Twitter as a Comparable Corpus to build Multilingual Affective Lexicons

Amel Fraisse, Patrick Paroubek

LIMSI-CNRS
Bât 508, Université Paris-Sud XI
fraisse@limsi.fr, pap@limsi.fr

Résumé

The main issue of any lexicon-based sentiment analysis system is the lack of affective lexicons. Such lexicons contain lists of words annotated with their affective classes. There exist some number of such resources but only for few languages and often for a small number of affective classes, generally restricted to two classes (*positive* and *negative*). In this paper we propose to use Twitter as a comparable corpus to generate a fine-grained and multilingual affective lexicons. Our approach is based in the co-occurence between English and target affective words in the same emotional corpus. And it can be applied to any number of target languages. In this paper we describe the building of affective lexicons for seven languages (en, fr, de, it, es, pt, ru).

Keywords: Affective Lexicon, Comparable Corpus, Sentiment Analysis

1. Introduction

Research in Sentiment Analysis and Opinion Mining, has flourished in the past years. The growing interest in processing emotions and opinions expressed in written text is motivated by the birth and rapid expansion of the Social Web that made it possible for people all over the world to share, comment or consult content on any given topic. In this context, opinions, sentiments and emotions expressed in Social Media texts have been shown to have a high influence on the social and worldwide economic behavior. In spite of the growing body of research in the area in the past years, dealing with affective phenomena in text has proven to be a complex and interdisciplinary problem that remains far from being solved.

As any emergent field, its challenges include the need to develop linguistic resources to perform computational tasks. In our case, we are interested in the sentiment classification task which is performed either with statistical approaches or with lexicon-based approaches. In the two cases, the lack and the scarcity of affective lexicons present a real issue for sentiment analysis system. Multilingual affective lexicons are central components for cross-lingual sentiment analysis systems. Their manual construction is a hard, long and costly process. While often it is impossible to consider for most under-resourced languages because of the scarcity or even lack of experts. Existing affective lexicons are always monolingual and often developed for English. Furthermore, many of these lexicons are very simple, i.e. they consist of a list of words divided into only two classes : *positive* and *negative*. To our knowledge, there is no fine grained affective and multilingual lexicons.

Most previous work addressing the problem of bilingual lexicon extraction are based on parallel corpora. However, despite serious efforts in the compilation of corpora (Armstrong and Thompson, 1995), (Church and Mercer, 1993), to our knowledge, there is no available affective parallel corpus for the field of sentiment analysis.

On the other hand, with the rapidly growing volume of resources on the Web, the acquisition of non-parallel texts is usually much easier. Thus, as mentioned by (Rapp, 1995) and (Rapp, 1999) it would be desirable to have an approach that can extract lexicons from comparable or even unrelated texts. In this paper, we propose to use Twitter as a comparable corpus to extract multilingual affective corpus. Our approach is motivated by the fact that, nowadays, social media user's and in particular twitter users' express and share their sentiments, opinions and emotions on a variety of topics and discuss current issues over the world. In fact, many people can talk about the same event and describe their emotional state triggered by this event in different languages. Hence, Twitter could be considered as a comparable corpus as we could group tweets (messages written by users) by emotion/opinion/sentiment expressed in different languages. We have tested our approach to build seven affective lexicons for English, French, German, Spanish, Italian, Portuguese and Russian.

2. Related Work

There are two ways to cover the lack of sentiment analysis resources. The first way is to create manually a lexicon in a source language as (Bradley and Lang, 1999) who developed the Affective Norms of English Words (ANEW) which is a set of normative emotional ratings for 1034 English words. And then localize the source lexicon into target languages.

(Redondo et al., 2007) have adapted the ANEW into Spanish, (Vo et al., 2009) localized it into German. This approach requires human translators to ensure the quality of the localized resource and therefore is cost expensive and not scalable.

(Strapparava and Valitutti, 2004) developed the WordNet Affect which is a manually created extension of the WordNet, including a subset of synsets suitable to represent affective concepts correlated with affective words. The second approach is automatic construction of a lexicon. The most common method is bootstrapping. This method starts with seed words with a known polarity (e.g. good, happy, wonderful for a positive class, bad, sad, terrible for a negative class). Next, the seed words are used to find related words and assign them the same class or estimate their po-

larity.
(Qadir and Riloff, 2013) present a bootstrapping algorithm to automatically learn English twitter hashtags that convey emotion. (Mohammad, 2012) use the pointwise mutual information to measure the association between a word and a given emotion. So he builds a word emotion association lexicons which are lists of words and associated emotions. For example, the word *victory* may be associated with the emotions of *joy* and *relief*.
(Pak and Paroubek, 2010) proposed to use Twitter to collect a dataset of emotional texts in French. Using the collected dataset, they estimated the affective norms of words present in the corpus and built a polarity classifier. Both for manual and automatic approaches, existing affective lexicons are always monolingual.

3. Word-Opinion/Sentiment/Emotion association lexicon

In a previous work (Fraisse and Paroubek, 2013), we have presented 20 semantic categories including all types of emotions, sentiments and opinions. Each semantic class correspond to one type of emotion/sentiment/opinion and is referred to by means of a multi-word label that regroups various subjective words generally associated to one of the various sentiments contained in the considered class (Table 1). For example the *Anger* label includes the *impatience, annoyance, irritation, nervousness, anger, exasperation* semantic categories. For each of the 20 Opinion/Sentiment/Emotion presented in the Table 1, our aim is to build the associated lexicon for each of the seven languages addressed in this paper.

#	Label	Dim.	uComp Semantic Category
1	NEGATIVE SURPRISE	e-	negative surprise / negative amazement
2	DISCOMFORT	e-	discomfort / disturbance / embarassment / guilt
3	FEAR	e-	shyness / worry / apprehension / alarm fear / terror
4	BOREDOM	e-	boredom
5	DISPLEASURE	e-	displeasure / deception / abuse
6	SADNESS	e-	sadness / resignation / despair / sorrow / hopelessness
7	ANGER	e-	impatience / annoyance / irritation / nervousness / anger / exasperation
8	CONTEMPT	e-	reluctance / contempts / disdain / blame / disgust / hate
9	DISATISFACTION	s-	disappointment / disatisfaction / discontent / shame
10	DEVALORIZATION	o-	disinterest / devalorization / depreciation
11	DISAGREEMENT	o-	disapproval / disagreement
12	VALORIZATION	o+	interest / valorization / appreciation
13	AGREEMENT	o+	understanding / approval / agreement
14	SATISFACTION	s+	satisfaction / contentment / pride
15	POSITIVE SURPRISE	e+	positive surprise / positive amazement
16	APPEASEMENT	e+	relief / appeasement / peacefullness forgiveness / thankfullness
17	PLEASURE	e+	pleasure / entertainment / enjoyment / joy / happiness / euphoria / play
18	LOVE	e+	love / affection / care / tenderness / fondness / kindness / attachment / devotion / passion / envy / desire
19	INFORMATION	i	information / announcement / news / demand / query / question
20	INSTRUCTION	i	recommandation / suggestion / instruction / order /command

TABLE 1 – uComp semantic categories of opinion/sentiment/emotion, e=emotion, s=sentiment, o=opinion, i=information, +=positive valence, -=negative valence.

For each label of the Table 1 and for each language, we whish to extract the associated lexicon. Table 2 illustrate an example of comparable tweets in four languages ; the four tweeter talked about the same topic *violence in ukraine* expressing the same emotion *Sadness* in different languages.
So, based on such data our approach aims to extract, across different languages, and for each affective label the associated lexicon (Table 3).

	**#Ukraine #death** toll rises as clashes continue **#sad #grief**.
	**#Ukraine** 60 **morts** aujourd'hui ! ! ! c'est vraiment **#triste #chagrin**
	Stop the **#violence** in **#Ukraine** 60 **#tod** heute **#traurig**
	Impresionantes imagenes de **#Kiev** que pasarian por fotogramas de una pelicula de guerra ! ! muchos **#muertos** ! ! estoy **#triste**

TABLE 2 – Example of comparable tweets

Affect. Label	Associated Words			
	English	French	German	Spanish
SADNESS	Sad Death Grief	Triste Mort Chagrin	Traurig Tod	Triste Muertos

TABLE 3 – Multilingual affective lexicon associated to the tweets described in the Table 2

4. Our approach for multilingual affective lexicon construction

Hashtags are a distinctive characteristic of tweets (Jackiewicz and Vidak, 2014). They are a community created convention for providing meta-information about a tweet. Hashtags are made by adding the hash symbol # as a prefix to a word. Thus, a hashtag is simply a way for people to search for tweets that have a common topic. In general, the tweeter (one who tweets) use emotion-word hashtag, to notify others of the emotions associated with the message he or she is tweeting. Consider the tweet bellow :

Oh okay all the people I fancy are taken ...that's cool watch them be happy as I sit in a corner and cry #sad

The tweeter has used the emotion word hashtag #sad, to convey that he or she is sad. And as English is considered the reference language on the Web, the tweeter use generally the emotion word hashtag in their native languages and give the corresponding English one as shown in the following French tweet :

Je suis vraiment #triste aujourd'hui #sad.

So, our approach is based on the co-occurence between the English and the target emotion word hashtags in the tweets. To achieve this, we proceed in two steps ; firstly we construct emotional corpora in the following seven languages : English, French, Spanish, German, Italian, Portuguese and Russian. Secondly, We extract affective lexicon

Anger	Fear	Love
#anger	#fear	#love
#rage	#terror	#affection
#irritation	#shyness	#care
#nervousness	#worry	#tenderness
#impatience	#apprehension	#fondness
#annoyance	#terrified	#kindness
#angry	#alarm	#attachment
#edgy	# scare	#devotion
#exasperated	#scared	#passion
#irritated		#envy
#annoyed		#desire

TABLE 4 – Seed Affective word for Anger, Fear and Pleasure affective classes

Affect.Cl.	En	Fr	De	It	Es	Ru	Pt
DISCOMFORT	551	232	65	33	157	10	63
FEAR	1677	156	123	15	488	35	124
DISPLEASURE	1617	645	13	7	74	6	15
SADNESS	283	211	204	209	459	110	272
ANGER	1690	73	9	16	198	102	43
CONTEMPT	506	606	53	43	310	69	68
PLEASURE	2414	1952	1639	1099	2082	664	1198
LOVE	2452	434	595	632	2251	1369	933

TABLE 5 – Number of document per affective class and per language.

from the collected corpora based on the co-occurence between English and target emotional hashtags in the same affenctive class.

4.1. Corpora collection

Data collection from the Web usually involves crawling and parsing of HTML pages which is a solvable but at the same time a consuming task. In our case, collecting data from Twitter is much easier since it provides an easy and well-documented API [1] to access its content. In this work, we selected from the Table 1 the 8 prominent affective classes that are frequent in tweets : *Negative surprise, Anger, Sadness, Fear, Displeasure, Boredom, Positive surprise, Pleasure and Love*. For each selected class we have defined a list of English seed emotional words that are commonly used by English speakers to express their affective state on Twitter.

Table 4 presents an extract of English seed emotional words that are used for the three affective classes *Anger*, *Fear* and *Love*. Then, we supplied the Twitter Search API with the English emotional hashtags queries and collected tweets written in their native languages and containing at least one hashtag of the English list. In fact, we noticed that when a user writes an affective tweet, he or she uses an emotional word hashtag in his or her native language and he or she, also, gives the corresponding English word.

The characteristics of the gathered corpus are presented in the Table 5.

4.2. Lexicon construction

In the preprocessing of the collected corpus, we discarded tweets with the prefix *Rt, RT, and rt*, which indicate that the tweet that follow are re-tweets (re-postings of tweets sent earlier by somebody else).

1. Twitter API : https ://dev.twitter.com/docs

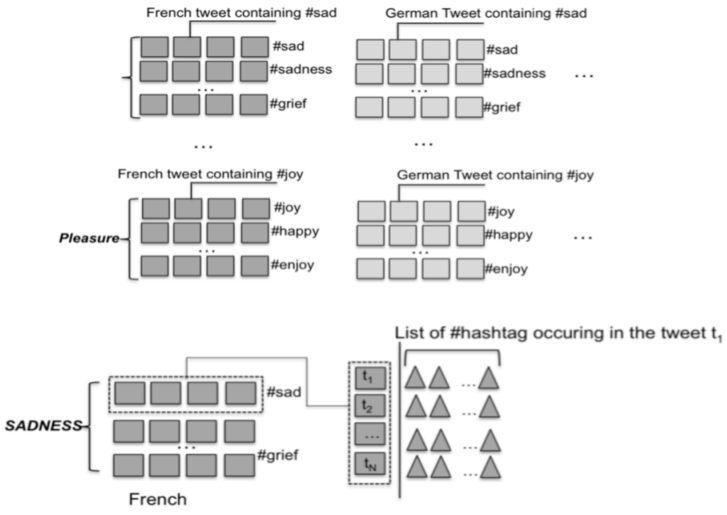

FIGURE 1 – Extraction of Hashtags from the French corpus

Second, we grouped the gathered tweets by language and by emotion (Figure 1). Then, for each emotion e i.e. SADNESS, PLEASURE, LOVE, etc., we extract all co-occurent hashtags and compute their correlation to e. In order to compute how much an hashtag h is correlated to an emotion e, we compute the Strength of Association the (SoA) between an hashtag *h* and an emotion *e* (Equation 1). We discarded short (less than 2 characters) and numerical hashtags.

$$\text{SoA}(h, e) = \log\left(\frac{freq(h, e)}{freq(h) \cdot freq(e)}\right) \quad (1)$$

Where the $freq(h, e)$ is the is the number of times h occurs in tweets belonging to the emotion e. And $freq(h)$, $freq(e)$ are the frequencies of h and e in the corpus.

If an hashtag appear in more than one emotion class, we associate it to the most correlated class. The size of the constructed lexicons is about 17.000 entries for the seven languages.

5. Conclusion

In this research we have presented a novel approach based on Twitter as a comparable corpus to extract automatically affective lexicons in seven langages (English, French, German, Italian, Spanish, Portuguese and Russian). Our approach was motivated by the fact, that non english speaker's, usually, use bilingual terms in their messages. So, we are based in the co-occurence between the English and the target affective terms to generate multilingual affective lexicons. The presented approach is generic as it could be applied for any language. Since the number of returned tweets is limited by the Twitter Search API, in a future work, we plan to use the Twitter Streaming API [2], in order to collect a larger corpus and then obtain larger lexicons. Obtained lexicons, contains not only purely emotio-

2. https ://dev.twitter.com/docs/streaming-apis

Affective Class	French	German
Anger	en colère	wütend
	fâcher	angepisst
	rage	Wut
	irriter	zerstoren
	rougir	Unterbrechung
	nervosité	Tollwut
	massacre	massaker
	énerver	Erregung
	exciter	schötteln
	furax	verärgert
Fear	peur	angst
	terreur	terror
	violence	befürchten
	trombler	gestrandet
	mort	Tod
	terrifié	erschrocken
	appréhender	achtgeben
	inquiétude	sorge
	timidité	eingeschüchtert
	anxiété	ängstlich
Love	amour	Liebe
	Valentin	verheiratet
	coeur	verpassen
	mariage	schön
	manquer	verpassen
	aimer	lieben
	adorer	leidenschaft
	envie	Neid
	gentillesse	freundlichkeit
	affection	zuneigung
Pleasure	heureux	Vergnügen
	content	glücklich
	génial	spielend
	bonheur	Musik
	plaisir	schön
	jouer	underschön
	vacances	Ferien
	podium	erstaunlich
	agréable	reizend
	amusant	lustig

TABLE 6 – The Top-10 entries of the French and German affective lexicons for the *Anger, Fear, Love and Pleasure* emotion classes.

nal words but also some common-sence words that are associated to an affective class ; such as the german word Tod which is associated to the Fear affective class or the french term coeur which is assiciated to the Love class. So, for each langiage, we plan to divide the obtained lexicon into two sub-categories : purely emotional words and common-sence words.

6. References

S. Armstrong and H. Thompson. 1995. A presentation of mlcc : Multilingual corpora for cooperation. *Linguistic Database Workshop*.

M. M. Bradley and P. J. Lang. 1999. Affective norms for english words (anew). University of Florida. Gainesville, FL. The NIMH Center for the Study of Emotion and Attention.

K. W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. volume 16 (1), pages 22–29.

K. W. Church and R. L. Mercer. 1993. Introduction to the special issue on computational linguistics using large

corpora. *Computational Linguistics*, 19(1) :1–24.

A. Fraisse and P. Paroubek. 2013. Toward a unifying model for opinion, sentiment and emotion information extraction. In *In proceedings of the The 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

A. Jackiewicz and M. Vidak. 2014. Etude sur les mots-dièse. In *Congrès Mondial de la Linguistique Française.*, Berlin.

S. M. Mohammad. 2012. Emotional tweets. In *In Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM)*.

A. Pak and P. Paroubek. 2010. Construction d'un lexique affectif pour le français à partir de twitter. In *In Proceedings of TALN (Traitement Automatique des Langues Naturelles) 2010*, Montréal, Canada.

A. Qadir and E. Riloff. 2013. Bootstrapped learning of emotion hashtags #hashtags4you. In *In the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Boston. Association for Computational Linguistics.

R. Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, Maryland, USA. Association for Computational Linguistics.

J. Redondo, I. Fraga, I. Padron, and M. Comesana. 2007. The spanish adaptation of anew (affective norms for english words). volume 39(3).

C. Strapparava and A. Valitutti. 2004. Wordnet-affect : an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

M. L.-H. Vo, M. Conrad, L. Kuchinke, K. Urton, M.J. Hofmann, and A. M. Jacobs. 2009. The berlin affective word list reloaded (bawl-r). volume 41(2).