Language evolution in social media: a preliminary study

Fabio Massimo Zanzotto & Marco Pennacchiotti

Abstract

Language, as a social phenomenon, is in constant evolution. New words are added, disused ones are forgotten, and some others change their morphology and semantics to adapt to a dynamic World. Today we are leaving a new "Social Media" revolution, that is changing many languages. The pace with which new words are created in social media is unprecedented. People from different demographic groups are often "speaking different languages", in that not only they use a different set of words, but also assign different meanings to the same words. In this paper, we investigate whether it is possible to lower the "linguistic barrier", by analyzing the phenomenon of language evolution in social media, and by evaluating to what extent the use of cooperative on-line dictionaries and natural language processing techniques can help in tracking and regulate the evolution of languages in the social media era. We report a study of language evolution in a specific social media, Twitter; and we evaluate whether cooperative dictionaries (specifically Urban Dictionary) can be used to deal with the evolving language. We discover that this method partially solves the problem, by allowing a better understanding of the behavior of new words and expressions. We then analyze how natural language processing techniques can be used to capture the meaning of new words and expressions.

Key words: Twitter language analysis, language evolution, natural language processing

1. Introduction

Language, as a social phenomenon, is in constant evolution. New words are added, disused ones are forgotten, and some others change their morphology and semantics to adapt to a dynamic World.

Radical changes in a language mostly happen when a social group moves from its native location or separates from an original and bigger social group. A clear example is the English language, that in the last three

centuries has evolved following the path of expansion of the British Empire, and giving birth to tenths of different dialects, including General American English, Australian English and Indian English [Crystal 2003]. Language evolution is also caused by the social impact of new scientific and technological discoveries. New words and new word meanings are the tools for better understanding and communicate the world around us.

Organizations such as the *Académie française* in France and the *Accademia della Crusca* in Italy and dictionary producers such as the *Oxford Dictionary*, have the goal of institutionalize and regulate the evolution of languages, by formally adding and removing words as they appear and disappear from common usage. Though, it is a rare event that words are added and their senses are ruled, making more than often news in the media, as in the case of the symbol of the heart included in the Oxford Dictionary in March 2011 [1]. The exponential growth of new scientific discoveries and techniques in the 19th century Industrial Revolution, and in the 20th centrury Electronic and Digital Revolution, has certainly put dictionary producers to the test, that more than once struggled to keep up with the rapid language evolution of a more and more sophisticated society. The job of producing and institutionalize new dictionaries is not a mere intellectual excercise. The American industrial worker of the 19th century and the English manufactures of his machineries had to share a common basic dictionary, in order to keep industry alive and functional. Producers of train carriages had to correctly and precisely understand the names and the measures of standard track components in the different target countries. Workers in nuclear power plants need to correctly understand words in technical manuals. To deal with these technical problems, terminology has been introduced as an important area of language studies to support and complement the work of dictionary producers [Wüster 1931].

Today we are leaving a new "Social Media" revolution, that is once again, and with a faster pace, changing many languages. Social media such as forums, blogs, Twitter, Facebook, Skype, and MSN Messenger, allow people to write their stories and ideas and share them with the Internet community. From a linguistic perspective, this is a much bigger and radical innovation than the Web itself. Indeed, the introduction of the Web in the

[1] Repubblica, 24/3/2011, *Quel cuoricino che dice tutto: Il segno "I love" entra nel dizionario (That little hearth says everything: The sign "I love" is included in the dictionary).*

early 90ies allowed people to read content from different sources, such as media organizations and companies. Most of the information flow was therefore one-way, with people acting as readers. On the contrary, Social media allows a two-way communication. Common people become content producer and, ultimately, language creators. Single individuals or small demographic groups rapidly coin and share new words and new meanings that can potentially and virally spread to larger groups, until they become of common usage and ultimately accepted into formal dictionaries.

The pace with which new words are created in social media is unprecedented. People from different demographic groups (e.g. hip-hop teenagers and their older parents) are often "speaking different languages", in that not only they use a different set of words, but also assign different meanings to the same words. In an extreme late-Wittgensteinian view, people may end up hardly communicating or understanding each other, building around themselves a "linguistic barrier" that inevitably isolates groups from each other. Dictionary producers and linguistic organizations cannot keep up with such a rapid evolution. Too many people and too many fractioned social groups have today the power of shaping the language. New methods and new resources for tracking and regulate languages' evolution are required.

In this paper, we investigate whether it is possible to lower the "linguistic barrier", by analyzing the phenomenon of language evolution in social media, and by evaluating to what extent the use of cooperative on-line dictionaries and natural language processing techniques can help in tracking and regulate the evolution of languages in the social media era.

The paper is organized as follows. In Section 2 we report a study of language evolution in a specific social media, Twitter; and we evaluate whether cooperative dictionaries (specifically Urban Dictionary) can be used to deal with the evolving language. We discover that this method partially solves the problem, by allowing a better understanding of the behavior of new words and expressions. In Section 3, we analyze how natural language processing techniques can be used to capture the meaning of new words and expressions. Finally, in Section 4, we conclude with ideas for future work.

2. Lexicon evolution and crowd-sourced dictionaries

In this section we investigate whether crowd-sourced dictionaries are valid tools to model the evolution of languages in social media. In particular, we are interested in understanding if new words introduced in the media are captured and stored in crowd-sourced dictionaries in a timely manner, i.e. as soon as the new words become of common usage. If this is true, crowd-sourced dictionaries could be used as prominent references for outsiders to a specific demographic group, to uderstand the language of that community.

We also explore automatic models to detect when a new linguistic entity in a social medium is actually promoted to a full fledged status of "new word", i.e. a linguistic entity with a specific meaning shared in a wide community.

In the rest of this section, we present an experiment that investigates the above issues. In detail, our experiment aims at answering the following questions: (1) Are crowd-sourced dictionaries good tools to support the undertanding of new words? (2) Can crowd-sourced dictionaries induce regularites of new words and expressions?

As social medium we experiment with **Twitter**, the second largest microblogging service available today. As for the crowd-sourced dictionary we use **Urban Dictionary**, which is to date the largest collaborative effort to build an up-to-date dictionary of new lingusitic expressions. We begin in Section 2.1 by describe the experimental set up for our study, and then comment on result in Section 2.2.

2.1. Twitter and Urban Dictionary: the experimental set-up

Twitter is a microblogging web service, where users are able to post short messages (called *tweets*)of a maximum length of 140 characters, and read the posts of all other users. Each user can also *follow* specific users he wants to be friend of. When a user logs into Twitter, a personalized "timeline" shows all his latest messages, and the messages of the users he follows.

Twitter is today one of the largest real-time microblogging service, having more than 200 millions users and more than 200 millions tweets per

day worldwide. People tweet about many different topics, from personal updates ("I am eating pizza") and conversation with friends, to breaking news ("Eartquake in Saf Francisco just now! ") and sending web links. According to a 2009 by Pear Analytics [2], 40% of tweets are personal updates, 37% are conversations, 9% are re-posting of other users (called *retweets*), 6% are ads, 4% are spam, and a last 4% are news. Despite these numbers, Twitter has recently played a prominent role in social and political happenings, such as the Arab Spring in 2011, and the riots in England in the summer of the same year. It has also been used to coordinate rescues during major eartquakes, such as those in Chile and Haiti in 2010.

From a demographic perspective, the latest US Quantcast study on Twitter released in September 2011 [3] shows that Twitter is mostly adopted by people between 18 and 34 years (45% of the total), while people under 18 years are only the 18% and over 35 years the 38%. Twitter is adopted by people with a diversified social status (30% earn more than 100K USD a year, 28% between 60K and 100K, 25% between 30K and 60K, 17% below 30K). Twitter is still mostly a American phenomenon, with 33% of the traffic localized in the USA [4], followed by India at 8%, Japan, Germany, United Kingdom and Brazil. English is overwelmingly the most used language: almost two third of the tweets are in English, followed by Portugese (11%) and Japanese (6%).

While Twitter has the form of a big connected graph [Cha et al. 2010], recent studies [Pennacchiotti&Popescu 2011] show that sub-communities exist. Twitter can be therefore seen as one of the meeting places in the web where different communities try to interact. In this study, we show that often standard language is not properly used, both because temd tend to adopt peculiar expressions proper of their own community, and because the short-lenght nature of tweets forces users to write in a succint style, with frequent use of acronyms, abbreviations and truncated words. Tweets

[2] http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf

[3] http://www.quantcast.com/twitter.com

[4] website-monitoring.com

like the following can easily appear: "è prp uno skifo l'hanno kiusa… allò dmn mattina confermato x le otto a piazza del popolo" [5]

Twitter is therefore the perfect medium for our study, as it is a place that can host potentially fast evoving languages of different communities. Our study is based on a corpus of tweets ranging from September 2009 to December 2010 written in any of the English dialects. From this corpus , starting from October 2009, we extracted the monthly frequencies of all words [6] that were not present in Twitter in the month of September 2009. We retain these expressions as potential "new words" that have been just introduced in the language. The final output of the corpus creation is therefore a list of potentially interesting new words along with their frequency for each month in the considered period (9/2009-12/2010).

Figure 1: Urban Dictionary: two definitions of Emo

Urban Dictionary is a crowdsourced web dictionary. Web crowdsourcing is a powerful way of producing resources, where common users can contribute to enrich, mantain and modify an on-line knowledge repositories. Crowdsourcing has emerged as a very succesfull paradigm in the last decades, producing resources such as Wikipedia, an on-line encyclopedia of human knowledge available in many different languages. The evolving version of Wikipedia is rivaling with the most important

[5] in an Italian of SMSs or tweets: it is really bad it has been closed… then tomorrow morning it's confirmed 8 o'clock in piazza del popolo.
[6] Words are extracted by a standard regular expression tokenizer.

encyclopedias for number of entries and, sometimes quality and quantity of content. All entries are written and modified exclusively by Wikipedia users without any reward. Crowdsourcing guarantees that many users can access and modify a specific entry, resulting in a balanced, objective and truthful description of the entry. Indeed, the revision system allow a fast control of content through a collaborative filtering of the knowlegde. The success of wikipedia proves that it is possible to solve many knowledge accumulation and encoding problems using crowsourcing methodologies.

The crowdsourcing approach is also used for dictionaries, e.g. Wikictionary and Urban Dictionary. In our study we use Urban Dictionary, because it is specifically dedicated to specific community languages and to the tracking of new verbal expressions, while Wikictionary aims at modelling standard language.

Urban Dictionary does not adopt a wiki approach, i.e. a site where users can change definitions. Instead, it prefers a more trivial model similar to a Web forum, where users post new words along with their definitions. As in many forums, votes are associated to each *dictionary entry* (that roughly correpsonds to a forum message). Urban Dictionary was created in 2003 as a sort of game, to collect definitions of new "street" words and colloquial language expressions. Today, Urban Dictionary has consitently grown up, becoming a solid reference for finding newly introduced colloquial words and expressions.

Entries in Urban Dictionaries are organized as follows (see example in Figure 1). Each entry has a set of definitions. Each definition is introduced by a user and it is striclty related to him. For example, the first definition of "Emo" (see Figure 1) is given by *7ThisIsWudie7*. Each definition is also given along with its introduction date. For each definition, other anonymous users can give a positive or a negative judgement. These judgements are used to sort definitions for a given word. In the example, the first definition has 62,243 positive and 18,625 negative judgements.

The organization of entries of Urban Dictionary makes this resource attractive for our study, for two main reasons. First, it is a source of colloquial words that are tipical in Twitter. Second, Urban Dictionary allows to easly find the date of introduction of the word in the dictionary, by looking at the date of the word's oldest definition. For our study we

created a repository of all words in Urban Dictionary with their associated date of introduction.

Input of our study are therefore two lists. The list of words in words in Urban Dictionary with their date of introduction; and the list of new words in Twitter with their monthly frequencies. By performing a time-sensitive comparisons these two lists we aim at investigating if (and when) Urban Dictionary captures the new words introduced in Twitter.

2.2. Results and analysis

2.2.1. Freshness of Urban Dictionary

In this first analysis we investigate the *freshness* of Urban Dictionary with respect to Twitter, i.e. whether Urban Dictionary adds new words before or after they emerge in Twitter. This analysis will therefore reveal if Urban Dictionary can provide a useful support to an outsider, for understanding the language of specific communities in the social network.

In order to provide an objective quantitative analysis, we define, for a given word, a *TimeShift* indicator. The *TimeShift* is defined as the difference in time between the introduction of a word in Twitter and the introduction of the word in Urban DIctionary. More formally, we define the following measures:

- *Month of Maximum Twitter Use (MMTU)*. Words in Twitter have a life: they appear, spread, have a period of high frequecy, and then stabilize or disappear. We define MMTU as the month in which a new word has its maximum frequency in Twitter. We consider this period as the landmark for the new word, i.e. the moment in which the word experiences its maximum success.

- *Month of Introduction in Urban Dictionary (MIUD)*. This measure indicates the month in which a word has been first introduced and defined in Urban Dictionary.

Given the two above definition, we further define the *TimeShift* of a words as follows:

$$TimeShift = MMTU - MIUD \quad (1)$$

For example, *TimeShift*=+1 indicates that a word has been first introduced in Urban Dictionary, and then a month later in Twitter. Conversely, *TimeShift*=−1 indicates that the word has been introduced in Twitter a month before than in Urban Dictionary. A *TimeShift*=0 indicates that the word has been introduced in Twitter and Urban Dictionary in the same month.

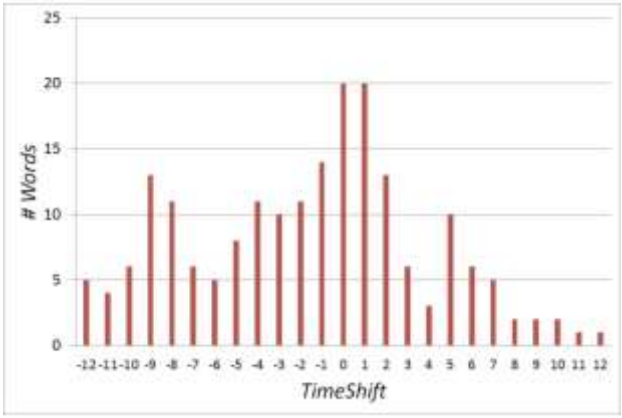

Figure 2: Time Shift between Urban Dictionary and Twitter

Figure 2 plots a summarizing analysis of the *TimeShift* across all words that have been introduced both in Urban Dictionary and Twitter. The figure shows that the TimeShift has a multimodal distribution that, we hypothise, should converge to a normal distrubtion with mean in 0, if more data was available for the experiment. It is interesting to note that the mode of the distribution (i.e. its most frequent value) is 0, which is also approximately the mean value of the distribution. This means that new words in Twitter should be expected with highest probability to be timely captured by Urban Dictionary in the same month of their introduction in Twitter. Urban Dictionary is therefore likely to support outsiders of a Twitter community in reading and understanding the tweets posted in that community.

The Figure also shows that the TimeShift distribution has a high variance, i.e. there are many words with postive or negative TimeShift. This result suggest that many words that are adopted by Twitter after they have been introduced via other media and fixed in Urban Dictionary (positive values of the TimeShift); and there are also many words that are

created in Twitter and then spread outside it (negative values of the TimeShift). We also observe that the distribution is skewed to negative values, i.e. it is more common that a word is first introduced in Twitter, and only after a few months added to Urban Dictionary.

2.2.2. Discovering novel words using frequency counts

With the previous experiment, we undestood that there is an important set of words that, even if covered by Urban Dictionary, their definitions are not timely given. We need then to envisage methods and models to capture the meaning of these words. For doing this, we need to focus on two issues:

First, we need to spot words that are relevantly new in streams like twitter. Not all the words that appear to be new are really novel words. There are many proper nouns or product nouns that gain fame for a short period of time. These are not novel words.

Second, we need to define methods to find the meaning and, then, the definition of these new words.

In this experiment we focus on the first issue. Possible ways to tackle the second issue are instead described in Section 3.

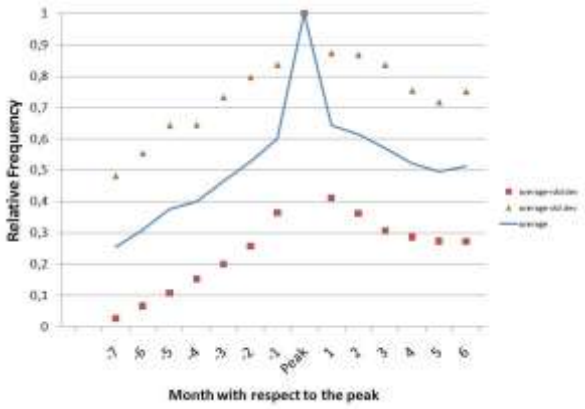

Figure 3: Word frequency in Twitter with respect to the peak of use

We want here to evaluate how simple models based on frequency analysis can be adopted for discovering novel words among words newly

introduced in Twitter. To develop these models we can exploit the data used for the previous experiments. We firstly observe the behavior of words in twitter and, then, we propose simple models to predict novelty observing the evolution of the frequency of words with respect to the time.

The first issue is observing the behavior of words: we took the set of new Twitter words that we used for the previous experiment. We analyzed all the words in this set and not only those in Urban Dictionary. Figure 3 plots the mean relative frequency and variance of all these words. Given a word, the relative frequency is the ratio between its actual frequency in a given month and its maximun frequency. We want to understand how words behave before and after their point of maximum spreading. Given this latter point, Figure 3 plots the relative frequencies of words with respect to the months before and the months after. We can observe that the average behavior of words in this set has a peak in time. Before and after this peak, words basically disappear. This seems to be the average behavior of words that have a peak of use and then are totally lost. These words cannot be novel words or expressions as their popularity last for a too short period. Words behaving averagely can be people names or product names. But, the analysis of the plot in Fig. 3 lead to an interesting conclusion. The standard deviation with respect to the average behavior is high. This implies that there are many words that are not know before their peak or they are steadly known and used after their peak. Words having these features are extremely interesting.

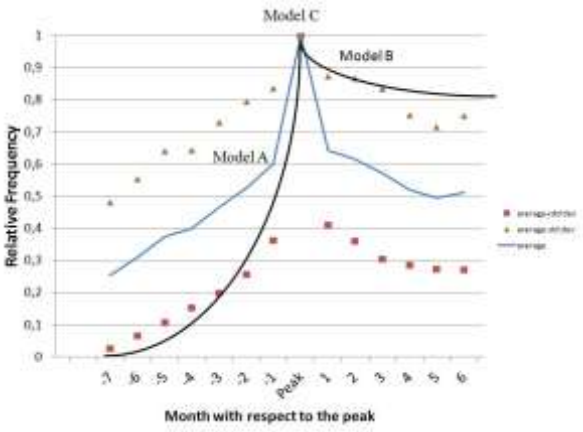

Figure 4: Simple methods for selecting novel words

Having the above analysis on the behavior of words with respect to their peak of popularity, we can propose simple models to spot novel words. This is the second issue we wanted to address. The idea is simple. We propose models based on this idea. Novel words should find thier space in new utterances. After a peak of use, these words should find a nearly constant distribution in the used language. Then, we should tend to prefer those words that have a steady frequency after the peak of use. Second, novel words should gain popularity in a short period of time. We should prefer candidate words that have a fast popularity. With these observations, we can define three different models for novel words. Models are presented in Fig. 4. We propose three models for the novelty of words:

Model A: novel words are words that, before thier peak of use, are less frequent than the average minus the standard deviation

Model B: novel words are words that, before after peak of use, are more frequent than the average plus the standard deviation

Model C: novel words ar words that have the properties of Model A and Model B

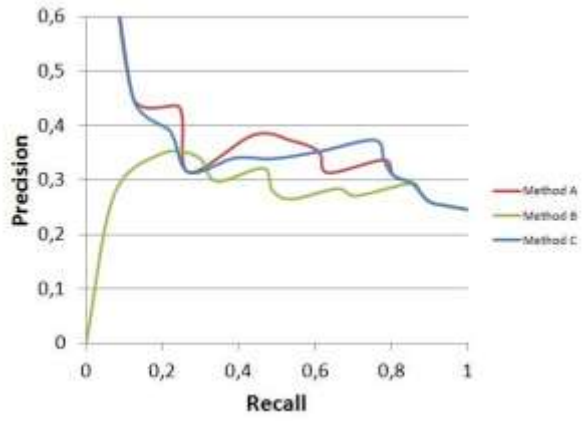

Figure 5: Simple methods for selecting novel words: recall vs. precision

We evaluate the results of the models proposed by using Urban Dictionary. Twitter words that are in Urban Dictionary are good novel words. We want then to evaluate how good these models are in capturing these good novel words. To evaluate these models we use the classical

information retrevial measure of precision and recall. Let's suppose that with a selection method we can find a set of words that we call *selected words*. Precision counts how many good novel words are in the list of *selected words*. This measure states how good is the method in deciding whether or not a word is a novel word. Recall counts how many words among the possible novel words are in *selected words*. This measure tends to say how good is the method in retreiving novel words. There is a strict correlation between precision and recall. Generally, when recall increases precision decreases. To increase recall, we need to have smaller threshold to have a bigger set of *selected words*. This bigger set can contain more words that are not novel words. This is why it is important to study recall and precision in combination. Figure 5 plots recall vs precision of the three methods. Tendentially higher curves represent better methods. Among the three methods, the best one seems to be Model A, i.e., the model that takes into consideration the behavior of candidate words before the peak. Novel words, that go into Urban Dictionary in the considered period, are those words that are basically not present in the period before the peak. Method B, that takes into consideration the behavior of the word after the peak period, is the worst method. The combined method, i.e., Method C, behaves similarly to Method A. The combination of the two methods does not add a considerable gain.

To conclude this section, we can say that simple frequency-based methods for selecting novel words are useful but these methods do not completely solve the problem.

3. Natural language processing and machine learning: basic techniques

We have shown that only a part of novel words are covered by crowd-sourced dictionaries. These dictionaries do not completely open the possibility to understand interacions on social media. We need different methods and models to help outsiders to understand the language of a social group.

In this section, we want to introduce basic natural language techniques that can help in solving the two issues expressed in Sec.

2.2.2: (1) "spotting novel words" task; (2) "give meaning to novel words" task. We will also report on how these basic techniques have been applied in Twitter and in Social Media.

We will focus on four problems: part-of-speech tagging, named-enity recognition, distributional semantics, and automatic classification. The combined use of these techniques can help in the two above tasks.

3.1. Part-of-speech tagging

Part-of-speech tagging is considered the first step for a syntactic analysis. It has been proposed as a separate task in early '90 [Church 1988, Brill 1992,

Abney 1996] when the big issue of natural language understanding (NLU) [Allen 1995] in a pool of tasks that can be independently solved by applying specific theories, models, and systems.

The task aims to assign part-of-speech tags to a sequence of words in a sentence. For each word, a part-of-speech (POS) tagger must state if the word is a noun, an adjective, a verb, etc. The task is formally defined as follows. Given a sentence $s=w_1...w_n$, a POS tagger is a function *POS* that assigns to *s* a sequence of POS tags:

$$\text{POS}(s)= t_1...t_n$$

Each word w_i should have only one interpretation, i.e., a tag t_i. For example, consider the sentence *"the boat sinks"*. The PoS tagger, after analysing the overall sentence $s= w_1w_2w_3$, assigns the POS-tags t_1=Article, t_2=Noun and t_3=Verb. The tagger has to disambiguate words performing a simple analysis and looking, for each word, at its close context. For example, *sinks* is both a noun and a verb. This decision should be taken using the context (i.e., *"the boat"*). Given this information, the tagger has to draw the most likely decision. However, the PoS tagger is not a word sense disambiguator. Homograph forms with the same PoS (e.g., the noun *bank* as *institution* or *river bank*) are not disambiguated with PoS taggers.

PoS taggers are important in a first step of analysis as these tools can help in better modelling later stage of analysis. These PoS taggers can be also used to focus the attention only on some word categories.

As social media have a language that it is not completely standard, some adaptation of existing and well estabished taggers has to be done. Similarly to the adaptation to historical languages [Pennacchiotti&Zanzotto 2008], studies have been carried out in porting techniques used for standard language to social media language [Gimpel et al., 2011].

3.2. Named entity extraction

Detecting named entities, i.e., named entity recognition (NER), in texts is one of the fundamental issue in the task of natural language processing called Information Extraction (IE) [MUC-7 1997]. Named Entity Recognition is the first step to discover more complex facts or relations between people, locations, date, companies, etc. Given a set of target classes (e.g., person and location), the task of named entity recognition in IE or semantic annotation in SW consists of detecting text fragments in documents or in web documents that represent an instance of a target class. For example, consider the following text fragment : *"Before Moscow! " repeated Napoleon, and inviting M. de Beausset, who was so fond of travel, to accompany him on his ride, he went out of the tent to where the horses stood saddled.* A named entity recognizer should extract the three named entities Moscow, Napoleon, and M. de Beausset and should determine the their class, i.e., Moscow is a location while Napoleon and M. de Beausset are two instances of the class person. Finding these bits of information are useful to determine more interesting facts such as the relation between Napoleon and M. de Beausset that, according to this piece of texts, know each other. A survey of the methods can be found in [Nadeau&Sekine 2007].

Named entity extraction is very relevant for social media and microblogging as twitter. Named entities can be products or brands. Monitoring opinons on brands and products is an attractive application for social media data. For this reason, named entity recognition has been adapted to social media [Ritter et al. 2011] and specific annontations have been done to help in building better named entity recognizers [Finin et al. 2010].

Named entity recognizers can be also useful for the problems presented in this paper as it can help in filtering out words that we do not have to analyze. For celebrities, products, and brands, we do not have to find a definition or a meaning.

3.2.1. Classifiers and machine learning

A well-assessed trend in natural language processing research is to design systems by combining linguistic theory and machine learning (ML). The latter is typically used for automatically designing classifiers. A classifier is a function:

$$C{:}I{\rightarrow}T$$

that assigns a category in T to elements of the set I. In supervised ML, the function C is learnt using a set of training instances Tr. Each training instance is a pair $(i,t){\in}Tr$, where $i{\in}I$ and $t{\in}T$, i.e. a class label subset.

ML algorithms extract regularities from training instances observing their description in feature spaces $F{=}F_1{\times}\ldots{\times}F_n$. Each dimension j of the space F is a feature and F_j is the set of the possible values of j. For example, if we want to learn a classifier that decides if an animal is a cat or a dog (i.e., the set $T{=}\{cat,dog\}$), we can use features such as the number of teeth, the length of the teeth, the shape of the head, and so on. Each of the features has values in the range defined with the set F_j. We can then define a function F that maps instances $i{\in}I$ onto points in the feature space, i.e.

$$F(i){=}(f_1,\ldots,f_n) \qquad (2)$$

Once F and Tr have been defined, ML algorithms can be applied for learning C, e.g., decision trees in [Quinlan 1993].

Classifiers are extremely important as these methods can help in automatically decide whether or not a candidate word is really a novel word. These techniques have been also used in the similar problem of deciding whether or not a novel expression is a term in a specific domain as medicine, space, physics, etc. [Basili&Zanzotto 2002].

3.3 Distributional semantics

	run	eat	window
Dog	1	1	0

Cat	1	1	0
Car	1	0	1

Table 1: Context vectors for the words "*dog*", "*cat*", and "*car*"

the	**car**	runs on the highway
she opened the window of the	**car**	
the	**cat**	eats the mouse
the	**dog**	eats the bone
the	**cat**	runs in the gardern
the	**dog**	runs in the gardern

Table 2: A small set of contexts for the words "*dog*", "*cat*", and "*car*"

Distributional semantics (DS) is a very important model to deal with word meaning. Its aim is to give models to determine similarity between words. It stems from the solid linguistic basis of Firth's principle, "*You shall know a word by the company it keeps.*" [Firth 1957], and Harris's Distributional Hypothesis, "*Words that occur in the same contexts tend to have similar meanings*" [Harris 1964]. Firth's principle justifies the idea that the meaning of words (or word sequences) can be modeled using contextual information and can be represented in vector spaces. Harris' Distributional Hypothesis suggests that the meaning of words can be compared through the vectors representing the context in which they occur. For example, Table 1 represents the vectors for "*dog*", "*cat*", and "*car*" derived from the set of sentences in Figure 2. Rows represent contextual vectors for words

and columns represent co-occuring words. "*Dog*" occurs once with "*run*" (see Fig. 2). Similarity between words is given by the similarity between vectors: simple distance measures between vectors such as dot product can be used. Then, "*dog*" and "*cat*" are more similar than "*dog*" and "*car*", as their distributional vectors are closer.

Different kinds of context can be considered to build the distributional vector representing a word:

a word occurring in a window of n tokens around the target word [Schutze 1997]

a lexicalized syntactic relation in which the target word participates [Pado&Lapata 2007]

a document in which the target word occurs [Deerwester et al. 1990]

Such contexts, co-occurring frequently with a target word, comprise its possibly salient attributes [Turney 2006].

This is a key model that can be used in assigning meaning to novel words. Similarity with existing and known words can help in better understanding novel ones.

4. Conclusions

In this paper, we studied the language evolution in a specific social media, Twitter and we evaluated whether cooperative dictionaries (specifically Urban Dictionary) can be used to deal with the evolving language. We discovered that this method partially solves the problem, by allowing a better understanding of the behavior of new words and expressions. We then analyze how natural language processing techniques can be used to capture the meaning of new words and expressions. Starting on these solid grounds, we can start studying to which extent we can use natural language techniques to lower language barriers in the social media era.

Fabio Massimo Zanzotto & Marco Pennacchiotti

fabio.massimo.zanzotto@uniroma2.it

marco.pennacchiotti@gmail.com

References

Abney 1996

Abney, Steven, "Part-of-speech tagging and partial parsing", in G. K.Church,S.Young (ed. by), *Corpus-based methods in language and speech*, Dordrecht, Kluwer Academic Publishers.

Allen 1995

Allen, James F., *Natural language understanding*, II ed., Redwood City, CA, USA, Benjamin-Cummings Publishing Co., Inc.

Basili&Zanzotto 2002

Basili Roberto, Zanzotto Fabio Massimo, "Parsing engineering and empirical robustness", Natural Language Engineering 8/2-3, 1245–1262.

Brill 1992

Brill Eric, "A simple rule-based part of speech tagger", in *Proceedings of the third conference on Applied Natural Language Processing*, ANLC '92, Stroudsburg, PA, USA, Association for Computational Linguistics, 152–155

Cha et al. 2010

Cha Meeyoung, Haddadi Hamed, Benevenuto Fabrício, Gummadi P. Krishna, "Measuring user influence in Twitter: The million follower fallacy", in W. W. Cohen & S. Gosling (eds.), *ICWSM*, The AAAI Press.

Church 1988

Church Kenneth W., "A stochastic parts program and noun phrase parser for unrestricted text", in *Proceedings of the second conference on Applied Natural Language Processing*, ANLC '88, Stroudsburg, PA, USA, Association for Computational Linguistics, 136-143.

Crystal 2003

Crystal David, *English as a global language*, II ed., Cambridge, Cambridge University Press.

Deerwester et al. 1990

Deerwester Scott C., Dumais Susan T., Landauer, Thomas K., Furnas George W., Harshman Richard A., "Indexing by latent semantic analysis", Journal of the American Society of Information Science 41/6, 391–407.

Finin et al. 2010

Finin Tim, Murnane Will, Karandikar Anand, Keller Nicholas, Martineau Justin, Dredze Mark, "Annotating named entities in twitter data with crowdsourcing" in *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's mechanical Turk*, CSLDAMT '10, Stroudsburg, PA, USA, Association for Computational Linguistics, 80–88.

Firth 1957

Firth John R., *Papers in linguistics.* London, Oxford University Press.

Gimpel et al., 2011

Gimpel Kevin, Schneider Nathan, O'Connor Brendan, Das Dipanjan, Mills Daniel, Eisenstein Jacob, Heilman Michael, Yogatama Dani, Flanigan Jeffrey, Smith Noah A., "Part-of-speech tagging for twitter: annotation, features, and experiments", in *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies: short papers - Volume 2*, HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics, 42-47.

Harris 1964

Harris Zellig, "Distributional structure", in J. J. Katz & J. A. Fodor (eds.), *The philosophy of linguistics*, New York, Oxford University Press.

MUC-7 1997

MUC-7, Proceedings of the seventh message understanding conference (MUC-7), in *Columbia, MD*: Morgan Kaufmann.

Nadeau&Sekine 2007

Nadeau David, Sekine Satoshi, "A survey of named entity recognition and classification", Linguisticae Investigationes 30/1, 3-26.

Pado&Lapata 2007
Pado Sebastian, Lapata Mirella, "Dependency-based construction of semantic space models", Computational Linguistics 33/2, 161–199.

Pennacchiotti&Popescu 2011
Pennacchiotti Marco, Popescu, Ana-Maria, "A machine learning approach to Twitter user classification", in L. A. Adamic, R. A. Baeza-Yates, & S. Counts (eds.), *ICWSM*, The AAAI Press.

Pennacchiotti&Zanzotto 2008
Pennacchiotti Marco, Zanzotto Fabio Massimo, "Natural language processing across time: an empirical investigation on Italian", in B. Nordström, A. Ranta (eds.), *GoTAL*, volume 5221 of *Lecture notes in Computer Science* Springer, 371-382.

Quinlan 1993
Quinlan, John Ross, *C4:5: programs for machine learning*, San Mateo, Morgan Kaufmann.

Ritter et al. 2011
Ritter Alan, Clark Sam, Etzioni Mausam, Etzioni Oren, "Named entity recognition in tweets: An experimental study", in *Proceedings of the 2011 conference on empirical methods in natural language processing*, Edinburgh, Scotland, UK., Association for Computational Linguistics, 1524-1534.

Schutze 1997
Schutze Hinrich, *Ambiguity resolution in language learning*, Stanford, CA, CSLI.

Turney 2006
Turney Peter D., "Similarity of semantic relations" Computational Linguis*tics* 32/3, 379–416.

Wüster 1931

Wüster Eugen, *Die Internationale Sprachnormung in der Technik besonders in der Elektrotechnik*, Berlin, VDI Verlag.