Syllabs Team at CLEF MC2 Task 1: Content Analysis

Olivier Hamon[1], Chloé Monnin[1] and Claude de Loupy[1]

[1] Syllabs, 35 rue Chanzy, 75011 Paris, France
{hamon,monnin,loupy}@syllabs.com

Abstract. This paper describes the participation of the Syllabs Team in the content analysis task of the CLEF MC2 Evaluation lab. In the current state of our work, we offer preliminary solutions to first detect the language of the microblogs used within the task, then extract the named entities that will be later used to recognize Wikipedia entities and finally, detect microblogs that deal with festivals.

Keywords: Language Detection, Named Entity Recognition, Clustering, Festival Detection, Wikipedia Entity Recognition.

1 Introduction

The presence of festivals on the social media is constantly increasing. An analysis[1] of the 30 biggest festivals in France made in 2015 was already showing that 97% of them were holding a Twitter account, while the number of followers was larger77% than in 2014. Tweets are a wonderful mean to be visible for events such as festivals. However, there are a few limitations about using Twitter: first of all, a single tweet may provide partial information due to the limitation of 140 characters; secondly, there is a lot of noise on Twitter (advertising, close events, etc.), or thirdly, important information may be drawn among the huge volume of tweets produced.

The first task – Content Analysis – of the MC2 evaluation [1] consists in analysing tweets so as to prepare their understanding by a festival participant. Therefore, building context is particularly important to help filtering relevant information.

Our goal[2] when participating in the MC2 evaluation task is threefold: firstly, by evaluating the Syllabs in-house technologies we hope to improve their performance, especially regarding linking data and Wikipedia recognition. Secondly, by experimenting festival detection we try to settle a method for event detection, based on data gathering, and that could lead to event analysis or specific information retrieval. Lastly, by working on microblogs we continue our exploration on limited content data that can be rich in relevant information and usable data.

In Section 2, we describe our current approaches regarding the different subtasks of CLEF MC2 Lab we participated in.

[1] http://www.socialband.fr/docs/les_festivals_et_les_reseaux_sociaux_en_2015.pdf
[2] http://www.agence-nationale-recherche.fr/?Project=ANR-14-CE24-0022

2 Objectives

We participated in three tasks of the MC2 Lab:

- Task 1.1: Filtering microblogs dealing with festivals;
- Task 1.2: Language Detection;
- Task 1.5: Wikipedia Entity Recognition.

Our main stream is starting with Task 1.2, since Language Detection is the precondition to the two other tasks as it helps us focus on the analysis of a specific language.

3 Data Preparation and Preprocessing

3.1 Used data

Specific data used for the training of Language Detection is detailed in Section 4.1.

Regarding the MC2 data, and due to time constraints, we only worked on text and kept aside metadata. Thus, we reduced the tuning work, but also the performance gains that metadata could have provided. Test data is of course processed to submit the results, and we also used the full stream of June 2016 for Task 1.1 to help the filtering of tweets dealing with festivals. While the test set is composed of 1,100 tweets, the June 2016 set is much bigger with more than 4.3 million tweets.

3.2 Preprocessing

Analyzing microblogs needs specific context and treatment due to the peculiar nature of the data. Information is reduced to its simplest concept and the shortness of each microblog makes this information hard to retrieve.

Noise is one of the most relevant characteristics to filter. In microblogs, noise usually prevails over relevance but the few pieces of useful information can have a strong impact.

Thus, data preparation, noise filtering and preprocessing are essential parts of the full process, and vital for the following subtasks.

Basically, we prepared the input data using the following steps:

- Removing hashtags and nicknames, specifically for language detection: although they convey meaning, we suppose that most of the hashtags and nicknames do not disambiguate languages (in French, for instance, there are plenty of them written in English or pseudo-English);
- Removing all symbols and punctuation marks: although some of those characters could help for language detection, such as "¿" used in Spanish, our tests showed that they were not relevant;
- Keeping one character when it is duplicated more than twice ("okkkkkk" for instance) so as to better fit the "traditional" language models;
- Not considering case.

4 Tasks and methodology

4.1 Language Detection (Task 1.2)

Many methods and studies exist to detect the language of textual contents, including microblog-oriented data. The specific nature of tweets, through their size and specific vocabulary, makes the task particularly complex. Microblogs are usually coming with metadata but language is generally not reliable, firstly because the language of a given writer is not always the same, secondly because the geolocalization cannot determine the language of a tweet. Therefore, we need to apply other methods to detect the language of a given microblog.

We used, adapted and tested four different methods (cf. Sections 4.1.1 to 4.1.4) that are not specific to microblogging. The main difficulty is to find a training corpus that would simulate microblogging. We describe the corpus in the following section and the results of our experimentations are detailed below.

4.1.1 N-gram distribution

In our first experiment we tested a very simple method using microblog n-grams, implemented from scratch. Microblogs being, by definition, short, we took away the possibility to use word n-grams, and thus we focused on letter n-grams.

Then, we first built a small microblogs training corpus, composed of 3 sources:

- The corpus from the SEPLN Workshop 2014 [2], representing 70k tweets written in Basque, Catalan, Galician, Spanish, English and Portuguese. At first, we used the full corpus for our tests, but finally we only kept Spanish, English and Portuguese, the size of the other languages being too small.
- More than 50k English tweets coming from the Crisilex corpus [3].
- An internal Syllabs corpus containing more than 80k tweets in French and English.

Thus, the whole corpus represents more than 200k annotated tweets. 95% of this corpus was arbitrarily used for the model training, the other 5% being used for testing purposes. Table 1 shows the results obtained using from 2-grams to 7-grams on the test corpus.

Table 1. Language Detection results on the test corpus, per n-gram set [% found]

2-grams	3-grams	**4-grams**	5-grams	6-grams	7-grams
73.66	82.70	**87.61**	86.19	81.85	73.00

After those tests, we determined that using 4-grams gives the best results.

4.1.2 LangID

LangID [4] uses a naïve Bayesian model that computes n-grams from one to four letters. Training is done on several corpora such as the JRC-Acquis, ClueWeb 09, Wikipedia, Reuters RCV2, Debian i18n and for 97 languages. Thus, there is no specific training corpus for microblogs. We did not make additional training.

4.1.3 Compact Language Detection 2 (CLD2)

CLD2 [5] is the available language detection tool coming from Google and using a naïve Bayesian classifier too. It is based on 4-grams and is available for 83 languages.

4.1.4 Guess Language

Guess Language [6] is a method using 3-grams for 60 languages.

4.1.5 Evaluation and results

The test corpus was filtered to keep 5 languages: French, Catalan, English, Portuguese and Spanish. Table 2 shows the associated figures.

Table 2. Language representation in the test corpus

Language	#Tweets
Catalan	1,493
Spanish	12,853
French	129
English	10,987
Portuguese	2,169

The low number of French tweets can be explained by both the lack of data and the need to keep part of the French tweets for training.
We computed precision so as to estimate the performance of the method. We also tracked the computation time for all the methods. Results can be found in Table 3.

Table 3. Evaluation of the different language detection methods

Method	Precision [0-100]		Time (s)	
	With cleaning	Without cleaning	With cleaning	Without cleaning
N-gram distribution	87.61	87.26	2.33	2.35
LangID	93.26	92.90	4.04	3.95
CLD2	86.63	86.02	0.49	0.55
Guess Language	73.71	74.70	65.89	103.17

Our basic method obtains good results despite of its simplicity, and we guess that with more training data, results could be comparable to other methods. On the other hand, our method is more adapted to the test since it was trained with the same limited languages while other methods were trained with more languages.

Higher results are obtained by LangID with more than 92% correctly identified. Considering the low quality of some tweets, this result is quite good.

The cleaning we apply to the tweets shows a slightly improvement of the performance, except for the Guess Language method, while the timing remains similar.

One last interesting result is the timing of our method, although CLD2 obtains way better results, with a comparable precision and more languages available.

In the context of the CLEF MC2 Lab, we use the LangID method with cleaning, since it obtains a higher performance.

4.2 (Wikipedia) Entity Recognition (Task 1.5)

We have used the named entity recognition system that was developed by Syllabs in this task. It is a rule-based system which has been used for years but not adapted to tweets. Since tweets are short, uppercase is not a relevant clue to find named entities. So, we had to improve our lexicon. We added a large number of organisms, places and persons. In the last few months, more than 1,5M entries have been added to our system, plus a few rules regarding new extractions for specific contexts, adaptation to new entry features.

Regarding the specific task of the CLEF MC2 Lab, we have simply used the Wikipedia API[3] to link our entities to the encyclopedia and translate them into other languages, for English, French, Portuguese and Spanish.

For each tweet in a given language, we provide the original entity extracted, and the translation for the other languages.

4.3 Filtering Microblogs dealing with Festivals (Task 1.1)

Filtering contents to a certain type of event is particularly complex especially when dealing with microblogs. Most of the time, in microblogs, the information given is not relevant enough to do that filtering. Hashtags may help, as well as specific keywords, but we cannot count on them for most of the microblogs.

Moreover, the concept of festival and how people write about a given festival are hard to define. A person who is happy to be in Cannes during the festival may be in Cannes for another reason. Depending on the context, we may, or may not, keep that kind of microblog.

Our method to filter microblogs works on a full microblog corpus (vs a single microblog) and is based on clustering the microblogs. It applies the following steps:

[3] https://www.mediawiki.org/wiki/API:Main_page

1. Removing duplicates and retweets, while keeping track of them: We then avoid biasing the clustering by using similar tweets, as well as giving very large weights to the information given;
2. Detecting the languages (from Task 1.2) on the remaining microblogs: The language detection is twofold, first by reducing the clustering process on a large data set, second by doing a first gathering of similar linguistic information;
3. Clustering for a first time the microblogs on a time frame (i.e. a month or a week), per language: A DBSCAN clustering is applied that, according to our experiments, gathers similar (i.e. very close to duplicates) tweets and thus reduces our data set again [7];
4. Clustering for a second time the microblogs on a time frame (i.e. a month or a week), per language: Another pass of a DBSCAN clustering is applied that tends to gather similar *topics*;
5. Using a lexicon to extract clusters which are supposed to be dealing with festivals: By this and the previous steps, we extend the context of single tweets and hope to find more tweets related to festivals;
6. Extracting the microblogs of the test set from the clusters: We only keep those tweets from the test corpus that deal with the festival topic.

By regrouping *topics* on the fourth step, we increase the possibility of finding tweets related to festivals that could not have been found by simply extracting festival-related ones.

5 Results on the test set

We submitted results on the three tasks presented above. Although evaluation results are not available yet, we provided a few statistics below.

5.1 Language Detection (Task 1.2)

The 1,100 tweets from the test corpus have been preprocessed, then analyzed using LangID. Results are shown in Table 4:

Table 4. Statistics on tweets per language

En	Es	Pt	Fr	It	Ja	Id	De	Nl	Tr	Ca	Tl	Ms	Ko
675	166	61	56	26	21	16	16	13	9	8	5	5	3
Hr	Th	Ru	Pl	La	Zh	Sw	Sv	Nb	Gl	Fi	Eu	Bn	An
3	2	2	2	2	1	1	1	1	1	1	1	1	1

Tweets are mainly written in English (more than half of the test corpus), then Spanish, Portuguese, French, Italian, Japanese, etc. More unusual languages are also present in the corpus, such as Aragonese, Bengalî or Bokmål.

5.2 (Wikipedia) Entity Recognition (Task 1.5)

The Wikipedia Entity Recognition has been processed on English, French, Portuguese and Spanish. Table 5 shows the entities found per language.

Table 5. Statistics on Wikipedia Entity Recognition per language

Language	#Tweets	#Entities	#Unique entities
English	318	524	389
French	3	3	3
Portuguese	13	18	17
Spanish	5	5	5

339 tweets contain at least one Wikipedia entity and a total of 550 entities (for 414 unique entities) have been found. Translation is possible only when the link in Wikipedia is available.

Unfortunately, our Entity Recognition system has found many more Entities that do not exist in Wikipedia. This is certainly due to the less well-known events or specific information included in the corpus.

5.3 Filtering Microblogs dealing with Festivals (Task 1.1)

By using our method, we have found that 734 tweets are dealing with festivals. Most of them are tweets in English (439), then Spanish (124), Portuguese (36), French (31), etc.

6 Conclusions

This article describes the first experimental results on three tasks of the MC2 evaluation Lab. We mainly enhanced our stream system by adding preprocessing that allowed us to improve the results of the Language Detection slightly.

Wikipedia Entity Recognition and Microblogs filtering are both basic methodologies and, although first results seem promising, they can be improved in many ways. Our rule-based system for Entity Recognition requires more lexicon and further work should be done on the rules, while the linking with Wikipedia could be done using disambiguation or word variation techniques. The clustering process shows limitations to filter microblogs, especially because we need a large amount of tweets, which leads to performance decrease.

We will continue this preliminary work within the “gallery of festival” project (GaFes) so as to extract proper content and to be able to represent festivals in a social context.

References

1. Ermakova L., Goeuriot L., Mothe J., Mulhem P., Nie J.-Y., and SanJuan E., CLEF 2017 Microblog Cultural Contextualization Lab Overview, International Conference of the Cross-Language Evaluation Forum for European Languages Proceedings, LNCS volume, Springer, CLEF 2017, Dublin (2017)
2. Zubiaga, A., San Vicente, I. n., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V.: Overview of TweetLID: Tweet language identification at SEPLN 2014. TweetLID@SEPLN (2014).
3. Olteanu A., Castillo C., Diaz F., Vieweg S.: CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In: Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM'14). AAAI Press, Ann Arbor, MI, USA (2014)
4. Marco, L and Baldwin, T.: Cross-domain Feature Selection for Language Identification. In: Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp. 553—561 (2011)
5. CLD2 Homepage, https://github.com/CLD2Owners/cld2
6. Guess Language Homepage, https://github.com/kent37/guess-language
7. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231 (1996)