Comparability measurement for terminology extraction

Fabien Poulard and **Béatrice Daille**
Christine Jacquin and **Laura Monceaux**
Emmanuel Morin
Université de Nantes – LINA / UMR CNRS 6241
first.last@univ-nantes.fr

Helena Blancafort
Syllabs
blancafort@syllabs.com

Abstract

In this paper we describe recent work carried out in the context of the TTC project[1] towards the automatic construction of comparable corpora for multilingual terminology extraction. We focus on the communicative intention as the variable of discourse analysis that is best suited to select Web documents valuable for terminology applications and propose a classifier based on language independent features to automatically cluster crawled documents sharing the same communicative intention. The results of our experiments indicate the need to consider more sophisticated features.

1 Introduction

The notion of comparability for a corpus is still under construction. Comparable corpora are pairs (or more) of monolingual corpora which are not necessarily translations of each others but share some characteristics (domain, genre, topic…). The degree of comparability is perceived as the amount of these common characteristics: on one extremity, we find parallel corpora and on the other extremity the independent corpora wich have nothing in common (Prochasson, 2010). The choice of the common characteristics which define the content of corpus depends on its application task. For multilingual terminology extraction, the monolingual corpora must share an important part of the vocabulary in translated forms (Déjean and Gaussier, 2002). Documents domain (including the sub-domain and the topic), genre, audience, language register, communicative intentions are also characteristics of interest.

The TTC project (Terminology extraction Translation tools and Comparable corpora) aims at leveraging machine translation tools (MT tools), computer-assisted translation tools (CAT tools) and multilingual content management tools by automatically generating bilingual terminologies from comparable corpora in five European languages (English, French, German, Spanish and one under-resourced language, Latvian), as well as in Chinese and Russian. One key objective of the project is to automate methods for building comparable corpora in specialized domains from the Web. We focus on the lexical quality of the documents as we want to select documents embedding a rich terminology.

In this paper, we report our work regarding the development of a system to automaticaly classify crawled Web documents according to several characteristics in order to ensure the monolingual comparability of automaticaly compiled corpora.

First, we present various methods used to categorize Web documents according to their genre, their discourse type or their communicative intention. Then, we present a corpus we built for this study composed of documents in seven languages from five different families, as well as the terminology we observed within. Thereafter we discuss our proposition of a classifier for communicative intentions based on language independent features. We finally discuss the results of our experiments and conclude.

2 Categorizing Web Documents

Genre is one of the various variables of discourse analysis together with domain, register, document typology, document structure, etc. It is a “social type of communicative actions, characterized by a socially recognized communicative purpose and common aspect of form” (Crowston and Williams, 2000). Kessler et al. (1997) argue that the categorization of documents should not be trained on genres as atomic entities given their heigh volatility. Instead they propose a classification of gen-

[1] http://www.ttc-project.eu/

res as “generic facets” to distinguish “a class of texts that answers to certain practical interests, and which is associated with a characteristic set of computable structural or linguistic properties”.

The genre is not the only characteristic to be considered to ensure monolingual comparability. The type of discourse (link between authors and audience, (Nakao et al., 2010; Ke and Zweigenbaum, 2009)) and the communicative intention may also be taken into consideration.

2.1 Webgenres

Deciding the genre of a Web document is a difficult task whether it must be done manually or automaticaly because the directory of webgenres is dynamic. Some genres are borrowed from traditional media, others derive from the formers, others again are emerging but are not yet well defined, others finally are spontaneous and have never been observed before. This evolutivity and the number of webgenres differenciates them from their traditional counterparts (Sharoff, 2011).

The attempts of automatic categorization of document in genre modelize the documents as “bags of words” (Dhillon et al., 2003) or combine dimension reduction (discriminative analysis, principal component analysis) and clustering (Poudat and Cleuziou, 2003) or classification (Cleuziou and Poudat, 2008). There has been several attemps to extend genre categorization to Web documents (Meyer-zu Eissen and Stein, 2004; Chaker and Habib, 2007; Dong et al., 2008; Mason, 2009; Waltinger et al., 2009). They usually combine various documents features with categorization algorithms based on machine learning techniques (support vector machines, clustering, neural networks...). Chaker and Habib (2007) group these features in four categories: metadata elements (URL, description, keywords...), presentation features (various HTML tags, links, images...), surface features (text statistics, function words, closed-class genre specific words, punctuation marks...) and structural features (parts-of-speech (POS), Tense of verbs...).

Experiments from Meyer-zu Eissen and Stein (2004) show that 70 % of the documents are assigned a correct genre.

2.2 Discourse

Goeuriot et al. (2008) have experimented the categorization of documents according to their type of discourse. They distinguished *scientific discourse* from *popular scientific discourse*. In the former, experts of a domain write for the same experts while in the latter experts or non experts write for non experts.

They propose a stylistic analysis on three levels implying deep linguistic analysis:

- The *structural* level consists of external criteria regarding the structure of the document and quantitative data (number of sentences and global size) ;
- The *modal* level consists of internal criteria caracterizing the position of the author in his writing. They considered allocutive [2] and elocutive modalities[3] inspired from Charaudeau (1992) ;
- The *lexical* level consists of internal criteria such as the presence of specific lexical units (specialized vocabulary, numbers, measure units), bibliographic elements, particular characters (brackets, other alphabet, symbols) and of quantitative data (size of the words, punctuation).

They obtain an average recall[4] of 87 % and an average precision[5] of 90 % for French documents and quite similar results for Russian (75 % recall and 87 % precision). The results on Japanese are lower with 46 % precision and 60 % recall.

2.3 Communicative intention

For Shepherd et al. (2004), the evolution of webgenres is also guided by the functional dimension of documents: browsing, emailing, searching, chatting, interacting, shopping, collaborating, etc. These communicative intentions may have a greater stability even if for annotators “the boundary between look’n’feel and communicative intentions is fuzzy” (Sharoff, 2011). Dong et al. (2008) consider the functionality of a Web document as part of its genre with its form and content. They associate for these three dimensions a particular kind of feature: stemmed terms for the content, HTML tags structuring the content (headings, tables, bullets...) for the form and HTML tags with

[2] Marks of the adressee presence.

[3] Marks of the author presence.

[4] Recall is a measure of completeness. It corresponds to the fraction of correct instances among all instances that actually belong to the relevant subset

[5] Precision is a measure of exactness. It corresponds to the correct instances among those that the algorithm believes to belong to the relevant subset.

content (applet, link, form. . .) for the functionality.

Sharoff (2011) experimented the classification of documents from the British National Corpus (BNC) according to their communicative intention (discussion, instruction, propaganda, recreation, regulation and reporting). He obtained an average precision of 83 % and an average recall of 80 %.

3 Corpus compilation

We built a multilingual corpus composed of German, English, Spanish, French, Latvian, Russian and Chinese Web documents. We present below our methodology to compile and annotate this corpus and its characteristics.

3.1 Crawled corpora

To compile the corpus, we used the first version of Babouk (de Groc, 2011), a focused web crawler (Chakrabarti et al., 1999) developed in the context of TTC to gather domain-specific corpora. To initialize the crawling, Babouk takes a list of seeds (terms or URLs) as input. During the first iteration of the crawling process, the given seeds are expanded to a large terminology using the BootCaT procedure (Baroni and Bernardini, 2004). Then, the generated lexicon is weighted automatically to build a thematic filter that is used by the categorizer in a second step to compute the relevance of webpages and filter non relevant documents. As a result, Babouk outputs a corpus consisiting of the retrived HTML files and two additional files for each HTML file:

- A Dublin Core[6] metadata file characterizing each crawled document retained for the corpus. It contains the file of the page, the seeds used for the crawling, the publisher, its original format as a mime-type, its geographic coverage, the language it is published in, the source url and the date of publication.
- A text file containing the plain text extracted from the corresponding web page.

To ensure the comparability of the corpus, we applied the same procedure to crawl the data using parallel term seeds (translation of seeds from English) in the domain of *wind energy*, a domain that is specific enough and for which corpora can be found on the web. Wind energy is one of the domains we deal with in TTC, as it is a new emerging domain for which little terminology resources exist. Other properties that may play a role in monolingual comparability, such as web genre, language register, authorship, communicative intentions and audience, are to be determined in a second step.

3.2 Inter-annotator Campain for the Annotation in English

In addition to the files and metadata produced by the crawler, we annotated other document features whose values are detailed in Table 1:

- the webpage type (consistent with the set of web page values from Montesi and Navarrete (2008)) ;
- the communicative intentions (Sharoff, 2004; Sharoff et al., 2007) ;
- the authorship (Sharoff, 2004) ;
- the audience (Sharoff, 2004) ;
- and the language register (Goeuriot et al., 2008).

Before annotating documents in the various languages, an annotation campaign was organized on a common language (English). The various annotators annotated in three phases the same 120 texts in English. After each phase, the results were analyzed and the annotation guide (Monceaux et al., 2011) was updated to improve the annotation. We measured the inter-annotator agreement (IAA) with the Kappa measure (Fleiss and others, 1971) to evaluate the reliability of the annotations.

Table 2 synthesizes the IAA rates obtained by the end of the campaign. While the agreement is moderate or fair for most of the annotations, no sufficient interannotator agreement could be reached on the author audience characteristic. In consequence, this characteristic has not been annotated in the final annotation process. It has to be noted that we do not obtain excellent agreement for the various annotations which gives an idea of the difficulty of the task.

3.3 Corpus characteristics

The webpage type, communicative intentions, authorship and language register features have been manually assigned to around 200 texts for seven languages (German, English, Spanish, French,

[6] http://dublincore.org/

Feature	Values
Webpage type	academic article, news article, adverts, legal text, expert report, report, guides, FAQs entries, catalog, glossary entries, announcement, encyclopedia entries, not text, blog entries, threads, homepages, reviews, warning, editorial, schedule, abstract, others
Communicative Intentions	information, discussion, instruction, list of something, regulation, promotion, reporting, unknown
Authorship	single author, multiple co-authors, corporate, unknown
Register	formal, informal

Table 1: Document features and their values as they are annotated on the corpus.

Annotation	Kappa	Interpretation
Web page type	0.472	Moderate agreement
Communicative Intentions	0.501	Moderate agreement
Authorship	0.513	Moderate agreement
Register	0.345	Fair agreement
Author Audience	0.097	Poor agreement

Table 2: Inter-annotator agreement for the annotated features measured with the Kappa measure and their interpretation.

Language	No. documents	No. words
German	200	285 286
English	210	209 150
Spanish	214	226 458
French	200	504 114
Latvian	225	388 098
Russian	193	318 966
Chinese	210	NA
	1 452	1 948 735

Table 3: Characteristics of the corpus.

Latvian, Russian and Chinese). These texts constitute our gold standard corpus.

Table 3 presents the main features of the corpus: the number of documents and the number of words for each language. This corpus is composed of almost two million words in seven languages. The texts have all been converted into utf-8 for convenience. Every document is stored in the corpus as an HTML file, a text file and an XML file containing the metadata and the annotations.

4 Corpus Analysis

After the corpus annotation task, we started to analyze the terminology that we extracted from the corpora. We observed a correlation between the kind of terminology and the communicative intentions.

The richest terminologies were found in the documents with informative, promotive and regulative intentions, each one whith a specific type of terminology. Informative documents i.e. show a rich technical terminology: *rotor bobiné*, *circuit rotorique* or even *multiplicateur de type planétaire épicycloïdal* for French, and *vertical axis turbines*, *Horizontal Axis Wind Turbines (HAWT)* or *Diffuser-Augmented Wind Turbines (DAWT)* for English.

The terminology of documents aiming at promotion make reference to products, such as named entities (name of products such as *Product Model:BF-H-500*), their constitutive element (*glass fiber reinforced plastic*) and their localizations (*parc éolien de Teterchen*).

As expected, documents aiming at regulation embed a legal terminology with terms such as *unacceptable harm*, *bienes inmuebles*, *impactos ambientales* and *planeamiento urbanístico*.

The documents with other communicative intentions show less numerous terms. Still, we found some terms in documents aiming at discussion, namely documents discussing the pros and contras of the installation of wind generators : *nuisances sonores* (*noise*) or *bruit mécanique* (*mecanical noise*).

Unfortunately, the various communicative intentions are not equally present and reachable on the Web as shows the Figure 1 representing their distribution among our corpora. Hence, discussion, information, reporting, promotion and list of something are the principal communicative intentions found in the corpus while regulation is mostly invisible. Therefore communicative intentions may be interesting features to choose documents relevant for terminology applications. They both allow the selection of documents with a rich terminology and enable to differentiate several kinds of terminology.

5 Classifying Web Documents Using Language Independent Features

We believe that the monolingual comparability of a corpus can be achieved by controling the domain and the communicative intention of the documents it is composed of. As we discussed in the previous section, it is possible to crawl documents belonging to the same domain. However, we do not have tools to predict the communicative intention of a document.

We face two main challenges to build a classifier for communicative intention in our context:

- we work with a relatively wide specialized domain with few resources and no scientific journals ;
- we must handle several distant languages with the same method and therefore are limited to features without any linguistic anchor.

5.1 Proposition

We propose to use supervised learning to predict the communicative intention of a document. Given the distant languages we deal with, we need a language independent method and therefore only use very shallow text features for the classification. Among the features experimented in the litterature, we selected the URL, the page layout, char ngrams and some other quantitative features.

We represent the URL as a bag of words by splitting it in sequences using special characters as delimiters (/, ., _, #, &...). The extracted sequences are normalized using unicode. For each document we obtain a vector of booleans indicating if any of the collected words is present in the URL of the document.

The page layout of the documents is constrained by the HTML tags. We compute the distribution, in terms of frequencies, of such tags. Preliminary experiments shown that it is preferable to only consider structuring tags (p, h1, ul, li...).

We also use bags of character ngrams. Like for URL we build vectors of booleans indicating if the associated ngram is present in the document. The best discrimination is offered by ngrams composed of four characters.

Finally, we used quantitative features such as the size of the document, the number of words[7] and their average size, the distribution of these words according to the unicode category of the characters they are composed of, the number and average size of sentences, ...

5.2 Experiments and results

We experimented two supervised learning approaches: a clustering one (k-Means) and a categorization one (SVM).

Using k-Means, we want documents to form cluster for each communicative intention. Therefore we compute a centroid for each communicative intention, using training data. Then communicative intention values are associated to documents depending on the centroid they are the closest to.

On the other side, SVM (Support Vector Machines) computes hyperplanes where the density of documents for each communicative intention is the highest while maximizing the margin between documents of different communicative intentions. Then communicative intention values are associated to documents depending on the hyperplane they belong to.

We experimented both learning algorithms with our language independent features. It results that the choice of the method has virtually no impact on the result and therefore we only present the results obtained with SVM in Table 4. A classifier is built for each language and evaluated with micro-precision, micro-recall and micro-f-score that is

[7] As we refuse the use of language specific tools, we consider as a word a sequence of characters sharing the same unicode category.

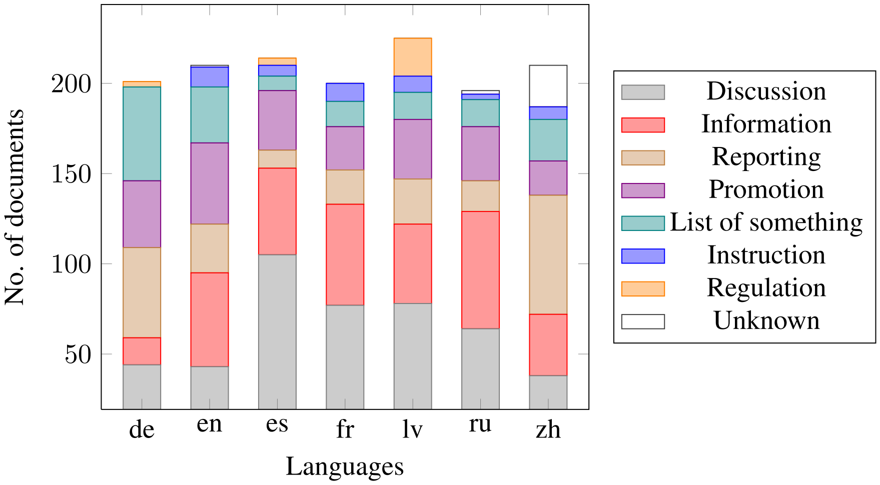

Figure 1: Distribution of the communicative intentions in terms of number of documents for each language composing our corpus.

Language	Precision	Recall	F1-score
English	25,2 %	25,8 %	13,3 %
French	39,8 %	39,8 %	24,9 %
German	6,8 %	25,8 %	10,8 %
Spanish	39,4 %	50,4 %	34,8 %
Latvian	52,2 %	41,0 %	30,6 %
Russian	32,2 %	33,4 %	20,8 %
Chinese	47,5 %	36,4 %	24,1 %

Table 4: Results obtained with SVM for each language.

the computation of these measures on the contingency table including all classes (all communicative intentions). As the various communicative intentions are not equaly distributed in the corpora, we run the evaluation with a 3-folds stratified cross-validation which preserve the same distribution of the communicative intentions among the various folds.

All the results are low which may indicate that the communicative intention is not language independent. The variations of the results between the languages mainly reflects the distribution of the communicative intentions among the documents as well as the lack of homogeneity between each monolingual corpus.

6 Conclusion

For comparable corpora extracted from the Web using a crawler for terminology oriented applications, it is important to categorize the documents with regards to terminology, named entities… Communicative intentions may be interesting features as they may allow to differentiate lexical items. Hence, informative documents should contain specific domain terminology, documents with promotion intentions should contain brand names, and regulative documents the legal terms.

In order to classify documents according to their communicative intention, in this paper we run an experiment with language independent features that seem relevant to other categorization tasks such as webgenre or discourse type. To classify documents written in seven languages belonging to five different families, we used features based on the URL, the page layout and characters ngrams. The experiments showed that these language independent features are not sufficient to distinguish communicative intentions.

More sophisticated features, including deeper linguistic features, should be considered and would require linguistic preprocessing. The best results on web genres classification make use of part-of-speech tagging while for discourse classifications very subtile features such as modality marks are used. Sharoff (2011) obtained better results in classifying English and Russian documents according to their communicative intentions using deeper linguistic features.

Another consideration is that maybe our hypothesis that the classification should be placed between the crawl process and the terminology extraction is not valid after all. Terminology may be necessary to predict the communicative intention and not the other way around.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no°248005.

We thank all the annotators (Ana Laguna, Maya Abboud, Somara Seng, Tatiana Gornostay, Iveta Keiša and others) for the time they have spent on the construction of the corpus as well as Jérôme Rocheteau for the development of the handy annotation tool OUSIA[8].

References

Marco Baroni and Silvia Bernardini, 2004. *BootCaT: Bootstrapping corpora and terms from the web*, volume 4.

Jebari Chaker and Ounelli Habib. 2007. Genre categorization of web pages. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 455–464, Oct.

Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640.

Patrick Charaudeau. 1992. *Grammaire du sens et de l'expression*. Hachette.

Guillaume Cleuziou and Céline Poudat. 2008. Classification de textes en domaines et en genres en combinant morphosyntaxe et lexique. In *Actes de TALN*, number 1, pages 9–13. ATALA.

Kevin Crowston and Marie Williams. 2000. Reproduced and emergent genres of communication on the world-wide web. *The Information Society*, 16(3):201–216.

Clément de Groc. 2011. Babouk - exploration orientée du web pour la constitution de corpus et de terminologies. In *22es Journées francophones d'Ingénierie des Connaissances (IC'2011)*, May.

Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. 2003. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3(7-8):1265–1287, Oct.

Lei Dong, Carolyn Watters, Jack Duffy, and Michael Shepherd, 2008. *An Examination of Genre Attributes for Web Page Classification*, pages 133–143. IEEE Computer Society, Jan.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Lorraine Goeuriot, Natalia Grabar, and Béatrice Daille. 2008. Characterization of scientific and popular science discourse in french, japanese and russian. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, number 1, pages 2933–2937. European Language Resources Association (ELRA).

Guiyao Ke and Pierre Zweigenbaum, 2009. *Catégorisation automatique de pages web chinoises*, pages 203–228. ARIA.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*.

Jane E. Mason. 2009. *An n-gram based approach to the automatic classification of web pages by genre*. Ph.D. thesis, Dalhousie University.

Sven Meyer-zu Eissen and Benno Stein. 2004. Genre classification of web pages. In S. Biundo, T. Fruhwirth, and G.Editors Palm, editors, *Advances in Artificial Intelligence (KI 2004)*, pages 256–269. Springer, Berlin Hedelberg New York.

Laura Monceaux, Christine Jacquin, and Béatrice Daille. 2011. Guidelines for monolingual annotation.

Michela Montesi and Trilce Navarrete. 2008. Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Inf. Process. Manage.*, 44:1410–1430, July.

Yukie Nakao, Lorraine Goeuriot, and Béatrice Daille. 2010. Multilingual modalities for specialized languages. *Terminology*, 16(1):51–76, May.

Céline Poudat and Guillaume Cleuziou. 2003. Genre and domain processing in an information retrieval perspective. In *Proceedings of the 3rd International Conference on Web Engineering (ICWE 2003)*, pages 399–402. Springer-Verlag Berlin Heidelberg.

Emmanuel Prochasson. 2010. Alignement multilingue en corpus comparables spécialisés.

Serge Sharoff, Bogdan Babych, and Anthony Hartley. 2007. "Irrefragable answers" using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

[8] `http://code.google.com/p/dublin-core-ousia/`

Serge Sharoff. 2004. Analysing similarities and differences between corpora. In *Proceedings of the 7th Conference of Language Technologies (Jezikovne Tehnologije)*, volume 83.

Serge Sharoff, 2011. *Chapter 7 - In the Garden and in the Jungle*, volume 42. Springer Netherlands.

Michael Shepherd, Carolyn Watters, and Alistair Kennedy. 2004. Cybergenre : Automatic identification of home pages on the web. *Journal of Web Engineering*, 3(3-4):236–251.

Ulli Waltinger, Alexander Mehler, and Armin Wegner. 2009. A two-level approach to web genre classification. In Joaquim Filipe and JoséEditors Cordeiro, editors, *Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, pages 689–692. INSTICC Press.