EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 144–158
CILC2016. 8th International
Conference on Corpus Linguistics
EPiC Language and Linguistics

A French weblog corpus for new insights on blog post tagging*

Ivan Garrido-Marquez, Laurent Audibert, Jorge Garcia Flores, François Lévy, and Adeline Nazarenko

LIPN–Université Paris 13
Sorbonne Paris Cité & CNRS
{garrido-marquez,audibert,jgflores,fl,nazarenko}@lipn.univ-paris13.fr

Abstract

The rapid evolution and informational growth of blogs requires enhanced functionality for searching, navigating and linking content. This paper presents the French Blog Annotation Corpus FLOG, intended to provide a research testbed for the study of annotation practices, and specifically tagging and categorizing blog posts. The corpus covers a ten year time span of blog posts on cooking, law, video games and technology. Statistical analysis of the corpus suggests that tag annotation of posts is more frequent than category attribution, but on the other hand categories provide a richer semantic structure for post classification and search. The review of the state of the art on automatic tag suggestion shows that tag suggestion tools are not of widespread use yet between bloggers, which might be a consequence of methods that do not take into account the past tagging history of the blog, the context of the post within the blog and the tagging pattern of each blog author.

1 Introduction

For almost twenty years, blogs have been popular platforms for online publications on almost any kind of subject. The word *blog* comes from a truncation of *weblog*: online journals that appeared on internet around 1998 [1]. Those first weblogs included personal notes, hyperlinks and reader's comments. With the development and widespread use of blogging platforms, the format was enriched with blog post tagging and categorizing possibilities. Both for their history and their particular text structure, blogs can be considered as a rich source for the study of text annotation and categorization over a long time span.

In this paper we present the French Blog Annotation Corpus (FLOG) : a research corpus covering almost ten years of blog posts in French language, with almost 25,000 topics and more than 11 millions of words. Given the blog's particular characteristic of having two different

*This work has been partially funded by the French Ministry of National Education, Higher Education, Training and Scientific Research and is supported by the French National Research Agency (ANR-10-LABX-0083) in the context of the Labex EFL.

levels of annotation (tag and category), the Flog corpus is intended to become a testbed for diachronic research on blog annotation[1].

The flog corpus was inspired on similar resources available for English language, like the Blog Authorship Corpus [7], the American political blog corpus [10] or the TREC blog track [4]. The blogs were selected from ranking lists of popular blogs in French. After having downloaded and collected the documents, they were imported to a relational database that allows a closer lexical, statistical and distributional analysis.

As the goal of flog corpus is to support the analysis of blogging activity and especially annotation practices over the long term, our statistic analysis focuses on the relation between posts and their annotations, *i.e.* tags and categories. We present statistics on tag and category frequency in relation with the number of posts and the number of authors. The analysis suggests a more frequent use of tags than categories for blog annotation and a high amount of content-based tags (that is, tag coming from a word included in the post text) as well, even if annotation policy changes from one blog to another and there is in fact a large variety of authors' practices.

Tag suggestion tools have been developed as a natural evolution of blog platforms, but their use is not widespread. As part of our study of blog annotation, we take into account tag suggestion methods in order both to answer the question of what an accurate annotation could be and to study the semantic relation that links an annotation to its blog posts.

The paper is organized as follows: in section 2 we review related work on blog corpora available for research. Section 3 describes the corpus collection method. Sections 4 and 5 respectively analyze blogging activity and annotation practices, while section 6 provides some insight on automatic blog annotation. Finally, section 7 closes the paper with some research perspectives opened by the flog corpus.

2 Related work

Blogging is a common source of information nowadays; from professional blogs to micro blogging in social networks. According to the web portal of statistics and studies Statista[2], in 2011, there were aproximately 173 million online blogs across various platforms such as Livejournal, WordPress or Blogger. The same source reports that in April 2016, the number of Tumblr[3] blog accounts has surpassed 291.7 million. Worldometers[4] estimates that the number of blog posts written in the world has increased by around 450,000, during the two and a half hours this section was written. Measuring precisely the size and rate of growth of the blogosphere is very complicated since it is tremendously large and diverse.

Over this uprising tide of information, blog readers need to locate posts relevant to their information needs and interests and navigate through them. Figure 1 shows an example of a blog post. It illustrates the current search and organizing functionality blog platforms provide: full text search, similar posts, tags and category labels of the post and cooccurrent tags.

Tagging activity in blogs is an easy and popular way to keep contents classified and organized. There are usually two types of annotation in blog posts: categories and tags. Categories are names for groups of posts with related topic, they can be seen as the main topics of the blog. Categories may be predefined but the bloggers can add new categories as needed. Sometimes categories have levels of sub-categories, which defines a taxonomy in a tree-like structure. Tags

[1] The Flog corpus will be be made publicly available both in XML format and as a database of metadata (see below). We are currently gathering all the necessary authorizations.

[2] www.statista.com/

[3] tumblr.com

[4] www.worldometers.info/

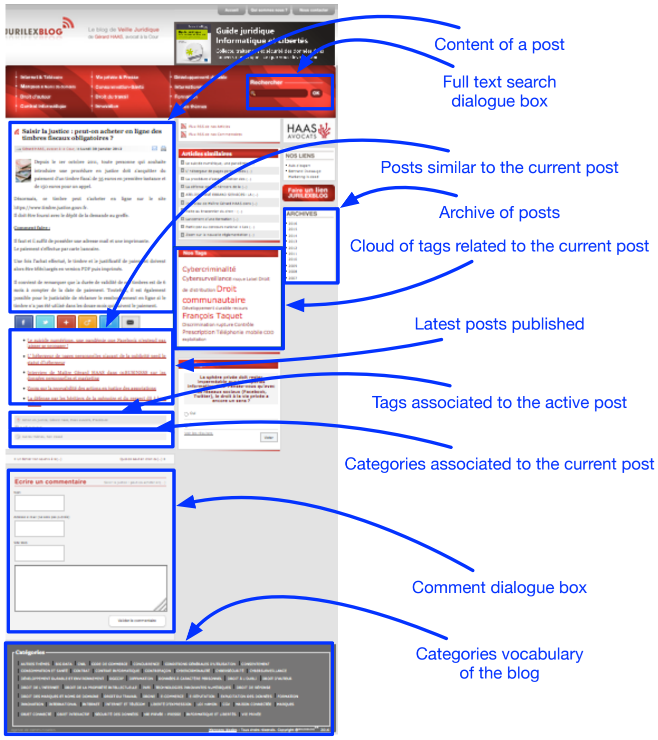

Content of a post
Full text search dialogue box
Posts similar to the current post
Archive of posts
Cloud of tags related to the current post
Latest posts published
Tags associated to the active post
Categories associated to the current post
Comment dialogue box
Categories vocabulary of the blog

Figure 1: Example of blog interface. The tags and categories associated to the post can used to navigate through the blogs and identify related posts.

are keywords authors freely add to their post summarizing the content of posts. Although tags can reappear in any number of posts, they are normally chosen by authors right after having written the post. Both categories and tags are helpful for searching a specific post, linking related posts, indexing blogs by Internet searching engines and so on.

2.1 Existing blog corpora

Social media have raised a lot of interest in the linguistics and natural language processing (NLP) communities, but collecting corpora is a first but critical step to support the description of ”social” special languages and the development of tools to process social unstructured data.

Just as there are corpora of tweets, there are corpora of blogs, such as:

Blog Authorship Corpus [7] It was collected in August 2004 from blogger.com. Its goal was to characterize blog authors in terms of gender or age. The corpus consists of 681,288 posts from 19,320 bloggers. It contains more than 140 million of words and it is presented as a collection of XML files, each one gathering the posts and comments of one author and being associated with the blogger’s id and self-provided information (gender, age, company and astrological sign).

Corpus of American political blogs [10] presents a corpus that has been built with the tokenized and standardized text and comments of blog posts from 40 blogs about American politics ranging from November 2007 to October 2008. This corpus was collected in order to test topic modeling for predicting users responses to future posts.

TREC blog corpus In 2006, the Text REtrieval Conference (TREC)[5] opened a blog track that lasted until 2009. It was intended to test and evaluate information retrieval systems and it included various tasks such as opinion retrieval, feed search, determination of opinion polarity, link-analysis and post retrieval. For this evaluation track, two versions of the blog corpus were provided [4] the first in 2006 and the second in 2008. The first corpus comprehended a short time span (few months from late 2005 to early 2006), while the second corpus had a longer time span of one year (from January 2008 to February 2009). They are both considerably large, with 3.2 and 28.5 million documents respectively. The corpus includes home pages, feeds, permalinks and even part of the spam.

Birmingham Blog Corpus [3] It is a 600 million word collection of blog posts and comments but it is only available through the WebCorp Linguists Search Engine interface.

These corpora are quite large but they are all in English and they cover a relatively small time span. None of them covers more than one year. For the sake of both language diversity and the development of NLP resources in French, we thought it was important to create a French blog resource. Furthermore, after 20 years of blogging activity, it seemed also important to analyze blogs on the long trend and therefore to collect blog corpora with a larger time span. Finally, despite their large size, some of these corpora are very specific: they are dedicated to a specific domain (*e.g.* American politics) or have been collected for a specific type of analysis (*e.g.* authors' characterization). In order to support a general analysis of the annotations of blogs, we decided to create a multitopic corpus with a more varied collection of blog posts.

2.2 Tool support for tagging blogs

Current blogging platforms provide some tools and plugins for helping bloggers to annotate their posts with keyword tags. Many of them use available services like Yahoo! Content Analysis[6], Open Calais[7], Alchemy API[8], Zemanta[9], for which APIs are provided. These services usually apply NLP techniques and shallow semantic analysis to extract elements from text content of a new post (named entities, keywords, concepts or even relations) and propose them as tags for it to the post author.

Let's take the example of the WordPress blogging platform. It proposes a list of plugins with several tagging tools[10] serving manifold purposes. Thoth's suggested tags scans the post text phrases and calculates a relevance value. Wiki CS Annotation exploits the computer science category of Wikipedia bahasa Indonesian as knowledge source. It selects some elements in the text of the post as tags and link them to those pages. Climate tagger scans unstructured content and identifies terms and concepts from a climate thesaurus. AlchemyTagger uses the entity and keyword extraction modules of the Alchemy API to suggest content based tags. Simple tags benefits of a combination of some of the above mentioned services to propose tags to the blog post author.

All these tools operate more or less in the same manner: they are meant to support bloggers manual annotation and not for automatic tagging; they mainly rely on the textual content of the post to annotate; some of them also exploit an external knowledge source but, to the best

[5] http://trec.nist.gov/

[6] https://developer.yahoo.com/contentanalysis/

[7] http://www.opencalais.com/

[8] http://www.alchemyapi.com/

[9] http://www.zemanta.com/api/

[10] https://wordpress.org/plugins/tags/tagging

of our knowledge, none of them exploits any blog internal taxonomy to propose categories in addition to content-based tags.

However, those tagging tools do not seem to be as widely used by bloggers as one could expect. In order to understand how NLP tools can actually help bloggers and enhance blogs, a thorough analysis of the traditional blogging activity would be necessary. Such an analysis requires in turn a large time span, multi-domain and multi-authors blog corpus. The FLOG corpus has been designed as a large time span, multi-domain and multi-authors blog corpus for that purpose.

3 Corpus collection

The following subsection explains how the corpus was gathered, processed and how it is presented. It also introduces a database created along with the corpus to be a side tool for working with it.

3.1 Collecting methodology

For building the corpus we relayed on the ranked list of popular blogs provided by the Teads Company[11]. Tead Labs explains that they maintain an up-to-date database of 2 million of blogs coming from 8 countries. Their ranking takes into account several factors, among which are the number of blogs pointing to the blog to rank, its relevance and the shares of the target blog in social networks like Facebook and Twitter. The ranking is automatically updated every 5 months.

We initially selected few topics of interest so as to allow for both inter-topic and intra-topic comparison. Cooking, technology, video-games and laws were chosen as strongly and permanently represented topics.

Among the top-ranked blogs of Teads, we selected the blogs that fulfilled the following requirements:

- be classified in one of the topics of interest;
- be annotated with categories or tags, and preferably both;
- be active at the time of fetching;
- support diachronic analysis (*i.e.* every post is associated with an explicit date).

The data was fetched via HTTP by using the GNU `wget` command. The complete sites were downloaded and then filtered to keep only HTML post files having some text within the `body` element, having a title and associated with at least one tag or category.

3.2 Corpus format

The blogs have therefore been split in posts, each post being stored in a separate file.

Every HTML post file was processed and transformed into an XML file. The XML templates were filled with the extracted `Title`, `Author`, `Date`, `Text`, `Tag list`, and `Class list`.

Figures 2 and 3 show the DTD[12] of the post XML format and an instance of a formatted post.

[11] http://fr.labs.teads.tv/top-blogs

[12] Document Type Description

```
<!DOCTYPE document [
<!ELEMENT document (date,title,author,tags_set,categories_set,text)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT categories_set (category)*>
<!ELEMENT tags_set (tag)*>
<!ELEMENT category (#PCDATA)>
<!ELEMENT tag (#PCDATA)>
<!ELEMENT text (#PCDATA)>
]>
```

Figure 2: Document model (DTD) for the XML format of posts

```
<document>
<date>2015-01-12</date>
        <title>[#CharlieJam] Une game jam pour permettre aux d&#233;veloppeurs
        de s&#8217;exprimer sur le sujet de Charlie Hebdo</title>
        <author>Kocobe</author>
        <tags_set>
                <tag>Charlie hebdo</tag>
                <tag>game jam</tag>
        </tags_set>
        <categories_set>
                <category>game jam</category>
        </categories_set>
        <text>
 La Charlie Jam invite les d&#233;veloppeurs du monde entier &#224; s&#8217;
exprimer sur le sujet de Charlie Hebdo &#224; travers la conception d&#8217;
un jeu. Si les &#233;v&#233;nements r&#233;cents ont d&#233;clench&#233
....
        </text>
</document>
```

Figure 3: Example of post in XML format

The `date` field indicates the date the post was published (in ISO 8601 format YYYY-MM-DD). The `author` gives the name of the user displayed in the post. The different tags (`tag` elements) are listed inside the `tags_set` one and, similarly, the `categories_set` element gives the list of the categories associated with the post, each one in a single `category` element. These lists can be arbitrarily long and either of them can be empty if the post is associated only with tags or with categories. The `text` field contains the text of the post without any link, image or any type of embedded element.

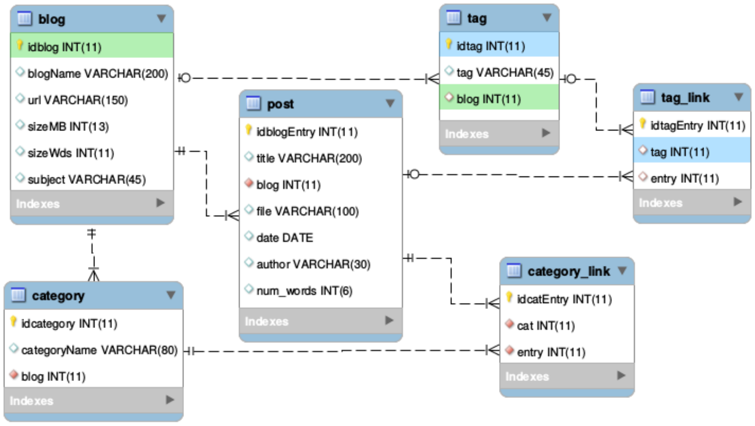


Figure 4: Schema of the metadata database structure (Enhanced Entity-Relationship diagram)

3.3 Post searching database

Along with the corpus in XML format, all the metadata information has been stored in a SQL relational database, which includes all the information of the XML files, except for the content of text field itself which is substituted with a link to the source file.

This database has been designed to help query, searching and exploration of the Flog corpus, which is quite large. The database allows to get straight forward statistics based on the blog structure or its attached metadata. For instance, one can analyze the content of a specific blog over a certain period, extract the posts of a given set of blogs or authors, or even analyze the distribution of tags through time or the activity of a certain blog author. This database has been designed to serve our purpose of analyzing the blogging activity over time. It would be distributed with the Flog corpus as a a corpus exploration tool.

Figure 4 shows the database schema. The blog table has row per blog containing all the general information attached to blogs. In the post table, there is a row for each post (or XML file). The `tag` and `category` tables respectively give the catalog of tags and categories that can be found in blogs. Finally, Tables `tag_link` and `category_link` record the information about which post is associated to which tag or category.

4 Analysis of blogging activity

The Flog corpus gathers 20 blogs related to 4 different domains: cooking (*cuisine*, 4 blogs & 1,280,000 words), technologies (*technologie*, 5 blogs & 303,000 words), Law (*Droit*, 5 blogs & 3,114,000 words) and videogames (*jeux vidéo*, 6 blogs, & 5,287,500 words)[13]

[13] Here are the actual names of the blogs, which are designated by domain and number in the rest of the paper: **Cuisine**: bcommebon.canalblog.com, www.beaualalouche.com, pechedegourmand.canalblog.com, www.la-gourmandise-selon-angie.com; **Technologie**: www.domadoo.fr, domotique34.com, gurau-audibert.hd.free.fr/josdblog, prendreuncafe.com/blog, www.maison-et-domotique.com; **Droit**: francoischarlet.ch, www.jurilexblog.com, www.paralipomenes.net, www.philippebilger.com; **Jeux vidéo**:, coupleofpixels.fr, www.johncouscous.com, www.journaldugamer.com, www.julsa.fr. The whole corpus counts around 11 millions of words and is made of 24,700 different posts. The mean size of collected blogs ranges from 75,000 words per blog for domotics to 880,000 words per blog for videogames.

Even if one cannot evaluate its representativity, both for its size and its heterogeneity we consider it a good corpus for analyzing bloggers' annotation practices. The twenty blogs have different structures, as shown on Table 1, which gives an overall description of the FLOG corpus.

On average, a blog contains 1,200 posts and 570,000 words in 1,200 posts (around 460 words per post) but there are significant differences. The smallest blog in size has 55,000 words with 225 words per post whereas the smallest blog in term of posts has a less but much longer posts (365 words per post). At the other end of the spectrum, the blog jeuxvideo3 is the largest both in word and post counts (1,600,000 words in 6600 posts - only 242 words per post). The most concise has 192 words per blog (technologie4: 110,000 words in 570 posts) and the more prolific counts has 1,296 per post (cuisine3: 366,000 words in 280 posts).

Blog	posts	authors	cats	tags	size in words
jeuxvideo6	184	6	18	556	66,991
technologie2	243	1	38	40	55,073
droit3	283	1	13	77	366,816
technologie5	305	1	16	295	177,034
technologie3	343	13	41	397	193,160
technologie5	374	2	25	358	317,551
cuisine3	474	1	50	243	152,377
droit1	485	2	4	84	466,702
cuisine1	514	1	60	460	133,063
technologie4	573	1	12	321	110,111
jeuxvideo5	1135	2	37	2467	387,632
cuisine2	1166	1	26	695	1,051,706
jeuxvideo1	1423	3	43	1772	868,019
technologie1	1423	17	56	1231	416,498
jeuxvideo4	1501	17	40	3146	698,151
cuisine4	1721	1	25	265	891,033
droit4	1752	1	15	0	1,333,494
droit2	1769	143	48	741	771,041
jeuxvideo2	2483	6	33	2978	1,349,318
jeuxvideo3	6587	67	91	4650	1,598,143
average	1236.9	14.35	34.55	1038.8	570,195.65
std dev	1426.12	33.77	20.57	1292.08	475,284.97
max	6587	143	91	4650	1,598,143
min	184	1	4	0	55,073
total	24738	287	691	20776	**11,403,913**

Table 1: Corpus description

While nearly half of collected blogs have a single author, the number of different authors grows up to 143 for droit2, which is quite an exception. Actually, only one over three blogs has more than 3 authors. Author prolixity varies from 12 posts per author in droit3 to 1,721 in droit4. Globaly, "the more authors, the less posts per author" is a clear tendency, even if it is not a strict law.

Blogs from the FLOG corpus have also very diverse historical profiles (see Table 2). Few of them have been rather regularly active during the whole period, such as cuisine3, while one shows a noticeable increasing activity (jeuxvideo3) and some others slow down gently such as cuisine2.

Blog	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
cuisine1			149	98	66	50	42	32	31	29	17
cuisine2		150	222	192	146	127	97	66	70	68	28
cuisine3							83	155	114	94	28
cuisine4	21	178	224	186	180	171	167	162	159	159	114
technologie1					93	162	132	201	316	398	121
technologie5								137	167	1	
technologie2			45	34	15	6	9	49	47	27	11
technologie3			3	5	4	4	30	32	44	96	125
technologie4	267	151	72	54	16	7					
droit1				2	2	28	44	69	150	116	74
droit2			153	507	425	385	40	52	50	101	56
droit3					10	48	75	64	40	27	19
droit4	23	107	177	183	165	186	192	194	193	192	140
jeuxvideo1							355	328	288	254	198
jeuxvideo2					8	41	303	421	566	658	486
jeuxvideo3	43	65	18	86	149	155	377	737	886	1655	2416
jeuxvideo4						257	436	236	194	180	198
jeuxvideo5						490	191	82	98	146	128
technologie5			2	20	81	55	46	37	62	52	19
jeuxvideo6							6	47	20	45	66

Table 2: Distribution of posts per year and per blog

Figure 5 gives an overview of the blogging activity in the four domains of the Flog corpus. It shows an impressive dynamism in the blogs related to video games (more than 3500 new posts have been recorded in 2015), the relative stability of annual number of cooking posts and fluctuating activity levels in Law and technology domains.

The number of authors is generally constant over a blog life. When it varies, it usually follows approximately the number of posts. However, a single author published 915 posts in one year and, in 7 cases (resp. 22 cases), one author published more than 500 posts (resp. 200). Regular authors actually play an important role: 71% of the posts (17,700 out of 24,700) are written by authors having produced at least 100 posts during the year.

5 Analysis of annotation practices

Categories and tags are often associated to posts (in colored and clickable fonts in the post presented on Figure 6). They form a sort of index that help readers to search information in the blogs but they also advertise the posts so that their authors can get more readers (or some other benefit). The choice of annotations therefore depends on the (known or hypothetical) standard behavior of the expected readers.

The Flog corpus has been designed to study the blog annotation practices. It contain large vocabularies of categories and tags:

- Categories have not been normalized, so that *Actualité/News* and *Actualités/News* are considered as distinct categories. There are 691 different categories but only 573 distinct category names, and few of them are common to different blogs (*e.g.* the word ”musique/-music” appears as a category name in 6 different blogs).

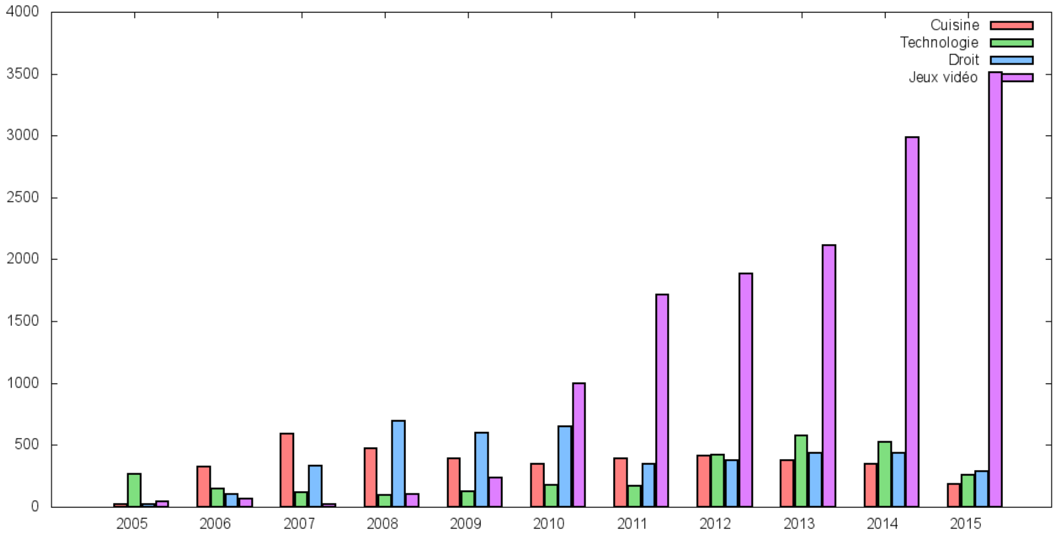
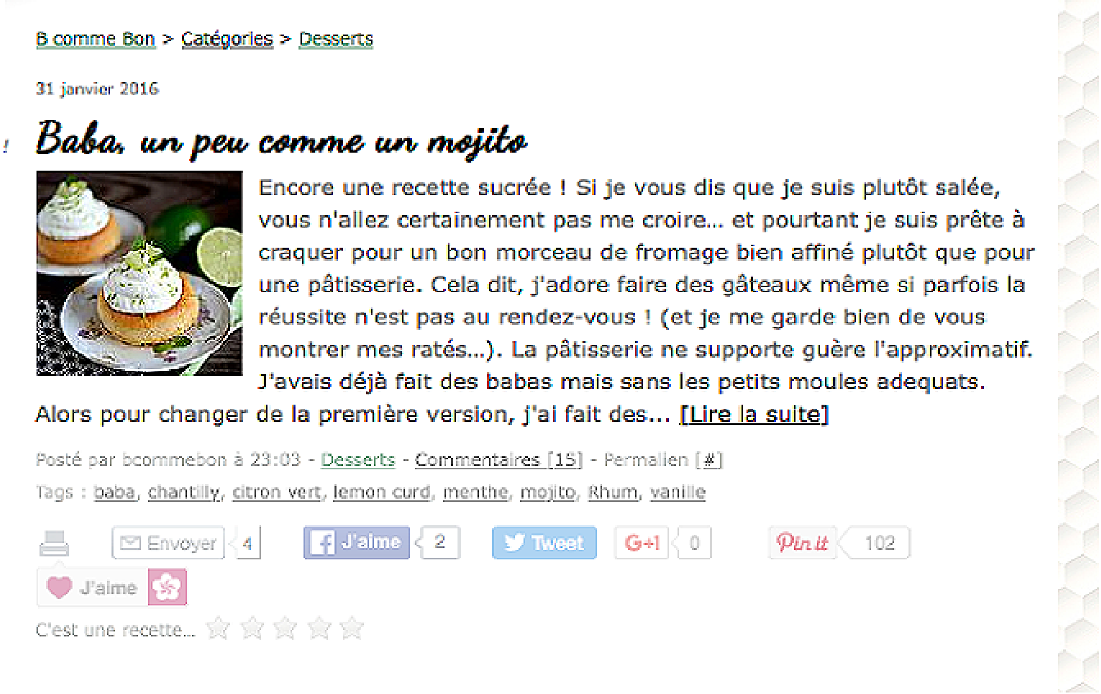


Figure 5: Distribution of posts per year in the different domains of the FLOG corpus.



B comme Bon > Catégories > Desserts

31 janvier 2016

Baba, un peu comme un mojito

Encore une recette sucrée ! Si je vous dis que je suis plutôt salée, vous n'allez certainement pas me croire... et pourtant je suis prête à craquer pour un bon morceau de fromage bien affiné plutôt que pour une pâtisserie. Cela dit, j'adore faire des gâteaux même si parfois la réussite n'est pas au rendez-vous ! (et je me garde bien de vous montrer mes ratés...). La pâtisserie ne supporte guère l'approximatif. J'avais déjà fait des babas mais sans les petits moules adequats. Alors pour changer de la première version, j'ai fait des... [Lire la suite]

Posté par bcommebon à 23:03 - Desserts - Commentaires [15] - Permalien [#]

Tags : baba, chantilly, citron vert, lemon curd, menthe, mojito, Rhum, vanille

Figure 6: Example of a post extracted from the cuisine1 blog. Its author proposes a *Baba* cake recipe. The post is categorized as *Desserts* (on top of the screen) and associated with different tags: the name of the cake (*Baba*) and 7 ingredients such as *chantilly* or *citron vert* (below the text).

- There are 20,776 different tags but only 15,159 distinct tag values. Almost 80% of these values (11,964) appear in one single blog, while the 3,195 remaining ones are usually used in 2-3 blogs and at most in 11 blogs. The tag shared by the more blogs is *iPhone*.

However, this system of tags and categories differs from one blog to another as shown on Table 3. The number of categories per blog ranges from 4 to 91. The relation between the number of categories and the number of posts is very loose: between 6 and 75 posts per category, 121 being an exception. Similarly, there is huge variations in the number of tags per blog (from 40 to 4600) and the number of posts is not related to the number of tags (from 0.33 to 6.5 posts per tag). 6 blogs have more tags than posts and 5 of which are in the video game domain. The respective roles of tags and categories actually puts apart the video game domain from the other domains: video games have between 30 and 90 tags per category, while others have between 1 and 26.

	Categories				Tags			
Blog	mean	min	max	σ	mean	min	max	σ
cuisine1	1	1	1	0	2.12	0	17	3.42
cuisine2	1	1	1	0	5.45	1	20	3.51
jeuxvideo1	3.41	1	11	1.87	4.95	0	19	1.69
technologie1	1.07	1	6	0.29	3.16	0	16	3.55
technologie5	2.31	1	8	1.22	4.20	0	24	4.34
droit1	1	1	1	0	2.41	0	6	1.31
jeuxvideo2	4.27	1	12	1.74	8.84	1	21	3.19
technologie2	1.88	1	5	0.96	0.79	0	5	1.09
jeuxvideo3	0.99	0	1	0.07	4.09	0	28	2.24
jeuxvideo4	2.22	1	3	0.71	6.07	0	45	3.92
droit2	1.72	1	31	2.81	3.19	0	19	2.82
cuisine3	1	1	1	0	5.20	1	14	1.88
technologie3	1.31	1	4	0.60	2.54	0	6	1.34
droit3	1.41	1	5	0.68	2.94	0	9	2.28
cuisine4	1	1	1	0	4.04	0	11	1.68
droit4	3.14	1	7	1.08	0	0	0	0
technologie4	1	1	1	0	3.13	0	13	2.1
jeuxvideo5	2.94	1	10	1.22	3.79	0	13	1.7
technologie5	4.18	1	12	2.03	6.72	0	18	3.17
jeuxvideo6	1.01	1	2	0.1	5.34	0	20	2.84

Table 3: Use of categories and tags per post

It is also interesting to analyze annotation from an historical perspective focused on tags only. Figure 7 gives an overview of tagging activity throughout the period of the corpus. By querying the database of corpus metadata, it is easy to extract detailed histograms and to compare the distribution of different tags over time. Figure 8 compares, for instance, the profiles of two cooking related tags (*chocolat/chocolate* and *citron/lemon*) and two video game tags (*ps3* and *ps4*). It shows an increasing popularity of the two latter and the relative stability of the former ones, which also vary but on an annual basis (the tag *chocolat* is gaining popularity every year for Christmas!).

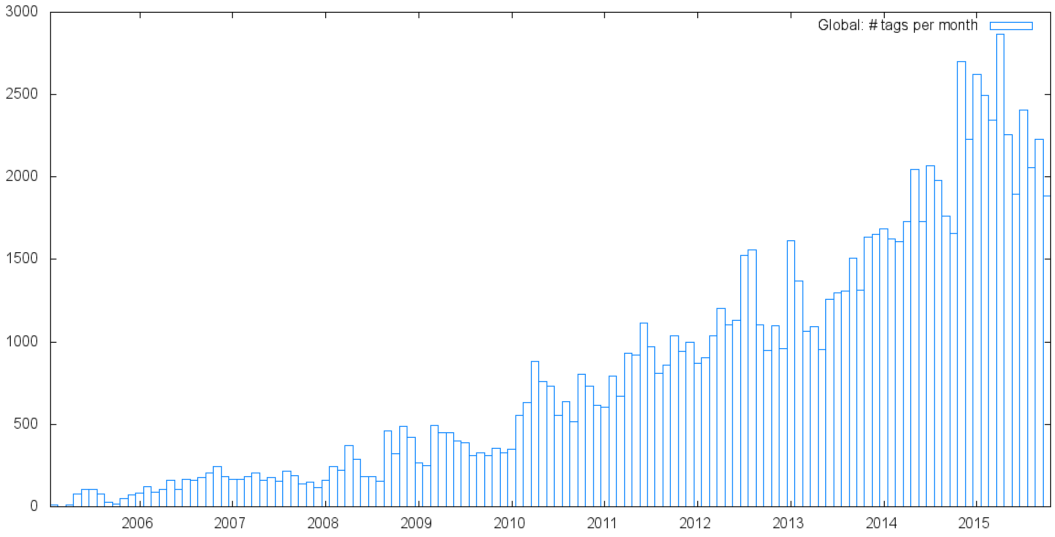


Figure 7: Monthly distribution of tags in the Flog corpus

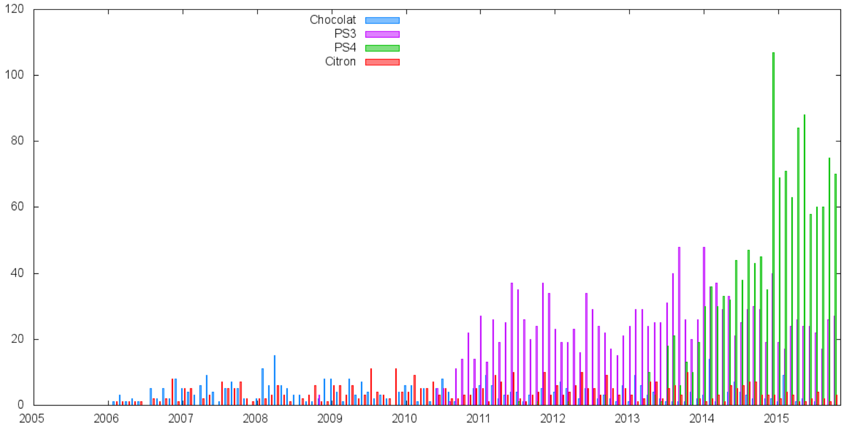


Figure 8: Monthly distribution of 4 tags in the Flog corpus: *chocolat*, *citron*, *ps3* and *ps4*

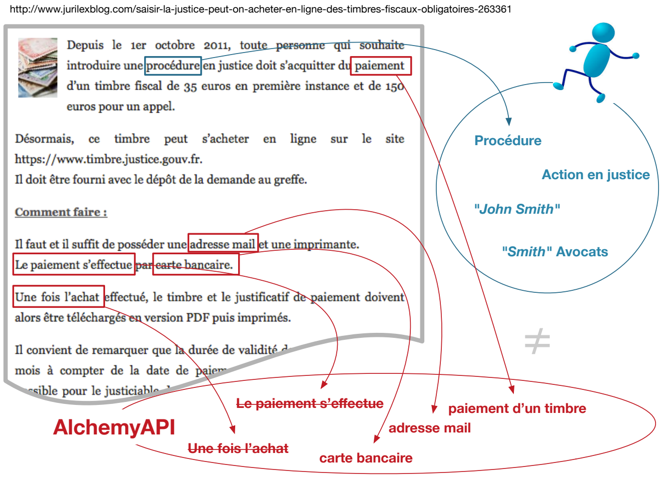
6 Automatic blog annotation

6.1 Content-based annotation

The comparison of the tags chosen by post authors and suggested by annotation tools show that they are often totally different. An example is given in Figure 9.

Apart from ill-formed tags, it appears that automatic annotation relies on textual content whereas authors sometimes choose annotations out of the vocabulary of the texts they want to annotate, as in the example of Figure 9.

Table 4 shows that the annotation strategy differs from one blog to another. In some blogs, the authors tend to choose some of the post's or blog's keywords as tags, but other authors



http://www.jurilexblog.com/saisir-la-justice-peut-on-acheter-en-ligne-des-timbres-fiscaux-obligatoires-263361

Depuis le 1er octobre 2011, toute personne qui souhaite introduire une procédure en justice doit s'acquitter du paiement d'un timbre fiscal de 35 euros en première instance et de 150 euros pour un appel.

Désormais, ce timbre peut s'acheter en ligne sur le site https://www.timbre.justice.gouv.fr.
Il doit être fourni avec le dépôt de la demande au greffe.

Comment faire :

Il faut et il suffit de posséder une adresse mail et une imprimante.
Le paiement s'effectue par carte bancaire.

Une fois l'achat effectué, le timbre et le justificatif de paiement doivent alors être téléchargés en version PDF puis imprimés.

Il convient de remarquer que la durée de validité d
mois à compter de la date de paiem
ssible pour le justiciable

Figure 9: Comparison of the tags suggested by Alchemy annotation tool and those of the post author.

prefer to enrich the content with new keywords.

Authors try at least to normalize the annotation vocabulary (always use the same tags, whatever form they may have in corpus) and whenever it is possible, they try to choose readers-oriented descriptors.

Blog	% CTags/post	# of posts	%CCats/post	# of posts
cuisine1	0.74	176	0.10	452
cuisine2	0.79	1166	0.06	927
jeuxvideo1	0.71	1421	0.47	1422
technologie1	0.74	705	0.28	1423
technologie5	0.76	235	0.12	132
droit1	0.64	450	0.62	243
jeuxvideo2	0.69	2483	0.35	1234
technologie2	0.84	108	0.45	243
jeuxvideo3	0.66	6424	0.30	5461
jeuxvideo4	0.67	1482	0.25	1501
droit2	0.56	1273	0.05	931
cuisine3	0.66	474	0.16	395
technologie3	1.00	321	0.27	343
droit3	0.47	202	0.23	283
cuisine4	0.82	1675	0.12	1561
technologie4	0.40	522	0.23	573
jeuxvideo5	0.79	1121	0.34	1134
technologie5	0.65	370	0.41	374
jeuxvideo6	0.50	175	0.19	184

Table 4: Ratio of content-based annotations: tags (CTags) and Categories (CCats) associated to a post in which their label occurs as a word or phrase, w.r.t. posts having tags or categories.

6.2 Fine-grained annotation

Annotation tools mainly extract keywords and phrases from the text of blog posts. They rely on various features to identify the elements to extract and to estimate their relevance with respect to the post to tag, but they all adopt the same extraction-based strategy.

For this reason, they usually propose only tags to authors. They do not offer different annotation grains, whereas a lot of blogs have a two level annotation system, with tags as fine-grained annotations and a (possibly) structured set of categories at a higher level of annotation.

In order to suggest categories to users, annotation system would have to detect coarse-grained topics or to learn a taxonomy based on the analysis of the blog content. This is currently out of reach of existing blog annotation tools. One of the main issue is that blogs provide only small data for each category.

6.3 Unstructured annotation

Blog annotation tools analyze the blogs as unstructured data. Most of them process the text of the blogs post by post.

This means that they do not make use of vocabulary of tags or categories which have already been used in past posts. Of course, annotations must highlight the specificity of the new post, but it is also important to show to which topics they belong to. In order to capture these properties, it seems sensible to rely on existing tags and categories and the way they have been used in the past: a simple experiments of category prediction show that tag-based approaches perform better than simple content based ones [2].

Moreover, blog annotation tools fail to take into account the inherent historical and dialogical structure of blogs. A post is not an isolated piece of text, it is a part of a sequence of posts. Novelty and position are important features for keyword extraction[14], but for blog annotation, novelty should be measured in the light of the whole blog not on from a single document basis. This is all the more important as the distribution of tags is strongly historically biased (see Section 4).

The blog context could also shed light on semantic specificity and relatedness, which are important features to take into account in keyword extraction. TF.IDF has long been acknowledged as an important metric for capturing the importance of a term in a document but also its specificity with respect to surrounding documents [8]. More recent works have shown that the centrality of a term in a cooccurrence graph reflects its importance [6]. In blog annotation, this semantic relatedness should be measured on a multi-document basis as suggested by Wan and Xiao [9].

When annotating a post, it is short-sighted to ignore the tags and categories than have already been used or the posts and comments surrounding it.

7 Conclusion

This paper presented the French Blog Annotation Corpus (FLOG) for text annotation study on the long term. It has been designed for analyzing how bloggers tags and categorize their blog posts. The corpus gathers ten years of blog posts, tags and categories (around 11 million words) on cooking, law, video games and technology. The resource includes a relational database that makes easier lexical, distributional and statistical analysis. This paper offers statistical insight

[14] In [5], they are measured through the distance of the first occurrence of the word from the beginning of the text.

on the tagging and categorization patterns of the blogging activity in French. The statistical analysis suggests that tag annotation is much more frequent than category attribution, but that categories provide a richer semantic structure for post classification and search. It also suggests that tags tend to be extracted from the blog content in a majority of cases, with some author-related exceptions. On the other hand, the review of the state of the art on automatic tag suggestion shows that tag suggestion tools are not of widespread use yet between bloggers, which might be a consequence of methods that do not take into account the past tagging history of the blog, the context of the post within the blog and the tagging pattern of each blog author. Further analysis on the FLOG corpus will include a diachronic study on tag and category evolution in order to grasp a better understanding of text annotation dynamics.

References

[1] Rebecca Blood. Weblogs: A history and perspective, 2000.

[2] Ivan Garrido-Marquez, Jorge Garcia Flores, Franqis Lévy, and Adeline Nazarenko. Blog annotation: From corpus analysis to automatic tag suggestion. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016))*, Konya, Turkey, 2015.

[3] A. Kehoe and M. Gee. Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. In *Studies in Variation, Contacts and Change in English Volume 12: Aspects of Corpus Linguistics: Compilation, Annotation, Analysis e-journal*, 2012.

[4] Craig. Macdonald and Iadh. Ounis. The trec blogs06 collection:creating and analysing a blog test collection. Technical report, University of Glasgow, Scotland, UK, 2006.

[5] Olena Medelyan, Eibe Frank, and Ian H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, volume 3, pages 1318–1327, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[6] Reda Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 404–411, July 2004.

[7] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, March 2006.

[8] Karen Sprck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[9] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI'08)*, volume 2, pages 855–860. AAAI Press, 2008.

[10] Tae. Yano, William. Cohen, and Noah A. Smith. Predicting response to political blog posts with topic models. In *Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference*, 2009.