miller, 2005; Bellomi and Bonato, 2005; Voss, 2005; Fissaha Adafre and de Rijke, 2005). This is probably due to the fact that different language versions of Wikipedia have different growth rates. Others describe its application in question answering and other types of IR systems (Ahn et al., 2005). We believe that currently, Wikipedia pages for major European languages have reached a level where they can support multilingual research.

On the other hand, there is a rich body of knowledge relating to multilingual text similarity. These include example-based machine translation, cross-lingual information retrieval, statistical machine translation, sentence alignment cost functions, and bilingual phrase translation (Kirk Evans, 2005). Each approach uses relatively different features (content and structural features) in identifying similar text from bilingual corpora. Furthermore, most methods assume that the bilingual corpora can be sentence aligned. This assumption does not hold for our case since our corpus is not parallel. In this paper, we use content based features for identifying similar text across multilingual corpora. Particularly, we compare bilingual lexicon and MT system based methods for identifying similar text in Wikipedia.

3 Wikipedia as a Multilingual Corpus

Wikipedia is a free online encyclopedia which is administered by the non-profit Wikimedia Foundation. The aim of the project is to develop free encyclopedias for different languages. It is a collaborative effort of a community of volunteers, and its content can be edited by anyone. It is attracting increasing attention amongst web users and has joined the top 50 most popular sites.

As of January 1, 2006, there are versions of Wikipedia in more than 200 languages, with sizes ranging from 1 to over 800,000 articles. We used the ascii text version of the English and Dutch Wikipedia, which are available as database dumps. Each entry of the encyclopedia (a page in the online version) corresponds to a single line in the text file. Each line consists of an ID (usually the name of the entity) followed by its description. The description part contains the body of the text that describes the entity. It contains a mixture of plain text and text with html tags. References to other Wikipedia pages in the text are marked using "[[" "]]" which corresponds to a hyperlink on the online version of Wikipedia. Most of the formatting information which is not relevant for the current task has been removed.

3.1 Links within a single language

Wikipedia is a hypertext document with a rich link structure. A description of an entity usually contains hypertext links to other pages within or outside Wikipedia. The majority of these links correspond to entities, which are related to the entity being described, and have a separate entry in Wikipedia. These links are used to guide the reader to a more detailed description of the concept denoted by the anchor text. In other words, the links in Wikipedia typically indicate a topical association between the pages, or rather the entities being described by the pages. E.g., in describing a particular person, reference will be made to such entities as country, organization and other important entities which are related to it and which themselves have entries in Wikipedia. In general, due to the peculiar characteristics of an encyclopedia corpus, the hyperlinks found in encyclopedia text are used to exemplify those instances of hyperlinks that exist among topically related entities (Ghani et al., 2001; Rao and Turoff, 1990).

Each Wikipedia page is identified with a unique ID. These IDs are formed by concatenating the words of the titles of the Wikipedia pages which are unique for each page, e.g., the page on Vincent van Gogh has "Vincent van Gogh" as its title and "Vincent_van_Gogh" as its ID. Each page may, however, be represented by different anchor texts in a hyperlink. The anchor texts may be simple morphological variants of the title such as plural form or may represent closely related semantic concept. For example, the anchor text "Dutch" may point to the page for the Netherlands. In a sense, the IDs function as the canonical form for several related concepts.

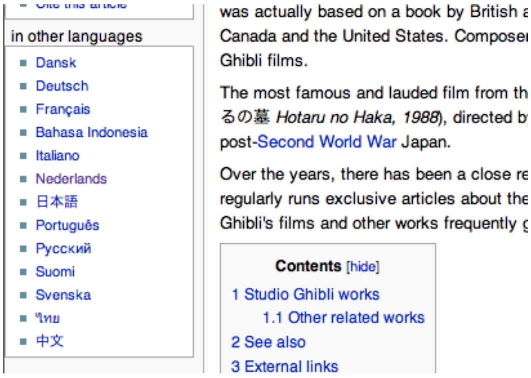
3.2 Links across different languages

Different versions of a page in different languages are also hyperlinked. For a given page, translations of its title in other languages for which pages exist are given as hyperlinks. This property is particularly useful for the current task as it helps us to align the corpus at the page level. Furthermore, it also allows us to induce bilingual lexicon consisting of the Wikipedia titles. Conceptual mismatch between the pages (e.g. *Roof* vs *Dakconstructie*) is rare, and the lexicon is generally of high quality. Unlike the general lexicon, this lexicon contains a relatively large number of names of individuals and other entities which are highly informative and hence are useful in identifying similar text. This lexicon will form the backbone of one of the methods for identifying similar text across different languages, as will be shown in Section 4.

4 Approaches

We describe two approaches for identifying similar sentences across different languages. The first uses an MT system to obtain a rough translation of a given page in one language into another and then uses word overlap between sentences as a similarity measure. One advantage of this method is that it relies on a large lexical resource which is bigger than what can be extracted from Wikipedia. However, the translation can be less accurate especially for the Wikipedia titles which form part of the content of a page and are very informative.

The second approach relies on a bilingual lexicon which is generated from Wikipedia using the link structure: pages on the same topic in different languages are hyperlinked; see Figure 2. We use the titles of the pages that are linked in this manner to create a bilingual lexicon. Thus, our bilingual lexicon consists of terms that represent concepts or entities that have entries in Wikipedia, and we will represent sentences by entries from this lexicon: an entry is used to represent the content of a sentence if the sentence contains a hypertext link to the Wikipedia page for that entry. Sentence similarity is then captured in terms of the shared lexicon entries they share. In other words, the similarity measure that we use in this approach is based on "concept" or "page title" overlap. Intuitively, this approach has the advantage of producing a brief but highly accurate representation of sentences, more accurate, we assume than the MT approach as the titles carry important semantic information; it will also be more accurate than the MT approach because the translations of the titles are done manually.



was actually based on a book by British
Canada and the United States. Compose
Ghibli films.

The most famous and lauded film from th
るの墓 *Hotaru no Haka, 1988*), directed b
post-Second World War Japan.

Over the years, there has been a close re
regularly runs exclusive articles about the
Ghibli's films and other works frequently

Figure 2: Links to pages devoted to the same topic in other languages.

Both approaches assume that the Wikipedia corpus is aligned at the page level. This is easily achieved using the link structure since, again, pages on the same topic in different languages are hyperlinked. This, in turns, narrows down the search for similar text to a page level. Hence, for a given text of a page (sentence or chunk) in one language, we search for its equivalent text (sentence or chunk) only in the corresponding page in the other language, not in the entire corpus.

We now describe the two approaches in more detail. To remain focused and avoid getting lost in technical details, we consider only two languages in our technical descriptions and evaluations below: Dutch and English; it will be clear from our presentation, however, that our second approach can be used for any pair of languages in Wikipedia.

4.1 An MT based approach

In this approach, we translate the Dutch Wikipedia page into English using an online MT system. We refer to the English page as *source* and the translated (Dutch page) version as *target*. We used the Babelfish MT system of Altavista. It supports a number of language pairs among which are Dutch-English pairs. After both pages have been made available in English, we split the pages into sentences or text chucks. We then link each text chunk or sentence in the *source* to each chuck or sentence in the *target*. Following this we compute a simple word overlap score for each pair. We used the Jaccard similarity measure for this purpose. Content words are our main features for the computation of similarity, hence, we remove stopwords. Grammatically correct translations may not be necessary since we are using simple word overlap as our similarity measure.

The above procedure will generate a large set of pairs, not all of which will actually be similar. Therefore, we filter the list assuming a one-to-one correspondence, where for each source sentence we identify at most one target sentence. This is a rather strict criterion (another possibility being one-to-many), given the fact that the corpus is generally assumed to be not parallel. But it gives some idea on how much of the text corpus can be aligned at smaller units (i.e., sentence or text chunks).

Filtering works as follows. First we sort the pairs in decreasing order of their similarity scores. This results in a ranked list of text pairs in which the most similar pairs are ranked top whereas the least similar pairs are ranked bottom. Next we take the top most ranking pair. Since we are assuming a one-to-one correspondence, we remove all other pairs ranked lower in the list containing either of the the sentences or text chunks in the top ranking pair. We then repeat this process taking the second top ranking pair. Each step results in a smaller list. The process continues until there is no more pair to remove.

4.2 Using a link-based bilingual lexicon

As mentioned previously, this approach makes use of a bilingual lexicon that is generated from Wikipedia using the link structure. A high level description of the algorithm is given in Figure 3. Below, we first describe how the bilingual lexicon is acquired and how it is used for enriching the link structure of Wikipedia. Finally, we detail how the

- Generating bilingual lexicon
- Given a topic, get the corresponding pages from English and Dutch Wikipedia
- Split pages into sentences and enrich the hyperlinks in the sentence or identify named-entities in the pages.
- Represent the sentences in these pages using the bilingual lexicon.
- Compute term overlap between the sentences thus represented.

Figure 3: The Pseudo-algorithm for identifying similar sentences using a link-based bilingual lexicon.

bilingual lexicon is used for the identification of similar sentences.

Generating the bilingual lexicon

Unlike the MT based approach, which uses content words from the general vocabulary as features, in this approach, we use page titles and their translations (as obtained through hyperlinks as explained above) as our primitives for the computation of multilingual similarity. The first step of this approach, then, is acquiring the bilingual lexicon, but this is relatively straightforward. For each Wikipedia page in one language, translations of the title in other languages, for which there are separate entries, are given as hyperlinks. This information is used to generate a bilingual translation lexicon. Most of these titles are content bearing noun phrases and are very useful in multilingual similarity computation (Kirk Evans, 2005). Most of these noun phrases are already disambuiguated, and may consist of either a single word or multiword units.

Wikipedia uses a redirection facility to map several titles into a canonical form. These titles are mostly synonymous expressions. We used Wikipedia's redirect feature to identify synonymous expression.

Canonical representation of a sentence

Once we have the bilingual lexicon, the next step is to represent the sentences in both language pairs using this lexicon. Each sentence is represented by the set of hyperlinks it contains. We search each hyperlink in the bilingual lexicon. If it is found, we replace the hyperlink with the corresponding unique identification of the bilingual lexicon entry. If it is not found, the hyperlink will be included as is as part of the representation. This is done since Dutch and English are closely related languages and may share many cognate pairs.

Enriching the Wikipedia link structure

As described in the previous section, the method uses hyperlinks in a sentence as a highly focused entity-based representation of the aboutness of the sentence. In Wikipedia, not all occurrences of named-entities or concepts that have entries in Wikipedia are actually used as anchor text of a hypertext link; because of this, a number of sentences may needlessly be left out from the similarity computation process. In order to avoid this problem, we automatically identify other relevant hyperlinks using the bilingual lexicon generated in the previous section.

Identification of additional hyperlinks in Wikipedia sentences works as follows. First we split the sentences into constituent words. We then generate N gram words keeping the relative order of words in the sentences. Since the anchor texts of hypertext links may be multiword expressions, we start with higher order N gram words (N=4). We search these N grams in the bilingual lexicon. If the N gram is found in the lexicon, it is taken as a new hyperlink and will form part of the representation of a sentence. The process is repeated for lower order N grams.

Identifying similar sentences

Once we are done representing the sentences as described previously, the final step involves computation of the term overlap between the sentence pairs and filtering the resulting list. The remaining steps are similar to those described in the MT based approach. For completeness, we briefly repeat the steps here. First, all sentences from a Dutch Wikipedia page are linked to all sentences of the corresponding English Wikipedia page. We then compute the similarity between the sentence representations, using the Jaccard similarity coefficient.

A sentence in Dutch page may be similar to several sentences in English page which may result in a large number of spurious pairs. Therefore, we filter the list using the following recursive procedure. First, the sentence pairs are sorted by their similarity scores. We take the pairs with the highest similarity scores. We then eliminate all other sentence pairs from the list that contain either of sentences in this pair. We continue this process taking the second highest ranking pair. Note that this procedure assumes a one-to-one matching rule; a sentences in Dutch can be linked to at most one sentence in English.

5 Experimental Evaluation

Now that we have described the two algorithms for identifying similar sentences, we return to our research questions. In order to answer them we run the experiment described below.

5.1 Set-up

We took a random sample of 30 English-Dutch Wikipedia page pairs. Each page is split into sentences. We generated candidate Dutch-English sentence pairs and passed them on to the two methods. Both methods return a ranked list of sentence pairs that are similar. As explained above, we assumed a one-to-one correspondence, i.e., one English sentence can be linked to at most to one Dutch sentence.

The outputs of the systems are manually evaluated. We apply a relatively lenient criteria in assessing the results. If two sentences overlap interms of their information content then we consider them to be similar. This includes cases in which sentences may be exact translation of each other, one sentence may be contained within another, or both share some bits of information.

5.2 Results

Table 1 shows the results of the two methods described in Section 4. In the table, we give two types of numbers for each of the two methods *MT* and *Bilingual lexicon*: *Total* (the total number of sentence pairs) and *Match* (the number of correctly identified sentence pairs) generated by the two approaches.

Overall, the two approaches tend to produce similar numbers of correctly identified similar sentence pairs. The systems seem to perform well on pages which tend to be alignable at sentence level, i.e., parallel. This is clearly seen on the following pages: *Pierluigi Collina*, *Marcus Cornelius Fronto*, *George F. Kennan*, which show a high similarity at sentence level. Some pages contain very small description and hence the figures for correct similar sentences are also small. Other topics such as *Classicism* (Dutch: *Classicisme*), *Tennis*, and *Tank*, though they are described in sufficient details in both languages, there tends to be less overlap among the text. The methods tend to retrieve more accurate similar pairs from person pages than other pages especially those pages describing a more abstract concepts. However, this needs to be tested more thoroughly.

When we look at the total number of sentence pairs returned, we notice that the bilingual lexicon based method consistently returns a smaller amount of similar sentence pairs which makes the method more accurate than the MT based approach. On average, the MT based approach returns 4.5 (26%) correct sentences and the bilingual lexicon based approach returns 2.9 correct sentences (45%). But, on average, the MT approach returns three times as many sentence pairs as bilingual lexicon approach. This may be due to the fact that the former makes use of restricted set of important terms or concepts whereas the later uses a large general lexicon. Though we remove some of the most frequently occuring stopwords in the MT based approach, it still generates a large number of incorrect similar sentence pairs due to some common words.

In general, the number of correctly identified similar pages extracted seems small. However, most of the Dutch pages are relatively small, which sets the upper bound on the number of correctly identified sentence pairs that can be extracted. On average, each Dutch Wikipedia page in the sample contains 18 sentences whereas English Wikipedia pages contain 65 sentences. Excluding the pages for *Tennis*, *Tank* (Dutch: *voertuig*), and *Tricolor*, which are relatively large, each Dutch page contains on average 8 sentences, which is even smaller. Given the fact that the pages are in general not parallel, the methods, using simple heuristics, identified high quality translation equivalent sentence pairs from most Wikipedia pages. Furthermore, a close examination of the output of the two approaches show that both tend to identify the same set of similar sentence pairs.

We ran our bilingual lexicon based approach on the whole Dutch-English Wikipedia corpus. The method returned about 80M of candidate similar sentences. Though we do not have the resources to evaluate this output, the results we got from sample data (cf. Table 1) suggest that it contains a significant amount of correctly identified similar

Title		MT		Bilingual Lexicon	
English	Dutch	Total	Match	Total	Match
Hersfeld Rotenburg	Hersfeld Rotenburg	2	_	3	2
Manganese nodule	Mangaanknol	5	2	1	1
Kettle	Ketel	_	_	1	1
Treason	Landverraad	2	_	1	_
Pierluigi Collina	Pierluigi Collina	14	13	13	11
Province of Ferrara	Ferrara (provincie)	7	1	1	1
Classicism	Classicisme	8	_	1	_
Tennis	Tennis	93	4	15	3
Hysteria	Hysterie	14	6	9	5
George F. Kennan	George Kennan	27	12	29	11
Marcus Cornelius Fronto	Marcus Cornelius Fronto	11	9	5	5
Delphi	Delphi (Griekenland)	34	2	8	1
De Beers	De Beers	11	5	10	5
Pavel Popovich	Pavel Popovytsj	7	4	4	4
Rice pudding	Rijstebrij	11	1	4	_
Manta ray	Reuzenmanta	15	3	7	2
Michelstadt	Michelstadt	1	1	1	1
Tank	Tank (voertuig)	84	3	27	2
Cheyenne(Wyoming)	Cheyenne(Wyoming)	5	2	2	2
Goa	Goa(deelstaat)	13	4	6	1
Tricolour	Driekleur	57	36	13	12
Oral cancer	Mondkanker	25	2	7	2
Pallium	Pallium	12	2	5	4
Ajanta	Ajanta	3	3	2	2
Captain Jack (band)	Captain Jack	16	3	2	2
Proboscis Monkey	Neusaap	15	6	4	1
Patti Smith	Patti Smith	6	2	4	2
Flores Island, Portugal	Flores (Azoren)	3	2	1	1
Mercury 8	Mercury MA 8	11	3	4	1
Mutation	Mutatie	16	4	6	3
Average		17.6	4.5	6.5	2.9

Table 1: Test topics (column 1 and 2). The total number of sentence pairs (column 3) and the number of correctly identified similar sentence pairs (column 4) returned by the MT based approach. The total number of sentence pairs (column 5) and the number of correctly identified similar sentence pairs (column 6) returned by the method using a bilingual lexicon.

sentences.

6 Conclusion

In this paper we focused on multilingual aspects of Wikipedia. Particularly, we investigated the potential of Wikipedia for generating parallel corpora by applying different methods for identifying similar text across multiple languages. We presented two methods and carried out an evaluation on a sample of Dutch-English Wikipedia pages. The results show that both methods, using simple heuristics, were able to identify similar text between the pair of Wikipedia pages though they differ in accuracy.

The bilingual lexicon approach returns fewer incorrect pairs than the MT based approach. We interpret this as saying that our bilingual lexicon based method provides a more accurate representation of the aboutness of sentences in Wikipedia than the MT based approach. Furthermore, the result we obtained on a sample of Wikipedia pages and the output of running the bilingual based approach on the whole Dutch-English gives some indication of the potential of Wikipedia for generating parallel corpora.

As to future work, the sentence similarity detection methods that we considered are not perfect. E.g., the MT based approach relies on rough translations; it is important to investigate the contribution of high quality translations. The bilingual lexicon approach uses only lexical features; other language specific sentence features might help improve results.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 640.-001.501, and 640.002.501.

References

D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. 2005. Using Wikipedia at the TREC QA Track. In E.M. Voorhees and L.P. Buckland, editors, *The Thirteenth Text Retrieval Conference (TREC 2004)*.

F. Bellomi and R. Bonato. 2005. Lexical authorities in an encyclopedic corpus: a case study with wikipedia. URL: `http://www.fran.it/blog/2005/01/lexical-authorities-in-encyclopedic.htm%l`. Site accessed on June 9, 2005.

A. Ciffolilli. 2003. Phantom authority, selfselective recruitment and retention of members in virtual communities: The case of Wikipedia. *First Monday*, 8(12).

S. Fissaha Adafre and M. de Rijke. 2005. Discovering missing links in Wikipedia. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*.

R. Ghani, S. Slattery, and Y. Yang. 2001. Hypertext categorization using hyperlink patterns and meta data. In Carla Brodley and Andrea Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 178–185.

D. Kirk Evans. 2005. Identifying similarity in text: Multi-lingual analysis for summarization. URL: `http://www1.cs.columbia.edu/nlp/theses/dave_evans.pdf`. Site accessed on January 5, 2006.

A. Lih. 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*.

D. Melamed. 1996. A geometric approach to mapping bitext correspondence. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–12, Somerset, New Jersey. Association for Computational Linguistics.

N. Miller. 2005. Wikipedia and the disappearing “Author”. *ETC: A Review of General Semantics*, 62(1):37–40.

U. Rao and M. Turoff. 1990. Hypertext functionality: A theoretical framework. *International Journal of Human-Computer Interaction*.

F. Viégas, M. Wattenberg, and D. Kushal. 2004. Studying cooperation and conflict between authors with history flow visualization. In *Proceedings of the 2004 conference on Human factors in computing systems*.

J. Voss. 2005. Measuring Wikipedia. In *Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*.

WiQA. 2006. Question answering using Wikipedia. URL: `http://ilps.science.uva.nl/WiQA/`. Site accessed on January 5, 2006.

Multilingual interactive experiments with Flickr

Paul Clough
Department of
Information Studies
University of Sheffield
Sheffield, UK
p.d.clough@sheffield.ac.uk

Julio Gonzalo
Departamento de Lenguajes
y Sistemas Informáticos
UNED
Madrid, Spain
julio@lsi.uned.es

Jussi Karlgren
Swedish Institute of
Computer Science
Stockholm
Sweden
jussi@sics.se

Abstract

This paper presents a proposal for iCLEF 2006, the interactive track of the CLEF cross-language evaluation campaign. In the past, iCLEF has addressed applications such as information retrieval and question answering. However, for 2006 the focus has turned to text-based image retrieval from Flickr. We describe Flickr, the challenges this kind of collection presents to cross-language researchers, and suggest initial iCLEF tasks.

1 Information Retrieval Evaluation by User Experiment

Information retrieval systems, especially text retrieval systems, have benefited greatly from a fairly strict and straight-laced evaluation scheme, which enables system designers to run tests on versions of their system using a test collection of pre-assessed data. These relevance-oriented experiments shed light on comparative system performance and enable both introduction of new algorithms and incremental optimization. However, batch-oriented system evaluation based on large amounts of data, abstracted away from situational constraints, variation in usage, and interactiveness issues only addresses some of the bottlenecks to build a successful system.

The CLEF[1] Interactive Track (iCLEF[2]) is devoted to the comparative study of user inclusive cross-language search strategies. Over the past 5 years, iCLEF has studied three cross-language search tasks: retrieval of documents, answers and annotated images (Gonzalo and Oard, 2002; Gonzalo et al., 2005). All tasks involve the user interacting with information systems in a language different from that of the document collection. Although iCLEF experiments continue producing interesting research results, which may have a substantial impact on the way effective cross-language search assistants are built, participation in this track has remained low across the five years of existence of the track. Interactive studies, however, remain as a recognized necessity in most CLEF tracks.

Therefore, to encourage greater participation in 2006 our focus has turned to FLICKR[3], a large-scale, web-based image database with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments. Our aim in selecting a primarily non-textual target to study textual retrieval is based on some of the multi-lingual and dynamic characteristics of FLICKR. We will outline them below.

2 The Flickr system

The majority of Web image search is text-based and the success of such approaches often depends on reliably identifying relevant text associated with a particular image. FLICKR is an online tool for managing and sharing personal photographs and currently contains over five million freely accessible images. These are available via the web, updated daily by a large number of users and available to all web users (users can access FLICKR for free, although limited to the upload of 20MB of photos per month).

[1] http://www.clef-campaign.org/

[2] http://nlp.uned.es/iCLEF/

[3] http://www.flickr.com/

2.1 Photographs in the collection

It is estimated that the complete FLICKR database contains 37 million photos with approximately 200,000 images added daily by 1.2 million members[4]. FLICKR provides both private and public image storage, and photos which are shared (around 5 million) can be protected under a Creative Commons (CC) licensing[5] agreement (an alternative to full copyright). Images from a wide variety of topics can be accessed through FLICKR, including people, places, landscapes, objects, animals and events. This makes the collection a rich resource for image retrieval research.

2.2 Annotations

In FLICKR, photos are annotated by authors with freely chosen keywords in a naturally multilingual manner: most authors use keywords in their native language; some combine more than one language. In addition, photographs have titles, descriptions, collaborative annotations, and comments in many languages. Figure 5 provides an example photo with multilingual annotations; Figure 5 shows what the query "cats" retrieves from the database, compared with what the query "chats" retrieves.

Annotations are used by the authors to organize their images, and by any user to search on. Keywords assigned to the images can include place names and subject matter, and photos can also be submitted to online discussion groups. This provides additional metadata to the image which can also be used for retrieval. An explore utility provided by FLICKR makes use of this user-generated data (plus other information such as Clickthroughs) to define an "interestingness" view of images[6].

3 Flickr at iCLEF 2006

Many images are accompanied by text, enabling the use of both text and visual features for image retrieval and its evaluation (Müller et al., 2006, see e.g.). Images are naturally language independent and often successfully retrieved with associated texts. This has been explored as part of ImageCLEF (Clough et al., 2005) for areas such as information access to medical images and historic photographs. The way in which users search for images provides an interesting application for user-centered design and evaluation. As an iCLEF task, searching for images from FLICKR presents a new multilingual challenge which, to date, has not been explored. Challenges include:

- Different types of associated text, e.g. keywords, titles, comments and description fields.
- Collective classification and annotation using freely selected keywords (known as folksonomies) resulting in non-uniform and subjective categorization of images.
- Annotations in multiple languages.

Given the multilingual nature of the FLICKR annotations, translating the user's search request would provide the opportunity of increasing the number of images found and make more of the collection accessible to a wider range of users regardless of their language skills. The aim of iCLEF using FLICKR will be to determine how cross-language technologies could enhance access, and explore the user interaction resulting from this.

4 Proposed tasks

For iCLEF, participants to this evaluation campaign will be provided with the following:

- A subset of the Flickr collection including annotations and photographs[7].
- Example (realistic) search tasks. Ideally these search tasks will reflect real user needs which could be derived from log files, studies or similar retrieval tasks.
- A framework in which to run an evaluation.

5 Summary

Flickr will allow us to create an extremely interesting interactive task based on truly heterogeneous annotations (that will in turn hopefully attract more participants). Using images from within a Web environment is a realistic and contemporary search challenge and allows many important research questions to be addressed from

[4] These figures are accurate as of October 2005: http://www.wired.com/news/ebiz/0,1272,68654,00.html

[5] http://creativecommons.org/image/flickr, http://flickr.com/creativecommons/

[6] http://www.flickr.com/explore/interesting

[7] We are currently in negotiations with Yahoo! (owners of Flickr) and Flickr to provide researchers with legitimate access to a subset of the collection.

a quickly developing field. User-centered studies are required within both text and image retrieval, but are often neglected as they require more effort and time from participating groups than a system-centered comparison that can often be run without human intervention. Still, user-centered evaluation cannot be replaced and the influence of the user on the results is in general stronger than the influence of the system itself.

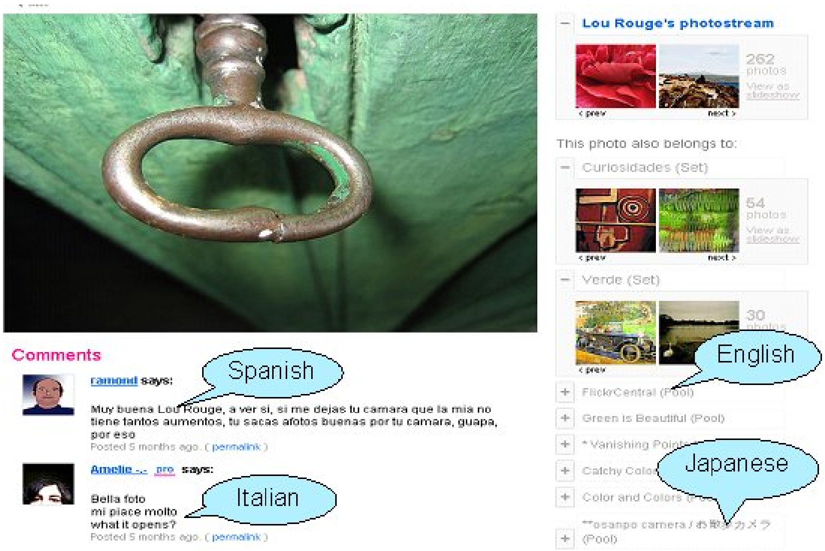
References

Paul Clough, Henning Müller, and Mark Sanderson. 2005. The clef 2004 cross language image retrieval track. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, number 3491/2005 in Lecture Notes in Computer Science, pages 597–613. Springer, Heidelberg, Germany.

Julio Gonzalo and Doug Oard. 2002. The clef 2002 interactive track. In *Advances in Cross-Language Information Retrieval*, number 2785 in Lecture Notes in Computer Science. Springer-Verlag, Berlin-Heidelberg-New York.

Julio Gonzalo, Paul Clough, and A Vallin. 2005. Overview of the clef 2005 interactive track. In *Working notes of the CLEF workshop*, Vienna, Austria, September.

Henning Müller, Paul Clough, William Hersh, Thomas Deselaers, Thomas Lehmann, and Antoine Geissbuhler. 2006. Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems. In *SPIE conference Photonics West, Electronic Imaging, special session on benchmarking image retrieval systems*, San Diego, February.



Lou Rouge's photostream

262 photos View as slideshow

‹ prev next ›

This photo also belongs to:

Curiosidades (Set)

54 photos View as slideshow

‹ prev next ›

Verde (Set)

30 photos

‹ prev

+ FlickrCentral (Pool)

+ Green is Beautiful (Pool)

+ * Vanishing Point

+ Catchy Colo

+ Color and Colors (P

+ **osanpo camera / お散歩カメラ (Pool)

Comments

ramond says:

Muy buena Lou Rouge, a ver si, si me dejas tu camara que la mia no tiene tantos aumentos, tu sacas afotos buenas por tu camara, guapa, por eso

Posted 5 months ago. (permalink)

Amelie -.- pro says:

Bella foto
mi piace molto
what it opens?

Posted 5 months ago. (permalink)

Figure 1: Example multilingual annotations in Flickr.

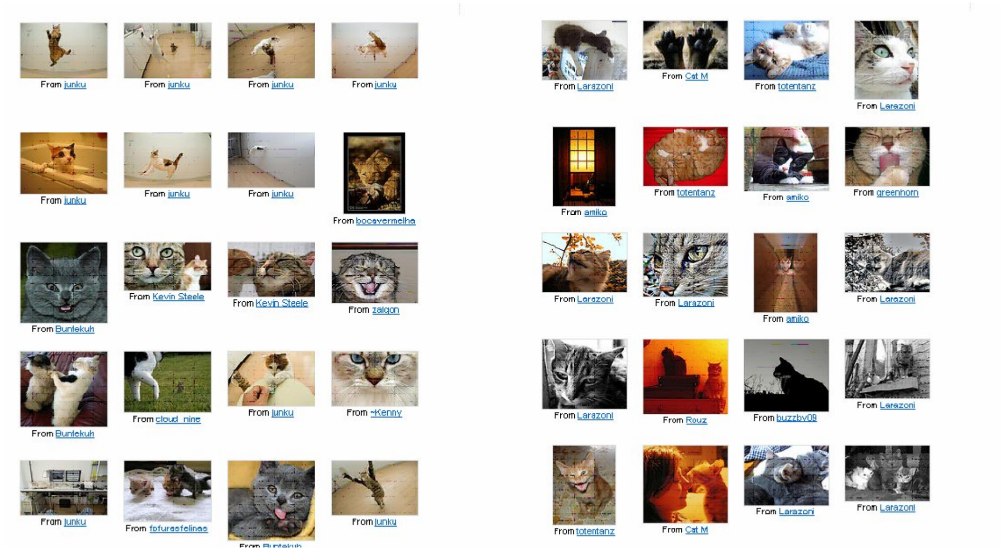


Figure 2: Retrieval of “cats” (left) and “chats” (right).