@NLP can u tag #user_generated_content ?! via lrec-conf.org

Workshop Programme

14:00 – 14:15 Welcome and introduction

14:15 – 14:40 Óscar Muñoz-García and Carlos Navarro (Havas Media), Comparing user generated content published in different social media sources

14:40 – 15:05 Mehdi Aminian, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, Menno van Zaanen (Tilburg University), Assigning part-of-speech to Dutch tweets

15:05 – 15:30 Diana Maynard, Kalina Bontcheva, Dominic Rout (University of Sheffield), Challenges in developing opinion mining tools for social media

15:30 – 16:00 Alejandro Mosquera and Paloma Moreda (University of Alicante), A Qualitative Analysis of Informality Levels In Web 2.0 Texts: The Facebook Case Study

16:00 – 16:30 Coffee break

16:30 – 16:55 Joan Codina and Jordi Atserias (Fundació Barcelona Media), What is the text of a Tweet?

16:55 – 17:20 Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee (Linguistic Data Consortium, University of Pennsylvania), Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT

17:20 – 18:00 Panel Discussion

Editors

Maite Melero	Barcelona Media Innovation Center (Spain)

Workshop Organizing Committee

Laura Alonso i Alemany	Universidad Nacional de Córdoba (Argentina)
Jordi Atserias	Yahoo! Research (Spain)
Maite Melero	Barcelona Media Innovation Center (Spain)
Martí Quixal	Barcelona Media Innovation Center (Spain)

Workshop Programme Committee

Toni Badia	Universitat Pompeu Fabra (Spain)
Rafael Banchs	Institute for Infocomm Research (Singapore)
Richard Beaufort	Université Catholique de Louvain(Belgium)
Steven Bedrick	Oregon Health & Science University
Louise-Amélie Cougnon	Université Catholique de Louvain(Belgium)
Jennifer Foster	Dublin City University (Ireland)
Michael Gamon	Microsoft Research (USA)
Dídac Hita	Infojobs (Spain)
Fei Liu	Bosch Research (USA)
Daniel Lopresti	Lehigh University (USA)
Ulrike Pado	VICO Research&Consulting GmbH
Lluís Padró	Universitat Politècnica de Catalunya (Spain)
Alan Ritter	CSE, University of Washington (USA)

Table of contents

Author Index

Preface

The Web 2.0 has transferred the authorship of contents from institutions to the people; the web has become a channel where users exchange, explain or write about their lives and interests, give opinions and rate others' opinions. The so-called User Generated Content (UGC) in text form is a valuable resource that can be exploited for many purposes, such as cross-lingual information retrieval, opinion mining, enhanced web search, social science analysis, intelligent advertising, and so on.

In order to mine the data from the Web 2.0 we first need to understand its contents. Analysis of UG content is challenging because of its casual language, with plenty of abbreviations, slang, domain specific terms and, compared to published edited text, with a higher rate of spelling and grammar errors. Standard NLP techniques, which are used to analyze text and provide formal representations of surface data, have been typically developed to deal with standard language and may not yield the expected results on UGC. For example, shortened or misspelled words, which are very frequent in the Web 2.0 informal style, increase the variability in the forms for expressing a single concept.

This workshop aims at providing a meeting point for researchers working in the processing of UGC in textual form in one way or another, as well as developers of UGC-based applications and technologies, both from industry and academia.

Comparing user generated content published in different social media sources

Óscar Muñoz-García, Carlos Navarro

Havas Media
Avenida General Perón 38, 28020- Madrid, Spain
oscar.munoz@havasmedia.com, carlos.navarro@havasmedia.com

Abstract

The growth of social media has populated the Web with valuable user generated content that can be exploited for many different and interesting purposes, such as, explaining or predicting real world outcomes through opinion mining. In this context, natural language processing techniques are a key technology for analysing user generated content. Such content is characterised by its casual language, with short texts, misspellings, and set-phrases, among other characteristics that challenge content analysis. This paper shows the differences of the language used in heterogeneous social media sources, by analysing the distribution of the part-of-speech categories extracted from the analysis of the morphology of a sample of texts published in such sources. In addition, we evaluate the performance of three natural language processing techniques (i.e., language identification, sentiment analysis, and topic identification) showing the differences on accuracy when applying such techniques to different types of user generated content.

1. Introduction

The rise of Web 2.0 technologies and social media have enabled users to author their own content. Such user generated content (UGC) is being used for many different purposes, such as opinion mining and market research.

Natural language processing (NLP) techniques are a key piece for analysing the content published in social media. Social media content presents the characteristics of non-editorially-controlled media, as opposite to the content published in traditional media. In this context, social media communication has moved from daily publications to real-time interactions. Thus, when applying NLP techniques to the UGC published in social media, we find issues on text quality that difficult the application of such techniques. Moreover, if we analyse social media sources by separate, we find that there are differences on language styles, expressiveness degrees, and levels of formalism that are conditioned by factors such as content length or publication pace. Namely, text length varies form short sentences posted in Twitter to medium-size articles published in blogs; very often the text published in social media contains misspellings, is completely written in uppercase or lowercase letters, or it is composed of set phrases; to mention a few characteristics that make social media content analysis challenging.

In this paper, we make use of a set of NLP tools, that we have at hand, to process and characterize corpora of UGC extracted from different social media sources. Specifically, we have studied differences of the language used in distinct types of social media content by analysing the distribution of part-of-speech (PoS) categories in such sources. In addition, we have measured the accuracy of specific techniques of language identification, sentiment analysis and topic identification when applied to the social media sources analysed.

The paper is structured as follows. Firstly, Section 2. characterises the sources from which we have extracted the content used in the experiments described in this paper. Secondly, Section 3. explains the distribution of PoS categories in the sources analysed. Section 4. shows the performance of a language identification technique applied to a Twitter corpus in comparison to applying the same technique to the rest of the sources, while Section 5. shows the performance of a sentiment analysis technique applied to content published in the same sources. Section 6. summarises the results of a previous experiment, showing the performance of a topic identification technique on the sources analysed. Finally, Section 7. presents the conclusions and depicts future lines of work.

2. Social media content analysed

The corpora used for evaluating different NLP techniques have been extracted from the sources detailed next.

Blogs. We have extracted the texts of the posts from the feeds of blog publishing platforms such as Wordpress and Blogger. Content published in these sites usually consists on medium-sized posts and small comments about such posts.

Forums. We have scrapped the text of the comments published in web forums constructed with vBulletin and phpBB technologies. Content published in these sites consists in dialogues between users in the form of a timely ordered sequence of small comments.

Microblogs (e.g., Twitter and Tumblr). We have extracted the short messages published in such sources by querying their APIs. Content published in this source consists on small pieces of text (e.g., maximum 140 characters for Twitter).

Social networks (e.g., Facebook, MySpace, LinkedIn and Xing). We have extracted the messages published in such sources by querying their APIs. Content published in this sites goes from small statuses or comments to medium-sized posts (e.g., Facebook notes).

Review sites (e.g., Ciao and Dooyoo). We have scrapped the text of the comments published in such sources.

The length of the content published in these sites is also variant.

Audiovisual content publishing sites (e.g., YouTube and Vimeo). We have extracted the textual comments associated to the audiovisual content. Textual content published in this sites takes the form of small textual comments.

News publishing sites. We have extracted the articles from the feeds published in such sources. Sites of this kind can be classified as traditional editorially-controlled media. However, comments posted by article readers can be catalogued as UGC. Thus, content published in news sites consists on articles and small comments about such articles.

Other sites not classified in the categories above (e.g., Content Management Systems) that publish their content as structured feeds, or that have a known HTML structure from which a scrapping technique can be applied. Content published in these sites is heterogeneous.

3. Distribution of PoS categories

For performing the study of the distribution of PoS categories in UGC, we have collected a corpora with $10,000$ posts written in Spanish, obtained from the sources described in the previous section. The post extracted are related to the telecommunication domain. We have performed the PoS analysis by implementing a Gate (Cunningham et al., 2011) pipeline, with TreeTagger (Schmid, 1994) as the PoS tagging component. Therefore, the PoS distributions obtained are based on an automatic tagger. A previous work (García Moya, 2008) includes an evaluation of TreeTagger with a Spanish parametrization when applied to a corpus of news articles. The precision, recall and F-measure obtained on such evaluation were 0.8831, 0.8733 and 0.8782 respectively.

Table 1 shows the distributions obtained. The PoS categories are determined by TreeTagger tag-set for Spanish[1].

As shown in the table, there are variations in the distribution of these categories with respect to the publication source. For example, in microblogs, determinants and prepositions are used to a lesser extent, because the limitation of post length (e.g., 140 characters in Twitter) requires that posts are written more concisely, and therefore meaningless grammatical categories tend to be used less.

The distribution of all PoS categories in news publishing sites and blogs is very similar, because the posts published in these sources have a similar writing style, as there are no limitations on the size of such posts.

In addition, the sources not classified (i.e, "other") have a similar distribution to the combination of all sources. This may be due to the heterogeneity of the publications contained in the web pages that have not been classified as specific content type.

Next, we discuss some relevant insights obtained from the distribution of each PoS category.

[1] ftp://ftp.ims.uni-stuttgart.de/pub/corpora/spanish-tagset.txt

3.1. Distribution of nouns

As shown in Table 1 the distribution of common and proper nouns is similar for all sources, with the exception of forums and reviews. It seemed strange to us that proper nouns, found in the sources where discussions about specific product models are raised, were less used than in the other sources. After examining a representative sample of texts, we noticed that in those sources, product names are often written in lower case, which lead to an incorrect PoS annotation. After reprocessing the corpus using gazetteers, including proper names in lower case, we found that this is a problem with TreeTagger precision. Such problem makes entity recognition less accurate, when such entity recognition requires a previous step of detecting proper nouns using PoS tagging. Although the use of gazetteers improves entity detection, this solution may be very domain-dependent.

In addition, foreign words are less used in news than in other sources, because the style rules of traditional media require avoiding such foreign words, as far as possible, whenever a Spanish word exists.

Finally, the relative big distribution of letters of the alphabet category is due to a TreeTagger accuracy error (overall when analysing short texts published in Twitter).

3.2. Distribution of adjectives

As shown in Table 1, the distribution of adjectives of quantity is near 50% for most of the sources. The adjectives of quantity commonly used are the cardinals and the less used are the ordinals, whose use is insignificant in all sources, except in news publishing sites. The rest of quantifying adjectives are used quite frequently in forums and reviews, because such sites include publications of quantitative evaluations and comparisons of products. Specifically, in these sites, we find multiplicative (e.g., *doble*, *triple*), partitive (e.g., *medio*, *tercio*), and indefinite quantity adjectives (e.g., *mucho*, *poco*, *bastante*).

3.3. Distribution of adverbs

The adverbs of negation (e.g., *jamás*, *nada*, *no*, *nunca*, *tampoco*) are used with more frequency in the sources with shorter publications. Moreover, there is an inverse correlation between the size of the texts and the use of adverbs of negation. The detection of such negations is essential when performing sentiment analysis, since they reverse the sentiment of the opinion about specific entities.

3.4. Distribution of conjunctions

With respect to conjunctions, the distribution of coordinating conjunctions is higher in sources where the texts are longer (i.e., news and blogs), and lower in sources were posts are shorter, especially in forums and reviews because these sources have a question-answer structure dominated by short sentences. Coordinating conjuntions are useful for opinion mining to identify opinion chunks, as well as punctiaction marks.

3.5. Distribution of pronouns

The distribution of personal pronouns (e.g., *yo*, *tú*, *mí*) is higher in microblogs, reviews, forums and audiovisual con-

Table 1: Distribution of PoS in different social media sources

	News	Blogs	Audiov.	Reviews	Micro.	Forums	Other	Social Net.	All
Noun	**30.9%**	**30.0%**	**29.0%**	**23.2%**	**33.7%**	**22.0%**	**26.6%**	**32.7%**	**27.4%**
Common	53.3%	56.9%	50.5%	71.5%	50.4%	68.8%	60.9%	50.2%	59.2%
Proper	42.3%	37.3%	42.9%	23.8%	36.1%	25.7%	34.1%	43.1%	34.6%
Foreign word	0.2%	0.5%	1.4%	0.5%	1.8%	0.9%	0.7%	1.0%	0.8%
Measure unit (e.g., GHz)	0.2%	0.8%	0.0%	0.6%	0.1%	0.2%	0.2%	0.2%	0.3%
Month name (e.g, Feb)	0.5%	1.1%	0.4%	0.1%	0.1%	0.3%	0.5%	0.4%	0.4%
Acronym (e.g., UN)	0.3%	0.5%	0.5%	0.1%	0.3%	0.5%	0.3%	0.5%	0.3%
Letter of the alphabet (e.g., b)	0.6%	1.1%	2.3%	1.0%	4.0%	1.7%	1.0%	1.9%	1.5%
Alphanumeric code (e.g., A4)	2.2%	1.5%	1.9%	0.9%	1.1%	1.2%	1.9%	1.1%	1.5%
Symbol (e.g, $, £)	0.4%	0.3%	0.1%	1.4%	6.1%	0.7%	0.5%	1.5%	1.3%
Adjective	**8.6%**	**8.3%**	**6.4%**	**8.2%**	**9.4%**	**7.1%**	**8.4%**	**6.2%**	**8.0%**
Quantity ordinal	4.6%	2.7%	1.4%	1.5%	0.4%	1.1%	1.7%	1.1%	1.9%
Quantity cardinal	34.7%	30.6%	28.5%	22.0%	33.0%	24.8%	34.3%	25.5%	29.6%
Quantity other	7.5%	12.0%	14.5%	23.6%	7.4%	23.3%	13.8%	19.3%	15.7%
Other	53.3%	54.8%	55.6%	53.0%	59.1%	50.8%	50.1%	54.1%	52.9%
Adverb	**2.5%**	**3.4%**	**3.2%**	**4.9%**	**3.9%**	**4.5%**	**3.7%**	**3.4%**	**3.8%**
Negation	18.2%	18.1%	29.7%	23.9%	36.2%	30.0%	30.6%	29.1%	27.4%
Other	81.8%	81.9%	70.3%	76.1%	63.8%	70.0%	69.4%	70.9%	72.6%
Determiner	**11.5%**	**9.8%**	**7.6%**	**8.0%**	**5.8%**	**8.0%**	**8.7%**	**7.5%**	**8.5%**
Conjunction	**6.1%**	**7.8%**	**6.6%**	**9.7%**	**6.2%**	**10.1%**	**8.7%**	**7.4%**	**8.3%**
Adversative coordinating	2.4%	3.1%	3.9%	5.7%	7.0%	5.7%	4.1%	3.7%	4.6%
Negative coordinating	0.3%	0.9%	0.7%	1.5%	1.0%	1.5%	1.3%	1.3%	1.2%
Other coordinating	44.3%	44.2%	36.6%	29.3%	36.6%	32.5%	38.9%	41.6%	36.7%
”que”	28.5%	26.9%	27.0%	34.4%	26.1%	31.7%	29.5%	26.7%	30.1%
Subord. (finite clauses)	2.2%	3.1%	1.6%	4.4%	1.4%	3.0%	2.9%	2.2%	3.0%
Subord. (infinite clauses)	10.6%	9.7%	18.7%	10.8%	10.7%	11.1%	10.2%	12.0%	10.8%
Other subordinating	11.7%	12.0%	11.5%	13.9%	17.2%	14.6%	13.1%	12.6%	13.5%
Pronoun	**1.9%**	**3.4%**	**5.0%**	**5.6%**	**4.7%**	**5.8%**	**4.3%**	**4.4%**	**4.4%**
Demonstrative	23.7%	24.3%	15.4%	20.2%	15.1%	13.9%	18.3%	16.2%	17.8%
Interrogative	0.7%	0.9%	0.0%	0.8%	1.8%	1.1%	0.6%	0.8%	0.9%
Personal (clitic)	17.1%	16.0%	11.4%	11.4%	16.3%	17.2%	14.6%	12.8%	14.6%
Personal (non-clitic)	15.7%	22.1%	37.3%	44.3%	42.9%	50.3%	39.0%	42.5%	40.8%
Posesive	38.4%	34.3%	33.0%	21.2%	22.0%	15.9%	24.8%	24.6%	23.4%
Relative	4.3%	2.4%	2.8%	2.1%	1.9%	1.6%	2.7%	3.1%	2.4%
Preposition	**15.2%**	**14.6%**	**11.8%**	**12.7%**	**8.2%**	**11.9%**	**12.9%**	**11.5%**	**12.6%**
Portmanteau word “al”	3.8%	3.1%	3.4%	2.8%	2.1%	2.8%	3.1%	3.0%	3.1%
Portmanteau word “del”	7.6%	4.2%	3.9%	4.5%	3.2%	3.9%	4.3%	4.8%	4.8%
Other	88.6%	92.7%	92.8%	92.6%	94.7%	93.3%	92.6%	92.3%	92.1%
Punctuation mark	**10.7%**	**8.5%**	**12.9%**	**9.4%**	**8.3%**	**9.2%**	**9.7%**	**10.5%**	**9.7%**
Full stop	4.9%	17.1%	41.5%	8.7%	29.8%	25.5%	13.2%	25.0%	16.8%
Comma	48.9%	54.5%	29.1%	50.1%	25.2%	44.1%	44.7%	33.8%	43.7%
Colon	3.8%	3.8%	2.4%	5.4%	13.9%	4.8%	5.2%	15.2%	6.6%
Semicolon	1.0%	0.9%	1.3%	0.5%	0.6%	0.5%	0.5%	0.6%	0.7%
Dash	2.5%	1.4%	3.4%	1.5%	0.7%	2.1%	3.6%	3.3%	2.4%
Ellipsis	2.9%	4.3%	7.7%	8.8%	16.3%	8.4%	6.2%	9.2%	7.4%
Slash	0.5%	0.0%	0.0%	0.6%	3.8%	0.1%	0.3%	0.1%	0.5%
Percent sign	1.3%	1.1%	0.0%	0.9%	0.0%	0.7%	1.3%	0.4%	0.9%
Left parenthesis	13.4%	6.2%	5.2%	8.8%	2.1%	5.1%	11.1%	4.1%	8.1%
Rigth parenthesis	13.4%	6.2%	4.7%	8.6%	4.1%	5.5%	11.0%	4.6%	8.3%
Quotation symbol	7.5%	4.5%	4.6%	6.2%	3.5%	3.2%	2.9%	3.6%	4.5%
Verb	**12.0%**	**13.8%**	**16.8%**	**17.8%**	**19.1%**	**20.5%**	**16.4%**	**16.0%**	**16.7%**
To be (“estar”)	1.6%	1.9%	0.5%	1.7%	1.1%	1.5%	1.5%	1.3%	1.5%
To have (“haber”)	5.8%	3.5%	2.4%	3.5%	2.0%	3.2%	3.9%	1.9%	3.4%
Lexical past participle	16.0%	13.4%	11.7%	10.2%	5.8%	10.0%	12.2%	8.9%	10.8%
Lexical finite	47.2%	48.8%	48.5%	46.8%	50.1%	50.2%	48.5%	51.8%	48.8%
Lexical gerund	1.0%	0.7%	0.3%	0.9%	0.4%	0.8%	0.8%	1.1%	0.8%
Lexical infinitive	20.4%	22.9%	28.1%	25.5%	32.0%	26.7%	25.0%	26.9%	26.0%
Modal	1.5%	1.8%	0.8%	1.4%	0.8%	1.9%	1.6%	1.9%	1.6%
To be (“ser”) past part.	0.6%	0.3%	0.6%	0.9%	0.1%	0.2%	0.4%	0.1%	0.4%
To be (“ser”) infinitive	0.4%	0.4%	0.4%	0.6%	0.3%	0.3%	0.5%	0.5%	0.4%
To be (“ser”) other	5.6%	6.4%	6.8%	8.7%	7.3%	5.3%	5.7%	5.6%	6.3%
“Se” (as particle)	**0.7%**	**0.6%**	**0.7%**	**0.5%**	**0.7%**	**0.7%**	**0.6%**	**0.6%**	**0.6%**

tent publishing sites because, in these sources, conversations between the users that generate the content are predominant, in contrast to the narrative style of news and blogs articles.
Generally, pronouns make it difficult to identify entities within opinions, because such entities are not explicitly mentioned when using pronouns.

3.6. Distribution of punctuation marks

Full stops are less used in news than in other sources, because longer sentences are published in news articles, in comparison to the rest of social media sources, where concise phrases are usually written.
The use of comma is lower in sources where there is less writing, that is, on Twitter and sites with comments on audiovisual content.
The heavy use of the colon and slash in microblogs is due to the inclusion of these characters in the emoticons and the sources cited through links embedded in tweets.
Ellipses are more used in microblogs than in the rest of the sources, because of the limitation of the size of the messages. In this source, unfinished messages are posted frequently, so ellipses are added to express that such messages are incomplete. Furthermore, some Twitter clients truncate messages longer than 140 characters, and automatically add the ellipsis.
Finally, parenthesis and other non-commonly used punctuation marks (e.g., percent sign) are less used in microblogs, because of the limited length of the tweets and the difficulty for introducing these characters on mobile terminals.

3.7. Distribution of verbs

With respect to verbs, in forums and microblogs its use is more extensive, in proportion to the rest of the PoS categories, than in the other social media sources. A reason for this may be that, intentions and actions are expressed more often in these sources.
In addition, there is less use of the past participle within microblogs than in other sources. This is because microblogs are used to transmit immediate experiences, so most of the posts are communicated in the present tense. Similarly, the infinitive is more used in microblogs for lexical verbs.

4. Performance of language identification

We have compared the performance of automatic language identification with the following kinds of UGC: (1) statuses posted on Twitter, and (2) content published in the other social media sources. For doing so, we have implemented a text categorization algorithm based on n-grams of characters (Cavnar and Trenkle, 1994). The category profiles, required by such algorithm, have been generated from a training corpus containing documents written in Spanish, Portuguese and English. Such corpus includes the Wikipedia articles about Europe continent for each language[2][3][4], as well as an article describing one city for each language (i.e., Madrid for Spanish[5], Lisbon for Portuguese[6], and London for English[7]).

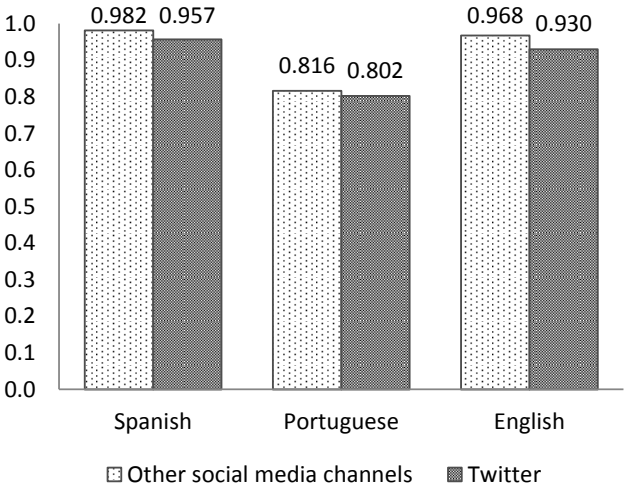

Figure 1: F-measure for the languages evaluated

We have evaluated the text identification algorithm with a corpus of $3,368$ tweets and a corpus of $2,768$ posts extracted from other social media sources different than Twitter. The languages of the documents contained in these corpora are the same as in the language profiles, and all the posts within the corpora have previously been manually annotated with their languages. Figure 1 shows the performance of the topic identification technique implemented.
The overall accuracy when identifying the language of the tweets is 93.02% and 96.76% for the rest of social media posts, while Kappa coefficient (Cohen, 1960) is 0.844 for Twitter and 0.916 for the rest of social media sources, both indicating an almost perfect agreement between human and machine language identification. Thus, the automatic language identification algorithm implemented behaves slightly better for social media sources different than Twitter.
After measuring the accuracy of the language identification technique applied to the Twitter corpus, we have preprocessed such corpus with the aim of improving the performance of language identification on Twitter. Specifically, we have implemented several rules for syntactic normalization of twitter messages. Some of such rules have been described in (Kaufmann and Jugal, 2010). The rules executed are the following: (1) delete references to users at the beginning of the tweet; (2) delete the sequence of characters "RT" followed by a reference to a Twitter user (marked by the symbol "@") and, optionally by a colon punctuation mark; (3) delete the hashtags found at the end of the tweet; (4) delete the "#" symbol from the hasthtags that are not at the end of the tweet; (4) delete the hyperlinks contained within the tweet; and (5) delete ellipses that are at the end of the tweet, followed by a hyperlink.
Once we applied these rules, we re-executed the evaluation for the Twitter corpus, without finding a significant

[2]http://es.wikipedia.org/wiki/Europa
[3]http://pt.wikipedia.org/wiki/Europa
[4]http://en.wikipedia.org/wiki/Europe
[5]http://es.wikipedia.org/wiki/Madrid
[6]http://pt.wikipedia.org/wiki/Lisboa
[7]http://en.wikipedia.org/wiki/London

gain in algorithm performance. Specifically, accuracy decreased from 93.02% to 93.01%, F-measure for Portuguese increased from 0.802 to 0.803, and F-measure for English decreased from 0.93 to 0.929. The rest of the values remained unaltered.

5. Performance of sentiment analysis

We have compared the performance of sentiment analysis with the same kinds of UGC as in language identification (i.e., Twitter statuses, and content published in other sources). For doing so, we used a sentiment analysis component that makes use of linguistic expressions.

Each linguistic expression is defined as a sequence of pairs (L, P), where L correspond with a lemma and $P \in \{Noun, Verb, Adjective, Adverb, Other\}$ with a PoS category, where *other* includes PoS categories different than $Noun$, $Verb$, $Adjetive$ and $Adverb$. In addition, L may correspond to the entity Σ for which the sentiment is to be calculated. An example of linguistic expression is $\langle(\Sigma, Noun), ('worth', Verb), ('it', Other)\rangle$, which matches any text like "Σ worth it", being the PoS of Σ a noun (e.g., "Twitter worth it"). The sentiment analysis algorithm makes use of an experimental proprietary dictionary of linguistic expressions consisting in the subsets defined next.

Expressions for detecting subjectivity, defined as the set B of linguistic expressions, in which verbs are usually included. Our dictionary includes 20 linguistic expressions of this type for the Spanish language.

Expressions for detecting sentiment of opinions, defined as the set E of linguistic expressions $E_i \in E$, each of them associated with a value $v(E_i) \in \{-2, -1\}$ for negative expressions, and $v(E_i) \in \{1, 2\}$ for positive expressions. Our dictionary includes 566 positive and 914 negative linguistic expressions for the Spanish language.

Expressions for reversing sentiment, defined as the set I of linguistic expressions that reverse the sentiment of opinions (i.e., negations). Our dictionary includes 22 linguistic expressions of this type for the Spanish language.

Expressions for augmenting or reducing sentiment, defined as the set U of linguistic expressions $U_j \in U$, each of them associated with a value $m(U_j) = 0.75$ for expressions that reduce sentiment and $m(U_j) = 1.5$ for expressions that augment sentiment. These expressions use to include adverbs. Our dictionary includes 32 expressions that augment sentiment, and 8 expressions that reduce sentiment for the Spanish language.

Listing 1 describes the algorithm implemented for sentiment analysis. Such algorithm receives a text F that mentions a given entity, the form S of the entity for which the sentiment is to be calculated, which is literally included in the text F, and the sets of expressions described previously. The algorithm returns a pair (v, V) containing the text annotated with its sentiment, where $v \in [0, 10]$ is the sentiment value and $V \in \{Negative, Neutral, Positive\}$ is the class of sentiment expressed in F about E. The steps executed by the algorithm are the following:

Listing 1: Algorithm implemented for analysing sentiment

```
1  function AnaliseSentiment(F, S, B, E, I, U)
2  begin
3     P ⇐ LemmatiseAndPoSTag(F)
4     G ⇐ ⟨⟩
5     for each P_i = (L_i, C_i) ∈ P do
6        if L_i = S then
7           G ⇐ concat(G, ⟨(Σ, C_i)⟩)
8        else
9           G ⇐ concat(G, P_i)
10       end if
11    end for
12    v ⇐ 0
13    if (∃E_i ∈ E · subseq(E_i, G))∧
14          ((∃B_j ∈ B · subseq(B_j, G))∨
15          (¬(∃L · subseq(⟨(L, Verb)⟩, G)))) then
16       for each E_i ∈ E · subseq(E_i, G) do
17          polarity ⇐ v(E_i)
18          for each I_j ∈ I · subseq(I_j, G)∧
19                pos(I_j, G) < pos(E_i, G) do
20             polarity ⇐ polarity · (−1)
21          end for
22          for each U_k ∈ U · subseq(U_k, G)∧
23                pos(U_k, G) < pos(E_i, G) do
24             polarity ⇐ polarity · m(U_K)
25          end for
26          v ⇐ v + polarity
27       end for
28    end if
29    v ⇐ (min{max{−10, v}, 10} + 10)/2
30    if v < 5 then
31       V ⇐ Negative
32    else if v = 5 then
33       V ⇐ Neutral
34    else
35       V ⇐ Positive
36    end if
37    return (v, V)
38 end
```

1. We extract the lemmas and the PoS category of every lexeme included in the text F, obtaining a sequence of pairs P containing the lemma and the PoS category of the corresponding lexeme. We have used Freeling (Padró et al., 2010) in our experiment for executing this task (line 3).

2. We look up the entity form S within the sequence P, replacing occurrences by a special token representing Σ and storing the resulting sequence in a new sequence, named G (lines 4-11).

3. We initialize the value of sentiment to the neutral value 0 (line 12).

4. After that, we determine if there is subjectivity in the text analysed. We consider that there is subjectivity, if at least one expression for detecting sentiment matches within the text and either fits some expression for detecting subjectivity, or there is not a verb within the text (lines 13-15). If subjectivity is detected, for

each expression of sentiment E_i that matches with the text we execute the following steps:

4.1. We assign the value of sentiment associated with the expression E_i to a temporal variable called $polarity$ (line 17).

4.2. We invert the sentiment by multiplying $polarity$ by (-1), each time an expression for reversing sentiment is found before the position of the expression E_i within the text(lines 18-21).

4.3. We increase or reduce the value of $polarity$, each time an expression for increasing or reducing sentiment is found before the position of the expression E_i within the text (lines 22-24).

4.4. We accumulate the sentiment obtained to the overall sentiment value (line 26).

5. The overall sentiment value obtained is transformed to a scale from 0 to 10 (line 29).

6. Finally we classify the text according to sentiment value (lines 30-36), and return the sentiment numerical value and the class of sentiment (line 37).

We have evaluated the sentiment analysis component for the Spanish language with a corpus of $1,859$ tweets and a corpus of $1,847$ posts extracted from other social media sources different than Twitter. All the posts within the corpora have been annotated manually previously with their sentiment class, and the form of the entity mentioned in the text, which is required by the algorithm. The accuracy of the algorithm/dictionary combination when analysing sentiment in Twitter is 66.92% and 80.17% in the rest of social media posts, while Kappa coefficient is 0.198 for Twitter and 0.31 for the rest of social media sources. Thus, the sentiment analysis algorithm implemented behaves significantly better for social media sources different than Twitter. As happened with language identification, pre-procesing the Twitter corpus by following the syntactic rules enumerated in the previous section, did not improved algorithm accuracy, which decreased from 66.92% to 66.918%.

6. Performance of topic identification

In a previous work (Muñoz-García et al., 2011), we described a technique for topic identification that uses DBpedia (Bizer et al., 2009) as a linguistic resource. Such technique was evaluated with corpora extracted from the sources described in Section 2..

Listing 2 describes the algorithm implemented for topic identification. Such algorithm receives the text of a post F, the output language l (i.e., language for which the topics must be defined), and a list of stop words θ (i.e., keywords that will be excluded for topic identification). The algorithm returns the set of topics identified T_F^l. The steps executed by this technique are the following:

1. PoS tagging and lematization (lines 3-9). This step receives the text of a social media post and returns a set of keywords that appear in such text. For doing so, we perform a PoS tagging and filter those lexical units that refer to fixed entities without meaning. More specifically, we only consider those words whose lexical category is a noun, including: common nouns, proper nouns, acronyms, foreign words and units of measures. In addition, in this process, each lexical unit is annotated with its lemma. In our experiment, PoS tagger settings were the same as those described in Section 3. (i.e., Gate and TreeTagger with a Spanish parametrization).

2. Topic recognition (lines 10-21). This step receives the keywords produced in the previous step and returns a set of topics, as semantic entities derived from such keywords. This step includes the execution of the following stages:

2.1. Context selection. This stage consists in selecting, for each keyword, a list of keywords in the same post that will help to disambiguate the meaning of the keyword. For executing this stage, the active context selection technique (Gracia and Mena, 2009) is used. Such technique consists in computing semantic relatedness, taking into account the co-occurrence of the keywords in a corpus of web pages.

2.2. Topic disambiguation. This stage receives the keywords and their contexts and returns the senses associated to each keyword in the form of DBpedia resources. For executing this stage, the techniquue described in (García-Silva et

Listing 2: Algorithm implemented for identifying topics

```
1  function IdentifyTopics(F, l, θ)
2  begin
3     P ⇐ LemmatiseAndPoSTag(F)
4     K_F ⇐ ∅
5     for each P_i = (L_i, C_i) ∈ P do
6        if L_i ∉ θ ∧ C_i = Noun then
7           K_F ⇐ K_F ∪ {L_i}
8        end if
9     end for
10    T_F ⇐ ∅
11    for each k_i in K_F do
12       k_i ⇐ PreProcessing(k_i)
13       if Ambiguous(k_i) then
14          S ⇐ DisambiguationLinks(k_i)
15          A ⇐ ActiveContext(k_i, K_F)
16          s ⇐ Disambiguate(k_i, A, S)
17          T_F ⇐ T_F ∪ {DBpediaResource(s)}
18       else
19          T_F ⇐ T_F ∪ {DBpediaResource(k_i)}
20       end if
21    end for
22    T_F^l ⇐ ∅
23    for each t_i in T_F do
24       if ∃b ∈ Labels(t_i) · language(b) = l then
25          T_F^l ⇐ T_F^l ∪ {t_i}
26       end if
27    end for
28    return T_F^l
29 end
```

al., 2010), is used. Such technique relies on Wikipedia redirection and disambiguation pages for performing disambiguation. The former are links between alternate titles and an article while the latter are lists of candidates articles defining the possible senses of an ambiguous term.

3. Language filtering (lines 23-28). This step receives the senses produced in the previous step and returns the set of topics that have been defined in DBpedia for a given language.

In (Muñoz-García et al., 2011), we measured how our topic identification technique performed with three variants of such technique. The first variant consists in identifying the topics without considering any context. Thus, we always assigned to keywords the sense that Wikipedia editors defined as the default sense for such keywords. The second variant consisted in identifying the topics by considering all the other keywords found in the same post as context. The third variant consisted in identifying the topics by applying the active context selection technique explained before.

The coverage of the technique was evaluated with a corpus of 10,000 posts in Spanish. Table 2 shows the coverage of the steps executed by our technique. Row 2 reflects the coverage of the PoS tagging step (i.e., the percentages of posts for which at least one keyword was found). Rows 4-6 show the coverage of the topic identification step (i.e., the percentages of posts for which at least one DBpedia resource was identified). Rows 8-10 show the coverage of the language filtering step (i.e., the percentage of the posts for which at least one DBpedia resource with a Spanish label was found).

The coverage of the PoS tagging step was nearly 100% for all the sources while the coverage of topic recognition step was over 90% for almost all the cases. However, when the language filtering step was executed, the overall coverage was reduced in about 10 points because not all the DBpedia resources are labeled with a Spanish term. By sources, the overall coverage for review sites was lower than for the rest of sources, because, in this kind of sites, there is information about specific product models whose commercial denomination is not necessarily translated to a Spanish label. The overall coverage for blogs and news publishing sites was the highest, because the length of the posts published in such sources is greater than in the other sources, what permits extracting more keywords from such posts. Generally, the coverage of the method is bigger when context is taken into account.

We evaluated the precision of the topic identification technique with a random sample of $1,816$ posts using 47 human evaluators. We showed to three different evaluators each post and the topics identified. For each topic, the evaluators selected one of the following options: (1) the topic is not related with the post, (2) the topic is somehow related with the post, (3) the topic is closely related with the post, or (4) the evaluator has not enough information for taking a decision. We applied Kappa test (Fleiss, 1971) to measure the agreement among the evaluators. The strength of agreement for 2 evaluators was very good (0.826). Such strength was moderate (0.493) when 3 evaluators agreed on the same answer. We considered an answer valid if at least two evaluators agreed on it.

Table 3 shows the results of the evaluation of the precision. The precision of the topic identification process depended on the source and its value ranged from 59.19% for social networks to 88.89% for review sites. One of the reasons that explain such variability is the specificity of the concepts mentioned in the posts of the different sources. As an example, in review sites the posts use to include references to specific brands or models, while in social networks such references are more ambiguous. Another reason is that some sources (e.g., social networks) include more misspellings than other sources (e.g., news publishing sites). With respect to the precision obtained by considering the context or not, there is not a general rule. While the first variant (without context) provides a better precision in most of the cases, the second variant (considering the other keywords in the post as context) is better for blogs, and the third variant (active context) is better for microblogs and review sites.

7. Conclusion

We have found differences among social media sources for every experiment executed.

Distribution of PoS categories vary across different sources. Since PoS tagging is a previous step for many NLP techniques, the performance of such techniques may vary according to the social media source from which the UGC have been extracted. As an example, our disambiguation strategy for topic identification uses nouns as context for performing disambiguation. Thus, sources with a higher distribution of nouns will provide more context than sources in which such distribution is smaller. The proportion of other categories may have impact over the performance of other techniques (e.g., adjectives and adverbs over sentiment analysis).

With respect to topic identification, we have found a slight difference in the performance of the technique applied between content extracted from Twitter (less accurate) and content extracted from other social media sources. Preprocessing tweets by applying syntactic normalization did not improve accuracy on Twitter since the use of n-grams of characters for language recognition is quite tolerant by itself to misspellings, or to the use of special symbols (e.g., "@" and "#").

Regarding sentiment analysis, we have found a significant difference when analysing tweets in comparison with analysing content extracted from sources different than Twitter. Sentiment analysis is much less accurate for content extracted from Twitter than for content extracted from other social media sources. As happened for language identification, performing syntactic normalization of tweets did not outperformed the accuracy of the sentiment analysis technique.

There are also differences in the precision of our topic identification technique, depending on the source in which the technique is applied. In addition, with respect to context selection criteria, from the three variants studied, there is not one that behaves better for all the sources.

	Blogs	Forums	Microblogs	Social Net.	Others	Reviews	Audiovisual	News	All
PoS tagged	99.63%	96.64%	99.01%	98.14%	98.77%	98.53%	97.20%	99.52%	98.38%
Topic Recognition									
Without context	96.7%	87.68%	94.22%	93.54%	92.71%	88.81%	90.29%	96.67%	92.35%
With context	96.64%	93.07%	95.54%	94.99%	95.13%	92.67%	97.41%	98.54%	95.02%
Active context	99.24%	89.71%	94.43%	96.4%	94.75%	93.81%	92.23%	97.4%	94.72%
Topic Recognition (after language filtering)									
Without context	91.21%	79.04%	87.54%	82.64%	86.93%	70.15%	82.52%	90.71%	82.74%
With context	88.43%	80.84%	86.31%	85.24%	88.72%	76.19%	89.66%	92.46%	84.85%
Active context	89.69%	80.51%	86.51%	86.78%	89.78%	75.59%	80.58%	90.54%	84.73%

Table 2: Coverage of the topic identification technique

	Blogs	Forums	Microblogs	Social Net.	Others	Reviews	Audiovisual	News	All
Without context	67.48%	66.67%	59.72%	72.32%	59.19%	79.17%	84.44%	71.95%	68.42%
With context	75.61%	59.35%	54.88%	65.71%	53.52%	83.87%	77.78%	64.37%	63.11%
Active context	67.71%	64.45%	65.58%	70.1%	49.15%	88.89%	79.07%	71.93%	66.59%

Table 3: Precision of the topic identification technique

To conclude, we have executed our experiments with a set of NLP tools that we have at hand. Although, other techniques may produce different results, according to our study of the distribution of PoS categories, there are differences in the language used across heterogeneous social media sources. Such differences alter the accuracy of the existing NLP techniques studied.

8. Acknowledgements

This work is supported by the Spanish CENIT project Social Media (CEN-20101037) .

9. References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the WWW*, 7:154–165, September.

William B Cavnar and John M Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*. University of Sheffield. Department of Computer Science, April.

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

Lisette García Moya. 2008. Un etiquetador morfológico para el español de Cuba. Master's thesis, Universidad de Oriente. Facultad de Matemática y Computación, Santiago de Cuba.

Andrés García-Silva, Oscar Corcho, and Jorge Gracia. 2010. Associating semantics to multilingual tags in folksonomies (poster). In *17th Int. Conference on Knowledge Engineering and Knowledge Management EKAW 2010 (poster), Lisbon (Portugal)*, October.

Jorge Gracia and Eduardo Mena. 2009. Multiontology semantic disambiguation in unstructured web contexts. In *Proc. of Workshop on Collective Knowledge Capturing and Representation (CKCaR'09) at K-CAP'09, Redondo Beach, California (USA)*. CEUR-WS, September.

Max Kaufmann and Kalita Jugal. 2010. Syntactic normalization of twitter messages. In *Proceedings of the International Conference on Natural Language Processing (ICON-2010)*.

Óscar Muñoz-García, Andrés García-Silva, Óscar Corcho, Manuel de la Higuera Hernández, and Carlos Navarro. 2011. Identifying Topics in Social Media Posts using DBpedia. In Jean-Dominique Meunier, Halid Hrasnica, and Florent Genoux, editors, *Proceedings of the NEM Summit 2011*, pages 81–86, Torino, Italy. Eurescom the European Institute for Research and Strategic Studies in Telecommunications GmbH.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Assigning part-of-speech to Dutch tweets

Mehdi Aminian, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, Menno van Zaanen

Tilburg University
Tilburg, The Netherlands
{ M.M.AminianJazi, T.P.C.Avontuur, E.Z.Azar, I.J.P.Balemans, L.M.W.Elshof, R.K.Newell, N.J.E.vanNoord, A.Ntavelos, M.M.vanZaanen }@uvt.nl

Abstract

In this article we describe the development of a part-of-speech (POS) tagger for Dutch messages from the Twitter microblogging website. Initially we developed a POS tag set ourselves with the intention of building a corresponding tagger from scratch. However, it turned out that the output of Frog, an existing high-quality POS tagger for Dutch, is of such quality that we decided to develop a conversion tool that modifies the output of Frog. The conversion consists of retokenization and adding Twitter-specific tags. Frog annotates Dutch texts with the extensive D-Coi POS tag set, which is used in several corpus annotation projects in the Netherlands. We evaluated the resulting automatic annotation against a manually annotated sub-set of tweets. The annotation of tweets in this sub-set have a high inter-annotator agreement and our extension of Frog shows an accuracy of around 95%. The add-on conversion tool that adds Twitter-specific tags to the output of Frog will be made available to other users.

1. Introduction

Social media sites provide people with an easy and accessible forum to collaborate and share information. Social media can be grouped in six types: collaborative projects, blogs and microblogs, content communities, social networking sites, virtual game worlds, and virtual social worlds (Kaplan and Haenlein, 2010). These social media are extremely popular nowadays. For instance, Twitter generates approximately 200 million tweets (140-character messages) per day[1].

Given that social media generate so much data, it is interesting to investigate the potential of extracting useful information from the data being shared through these social media channels. In order to do so, some enabling technologies are essential. In the area of natural language processing, many tools rely on part-of-speech (POS) information. POS taggers (Voutilainen, 2003) assign tags that provide information on syntactic or morphological properties to words.

In this paper, we focus on the development of a POS tagger specifically for texts generated in a microblogging context. Microblogging services, such as Twitter, allow people to share information in the form of short messages. In the case of Twitter, a maximum of 140 characters are allowed per tweet or message. This small size has caused people to be very brief, sometimes even omitting words that may be obvious to human readers from the context.

The idea of developing a POS tagger for microblogging posts is based on the work by Gimpel et al. (2011), which describes the development of a POS tagger for English tweets. More information about this project can be found in section 2. Similarly to Gimpel et al. (2011), who worked on their project with 17 people, the project discussed in this paper has been accomplished by a group of students. More specifically, the group consisted of eight Master's students from Tilburg University who had just completed a Master's course in natural language processing. The authors not only come from varying scientific backgrounds (such as linguistics and computer science) but the group also had a variety of native tongues. In addition to the theoretical knowledge the students acquired during the natural language processing course, this project, which took approximately a week in person-hours, offered them a hands-on experience and insight into the practical decisions that need to be made when working on real-world natural language processing projects.

2. Background

This project is based on a similar project by Gimpel et al. (2011). They address the problem of POS tagging for English data from the microblogging service, Twitter. They develop a tag set, annotate data, develop features and conduct experiments to evaluate these features. The evaluation is designed in such a way to make it possible to test the efficacy of the feature set for POS tagging given limited training data. The features relate to Twitter orthography, frequently-capitalized tokens, the traditional tag dictionary, distributional similarity and phonetic normalization. The tagger with the full feature set leads to 89.37% accuracy on the test set. The project of Gimpel et al. (2011) was accomplished in 200 person-hours spread across 17 people and two months. With the results of their project, they want to provide richer text analysis of Twitter and related social media datasets. They also believe that the annotated data can be useful for research into domain adaptation and semi-supervised learning.

The effectiveness of the large amounts of data is shown in several studies. Keep in mind that while microblogging services generate large amounts of data, this also includes a large amount of "useless" data if one considers using the data for a particular purpose or when searching for information on a particular subject. Recently, there have been studies on the use of Twitter information in the area of sentiment analysis. In these cases, English POS tags are being used increasingly to analyze different aspects of social networks and Twitter in particular.

[1] http://blog.twitter.com/2011/08/your-world-more-connected.html

In the research paper, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", English POS tags for Twitter are used for the task of sentiment analysis (Pak and Paroubek, 2010). In this work, the researchers show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain the discovered phenomena. Using the corpus, they build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. The researchers use TreeTagger for POS tagging and observe the difference in distributions of POS tags among positive, negative and neutral sets. Results show that some POS tags might be strong indicators of emotional text.

Other research that uses Twitter information focuses on a combination of sentiment and event detection on Twitter. An example of a study in this field is from Thelwall et al. (2011), who assess whether popular events are typically associated with increases in sentiment strength. They find strong evidence that popular events are normally associated with increases in negative sentiment strength and some evidence that the same is true for positive sentiment. However, the negative sentiment seems to be more central than the positive one.

Another example is the study about real-time event detection by social sensors by Sakaki et al. (2010). The authors devise a classifier of tweets based on features as keywords in a tweet, number of words, and their context. Secondly, they produce a probabilistic spatiotemporal model which can detect the center and the trajectory of the event location. In this work, every Twitter user is seen as a sensor. Filters are used to estimate location. Using this approach, they construct an earthquake reporting system in Japan. Because of the numerous earthquakes and Twitter users throughout Japan, they are able to detect earthquakes with a probability of 96%.

3. System overview

3.1. Tag set

Initially, we followed the process from Gimpel et al. (2011) in the development of a POS tag set for Dutch Twitter data. We started from their tag set and for each of the tags checked whether the tag made any sense in Dutch. It turned out that some of the English POS tags are not relevant in Dutch. For instance, the situation of nominal and verbal glued together, which is described by the 'L' POS tag, does not occur in Dutch.

To come up with better (non-Twitter specific) POS tags, we considered existing POS tag sets for Dutch, with the intention of extending these with Twitter specific tags. In this context, we looked at the POS tag set that is used in the SoNaR project.

SoNaR, which stands for Stevin Nederlandstalig Referentiecorpus is a corpus building project aiming at compiling a large corpus containing contemporary written Dutch (and Flemish). It is currently under development by Radboud University Nijmegen, Tilburg University, University of Twente, Utrecht University and KU Leuven. This project is financed within the Dutch-Flemish Stevin project[2] and is an extension of the D-Coi (Dutch Language Corpus Initiative) project (Oostdijk et al., 2008).

The tag set used in the SoNaR project is originally developed in the D-Coi project. The D-Coi tag set is described in more detail in Van Eynde (2005). This is an extensive tag set consisting of a total of 320 distinct tags. The tags are grouped by main tag, of which there are 13. Many specific tags are specializations of the main tag. For instance, the 'N' tag specifies nouns, which can be made more specific by adding arguments: 'N(soort,ev,basis,onz,stan)', which is a singular ('ev'), neuter ('onz'), common noun ('soort') in a non-diminutive ('basis') and nominal ('stan') form.

When analyzing the D-Coi tag set, it became clear that Twitter data requires some additional tags that are not present in the tag set used to annotate "regular" linguistic texts. Hence, we needed to extend the D-Coi POS tag set. When making decisions on which POS tags to select from the D-Coi tag set or to add, we took two factors into consideration:

1. the variety of parts-of-speech that can be encountered in Dutch tweets, and

2. the ease of user who will utilize our POS tagger.

In this sense we aim at choosing POS tags which will give enough information to discriminate POS of importance to the user.

Given the consideration and combining this with our aim for compatibility of the SoNaR project, we chose to base our tags on the main tags taken from the D-Coi tag set. The reason for not using the full D-Coi tag set is that we expected problems with manual annotation.

To incorporate Twitter-specific information, we had to add some Twitter-specific tags to the tag set. We chose to use the same Twitter-specific tags as Gimpel et al. (2011) used in their study. This led to the tag set that can be found in table 1. Two of these tags have more specific variants that deal with the more detailed linguistic aspects of the token it describes. 'N' has two sub-types: 'N(eigen)': proper nouns and 'N(soort)': common noun. 'SPEC' has seven sub-types to deal with symbols, incomprehensible words, abbreviations, etc.

However, during the development of the implementation of the POS tagger, we came across Frog[3], a POS tagger that can handle Dutch text and assigns tags according to the full D-Coi tag set. Initial experiments showed that the output of this tool is of such quality that it can also be used on Twitter data.

In the end, the availability of Frog, combined with the support of the POS tag set used in SoNaR and D-Coi allowed us to use the full D-Coi tag set. The only modification required was the addition of the Twitter-specific tags as shown in the rightmost column of table 1. These tags are based on the work of Gimpel et al. (2011).

[2] http://lands.let.ru.nl/projects/SoNaR/

[3] A more extensive description of the system is presented in section 3.2.

Table 1: Initial POS tag set.

Generic		Twitter	
ADJ	Adjective	AT	@ mention
BW	Adverb	DISC	Discourse marker
LET	Punctuation	EMO	Emoticon
LID	Determiner	HASH	# tag
N	Noun	URL	URL
SPEC	Special token		
TSW	Interjection		
TW	Number/ordinal		
VG	Conjunction		
VNW	Pronoun		
VZ	Preposition		
WW	Verb		

3.2. System implementation

The tweets were initially tagged using the POS tagger Frog, formerly known as Tadpole. Frog is a complete system that comes with the UCTO[4] tokenizer incorporated. Frog produces tab-delimited column-formatted output, one line per token. An example of such output can be found in table 2. The nine columns contain the following information (in order from left to right):

1. Token number (resets every sentence);
2. Token;
3. Lemma (according to the memory-based lemmatizer MBLEM[5]);
4. Morphological segmentation (according to the memory-based morphological analyzer MBMA[6]);
5. POS tag (D-Coi tag set; according to the memory-based tagger MBT[7]);
6. Confidence in the POS tag, which is a number between 0 and 1. This represents the probability mass assigned to the best guess tag in the tag distribution;
7. Chunker or shallow parser output on the basis of MBT;
8. Token number of head word (according to the constraint-satisfaction inference-based dependency parser, CSI-DP);
9. Type of dependency relation with head word.

From the Frog output we extract the token (2) and POS tag (5) columns and then automatically convert it to a Twitter-specific format. The conversion is based on a collection of regular expressions modifying the Frog output. This means that when needed we add Twitter-specific tags: 'HASH', 'AT', 'DISC', 'URL' or 'EMO'. In certain cases, this requires retokenization of the input. For instance, this is required when '#' or '@' tokens are found. In the cases of discourse markers or URLs, we changed the tag to DISC and URL respectively. As a URL we considered every token that begins with 'http://' or 'www.'. Moreover, URLs like 'http://www.youtube.com/watch?v=lRzFqW4Xh2k' which were separated by Frog at punctuation characters such as '=' in this case, are also retokenized.

Regarding the emoticons, we manually created a list of 156 emoticons that were found in the collection of tweets. We also included cases of big emoticons like: ':-))))))))'. Additionally, emoticons formed in a reversed fashion were added in the list because there are users that use emoticons in the way around (from right to left). This list covers the vast majority of the emoticons that are found in tweets.

Finally, the processing was done in parallel with the actual texts in order to avoid wrong conversion in cases similar to e.g. 'C# programming' which otherwise would lead to a tagging like '#programming' with a 'HASH' tag. Table 3 provides an example depicting the conversion of the Frog output to the Twitter-specific format. Note that the empty lines in the Twitter column are not in the output, but merely illustrate the alignment with the Frog column.

4. Experiments

To evaluate the quality of the output of the Frog POS tagger combined with the addition and modification of the output into Twitter-specific tags, we apply the tool to the collection of tweets that was provided by the SoNaR project. The output of this automatic annotation serves as an input for manual correction of the annotated tweets.

To perform the manual checking of the automatically annotated tweets we first tried to use the annotation tools Callisto[8] and MMAX2[9]. However, both systems turned out to be user unfriendly. Callisto cannot handle large amounts of tags (our POS tag set consists of 325 distinct tags). Changing tags using MMAX2 turned out to be difficult. In the end, we decided to use Gate[10]. Gate's annotation tool also had a minor disadvantage; it allows annotators to change the actual text (of the tweets), which is undesirable. Furthermore, it allows editing of the POS tags themselves, which can lead to inconsistencies.

We then evaluated the performance of the Twitter POS tagger by comparing the manually corrected output against the POS tagger output. In section 4.2., we provide information on the consistency of manual tagging/checking in the form of inter-annotator agreement and we will discuss the performance of the full system in the form of accuracy and F-score.

4.1. Dataset

The dataset that has been used in the experiments consists of 1,074,360 tweets. The large majority of these are tweets in Dutch, but we managed to identify a few non-Dutch tweets in the corpus. As mentioned earlier, the collection comes from the SoNaR corpus.

The original format of the tweets in the collection included among others timestamp, re-tweet information and any

[4] http://ilk.uvt.nl/ucto/
[5] http://ilk.uvt.nl/mbma/
[6] http://ilk.uvt.nl/mbma/
[7] http://ilk.uvt.nl/mbt/
[8] http://callisto.mitre.org/
[9] http://mmax2.sourceforge.net/
[10] http://gate.ac.uk/

Table 2: Frog column output.

1	Ze	ze	[ze]	VNW(pers,pron,stan,red,3,ev,fem)	1.000000	B-NP	2	su
2	vroeg	vragen	[vraag]	WW(pv,verl,ev)	0.532544	B-VP	0	ROOT
3	zich	zich	[zich]	VNW(refl,pron,obl,red,3,getal)	0.999740	B-NP	2	se
4	af	af	[af]	VZ(fin)	0.996853	O	2	svp
5	of	of	[of]	VG(onder)	0.733333	B-SBAR	2	vc
6	hij	hij	[hij]	VNW(pers,pron,nomin,vol,3,ev,masc)	0.999659	B-NP	8	su
7	nog	nog	[nog]	BW()	0.999930	B-ADVP	8	None
8	zou	zullen	[zal]	WW(pv,verl,ev)	0.999947	B-VP	5	body
9	komen	komen	[kom][en]	WW(inf,vrij,zonder)	0.861549	I-VP	8	vc
10	.	.	[.]	LET()	0.999956	O	9	punct

Table 3: Conversion from Frog to Twitter-specific output.

Frog		Twitter	
RT	SPEC(symb)	RT	DISC
@	SPEC(symb)		
nilicule	ADJ(prenom,basis,met-e,stan)	@nilicule	AT
#	SPEC(symb)		
sdgeld	WW(vd,vrij,zonder)	#sdgeld	HASH
http://t.co/74h22oo	SPEC(deeleigen)	http://t.co/74h22oo	URL
:	LET()		
-	LET()		
)	LET()		
)	LET()		
)	LET()	:-)))	EMO

URLs that are found in the tweet. In our project, we only considered the actual text of the tweets for further processing. All other information was discarded (but it is trivial to link the additional information with the POS tagged version of the tweets).

Going over the tweets manually, we identified specific aspects of the special nature of the tweets as texts in contrast to "regular text". Based on our qualitative analysis of Dutch tweets, we summarize those differences as follows:

Discourse markers Tweets may contain discourse markers like RT which is used when someone re-tweets another user's tweet. These types of discourse markers are typically not found in regular text.

@ mentions When a user wants to refer to another Twitter user, they use the character '@' before their Twitter user name;

 # tags People use the hash tag symbol '#' before relevant keywords in their tweet to categorize those tweets so that they are returned more easily as results of a Twitter search;

Alternative grammar and spelling Probably due to the limited length of a tweet (of at most 140 characters), tweets usually lack coherence. Also, they are sometimes written with limited grammar and non-perfect spelling.

An example of a typical Dutch tweet is: "RT @JoelSerphos: Kunnen de jongeren van #Iran de wereld net zo inspireren als hun leeftijdsgenoten in Egypte." (which translates to "RT @JoelSerphos: Can the youth of #Iran inspire the rest of the world just like their peers in Egypt." This tweet contains a discourse marker (RT), followed by an @ mention. Furthermore, "Iran" is used with a # tag.

4.2. Quantitative results

To conduct an evaluation of the generated output we need to build a gold standard dataset that contains POS tag annotation. We can then compare the output of the system against this gold standard dataset. For this purpose, we manually corrected the generated output of 1,056 tweets. This task has been done by three (Dutch) annotators who all manually corrected the POS tags of all of the approximately one thousand tweets.

To investigate the consistency with which the annotators agreed to the tags, we considered inter-annotator agreement. To measure inter-annotator agreement, we chose to use two measures: Cohen's Kappa and Fleiss' Kappa. Cohen's Kappa measures inter-annotator agreement between two annotators. Since we have three annotators, we compute this measure in pairs at a time, which leads to three results. We provide the pair-wise results in table 4 and also show the average inter-annotator agreement. Furthermore, to reach an overall inter-annotator agreement score, we also computed the Fleiss' Kappa which can compare multiple annotators at once. The results of the inter-annotator agreement can be found in table 4. As can be seen from this table, the inter-annotator agreement is very high. Note that the average Cohen's Kappa and the Fleiss' Kappa are only the same due to rounding.

Even though there may be some discussion on how to interpret these values, the inter-annotator agreement measures show consistently high values, which leads us to conclude

Table 4: Inter-annotator agreement of gold standard POS tags.

Measure	Annotators	Score
Cohen's Kappa	A vs. B	91.20
	A vs. C	92.07
	B vs. C	93.73
Average Cohen's Kappa		92.33
Fleiss' Kappa		92.33

Table 5: Evaluation results.

	Accuracy	F-Score
Complete tag set	92.87	92.61
Complete, simplified tags	94.12	93.94
Modified tokens	51.11	35.29
Modified tokens, simplified tags	50.57	34.43

that there was near complete agreement amongst the annotators. Note however, that the annotators corrected the POS tags and did not annotate the tweets from scratch, which would likely have led to a lower inter-annotator agreement.
During the process of manually correcting the POS tags of the tweets, the annotators noticed that the language used in the tweets corresponds highly with "regular" Dutch. As mentioned earlier, alternative spelling and grammar in tweets does occur, but not very frequently. Because of this, the quality of the output of Frog is already expected to be high. More research into the portion of creative use of language in tweets needs to be conducted to get a better idea on the impact of this phenomenon.
In table 5 the results of four evaluations are shown. Firstly, an evaluation is performed on the entire gold standard dataset with detailed POS tags (in other words, the full D-Coi tag set extended with the Twitter-specific tags). Secondly, the same evaluation is performed, but on a simplified tag set. For each of the complex POS tags, such as 'N(soort,ev,basis,zijd,stan)', only the main POS tag is used. In this case, the tag would be 'N'.
For the third and fourth evaluation only the tokens that are tagged differently by at least one of the annotators are taken into consideration. In table 5 these results are referred to as modified tokens. This comes down to 1,981 out of a total of 16,881 tokens in the gold standard dataset. Similarly to the first and second evaluation, the third evaluation makes use of the detailed tags while the fourth measures using the simplified tags.
Note that the modified tokens are the difficult tokens. The annotators did not necessarily agree in these cases. From the 1,981 tokens, there are 272 tokens for which the tags selected by the annotators did not lead to a majority vote. In these cases, a random selection was made.
Additionally, since the modified tokens are exactly the tokens where (at least one of) the annotators did not agree with the system output, we can expect that these results are much lower than the overall result. The fact that the accuracy for these cases is still around 50% means that the majority vote over all annotators still lead to the system output half of the time.
The results show that overall the performance of the POS tagger (Frog output converted into a Twitter-specific tag set) performs very well. Considering the complete (manually annotated gold standard) data set, accuracy and F-score are both over 90%.
Unfortunately, we cannot compare the output of our modified Frog against the output of plain Frog (without the conversion module). This is due to the retokenization of emoticons and URLs.

4.3. Qualitative results

During the manual annotation, the annotators encountered some consistent problems in the system output. URLs, for example, are hard to identify correctly because of tokenization problems. UCTO tokenizes parts of URLs, which leads to whitespace between parts of URLs, such as "echtbroodjeaap␣.␣nl". As a result of tokenization, the different tokens are annotated separately instead of as part of the URL.
Another difficulty is found with the tag that is used to annotate names ('SPEC(deeleigen)'). Although in most of the cases this tag is assigned to tokens correctly, sometimes this tag is too generic and a more specific tag would have been more appropriate. The tag set contains more specific tags for names, providing more information about the token such as gender, number, etc.
Another case deals with imperatives and interjections, which are also often tagged incorrectly. The latter, for example, is sometimes tagged as a verb instead of an interjection: "zeker", for instance, in the context of "ja, zeker!" (which translates to "yeah, sure"), is tagged as an adjective in its basic form. In this context, however, the token should obviously be tagged differently.
Finally, sometimes the system fails to recognize emoticons correctly. In some cases emoticons are not recognized where they should be recognized (false negatives). This is due to the fact that emoticons are used very creatively in tweets, which implies that a rather long list of emoticons is required in the system. In other cases, the system identifies an emoticon which is not a true emoticon (false positive). For instance, emoticons are found in places that do not practically allow for emoticons, such as within URLs, such as "(http://...=)".

5. Conclusion

Social media, Twitter in particular, is growing rapidly worldwide. In 2011 The Netherlands ranked #1 worldwide in penetration for Twitter users[11]. This rapid growth of Dutch tweets provides a great source of user-created contents in the Dutch language which can serve as an informal basis of information. However, to tap into this source of information, the data needs to be analyzed and understood. The POS tagger developed and presented in this paper can be applied to many linguistic analysis studies that involve Dutch tweets. This study provides a tool that enables a richer linguistic analysis of Dutch tweets.

[11] http://www.comscore.com/Press_Events/Press_Releases/2011/4/The_Netherlands_Ranks_number_one_Worldwide_in_Penetration_for_Twitter_and_LinkedIn

In this study we have modified the Frog POS tagger for Dutch to annotate Dutch tweets by adding a set of Twitter-specific tags. The results showed that it is possible to annotate Twitter-specific language. However, some problems remain. For instance, Frog finds it hard to identify URLs. This is partially solved by adjusting the conversion script, however, a modification of the UCTO tokenizer may be a more consistent solution. Furthermore, in this research we used a static list to recognize emoticons. This might pose a problem since emoticons are used creatively. A dynamic emoticon recognizer might help to deal with this creativity. Future work should include a deeper analysis of system errors and a possible modification of the conversion scripts to handle errors that are made consistently by the current system. Finally, to improve usability, the system should be build as a direct extension of Frog or perhaps even be included in the Frog distribution.

6. References

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers; Portland, OR, USA*, pages 42–47, New Brunswick:NJ, USA, June. Association for Computational Linguistics.

A.M. Kaplan and M. Haenlein. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, January–February.

N. Oostdijk, M. Reynaert, P. Monachesi, G. van Noord, R.J.F. Ordelman, I. Schuurman, and V. Vandeghinste. 2008. From d-coi to sonar: A reference corpus for dutch. In *Proceedings on the sixth international conference on language resources and evaluation (LREC 2008); Marrakech, Marokko*, pages 1437–1444. ELRA. ISBN=2-9517408-4-0.

A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010); Valletta, Malta*, pages 1320–1326.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web; Raleigh, NC, USA*, pages 851–860. ACM, April.

M. Thelwall, K. Buckley, and G. Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

F. Van Eynde. 2005. Part of speech tagging en lemmatisering van het D-COI corpus.

A. Voutilainen. 2003. Part-of-speech tagging. In R. Mitkov, editor, *The Oxford handbook of computational linguistics*, pages 219–232. Oxford University Press, New York:NY, USA.

Challenges in developing opinion mining tools for social media

Diana Maynard, Kalina Bontcheva, Dominic Rout

Department of Computer Science
University of Sheffield
Regent Court, Sheffield, S1 4DP, UK
diana@dcs.shef.ac.uk

Abstract

While much work has recently focused on the analysis of social media in order to get a feel for what people think about current topics of interest, there are, however, still many challenges to be faced. Text mining systems originally designed for more regular kinds of texts such as news articles may need to be adapted to deal with facebook posts, tweets etc. In this paper, we discuss a variety of issues related to opinion mining from social media, and the challenges they impose on a Natural Language Processing (NLP) system, along with two example applications we have developed in very different domains. In contrast with the majority of opinion mining work which uses machine learning techniques, we have developed a modular rule-based approach which performs shallow linguistic analysis and builds on a number of linguistic subcomponents to generate the final opinion polarity and score.

1. Introduction

In this new information age, where thoughts and opinions are shared so prolifically through online social networks, tools that can make sense of the content of these networks are paramount. In order to make best use of this information, we need to be able to distinguish what is important and interesting. There are obvious benefits to companies, governments and so on in understanding what the public think about their products and services, but it is also in the interests of large public knowledge institutions to be able to collect, retrieve and preserve all the information related to certain events and their development over time. The spread of information through social networks can also trigger a chain of reactions to such situations and events which ultimately lead to administrative, political and societal changes.

Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, rapidly changing to reflect the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. The social networks are pools of a wide range of articulation methods, from simple "I like it" buttons to complete articles, their content representing the diversity of opinions of the public. The user activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles, persons, locations) and topics (e.g. global warming, financial crisis, swine flu). In order to include this information, a semantically-aware and socially-driven preservation model is a natural way to go: the exploitation of Web 2.0 and the wisdom of crowds can make web archiving a more selective and meaning-based process. The analysis of social media can help archivists select material for inclusion, providing content appraisal via the social web, while social media mining itself can enrich archives, moving towards structured preservation around semantic categories.

Within this work, we focus on the challenges in the development of opinion mining tools which, along with entity, topic and event recognition, form the cornerstone for social web analysis in this respect. We discuss a variety of issues related to the adaptation of opinion mining tools to social media, and the challenges they impose on a Natural Language Processing (NLP) system, along with two example applications we have developed in very different domains: socially aware federated political archiving (realised by the national parliaments of Greece and Austria), and socially contextualized broadcaster web archiving (realised by two large multimedia broadcasting organizations based in Germany: Sudwestrundfunk and Deutsche Welle). The approach we have developed forms part of a set of tools for the archiving of community memories and the long-term preservation of (multilingual) Social Web content. Based around a number of use cases in various domains, ultimately we aim to answer questions such as:

- What are the opinions on crucial social events and on the key people involved?
- How are these opinions distributed in relation to demographic user data?
- How have these opinions evolved over time?
- Who are the opinion leaders?
- What is their impact and influence?

There are many challenges inherent in applying typical opinion mining and sentiment analysis techniques to social media. Microposts such as tweets are, in some sense, the most challenging text type for text mining tools, and in particular for opinion mining, since they do not contain much contextual information and assume much implicit knowledge. Ambiguity is a particular problem since we cannot easily make use of coreference information: unlike in blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also exhibit much more language variation, tend to be less grammatical than longer posts, contain unorthodox capitalisation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect. On the other hand, their

terseness can also be beneficial in focusing the topics more explicitly: it is very rare for a single tweet to be related to more than one topic, which can thus aid disambiguation by emphasising situational relatedness.

Most opinion mining techniques make use of machine learning, but this is problematic in applications such as ours where a number of different domains, languages and text types are involved, because models have to be trained for each one, and large amounts of training data are required for good results. Typically, classifiers built using supervised methods, e.g. (Boiy et al., 2007), perform well on polarity detection tasks, but when used in new domains, their accuracy reduces disastrously (Aue and Gamon., 2005). While some work has focused on adapating ML methods to new domains (Blitzer et al., 2007), this only really focuses on the use of different keywords in similar kinds of text, e.g. product reviews about books vs. reviews about electronics. Our entity-centric approach, on the other hand, makes use of rule-based NLP techniques, but in contrast to more traditional NLP approaches involving full parsing, we use a much shallower but more focused approach based around entity and event recognition, which lends itself better to non-standard text.

In the following section, we discuss some related work in the field of opinion mining and more generally, in the field of text mining from social media. We then describe in Section 3 the approach we have adopted, and some ofthe challenges faced in Section 4. In Section 5 we discuss evaluation issues and give some preliminary results, and finish with an outlook to the future in Section 6.

2. Related Work

(Pang and Lee, 2008) present a wide-ranging and detailed review of traditional automatic sentiment detection techniques, including many sub-components, which we shall not repeat here. In general, sentiment detection techniques can be roughly divided into lexicon-based methods (Popescu and Etzioni, 2005; Scharl and Weichselbraun, 2008; Taboada et al., 2011) and machine-learning methods, e.g. (Boiy and Moens, 2009). Lexicon-based methods rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. Machine learning approaches make use of syntactic and/or linguistic features (Pak and Paroubek, 2010b; Go et al., 2009), and hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods, e.g. (Diakopoulos et al., 2010). For example, (Moghaddam and Popowich, 2010) establish the polarity of reviews by identifying the polarity of the adjectives that appear in them, with a reported accuracy of about 10% higher than pure machine learning techniques. However, such relatively successful techniques often fail when moved to new domains or text types, because they are inflexible regarding the ambiguity of sentiment terms. The context in which a term is used can change its meaning, particularly for adjectives in sentiment lexicons (Mullaly et al., 2010). Several evaluations have shown the usefulness of contextual information (Weichselbraun et al., 2010; Wilson et al., 2009), and have identified context words with a high impact on the polarity of ambiguous terms (Gindl et al., 2010). A further bottleneck is the time-consuming creation of these sentiment dictionaries, though solutions have been proposed in the form of crowdsourcing techniques[1].

Recently, techniques for opinion mining have begun to focus on social media, combined with a trend towards its application as a proactive rather than a reactive mechanism. Understanding public opinion can have important consequences for the prediction of future events. One of the most obvious applications of this is for stock market predictions: (Bollen and Mao, 2011) found that, contrary to the expectation that if the stock markets fell, then public mood would also become more negative, in fact a drop in public mood acts as a precursor to a fall in the stock market.

Almost all the work on opinion mining from Twitter has used machine learning techniques. (Pak and Paroubek, 2010b) aimed to classify arbitrary tweets on the basis of positive, negative and neutral sentiment, constructing a simple binary classifier which used n-gram and POS features, and trained on instances which had been annotated according to the existence of positive and negative emoticons. Their approach has much in common with an earlier sentiment classifier constructed by (Go et al., 2009), which also used unigrams, bigrams and POS tags, though the former demonstrated through analysis that the distribution of certain POS tags varies between positive and negative posts. One of the reasons for the relative paucity of linguistic techniques for opinion mining on social media is most likely due to the difficulties in using NLP on low quality text, something which machine learning techniques can – to some extent – bypass with sufficient training data. For example. the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets (Liu et al., 2010). (Ritter et al., 2011) also demonstrate some of the difficulties in applying traditional POS tagging, chunking and Named Entity Recognition techniques to tweets, proposing a solution based on LabeledLDA (Ramage et al., 2009).

There also exists a plethora of commercial search-based tools for performing sentiment analysis of tweets. Generally, the user enters a search term and gets back all the positive and negative (and sometimes neutral) tweets that contain the term, along with some graphics such as pie charts or graphs. Typical basic tools are Twitter Sentiment[2], Twends[3] and Twitrratr[4]. Slightly more sophisticated tools such as SocialMention[5] allow search in a variety of social networks and produce other statistics such as percentages of Strength, Passion and Reach, while others allow the user to correct erroneous analyses. On the surface, many of these appear quite impressive, and have the advantage of being simple to use and providing an attractive display with copious information about trends. However, such tools mostly aim at finding public opinion about famous people, sports events, products, movies and so on, but do not lend themselves easily to more complex kinds of opinion or to more abstract kinds of searches. Furthermore, their analy-

[1] http://apps.facebook.com/sentiment-quiz
[2] http://twittersentiment.appspot.com/
[3] http://twendz.waggeneredstrom.com/
[4] http://twitrratr.com/
[5] http://socialmention.com/

sis tends to be fairly rudimentary, performance can be quite low, and many of them do not reveal the sources of their information or enable any kind of evaluation of their success: if they claim that 75% of tweets about Whitney Houston are positive, or that people on Facebook overwhelmingly believe that Greece should exit the eurozone, we have no proof as to how accurate this really is.

Our approach to opinion mining takes inspiration from a number of sources. It is most similar to the work of (Taboada et al., 2011) in terms of technique, but because we focus on social media, we need to employ some different strategies to deal with the linguistic issues imposed. For example, we incorporate detection of swear words, sarcasm, questions, conditional statements and so on, while our entity-centric approach focuses the opinions on specific topics and makes use of linguistic relations.

3. Opinion mining

We have developed a series of initial applications for opinion mining from social media using GATE (Cunningham et al., 2002), a freely available toolkit for language processing. Based on the work described in (Maynard and Funk, 2011), which focused on identification in tweets of sentiments about political parties, we have extended this to a more generic analysis of sentiment about any kind of entity or event mentioned, within two specific domains: the current Greek financial crisis and the Rock am Ring rock festival in Germany in 2010. In both cases, we perform first a basic sentiment analysis by associating a positive, negative or neutral sentiment to each relevant opinion target, together with a polarity score. In the current scenarios, this could be any entity or event which is pertinent to the domain and use case. In the Rock am Ring corpus, this might be the overall event, a band or a band's particular performance at the concert, or some sub-event such as a light show that occurred during the performance. In the Greek crisis corpus, this might be a politician, an organisation, or an event such as a general strike or a relevant meeting that took place.

3.1. Entity extraction

The opinion mining application first requires that the corpus be annotated with entities and events. For this we have also developed a series of applications in GATE. We use a modified version of ANNIE (Cunningham et al., 2002), the default Named Entity (NE) recognition system in GATE, to find mentions of Person, Location, Organization, Date, Time, Money and Percent (though we only use the first three of these as potential opinion targets – the other entity types are used as additional indicators and, in some cases, feature values, in the linguistic patterns for opinion mining. We include some extra subtypes of Organization such as Band (for the Rock am Ring domain) and Political Party (for the Greek crisis domain), and have relaxed some of the settings to deal with the incorrectness of the English, though this has important ramifications. Detecting NEs in tweets, in particular, is challenging and we are currently performing some separate experiments about this. Enabling gazetteer lists to match against lowercase versions of proper nouns, for example, entails much greater ambiguity with common nouns. For example, the month "May" would be matched with the verb "may" – and even though we can also use a version of the POS tagger specially trained to deal with case-insensitive text, this is by no means guaranteed to work accurately all the time.

In addition to named entities, we also acquire a set of domain-specific terms using TermRaider[6]. This considers all noun phrases (NPs) – as determined by linguistic processing tools in GATE – as candidate terms, and then ranks them in order of termhood according to three different scoring functions: (1) basic tf.idf (Buckley and Salton, 2009) (2) an augmented tf.idf which also takes into account the tf.idf score of any hyponyms of a candidate term, and (3) the Kyoto score based on (Bosma and Vossen, 2010), which takes into account the number of hyponyms of a candidate term occurring in the document. All are normalised to represent a value between 0 and 100. We have not yet formally evaluated the three methods, though this is part of our planned future work, and indeed, it is possible that this may differ for differing domains or text types. Two further restrictions are placed. First, a candidate term is not considered as an entity if it matches or is contained within an existing Named Entity. Second, we set a threshold score above which we consider a candidate term to be valid. This threshold is a parameter which can be manually changed at any time – currently it is set to an augmented score of 45, i.e. only terms with a score of 45 or greater will be annotated as an Entity and used as input for the opinion mining and other tools.

3.2. Event recognition

In addition to entities, we also identify events to be used as possible targets for the opinions, and as input for other processes such as topic extraction (which fall outside the scope of this paper). Events can be expressed by text elements such as verbal predicates and their arguments ("The committee dismissed the proposal"), noun phrases headed by nominalizations ("economic growth"), adjective-noun combinations ("governmental measure"; "public money") and event-referring nouns ("crisis", "cash injection").

The pattern-based method we adopt involves the recognition of entities and the relations between them in order to find domain-specific events and situations, and is described more fully in (Risse et al., 2011). Currently we use only the events recognised by the top-down template-based approach, which consists of identifying a number of important events in advance, based on analysis of the user needs and manual inspection of the corpora. The template slots are pre-defined, and the values are entities extracted from the text as described in Section 3.1. In a semi-closed domain, this approach is preferable over the bottom-up approach, because it generates much higher precision results, while recall is not affected as significantly as in an open domain scenario.

Work on the event recognition is still very much in progress, though preliminary experiments showed very high precision (98% on a corpus of 1474 extracted events in the Greek

[6] http://gate.ac.uk/projects/neon/termraider.html

crisis dataset). We have not yet applied the event recognition to our Twitter or German datasets, where we expect to get somewhat lower results; however, these will be highly dependent on the quality of the entities extracted. Actually, we expect the quality of the event recognition (assuming correct entity detection) to be affected less by the typical problems associated with social media than the quality of the opinion mining and entity recognition tools, because we use such a shallow approach.

3.3. Sentiment Analysis

The approach we take for sentiment analysis is a rule-based one which is quite similar to that used by (Taboada et al., 2011), focusing on building up a number of subcomponents which all have an effect on the score and polarity of a sentiment. The main body of the opinion mining application involves a set of JAPE grammars which create annotations on segments of text. JAPE is a Java-based pattern matching language used in GATE (Cunningham et al., 2000). The grammar rules use information from gazetteers combined with linguistic features (POS tags etc.) and contextual information to build up a set of annotations and features, which can be modified at any time by further rules. The set of gazetteer lists contains useful clues and context words: for example, we have developed a gazetteer of affect/emotion words from WordNet (Miller et al., 1980). These have a feature denoting their part of speech, and information about the original WordNet synset to which they belong. The lists have been modified and extended manually to improve their quality: some words and lists have been deleted (since we considered them irrelevant for our purpose) while others have been added.

Once sentiment-bearing words have been matched, an attempt is made to find a linguistic relation between an entity or event in the sentence or phrase, and one or more sentiment-bearing words, such as a sentiment-bearing adjective modifying an entity or in apposition with it, or a sentiment-bearing verb whose subject or direct object is an entity. If such a relation is found, a Sentiment annotation is created for that entity or event, with features denoting the polarity (positive or negative) and the polarity score. The initial score allocated is based on that of the gazetteer list entry of the relevant sentiment word(s). The concept behind the scoring (and final decision on sentiment polarity) is that the default score of a word can be altered by various contextual clues. For example, typically a negative word found in a linguistic association with it will reverse the polarity from positive to negative and vice versa. Similarly, if sarcasm is detected in the statement, the polarity is reversed (in the vast majority of cases, sarcasm is used in conjunction with a seemingly positive statement, to reflect a negative one, though this may not necessarily be true of other languages than English). Negative words are detected via our Verb Phrase Chunker (e.g. “didn’t”) and via a list of negative terms in a gazetteer (e.g. “not”, “never”.). Adverbs modifying a sentiment adjective usually have the effect of increasing its intensity, which is reflected by multiplying the intensity factor of the adverb (defined in a gazetteer list) by the existing score of the adjective. For example, if “brilliant” had a score of 0.4, and “absolutely” had an intensity factor of 2, then the score of “brilliant” would increase to 0.8 when found in the phrase “absolutely brilliant”. Currently, the intensity factors are defined manually, but some of these could also be generated automatically where they are morphologically derived from an adjective (e.g. we could use the sentiment score of the adjective “brilliant” defined in our adjective list to generate an intensity factor for the adverb “brilliantly”).

Swear words, on the other hand, have a slightly more complex role. These are particularly prolific on Twitter, especially in the Rock am Ring corpus and on topics such as politics and religion, where people tend to have very strong views. First, we match against a gazetteer list of swear words and phrases, which was created manually from various lists found on the web and from manual inspection of the data, including some words acquired by collecting tweets with swearwords as hashtags (which also often contain more swear words in the main text of the tweet). The following rules are then applied:

- Swear words that are nouns get treated in the same way as other sentiment-bearing words described above. For example, in the tweet "Ed Miliband the world’s first talking garden gnome #f***wit", the word "f***wit" is treated as a sentiment-bearing word found in association with the entity "Ed Milliband".

- Swear words that are adjectives or adverbs are treated in the same way as regular adverbs, increasing the strength of an existing sentiment word. For example, if "awesome" scores 0.25, "fricking awesome" might score 0.5.

- Finally, any sentences containing swear words that have not been previously annotated are awarded a Sentiment annotation on the whole sentence (rather than with respect to an entity or event). For example, "Imagine saying how accepting of religions you are one day and the next writting a blog about how f***ed religions are" has no sentiment-bearing words other than the swear word, so the whole sentence is just flagged as containing a swearing sentiment. In this case, it is not easy to establish whether the sentiment is positive or negative – in the absence of any other clues, we assume such sentences are negative if they contain swear words and no positive words.

Finally, emoticons are processed like other sentiment-bearing words, according to another gazetteer list, if they occur in combination with an entity or event. For example, the tweet "They all voted Tory :-(" would be annotated as negative with respect to the target "Tory". Otherwise, as for swear words, if a sentence contains a smiley but no other entity or event, the sentence gets annotated as sentiment-bearing, with the value of that of the smiley from the gazetteer list.

Once all the subcomponents have been run over the text, a final output is produced for each sentiment-bearing segment, with a polarity (positive or negative) and a score.

3.4. Multilingual issues

Another artefact of social media is that corpora consisting of blogs, forums, Facebook pages, Twitter collections and so on are often multilingual. In our Rock am Ring corpus, comments and tweets can be in either English or German, while in the Greek financial crisis corpus, they can be in English or Greek, but also sometimes in other languages such as French. We therefore employ a language identification tool to determine the language of each sentence. The tool we use is a GATE plugin for the TextCat language identifier[7], which is an implementation of the algorithm described in (Cavnar and Trenkle, 1994). Each sentence is annotated with the language represented, and the application in GATE then calls one of two further applications, for English and German respectively, for each sentence being processed. If other languages are detected, then the sentence is ignored by the application and is not further analysed.

Language identification in tweets is a particular problem, due to their short length (140 characters maximum) and the ubiquity of language-independent tokens (RT (retweet), hashtags, @mentions, numbers, URLs, emoticons). Often, once these are removed, a tweet would contain fewer than 4 or 5 words, some would even have no "proper" words left. For English and German, we are currently achieving best results with the multinominal Naive Bayes language identifier by (Lui and Baldwin, 2011).

3.5. Adapting the tools for German

The approach we follow for processing German is very similar to that for English, but makes use of some different (though equivalent) processing resources in GATE. We have adapted the English named entity and term recognition tools specifically for German, using different POS taggers and grammars, for example. We also use the SentiWS dictionary (Remus et al., 2010) as the basis for our sentiment gazetteer. Currently, we do not perform event recognition in German (though this will be developed at a later stage), so opinions relate only to entities or to entire sentences and tweets.

4. Challenges imposed by social media

In addition to the factors already discussed, social media imposes a number of further challenges on an opinion mining system.

4.1. Relevance

Even when a crawler is restricted to specific topics and correctly identifies relevant pages, this does not mean that every comment on such pages will also be relevant. This is a particular problem for social media, where discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand. For example, in the Rock am Ring forum, we also found comments relating to a televsion program that had been shown directly after the Rock am Ring event. Similarly on Twitter, the topics in which a user is interested can be very diverse, so it makes little sense to characterise "interesting" tweets for all users with a single lexical model.

[7] http://www.let.rug.nl/vannoord/TextCat/

There are a number of ways in which we can attempt to deal with the relevance issue. First, we could try to train a classifier for tweets or comments which are relevant, e.g. we might want to disregard tweets if they contain certain terms. Second, we can make use of clustering in order to find opinionated sentences or segments related to certain topics, and disregard those which fall outside these topics. This is probably the most promising approach, especially since we already make use of topic clustering algorithms within the wider project, although it does risk that some relevant comments might be left out.

4.2. Target identification

One problem faced by many search-based approaches to sentiment analysis is that the topic of the retrieved document is not necessarily the object of the sentiment held therein. This is particularly true of the online sentiment analysers discussed in Section 2, which make no connection between the search keyword and the opinion mentioned in the tweet, so that in fact while the polarity of the opinion may be correct, the topic or target of the opinion may be something totally different. For example, the day after Whitney Houston's death, TwitterSentiment and similar sites all showed an overwhelming majority of tweets about Whitney Houston to be negative; however, almost all these tweets were negative only in that people were sad about her death, and not because they disliked her. So the tweets were displaying dislike of the situation, but not dislike of the person. One way in which we deal with this problem is by using an entity-centric approach, whereby we first identify the relevant entity and then look for opinions semantically related to this entity, rather than just trying to decide what the sentiment is without reference to a target, as many machine learning approaches take. We use linguistic relations in order to make associations between target and opinion (for example, a target may be linked to a verb expressing like or dislike as its direct object, as in "I like cheese", or the opinion may be expressed as an adjective modifying the target "the shocking death of Whitney"). There are a number of ways in which sentences containing sentiment but which have no obvious target-opinion link can be annotated. Currently, we simply identify the sentence as "sentiment-containing" but make no assumption about the target. Future work will investigate further techniques for assigning a topic in such cases.

4.3. Negation

The simpler bag-of-words sentiment classifiers have the weakness that they do not handle negation well; the difference between the phrases "not good" and "good" is somewhat ignored in a unigram model, though they carry completely different meanings. A possible solution is to incorporate longer range features such as higher order n-grams or dependency structures, which would help capture more complete, subtle patterns, such as in the sentence "Surprisingly, the build quality is well above par, considering the rest of the features." in which the term "surprisingly" should partially negate the positive overall sentiment (Pang and Lee, 2008). Another way to deal with negation, avoiding the need for dependency parsing, is to capture simple

patterns such as “isn’t helpful” or “not exciting” by inserting unigrams like “NOT-helpful” and “NOT-exciting” respectively (Das and Chen, 2001). This work-around was implemented for tweets by Pak and Paroubek (Pak and Paroubek, 2010a).
For a rule-based system such as ours, we believe that the approach adopted, similar to that of (Taboada et al., 2011), is sufficient to capture most aspects of negation: indeed, Taboada’s evaluation appears to support this.

4.4. Contextual information

Social media, and in particular tweets, typically assume a much higher level of contextual and world knowledge by the reader than more formal texts. This information can be very difficult to acquire automatically. For example, one tweet in the political dataset used in (Maynard and Funk, 2011) likened a politician to Voldemort, a fictional character from the Harry Potter series of books. While the character is sufficiently well known to have its own Wikipedia entry, assimilating the necessary information (that Voldemort is considered evil) is a step beyond current capabilities, and we may have to just accept that this kind of comment cannot be readily understood by auomatic means.
One advantage of tweets, in particular, is that they have a vast amount of metadata associated with them which can be useful, not just for opinion summarisation and aggregation over a large number of tweets, but also for disambiguation and for training purposes. Examples of this metadata include the date and time, the number of followers of the person tweeting, the person’s location and even their profile. For example, we may have information about that person’s political affiliation mentioned in their profile, which we can use to help decide if their tweet is sarcastic when they appear to be positive about a particular political figure. Because each person registered on Twitter has a unique ID, we can disambiguate between different people with the same name – something which can be problematic in other kinds of text.

4.5. Volatility over Time

Social media, especially Twitter, exhibits a very strong temporal dynamic. More specifically, opinions can change radically over time, from positive to negative and vice versa. Within another project, TrendMiner[8], we are studying two highly dynamic opinion- and trend-driven domains: investment decisions and tracking opinions on political issues and politicians over time, in multiple EU states and languages. Since there is also correlation between the two domains, joint models of political opinions and financial market opinions also need to be explored.
To address this problem, the different types of possible opinions are associated as ontological properties with the classes describing entities, facts and events, discovered through information extraction techniques similar to those described in this paper, and semantic annotation techniques similar to those in (Maynard and Greenwood, 2012) which aimed at managing the evolution of entities over time. The extracted opinions and sentiments are time-stamped and stored in a knowledge base, which is enriched continuously, as new content and opinions come in. A particularly challenging question is how to detect emerging new opinions, rather than adding the new information to an existing opinion for the given entity. Contradictions and changes also need to be captured and used to track trends over time, in particular through opinion merging, which we turn to next.

[8] http://www.trendminer-project.eu

4.6. Opinion Aggregation and Summarisation

Another novel aspect to our work concerns the type of aggregation that can be applied to opinions to be extracted from various sources and co-referred. In classical information extraction, this can be applied to the extracted information in a straightforward way: data can be merged if there are no inconsistencies, e.g. on the properties of an entity. Opinions behave differently here, however: multiple opinions can be attached to an entity and need to be modelled separately, for which we advocate populating a knowledge base. An important question is whether one should just store the mean of opinions detected within a specific interval of time (as current opinion visualisation methods do), or if more detailed approaches are preferable, such as modelling the sources and strength of conflicting opinions and how they change over time. Effectively, we advocate here a form of opinion-based summarisation, e.g. displaying positive/negative opinion timelines, coupled with opinion holders and key features.
A second important question in this context involves finding clusterings of the opinions expressed in social media, according to influential groups, demographics and geographical and social cliques. Consequently, the social, graph-based nature of the interactions requires new methods for opinion aggregation.

5. Evaluation

Evaluation of opinion mining can be tricky, for a number of reasons. First, opinions are often subjective, and it is not always clear what was intended by the author. For example, we cannot necessarily tell if a comment such as "I love Baroness Warsi", in the absence of further context, expresses a genuine positive sentiment or is being used sarcastically. Inter-annotator agreement performed on manually annotated data therefore tends to be low, which affects the reliability of any gold standard data produced. While Amazon Mechanical Turk has been used for producing such gold standard annotated corpora, similar problems apply with respect to inter-annotator agreement, even if multiple annotations are produced for each document. Second, it is very hard to evaluate polarity scores such as the ones we produce: for example, we cannot really say how correct the score of 0.6012 awarded to a comment in the Rock am Ring forum about the band "In Flames" being the person’s favourite band is, or whether a score of 0.463 would be better. However, while these scores technically represent strength of opinion, we can view them instead as an indicator of confidence. So we would therefore expect the sentiments expressed with high polarity scores to have higher accuracy, and can tailor our evaluation accordingly, looking for higher accuracy rates as the polarity score increases.
As mentioned in Section 4, much of the success of an entity-centric opinion mining tool depends on the quality

of the entities and events extracted. Because we adopt a high precision strategy, at the potential expense of recall, we aim to minimise this effect. Because we risk missing some opinions, we also have a backoff strategy of identifying opinionated sentences which do not specifically map to an extracted entity or event. These give us some extra opinions, but risk being irrelevant or outside the scope of our interest.

We have not yet formally evaluated the opinion mining tools, other than for the political tweets dataset, whose results are reported in (Maynard and Funk, 2011). However, initial results look promising. We manually annotated a small corpus of 20 facebook posts (in English) about the Greek financial crisis (automatically selected according to certain criteria by our crawler) with sentiment-containing sentences, and compared these with our system generated sentiment annotations. Our system correctly identified sentiment-containing sentences with 86% Precision and 71% Recall, and of these correctly identified sentences, the accuracy of the polarity (positive or negative) was 66%. While the accuracy score is not that high, we are satisfied at this stage because some of the components are not fully complete – for example, the negation and sarcasm components still require more work. Also, this accuracy score takes into account both incorrect and correct sentiment-bearing sentences, since the two tasks are not performed independently (i.e. we are not assuming perfect sentiment sentence recognition before we classify the polarity of them). On the other hand, the named entity recognition is very accurate on these texts - our evaluation showed 92% Precision and 69% Recall. Since we aim for high Precision at the potential expense of Recall, and since we have further plans for improving the recall, this is most promising. Clearly, further and more detailed evaluation is still necessary.

6. Prospects and future work

While the development of the opinion mining tools described here is very much work in progress, initial results are promising and we are confident that the backoff strategies inherent in the incremental methodology will enable a successful system. We advocate the use of quite shallow techniques for much of the linguistic processing, using chunking rather than full parsing, for instance. While we could incorporate the Stanford parser to give us relational information, previous experience shows that the performance of such tools is dramatically reduced when used with degraded texts such as tweets. Furthermore, our methodology enables the system to be easily tailored to new tasks, domains and languages. On the other hand, the linguistic sub-components can also be used as initial pre-processing to provide features for machine learning, where such data is available, and we are currently experimenting with such techniques.

In previous work we have obtained good results using SVM-based machine learning (ML) from linguistic features for opinion classification (Funk et al., 2008; Saggion and Funk, 2009). We plan to experiment with similar data-driven techniques on tweets, although we would probably use the Perceptron algorithm instead, since it is faster and (in our experience) about as accurate for NLP. Our previous experiments were carried out on longer, somewhat more consistently edited texts (film, product and business reviews), which were quite unlike the highly abbreviated and inconsistent styles found in tweets. However, we obtained good results with unigrams of simple linguistic features, such as tokens and their lemmas, as well as with features derived from SentiWordNet values. With the additional features we already identify using our rule-based techniques, such as negative and conditional detection, use of swear words and sarcasm, we would expect to have some reasonable results. To carry out such experiments successfully on tweets, however, we would need a larger manually annotated corpus than the one previously used

As discussed earlier, there are many improvements which can be made to the opinion mining application in terms of using further linguistic and contextual clues: the development of the application described here is a first stage towards a more complete system, and also contextualises the work within a wider framework of social media monitoring which can lead to interesting new perspectives when combined with relevant research in related areas such as trust, archiving and digital libraries.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1) and the European Union under grant agreements No. 270239 (Arcomem[9]) and No. 287863 (TrendMiner[10]).

7. References

A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proc. of the International Conference on Recent Advances in Natural Language Processing*, Borovetz, Bulgaria.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, page 440.

E. Boiy and M-F. Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12(5):526–558.

E. Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. 2007. Automatic sentiment analysis of on-line text. In *Proc. of the 11th International Conference on Electronic Publishing*, Vienna, Austria.

Johan Bollen and Huina Mao. 2011. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10):91–94.

W. Bosma and P. Vossen. 2010. Bootstrapping language-neutral term extraction. In *7th Language Resources and Evaluation Conference (LREC)*, Valletta, Malta.

C. Buckley and G. Salton. 2009. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

W.B. Cavnar and J.M. Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113:4001.

[9] http://www.arcomem.eu

[10] http://www.trendminer-project.eu/

H. Cunningham, D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pages 115–122.

A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold. 2008. Opinion analysis for business intelligence applications. In *First Int. Workshop on Ontology-Supported Business Intelligence*, Karlsruhe, October. ACM.

S. Gindl, A. Weichselbraun, and A. Scharl. 2010. Cross-domain contextualisation of sentiment lexicons. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 771–776.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2010. Recognizing Named Entities in Tweets. *Science And Technology*, 2008.

M. Lui and T. Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, November.

D. Maynard and A. Funk. 2011. Automatic detection of political opinions in tweets. In Dieter Fensel Raúl García-Castro and Grigoris Antoniou, editors, *The Semantic Web: ESWC 2011 Selected Workshop Papers, Lecture Notes in Computer Science*. Springer.

D. Maynard and M. A. Greenwood. 2012. Large Scale Semantic Annotation, Indexing and Search at The National Archives. In *Proceedings of LREC 2012*, Turkey.

G. A. Miller, R. Beckwith, C. Felbaum, D. Gross, G. A. Miller, C.Miller, R. Beckwith, C. Felbaum, D. Gross, and M. Miller, C.Minsky. 1980. Five papers on WordNetk-lines: A theory of memory.

S. Moghaddam and F. Popowich. 2010. Opinion polarity identification through adjectives. *CoRR*, abs/1011.4623.

A.C. Mullaly, C.L. Gagné, T.L. Spalding, and K.A. Marchak. 2010. Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. *The Mental Lexicon*, 5(1):87–114.

A. Pak and P. Paroubek. 2010a. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC 2010*.

A. Pak and P. Paroubek. 2010b. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1).

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 339–346, Vancouver, Canada.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Remus, U. Quasthoff, and G. Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

T. Risse, S. Dietze, D. Maynard, and N. Tahmasebi. 2011. Using Events for Content Appraisal and Selection in Web Archives. In *Proceedings of DeRiVE 2011: Workshop in conjunction with the 10th International Semantic Web Conference 2011*, Bonn, Germany, October.

A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.

H. Saggion and A. Funk. 2009. Extracting opinions and facts for business intelligence. *RNTI Journal*, E(17):119–146, November.

A. Scharl and A. Weichselbraun. 2008. An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics*, 5(1):121–132.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41.

A. Weichselbraun, S. Gindl, and A. Scharl. 2010. A context-dependent supervised learning approach to sentiment detection in large textual databases. *Journal of Information and Data Management*, 1(3):329–342.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

A Qualitative Analysis of Informality Levels In Web 2.0 Texts: The Facebook Case Study

Alejandro Mosquera, Paloma Moreda

University Of Alicante
DLSI. Ap.de Correos 99. E-03080 Alicante, Spain
amosquera@dlsi.ua.es, moreda@dlsi.ua.es

Abstract

The study of the language used in Web 2.0 applications such as social networks, blogging platforms or on-line chats is a very interesting topic and can be used to test linguistic or social theories. However the existence of language deviations such as typos, emoticons, abuse of acronyms and domain-specific slang makes any linguistic analysis challenging. The characterization of this informal writing can be used to test the performance of Natural Language Processing tools when analysing Web 2.0 texts, where informality can play an important role. By being one of the most popular social media websites, Facebook handles an increasing volume of text, video and image data within its user profiles. In this paper, we aim to perform a qualitative analysis of informality levels in textual information publicly available on Facebook. In particular, this study focus on developing informality dimensions, a set of meaningful and comparable variables, discovered by mapping textual features by affinity and using unsupervised machine learning techniques. In addition, we explore the relation of informality and Facebook metadata such as received likes, gender, time range and publication type.

Keywords: Web 2.0, Informality, Facebook

1. Introduction

Over the last few years Web 2.0 sites have become the most popular Internet services. Social networks (Facebook), video sharing tools (Youtube), blogging platforms (Blogger), collaborative encyclopaedias (Wikipedia) and microblogging applications (Twitter) are in the top ten most visited websites on the Internet [1] nowadays. These technological tools focus on user-generated content (UGC), where users provide, share and use information. This paradigm shift have been able to change the way information is generated and consumed.

The characteristics present in these publications can be considerably different in comparison with traditional texts such as abbreviations, emoticons or non-standard spellings (Crystal, 2001). Moreover, Internet users have popularised special ingroup dialects, such as Internet slang,
textitleetspeak or acronyms. Besides all these features, the informal language usually present in web genres takes one step further in UGC (Mehler et al., 2009), which is usually written in a more informal context.

Many questions arise in this situation: how state-of-the-art Natural Language Processing (NLP) applications can handle the informal nature of UGC and, in case of deficiencies are present, how we can solve this inaccuracies, adapting NLP tools or normalising non-standard language features.

Studies following the normalisation strategy are usually involved in a translation approach, performing a conversion of language deviations to their normalised form (Gouws et al., 2011). However, not all Web 2.0 texts present the same level of informality (Mosquera and Moreda, 2011a) and therefore the optimal solution could depend on that level. So in both cases, an informality analysis is the first step in order to develop new language technologies for UGC. Then we can obtain enough information to decide the most appropriate strategy for each informality level and perform additional actions only when necessary.

For this reason, we are going to perform a qualitative analysis of informality levels in Web 2.0 texts. In particular, this study will be focused on developing informality dimensions, a set of meaningful and comparable variables. These informality dimensions are discovered by mapping textual features by affinity and using unsupervised machine learning techniques

In order to do this, we analyse texts extracted from the Facebook social network. We chose Facebook by being one of the most relevant social media websites by number of users and volume of information[2], however our methodology can be applied to any Web 2.0 application. Facebook hosts not only texts, images and videos but also interesting metadata such as publication type, date, number of likes or gender. The relation between the obtained informality levels and this additional information will be also analysed.

This article is organised as follows: In Section 2 we review the state of the art. Section 3 describes our methodology. In Section 4, the obtained results are analysed. Finally, our main conclusions and future work are drawn in Section 5.

2. Related Work

Within our area of interest, the most works to date focus on the analysis of text formality rather than informality. A distinction can be made between three basic approaches taking into account their smallest analysis unit: analysis at document-level, sentence-level and word-level.

Regarding formality at document-level, one of the first studies was performed with the F-Measure (Heylighen and Dewaele, 1999), a score based on Part-of-Speech (POS) tags using the concepts of deixis and lexical density (Ure, 1971), whereby the frequency of deictic words is expected

[1] http://www.alexa.com/topsites

[2] http://www.facebook.com/press/info.php?statistics

to increase with the informality of a text and, conversely, the frequency of non-deictic words should increase with text formality.

Another formality document-level classification methods have been used to exploit the concept of social distance and politeness, detecting two formality levels (formal/informal) in email documents (Peterson et al., 2011)

Regarding sentence-level formality approaches, it has been shown that readability metrics correlate with formality (Lahiri et al., 2011). This relation was used to explore the formality of Internet news sites with the F-Measure and readability indexes obtaining a formal/informal binary classification. While the F-Measure score can be used to detect deep formality this approach have issues quantifying stylistic or grammatical deviations. For this reason, new efforts were performed to obtain a new formality measure: a five-point Likert scale has been experimented to explore the inter-rater agreement for assessing sentence formality (Lahiri and Lu, 2011).

Finally, word-level approaches can make use of corpora and formality lexicons to quantify word-level formality (Brooke et al., 2010).

Taking into account document-level informality, the I-Measure (Mosquera and Moreda, 2011a) has been used to classify Web 2.0 texts into three informality levels (very informal, moderately informal and slightly informal) using unsupervised machine learning and Principal Component Analysis (PCA) (Jolliffe, 2002) techniques. The value of this obtained variable (I-Measure) was used to measure and compare the informality level of each cluster.

The use of quantitative formality scores such as the F-Measure provides a direct metric, capable to differentiate between genres (Teddiman, 2009) as each genre will have its own formality spectrum. Nevertheless, there are significant differences between Internet and traditional genres due the informal nature of UGC (Santini, 2006). On the one hand, analysing informality instead formality usually produces a more accurate text classification (Mosquera and Moreda, 2011b). On the other hand, existing informality quantitative metrics can be very genre-specific, thus being difficult to compare results among different text types.

We hypothesize that a qualitative classification based on text informality can help to augment and improve quantitative approaches for UGC such as the I-Measure. For this reason, in this paper we propose a new qualitative analysis with more emphasis on discriminating and understanding the nature of the different informality types than obtaining a classification based on a single measure. Therefore, this approach proposes the use of informality dimensions, a genre-independent layer that allows the direct comparison of different informality models.

3. Methodology

In this study, we are going to analyse informality levels in publicly available English texts from Facebook publications in an unsupervised manner, using machine learning and cross-validation techniques. In addition, the relation between several Facebook metadata and the obtained informality levels will be explored. This section describes in two steps the analysis process. First, section 3.1 describes the used corpora. Secondly, in section 3.2 the selected text features used in the classification algorithm are introduced and justified. In section 3.3, we explain our unsupervised classification step based on clustering techniques used to discover informality levels. Finally, the identified informality dimensions are shown in section 3.4.

3.1. Corpus

We crawled and processed 9887 random English texts from public post (wall updates, photo comments, video comments and link comments) on the Facebook social network using the Graph API. In addition, several linked metadata such as gender, publication type, time range and received likes was extracted. The resulting dataset of about 350.000 words was anonymised by removing user names to avoid data privacy issues.

3.2. Text Characteristics

The use of POS, word-length or sentence-length features have been used extensively in the literature to characterize formal or informal writing. This being more adequate than an exclusive n-gram analysis (Evans et al., 2004) (Thayer et al., 2010). Taking this into account, from an initial set of 59 features with forward, backward and greedy forward correlation feature selection (CFS) (Hall, 1998) algorithms a reduced subset of 11 features was selected:

(F1) RIX: This index measures text readability (Anderson, 1983). It is based on two factors both related to text formality, the length of words and the sentence length:

$$RIX = LW/S$$

Where *LW* the number of words with more than 7 letters and *S* is the number of sentences.

(F2) Entropy: Making use of information theory, texts with high entropy would imply a higher amount of expressed information and more quality, otherwise texts with low entropy would communicate less information and can be considered of lower quality. This feature calculates de Shannon entropy (Shannon, 1951) of each text.

(F3) Emotional distance: From a marketing and sociologically point of view, understanding user emotions would allow more effective ad-targeting and knowledge about actual trends. Otherwise, the use of emotional words implies closeness to the reader, that is a characteristic of informal writing. The field of sentiment and emotion analysis in Web 2.0 informal texts is not a novel topic, helping to determine authority and truthfulness (Malouf and Mullen, 2007).

Using a similar approach than in (Park et al., 2011), we measure the emotion with its corresponding strength calculating the path distance with WordNet (Fellbaum, 1998) and an emotion lexicon based on 6 primary emotions: *Love, Joy, Surprise, Anger, Sadness and Fear* (Parrott, 2001). This method is capable of detecting terms which are associated with emotions even

when they are not present in WordNet, those are expanded using Roget's Thesaurus (American Psychological Association, 2011) definitions.

(F4) Wrong-written sentences: We use 3 heuristic rules to determine ill-formed sentences (Lloret, 2011): Each sentence must contain at least three words, each sentence must contain at least one verb and the sentences can not end in articles, prepositions or conjunctions.

(F5) Wrong-typed words: Simple heuristics were used to detect common wrong-typed words taking into account their case, punctuation symbols and position in the sentence (*YoU, How u DARE!*)

(F6) Frequency of contractions: The use of contractions or word shortening is a common feature of spoken English. However in written English they are usually present in a more informal context. (*can't, It's...*) .

(F7) Frequency of repetitions: The repetition of vocal or consonants within a word to add expressive power is used in informal speech and writing (*sooooo, yesssssss*).

(F8) Frequency of emoticons: The presence of facial expression represented by punctuation and letters is frequently used to express moods and sentiments in Web 2.0 applications.

(F9) Frequency of interjections: We use TreeTagger (Schmid, 1994) to extract this part of speech capable to express an emotion or sentiment. Although they can not be interpreted out of context (For example: *Ah!* can express pleasure, surprise or even resignation) its presence is common in informal writing styles.

(F10),(F11) Frequency of slang and offensive words: They are obtained by querying online dictionaries like Wiktionary [3], Online Slang Dictionary [4] and Advanced Learner Cambridge Online Dictionary [5] looking for special tags like into word definitions.

3.3. Classification Algorithm

Unsupervised learning have the advantage of not depending on manually-annotated corpora. For this reason, we clustered our reduced multi-dimensional set with the Expectation-Maximization (EM) (Dempster et al., 1977) unsupervised machine learning algorithm. EM finds maximum likelihood estimates of parameters in probabilistic models, without the need to use distance measures like K-Means (Hartigan and Wong, 1979). Instead it computes probabilities of cluster memberships based on one or more probability distributions. Estimating the means and standard deviations for each cluster to maximize the likelihood of the observed data. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. A ten-fold cross-validation algorithm was used for discovering the most optimal number of clusters without requiring a separate dataset.

[3] http://en.wiktionary.org

[4] http://onlineslangdictionary.com

[5] http://dictionary.cambridge.org

3.4. Informality Dimensions

The use of dimensionality reduction techniques such as PCA, can improve the analysis and classification of texts by their informality level providing a meaningful score. Nevertheless, the comparison between scores obtained from different datasets is not always possible, because they depend on corpora genres and text types. For this reason, in this study we chose to perform a qualitative analysis by clustering our dataset in four partitions (C1, C2, C3 and C4), each one mapped to an informality level. Then, instead of developing a hierarchical classification where the top level is more informal than its immediate neighbour, the classification was enhanced by grouping text features into informality dimensions using a feature affinity criterion. We identified four informality dimensions in each discovered informality level with normalised values in a 0-100 range (see Table 1). Using this flexible approach the cluster hierarchy can vary depending on the considered dimensions. This method not only shows information about what texts are more informal but it also allows the comparison of texts from other corpora, genres or with different number of informality levels:

(Dimension 1), Complexity: This dimension measures text complexity. Covering F1 and F2 features, informality can be correlated with both readability and word/sentence length.

(Dimension 2), Emotiveness: As we stated before, the direct or indirect expression of emotions lowers the distance with the reader and consequently the formality. This dimension aggregates F3, F8 and F9 features for measuring emotion strength.

(Dimension 3), Expressivenes: This dimension measures text expressiveness grouping F7, F10 and F11 features. Expressive texts would contain specific domain words like slang or offensive words and repetitions.

(Dimension 4), Incorrectness: Word and sentence typos or misspellings are reflected in this dimension. Covering F4, F5 and F6 features, this dimension measures language deviations that are directly correlated with text informality.

4. Analysis and Results

In this study we obtained 4 clusters using unsupervised machine learning techniques. Additionally, four informality dimensions were identified and the relation between informality and Facebook metadata was analysed. In section 4.1 we introduce our informality analysis approach. The analysis of Facebook metadata is described in section 4.2.

4.1. Informality Analysis

In our analysed dataset, the results pointed that texts corresponding to the first informality level (C1), scored the maximum value (100) in the 2 and 3 dimensions and they have a high complexity (79) and moderate incorrectness (56). On the other side, texts assigned to the C2 level are characterised by wrong-written small words and sentences, reflected in the fourth dimension (76). Finally, levels C3

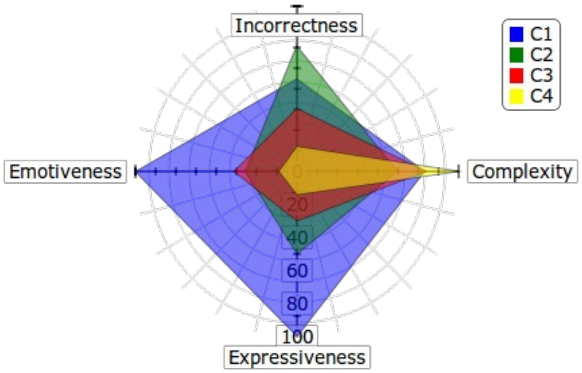

Figure 1: Clustering results of Facebook publications by dimension.

and C4 show a continuous decrease of Incorrectess, Emotiveness and Expressiveness dimensions with an increase of Complexity (see Figure 1 and Table 2).
With the informality continuum explained in a multidimensional level we can model subjective and problematic situations. Using this model we can perform the comparison of two texts in the same informality level but contained within different informality dimensions, something that is not possible with a single measure.

Dimension	Features
Complexity	RIX, Entropy.
Expressiveness	Freq. Slang, Offensive words, Repetitions.
Emotiveness	Freq. Interjections, Emoticons, Emotional distance.
Incorrectness	Freq. Wrong typed words, Contractions, Wrong written sentences.

Table 1: Features and discovered dimensions.

Cluster	C1	C2	C3	C4
N^o Instances	659	1238	2243	5747
Complexity	79	62	80	100
Expressiveness	100	50	30	14
Emotiveness	100	32	39	11
Incorrectness	56	76	38	15

Table 2: Normalised dimensions by informality level.

4.2. Metadata

We explored the relation between interesting Facebook metadata and the obtained informality levels.

4.2.1. Gender

The gender of the user who authored the text can be an interesting feature to explore social variables. The F-measure

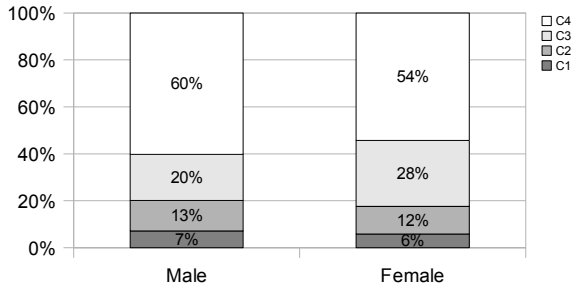

Figure 2: Gender distribution by cluster.

was applied to texts of known gender obtaining differences between the two sexes. The results showed that females scored lower, preferring a more contextual style, while men prefer a more formal style (Heylighen and Dewaele, 1999). Biber (1988) also found strong differences between male and female authors along their multidimensional analysis of English language, where female authors tend to be more ”involved” and male authors to be more ”informational”. In our study of Facebook texts we have discovered slight differences between males and females in the C3 and C4 levels (see Figure 2), were females tend to write less complex and more informal texts than males in the lower informality levels.

4.2.2. Time range

The time stamp of each post was extracted from Facebook metadata, developing six 4-Hours groups. We can appreciate that the amount of posts corresponding to the most informal level (C1), remain almost constant along all the day, otherwise we discovered an increase of the number publications of the most formal level (C4) between 16h-24h (see Figure 3). Without an age/country analysis we cannot extract direct conclusions, like informal writing at work/college hours or by day of the week, being this an interesting topic to explore further.

4.2.3. Publication type

Currently there are four different main publications types in the Facebook social network: wall posts, videos, images and links. We can appreciate a significant increase of informality amount in photo comments, being highly emotive and expressive texts in comparison with link and status comments, that range from slightly informal to neutral (see Figure 4). On the other hand, the language used in video comments can be considered mostly informal but with a lesser amount of typos, slang and offensive words.

4.2.4. Likes

The like button lets an user add a like to any post or comment, but one user can like each post or comment only once. We hypothesize that emotive and expressive comments tend to receive more feedback that normal ones (see Table 3). Observing the average received likes per text for each cluster, we can notice that the C1 level has the higher like-ratio. However the values scored by the less informal level (C4) suggests that informality and Facebook likes are not directly correlated.

Cluster	C1	C2	C3	C4
Avg. Likes	0,53	0,18	0,12	0,34

Table 3: Average likes per text for each cluster.

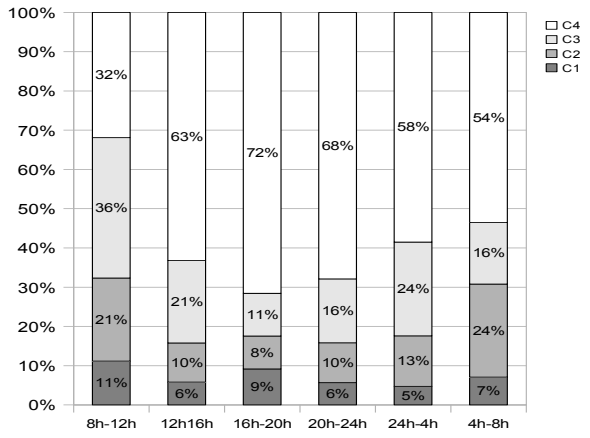

Figure 3: Time range distribution by cluster.

5. Conclusions and Future Work

In this study we performed a qualitative informality analysis using Facebook texts, identifying four informality dimensions in each informality level. We prove that this is a more complete an detailed classification than our previous works, improving quantitative-only analysis. The proposed analysis framework address the gaps identified in the baseline, such as the difficulty of performing a comparison of texts between informality levels or even between the same informality level, something impossible with a quantitative approach.

The analysis of Facebook metadata proved interesting, in particular the gender and post-type variables, but the relation of likes and time range with the informality spectrum was not conclusive.

We were not able to collect enough profiles with public birth date and location due to new changes in Facebook privacy settings, this being other interesting variables to contempt in a future work. Another domains that can benefit from this study would be reliability and credibility, regarding of their relation with the different informality levels. Other future work would explore our initial proposal of adapting NLP tools, analysing if our secondary hypothesis about the need to rewrite only in the more informal levels is correct. Finally, we plan to apply our analysis to another Web 2.0 genres and applications.

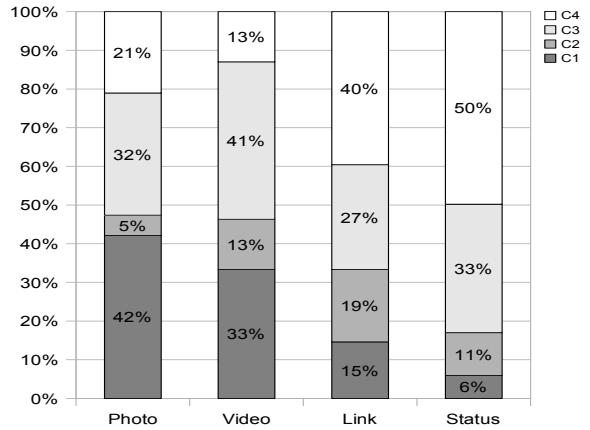

Figure 4: Post type distribution by cluster.

Acknowledgements

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001)

6. References

APA American Psychological Association. 2011. Roget's 21st century thesaurus, third edition. Jul.

Jonathan Anderson. 1983. Lix and rix: variations on a little-known readability index. *Journal of Reading*, 26(6):490–497.

Douglas Biber. 1988. *Linguistic features: algorithms and functions in Variation across speech and writing*. Cambridge University Press.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 90–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Crystal. 2001. *Language and the Internet*. Cambridge Univ. Press.

Arthur P. Dempster, M. Nan Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22.

Mary B. Evans, Alice A. McBride, Matt Queen, Alexander Thayer, and Jan H. Spyridakis. 2004. The effect of style of typography on perceptions of document tone. *Proceedings of IEEE International Professional Communication Conference*, pages 300–303.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual Bearing on Linguistic Variation in Social Media. In *ACL '11*.

Mark Hall. 1998. Correlation-based feature selection for machine learning. *PhD dissertation Hamilton NZ Waikato University Department of Computer Science*.

John A. Hartigan and M. Anthony Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels.

I. T. Jolliffe. 2002. *Principal Component Analysis*. Springer, second edition, October.

Shibamouli Lahiri and Xiaofei Lu. 2011. Inter-rater agreement on sentence formality.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 446–457. Springer-Verlag.

Elena Lloret. 2011. Text summarisation based on human language technologies and its applications. *Ph.D. dissertation. University of Alicante*.

Rob. Malouf and Tony Mullen. 2007. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web*.

Alexander Mehler, Serge Sharoff, Georg Rehm, and Marina Santini, editors. 2009. *Genres on the web: Computational Models and Empirical Studies*. Springer.

Alejandro Mosquera and Paloma Moreda. 2011a. Enhancing the discovery of informality levels in web 2.0 texts. *Proceedings of the 5th Language Technology Conference (LTC'11)*.

Alejandro Mosquera and Paloma Moreda. 2011b. The use of metrics for measuring informality levels in web 2.0 texts. *Proceedings of 8th Brazilian Symposium in Information and Human Language Technology (STIL)*.

Seung-Bo Park, Eunsoon Yoo, Hyunsik Kim, and GeunSik Jo. 2011. Automatic emotion annotation of movie dialogue using wordnet. In *ACIIDS (2)*, pages 130–139.

W. Gerrot Parrott. 2001. Emotions in social psychology.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Marina Santini. 2006. Web pages, text types, and linguistic features: Some issues. *ICAME Journal*, 30.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Claude E. Shannon. 1951. Prediction and entropy of printed english. (30):50–64.

Laura Teddiman. 2009. Contextuality and beyond: Investigating an online diary corpus. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA*.

Alexander Thayer, Mary B. Evans, Alicia A. McBride, Matt Queen, and Jan H. Spyridakis. 2010. I, pronoun: A study of formality in online content. *Journal of Technical Writing and Communication*, 40:447–458.

Jean Ure. 1971. Lexical density and register differentiation. in g. perren and j.l.m. trim (eds), applications of linguistics. pages 443–452.

What is the text of a Tweet?

Jordi Atserias, Joan Codina

Fundació Barcelona Media, Roc Boronat 138, 08018 Barcelona
{jordi.atserias,joan.codina}@barcelonamedia.org

Abstract

Twitter is a popular micro blogging/social medium for broadcasting news, staying in touch with friends and sharing opinions using up to 140 characters per message. In general, User generated Content (e.g. Blogs, Tweets) differs from the kind of text the traditional Natural Language Processing tools has been developed and trained. The use of non-standard language, emoticons, spelling errors, letter casing, unusual punctuation makes applying NLP to user generated content still an open issue. This work will focus on the effect of the Twitter metalanguage elements in the text processing, specifically for PoS tagging. Several different strategies to deal with twitter specific metalanguage elements are presented and evaluated. The results shows that it is necessary to remove metalanguage elements. However some text normalisation or PoS tagger adaptation is needed in order to have a clear evaluation about which of the different methods to treat twitter metalanguage elements is better.

Keywords: twitter, PoS tagging, UGC, text normalization

1. Introduction

Twitter is a popular micro blogging/social medium for broadcasting news, staying in touch with friends and sharing opinions using up to 140 characters per message. In general, User generated Content (e.g. Blogs, Tweets) differs from the kind of text the traditional Natural Language Processing tools has been developed and trained. The use of non-standard language, emoticons, spelling errors, letter casing, unusual punctuation, etc makes applying NLP to user generated content still an open issue (Kobus et al., 2008), (Simard and Deslauriers, 2001), CAW2.0 workshop[1]. In addition, text normalisation in Spanish has to address diacritic marks (accents) (Yarowsky, 1994) (Atserias et al., 2012) since few users place when writing twits.

In Twitter, all these differences of the user generated content are magnified by the message length restriction and the use of several particular conventions of the twitter framework (user references, hashtags, etc). Although previous works have proposed some methodology, e.g. (Kaufmann and Kalita, 2010), as far as we know no evaluation has been carried out to measure the impacts of these heuristics on the text processing and no substantial work on this subject has been conducted in Spanish.

There are many different efforts on improving, adapting the POS set, several text normalisation strategies, that try to address many of this issues. However, in this work will focus on the effect of the Twitter metalanguage elements in the text processing. With Twitter metalanguage elements we are referring to the special set of words-tags than have an special meaning on twitter.

The most important "metalanguage" elements on Twitter are:

- **Hashtags:** Allow to explicitly mark the topic of a tweet. Hashtags starts with character '#" and can be common words or concatenation of several words (using capitals for word boundaries) e.g. CamelCase. A tweet can contain any number of hashtags and these hashtags can be placed at any position in the text.

- **User References:** Users are identified by their names prefixed with the character "@" Twitter users can make references to other users in different ways:

  - As a reference to a person inside a sentence:
    muy fan del "Shakiro", ojalá se cruzara con **@3gerardpique***, le cantará y se lo acabara zumbando*
  - When Tweeting to an specific user, usually at the beginning of the tweet:
    @Buenafuente *El programa sin Joan Eloy no sería lo mismo*
  - When resending a tweet wrote by another user. In this case the original author of the tweet is kept and receives a notification of the retweet:
    RT **@SSantiagosegura***: Y esos seguidores ?!!!*

- **Vía:** Is used to reference the source of the information usually also adding a link
  El Arte de Presentar http://bit.ly/hKjSdd / **vía** *@loogic*

- **URLs:** Due to the length limitation of the tweet the original links are usually replaced by a shorten version.

- **Truncated Tweets:** The space limitation of the twitter messages trunks some of the tweets when they are sent through a 3rd party application (e.g. vía) or retweeted. In this case the text is trunkated, adding "...". Notice that this phenomena not only trunkates sentences but also the last word.
  El TUE prohíbe discriminar en los seguros: El Tribunal de Justicia de la Unión Europea (TUE) acaba de dictar u... http://bit.ly/fnPGsr

Although the techniques presented are language independent we will focus on the processing of Spanish Tweets and specifically on PoS tagging which is a basic previous step to more complex NLP tasks, such Named Entity Recognition or parsing.

[1] http://caw2.barcelonamedia.org/

2. Freeling Text Processing

FreeLing (Padró et al., 2010) includes different text processing tools (Tokenization, Sentence Splitting and PoS tagging, etc). Freeling models (e.g. PoS tagger) are trained on general well written text, which differs from the type of language used in twitter. In the experiments spellchecking, date and quantities Freeling modules were disabled while the tokenizer was modified to correctly tokenize: user references (@USER), hashtags (#HASHTAG) and retweets (RT:).

It is also important to notice that the PoS tagger chooses the PoS assignment for a word from a closed list of possible tags (with the exceptions of proper noun). That means that from words appearing in this list the possible PoS assignments are restricted. This tagger feature usually brings more robustness to the PoS tagging process but can also be misleading when the word is misspelled.

Table 1 shows the closed list of possible PoS tags associated with *mas* and *más*. Notice that both are correct wordforms in Spanish but the possible PoS that can be assigned are different, if we do not add the diacritic when writing the twit the PoS tagger will certainly assign a wrong PoS Label.

Word	Possible PoS
mas (but)	CC, NC
más (more)	RG

Table 1: Possible PoS closed list

Next section will present three different techniques to preprocess text before applying a PoS tagger: Synonym substitution for non-standard text normalization and Text and PoS Filtering strategies for removing Twitter metaelements.

3. Twitter Metaelement Text Filter

This method consists in applying a basic normalization of the metalanguage of the tweet based only on the text:

- Remove "RT @*user*" and "@*user*" tag at the beginning of the tweet.
- Remove "#" from the hashtags.
- Remove "@" from the remaining "@*user*" tags and uppercase the first letter of the username.

4. Twitter Metaelement PoS Filter

In order to decide whether a user or hashtag are part of the syntactic structure of the sentence we need to contextualize them with the PoS around. Freeling is used to obtain the first proposals of PoS.

In a second step these PoS are used to determine whether the user or hashtags are syntacticaly part of the sentence or are metainformation. Once the metalanguage elements are removed, the new resulting text is re-processed.

(Kaufmann and Kalita, 2010) proposes the following set of rules to establish what can be considered part of the sentences with syntactic relation with the rest of the words:

- **@user**
  - A user reference appearing at the beginning as part of a retweet "retweet": (RT @user:) should be removed as well as the word retweet.
  - When an username is followed by a coordinating conjunction, subordinating conjunction, preposition or a verb, it is part of the sentence (and should be kept)
- **#hashtags:** usually identifying the topic of the tweet.
  - hashtags that are not at the end are kept but removing the initial "#"
  - hashtags at the end are considered topic marks and thus removed
- The rest of the metalanguage elements are always removed when detected:
  - **URL**: All URLs are removed.
  - **Vía** at the end of the tweet is removed
  - **...** It can appear close to the end of the tweet indicating that the tweet is being truncated. It is removed if it is the last word of the tweet or the one before last and followed by an URL.

Basically these rules do their decision based on the previous word and its PoS and the PoS of the following word. Notice that the PoS information we will be considering was obtained in the processing of the original text that contains the errors induced by these particular elements.

5. Synonym Substitution Text Normalization

One of the main issues on the text derived from user generated content is the non-standard uses of language. The use of non standard wordforms forces the techniques to deal with many unknown words in a sentence, which is still an open issue, even at the PoS level.

In standard text unknown words tend mostly to be classified as proper-nouns, So that using "ke" instead of "que" although phonologically equivalent can make the PoS tagger to misclassify it as proper noun instead of a subordinate conjunction, and the error can be propagated to the rest of PoS in the sentences and to other NLP task that relies on PoS information.

In order to easy that phenomena, a list of frequent unknown words was replaced by its equivalent standard wordforms. The list is composed by 30 words that where manually detected among the words that appear more than 200 times in the corpus of 100,000 tweets, Table 5. shows the first ones. By replacing these wordforms with their standard equivalent we will do a normalization of the text.

Even that this approach may seem too simple, the normalization of text using classical spell checkers does not offers good results (Clark and Araki, 2011) and many authors

like (Henriquez and Hernandez, 2009) or (Pennell and Liu, 2011) use machine translation systems to perform the normalization, which has a cost out of the scope of this paper. The direct substitution of some misspelled words is efficient in the way that the chat-speak does not change the order of the words, but replaces some of them.

Word	Standard Synonym
D	de (of)
finde	fin de semana (weekend)
x	por (for or times)
Xq	porque (why)
ke	que (what)

Table 2: Examples of wordforms and its standard equivalent

6. Experiments

In order to evaluate the impact of different techniques-steps in the PoS tagging, a corpus of Tweets in Spanish was collected and PoS tagged using different combination of these techniques-steps. The results of the different PoS tagging was later indirectly evaluated.

6.1. Twitter Corpus

The twitter corpus consists in a selection of about 100,000 tweets (1,693,407 tokens) dated from the first to 7th of May 2011. Table 6.1. shows some approximated counting of the twitter metaelements.
The Tweets were previously filtered by several criteria to assure they where written in Spanish: some about the authors (users whose profiles states that they are Spanish speakers and whose timezone is Madrid) and others about the text (high confidence score of being in Spanish by language identification tool).
Filtering using a language identification score may bias our corpus to longer tweets and probably better written, which it is not an issue since we want to focus on the Twitter particular phenomena more than on the general user generated content uses of non-standard language.

metalement	aprox. #
#hashtags	22,349
@user	43,459
RT	14,585
URL	26,116

Table 3: Twitter corpus

The aim of these experiments is to evaluate the effect of the different pre-processing in the linguistic analysis of the Tweets and specifically at the PoS level (notice that more complex levels of annotation, such as parsing, relies heavily on PoS) .

We are currently carrying out a small evaluation by manually evaluating the differences in the PoS tagging between the different strategies. That will allow us to present some qualitative and quantitative evaluation of the results.

6.2. Evaluation of the Impact on the PoS

All the methods used Freeling with our twitter-adapted tokenization rules (as explained Sec 2.). Thus following the different strategies explained in the previous sections we built the following systems:

- **Raw**: Applying directly Freeling
- **Fil**: Removing metaelements using regular expressions (Sec 3.)
- **RPos**: Applying the heuristical Substitutions of metaelements (Sec 4.)
- **SynP**: Applying synonym substitution (Sec 5.) and then the heuristical Substitutions of metaelements using Freeling (PoS)

Table 4 shows the different outcomes of the methods for the twit "#RedesSociales para encontrar #trabajo URL"[2]

Raw	#RedesSociales AQ	para SP	encontrar VM	#trabajo NC	URL Z
Fil	RedesSociales NP	para SP	encontrar VM	trabajo NC	URL Z
RPos	RedesSociales NP	para SP	encontrar VM	trabajo NC	-
SynP	RedesSociales N P	para SP	encontrar VM	trabajo NC	- -

Table 4: Example of different processing methods

In order to grasp the impact of these different strategies in the PoS Tagger, we align the tokens (if the tokens cannot be aligned we discart the tweet. Tokens with synonym substitution or the twitter metaelements are not taken into account).
Table 5 shows the number of different PoS between methods of the 1,140,151 tokens we were able to align (corresponding to 99,898 twits we were able to automatically align). It can be observed that about a 6% of the PoS assignments changes when filtering twitter's specific codings (users, hashtags and url's) while the way this filtering is performed represents less than 0.2%.
So while removing metalanguage elements have a big impact on the PoS tagging there seems to not be much difference on the PoS tagging between the different proposed methods. We think that this phenomena can explained by two main factors, the PoS tagger is base on a dictionary and normalisation issues are dominating the tagger (e.g., the lack of accents, wrong spelling, etc.). For instance in twits like *yo mu bien pero logo platicams mas y t cuento mjor xq aora voy a spining y seguro q Dios y l (cont)* where almost half of the words are misspelled.

[2]We have substituted the real url by URL

	Raw	SynPoS	Fil	PoS
Raw	-	79,331	79,331	79,207
SynPoS	79,331	-	2,050	1,298
Fil	79,331	2,050	-	763
PoS	79,207	1,298	763	-

Table 5: Variation of PoS

6.3. Evaluation on the improvement of PoS

In the next sub-sections two ways to evaluate the quality of the PoS are exposed. The first one was done by manually annotating some specific tokens where the different tools diverge. The second one uses global statistics of the POS assigned to each word to measure how plausible they are.

6.3.1. Evaluation of the different PoS assignments

In order to evaluate if the removing metalanguage elements process improves the PoS tagging we decide to randomly sample words where Raw and the other methods differs. Since the other three methods does not differ greatly we decide to evaluate Raw against the most complex strategy, i.e SynP. Thus 175 words which Raw and SynP PoS assignments differs where randomly selected and annotated by two volunteers. Table 6 shows that SynP improves over the Raw. We removed 11 words from the sample because the twitter was so badly written that it was unclear which could be the right PoS for the chosen target word.

Method	OK	KO	P/R/F1
Raw	27	123	19.28
SynP	105	45	75.00

Table 6: Evaluation on differences between Raw and SynP

To evaluate the effect of the synonym expansion, we also perform an evaluation on a 175 words randomly sample on words where the method RPos and SynP gives a different result. Table 7 shows that the synonym substitution improves the accuracy on the PoS tagging. Like in the previous corpus, 34 words were removed from the sample because it was unclear which could be the right PoS for the chosen target word.

Method	OK	KO	P/R/F1
RPos	57	114	33.33
SynP	83	88	48.53

Table 7: Evaluation on differences between RPos and SynP

Given the different nature of the text in twitter, we also wanted to estimate the interagreement on the PoS assignments. 50 twits of the above samples were annotated by both annotators, the interagreement between the two annotators is substantial, a Cohen's Kappa of 0.78, which is quite close to perfect agreement (above 0.81)[3]

[3]there were 10 disagremments out of 50, and 3 of them about whether the word was suitable for PoS tagging or not

6.3.2. Evaluating NP-unknown words

Since the Freeling PoS chooses among closed list of possible PoS for the known words (plus the Proper Noun label). We decide to perform a specific evaluation for the words that are not present in the dictionary (including known words but tagged as proper nouns). Due to the cost of hand annotating a PoS corpus to carry out a significant evaluation, a context-less statistical indirect evaluation was carried out to evaluating the impact on the unknown words and the Proper Noun category. This evaluation method consists in collecting and reviewing (independent of the context) all the PoS assigned to the most frequent words, using all different strategies. For each word statistics of PoS are collected .

Table 8 shows an example of the annotation task: "me" is a Spanish personal pronoun (pp or p0) but it is not likely to be a proper noun (np). See http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html for a description of the PoS tagset.

Word	PoS	Possible
me	pp	Yes
	np	No
	p0	Yes

Table 8: Word-PoS pairs along with its linguistic plausibility

Following this method 6,019 pairs word-PoS were annotated (corresponding to 3,322 different words). Table 9 shows the results using this ground truth to evaluate the PoS annotations made automatically using the different strategies.

Raw	Fil	RPos	SynP
86.02%	87.43%	87.60%	90.04%

Table 9: Word-PoS assignment precision

The best results are obtained using the PoS filter after the synonyms substitution. The improvements increase up to 4% . While the increase due to the filtering is small (1.58%), only 0.17% better that the straight forward Text filter.

It is likely that the most of errors are due to incorrect words and that they are corrected by the synonym substitution. Table 10 shows the results without taking into account those words (both the originals and their correct synonym) in order to measure the effect of these substitutions in the analysis of the sentence. The improvements in this case are smaller (between 0.5 and 1) and the PoS filter gives the best results, although the differences are small. Thus synonym substitution impact on the determination of the PoS on other words is small.

Raw	Fil	RPos	SynP
89.86%	90.60%	90.82%	90.59%

Table 10: Word-Pos assignment precision excluding the synonyms substituted words

7. Conclusions and Future Work

We can conclude that the pre-processing of Twitter is key to improve the quality of the linguistic analysis using NLP tools trained with standard text. We have empiricaly evaluated the set of rules proposed by (Kaufmann and Kalita, 2010) to remove twitter metalanguage elements. Removing metalanguage is important factor (about 6%) but the different methods seems to have a small impact (less than 0.1%).

The best results seems to be obtained applying this set of rules to remove metalanguage elements after normalizing the text (synonym substitution). The major contribution of this text normalization (synonym substitution) seems to be on correctly assigning the PoS of the replaced words which helps the PoS Filtering rules. The overall improvement on the indirect evaluation is small. But any small improvement on real PoS tagging is fundamental since PoS errors will be propagated and probably magnify in other text analysis task which are based on PoS, such as NERC or parsing.

The evaluation seems to suggest that even it is necessary to remove metalanguage elements, some text normalisation or PoS tagger adaptation is needed in order to have a clear evaluation about which of the different methods to treat twitter metalanguage elements is better.

The text normalization performed in this work is effective but restricted to these words that can be easily identified and replaced by the correct one. To perform a more extensive text normalization, like to detect if "que" has a missing diacritic accent to indicate that is a pronoun, a deeper analysis of the tweet is needed. Maybe another strategy is to develop a Pos tagger trained with tweets, with an internal dictionary based on phonemes instead of words and ignoring capitalization. A compromise between these two techniques may be a good strategy for future work.

8. Acknowledgments

We want to thank the anonymous reviews reviewer for their thoughtful comments. This work is part of the Holopedia Project (TIN2010-21128-C02-02) Social Media Project (CEN 2010 1037) and was partially funded by the Centro para el Desarrollo Tecnológico Industrial (CDTI)

CDTI Centro para el Desarrollo Tecnológico Industrial

9. References

Jordi Atserias, Maria Fuentes, Rogelio Nazar, and Irene Renau. 2012. Spell checking in spanish: The case of diacritic accents. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Eleanor Clark and Kenji Araki. 2011. Text normalization in social media: Progress, problems and applications for a Pre-Processing system of casual english. *Procedia - Social and Behavioral Sciences*, 27(0):2–11.

CA Henriquez and A. Hernandez. 2009. A n-gram-based statistical machine translation approach for text normalization on chatspeak style communications. *CAW2 (Content Analysis in Web 2.0)*.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of Twitter messages. In *Proceedings of the 8th International Conference on Natural Language Processing (ICON 2010)*, Chennai, India. Macmillan India.

C. Kobus, F. Yvon, and G. Damnati. 2008. Normalizing sms: are two metaphors better than one? In *COLING '08*.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

D.L. Pennell and Y. Liu. 2011. A Character-Level machine translation approach for normalization of SMS abbreviations. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, page 974–982, November.

Michel Simard and Alexandre Deslauriers. 2001. Real-time automatic insertion of accents in french text. *Natural Language Engineering*, 7(2):143–165.

D. Yarowsky. 1994. A comparison of corpus-based techniques for restoring accents in spanish and french text. In *Proceedings of the 2nd Annual Workshop on very large Text Corpora*.

Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
E-mail: garjen@ldc.upenn.edu, strassel@ldc.upenn.edu, safa@ldc.upenn.edu, zhiyi@ldc.upenn.edu, haejoong@ldc.upenn.edu

Abstract

We describe an ongoing effort to collect and annotate very large corpora of user-contributed content in multiple languages for the DARPA BOLT program, which has among its goals the development of genre-independent machine translation and information retrieval systems. Initial work includes collection of several hundred million words of online discussion forum threads in English, Chinese and Egyptian Arabic, with multi-layered linguistic annotation for a portion of the collected data. Future phases will target still more challenging genres like Twitter and text messaging. We provide details of the collection strategy and review some of the particular technical and annotation challenges stemming from these genres, and conclude with a discussion of strategies for tackling these issues.

Keywords: Linguistic resources, collection, annotation, data centers

1. Introduction

The DARPA BOLT (Broad Operational Language Translation) Program has among its goals the development of genre-independent machine translation and information retrieval systems. While earlier DARPA programs including GALE (Olive, 2011) made significant strides in improving natural language processing capabilities in structured genres like newswire and broadcasts, performance degrades rapidly when systems are confronted with data that is less formal or whose topics are less constrained that what is typically found in news reports. BOLT is particularly concerned with improving translation and information retrieval performance on informal genres, with a special focus on user-contributed content in the early phases of the program. In the first phase of BOLT, currently underway, Linguistic Data Consortium is collecting and annotating threaded posts from online discussion forums, targeting at least 500 million words in each of three languages: English, Chinese and Egyptian Arabic. A portion of the collected data is manually "triaged" for content and linguistic features, with an optional annotation pass to normalize orthographic and linguistic variation that may prove particularly challenging for downstream (human or automatic) annotation processes. The triage process results in a selection of approximately one million words per language; this data is then tokenized and segmented into sentences with English translations produced where required. The resulting parallel text is manually aligned at the word level, and approximately half of the source data selected for translation is further annotated for morphological and syntactic structure (via Treebanking) for predicate argument structure (via PropBanking), and for entity co-reference.

Later phases of the program target similar data volumes in still more challenging genres including text messaging, chat and micro-blogs like Twitter. The data goals and performance targets for BOLT pose intensive demands, with several key factors that add appreciable risk to the endeavor, most notably an aggressive schedule for collection and annotation combined with the need to develop robust collection and annotation methods to address the inherent variation and inconsistency reflected in the informal genres that are targeted. In this paper we describe the current collection effort, review several of the linguistic and content challenges that are pervasive in this data, and discuss some of the solutions we have adopted.

2. Collection

2.1 Data Scouting

In order to create a corpus with both a high volume of data and a reasonable concentration of threads that meet content and language requirements, we are pursuing a two-stage collection strategy: manual data scouting seeds the corpus with appropriate content, and a semi-supervised harvesting process augments the corpus with larger quantities of automatically-harvested data.

Collection of discussion forums begins with native speaker annotators who are trained in the BOLT data scouting process. These trained data scouts search for individual threads that meet BOLT requirements. Formal guidelines define basic concepts and provide detailed instructions for evaluating the appropriateness of candidate threads. For BOLT, appropriate threads contain primarily original content (as opposed to copies of a published news article, for instance); primarily informal discussion in the target language; and a primary focus on discussion of dynamic events or personal anecdotes. The data scouting guidelines also specify what types of threads or forums should be avoided.

In addition to formal guidelines, data scouting is facilitated through BScout, a customized user interface developed by LDC for BOLT. BScout is a Firefox browser plug-in that records judgments for each scouted thread, including the thread URL, a brief synopsis and an assertion that the thread contains no sensitive personal identifying information or other problematic content. Data scouts also record additional information about thread and forum properties including the level of formality and (for Egyptian scouts) the use of Egyptian Arabic versus Modern Standard Arabic. This meta-information informs the automatic harvesting process.

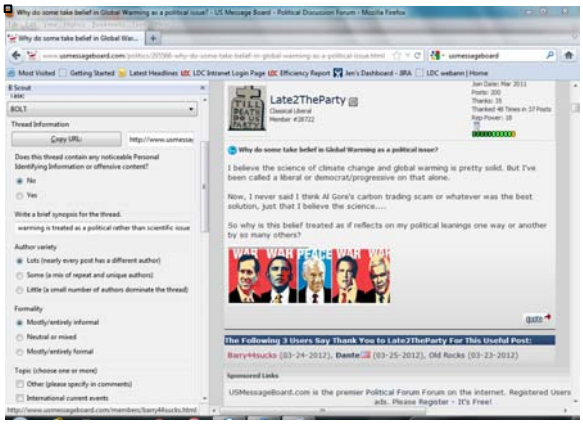

Figure 1: Data Scouting with BScout

The resulting URLs and their corresponding annotations are logged to the BScout database and added to a whitelist for harvesting. When multiple threads are submitted from the same forum that entire forum is targeted for harvesting. Similarly, when multiple forums are targeted from a single host site, that entire site is added to the harvesting whitelist.

2.2 Intellectual Property and Privacy Issues

The type of data targeted presents particular challenges in the domains of copyright and contract law, privacy and objectionable content. Although web content may originate from anywhere in the world, our conservative default assumption is that all content is copyrighted, and we take additional steps to ensure that collected data can be redistributed for research, education and technology development. To further protect the privacy of data creators and to ensure that the corpus does not contain problematic content, data is manually screened for sensitive personal identifying information or other sensitive content prior to inclusion in the annotated corpus. For instance, discussion forums contain numerous credited and uncredited copies of published materials such as newspaper articles. Data scouts are instructed to exclude such content.

2.3 Triage and Segmentation

While our data scouting and automated harvesting approach supports the data volume requirements for BOLT, it also results in a certain amount of unsuitable material making its way into the corpus. While all harvested data is made available to BOLT performers, only a small subset is selected for manual translation and annotation to create BOLT training, development and evaluation sets. It is important that the data selected for annotation meets requirements for language and content; it is also highly desirable that the selected data is high-value; i.e. that it does not duplicate the salient features of existing training data. For these reasons data scouting is followed by a manual triage process. Threads are selected for triage based in part on the results of data scouting, with manually scouted threads and threads from whitelisted forums having highest priority. Additional threads may be selected for triage based on meta-information provided by data scouts as well as other factors like number of posts, average post length and the like.

The triage task has two stages: post selection and sentence segmentation/labeling. During post selection, a native speaker annotator first confirms that the candidate thread generally meets content and language requirements and that it does not contain offensive material or sensitive personal identifying information; problematic threads are discarded from subsequent stages. The annotator then selects individual posts from the thread that are suitable for translation and downstream annotation, following selection guidelines developed with input from BOLT research sites, evaluators and sponsors. For instance, a post that consists solely of the poster agreeing or disagreeing with a previous poster, or a post that contains primarily quoted text, adds little novel content to translation training models and is therefore less appropriate for translation when compared to a post that contains novel linguistic content about an event or entity.

LDC's customized BOLT data triage user interface displays each thread in its entirety, with posts clearly separated and quoted text displayed in blue font. Annotators click on a post to select it; the list of selected posts and associated post metadata appears on the right side of the interface.

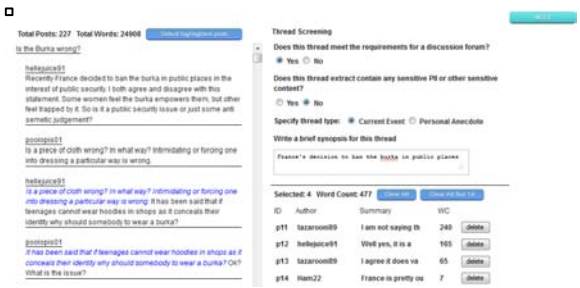

Figure 2: Selecting Posts for Annotation

The second stage of data triage, sentence segmentation/labeling, requires the annotator to identify and label individual Sentence Units within each selected post. A Sentence Unit (SU) is a natural grouping of words written or spoken by a single person. SUs have

semantic cohesion—that is, they can have some inherent meaning when taken in isolation; and they have syntactic cohesion—that is, they have some grammatical structure. The goal of SU annotation is to provide a stable basis for later linguistic annotation activities including translation and syntactic analysis. Annotators first identify SU boundaries by marking the last word of each sentence in the post; they then classify each SU as Keep or Exclude, to indicate which sentences should be excluded from subsequent translation and annotation tasks. Excluded content may include sentences that consist entirely of quotes, sentences that are not in the target language, and segments that consist of formulaic greetings, hyperlink text, image labels, or other undesirable material. Sentence Units marked Exclude are dropped from further annotation but are not deleted from the source corpus.

Where possible, annotators correct automatic segmenter output rather than generating Sentence Unit boundaries from scratch. While automatic sentence segmentation is fairly accurate for more formal genres like newswire, discussion forums and other user-generated content is much more challenging. Use of punctuation and white space is highly variable; for Arabic in particular even long posts may lack punctuation entirely. This makes manual SU segmentation, let alone automatic segmentation, quite challenging. Formal SU annotation guidelines provide specific rules for locating sentence boundaries, and for handling common features like strings of emoticons.

2.4 Automatic Harvesting and Processing

In addition to the front end user interfaces designed to support manual data scouting and triage, LDC has developed a backend framework for BOLT to enable efficient harvesting, processing and formatting of large volumes of discussion forums and other user-generated web data. Each forum host site presents its own unique challenges for automatic harvesting in terms of structure and formatting, so the framework assumes a unique configuration for each site.

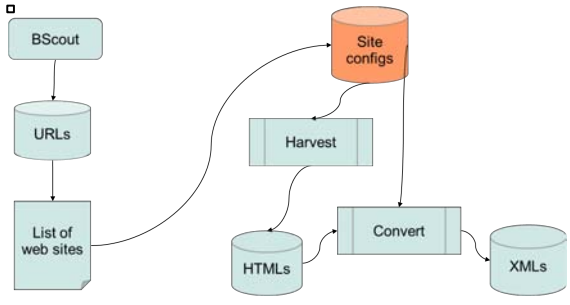

Figure 3: Harvesting and Conversion Process

URLs submitted by data scouts using BScout are first grouped by host site. For each site, a configuration file is written for both the harvester and converter, consisting of a dozen or more XPath expressions and regular expressions. For example, given a home page for a particular forum, an XPath expression is written to identify individual thread URLs contained within that page. Similarly, given a thread page, an XPath expression is written to identify the specific HTML element that contains the body text of posts. Regular expressions are used to clean up target strings. For example, when extracting the post date from the byline, extraneous strings such as "*This post was written on"* are cleaned up using regular expressions.

Once site configuration files have been developed, a harvester processes downloads individual threads, and a converter processes transforms the downloaded HTML files to an XML format. The XML format for BOLT was designed with input from research sites, and consists of a series of post elements including author, post date and post body, with additional markup to identify quoted material (to the extent that such material is consistently marked in the source HTML).

Figure 4: XPath Expressions in Harvesting

Site configuration is often quite challenging. Many site configuration difficulties require a careful examination of the source HTML file in order to identify the problem and achieve the correct configuration. For example, URL navigation (next forum, next thread) may need to be computed from a snippet of Javascript code. Illegal characters, control characters and poorly-rendered HTML can cause parse errors, requiring manual review to diagnose and correct problems.

A particularly difficult (and increasingly common) challenge is harvesting host sites that use AJAX. For such sites, the downloaded HTML contains no content; i.e., there is no body text. Instead, the contents are downloaded dynamically to the web browser when the Javascript code embedded or linked on the HTML page is executed. The use of AJAX among host sites appears to be increasing over time. So far in BOLT, these sites have been dealt with outside of the standard site configuration and harvesting framework, but work is in progress to account for this emerging pattern in the generalized framework.

3. General Challenges

3.1 Quoted Text

The prevalence of quoted material in discussion forums poses challenges in both formatting and content. Quotes in discussion forums often consist entirely of content copied directly from a third party data provider, e.g. an entire newspaper article. It is also very common for forum posts to quote content from prior posts within the

same thread. Setting aside issues of copyright, external quotes are undesirable for BOLT annotation because the language is primarily formal and non-interactive, while internal quotes are undesirable because the same content is likely to have been annotated previously, as part of the original post. As such, the presence or absence of quoted text is an important consideration during data triage. While quoted text is not itself an annotation target, quotes can nonetheless provide important context during annotation. Accurate representation quoted text is also important when establishing provenance during information retrieval tasks.

Posters themselves exhibit considerable variety in choosing to quote entire posts from earlier in the thread or only relevant portions. Additionally, posters may engage in complex quoting in which Poster A quotes a post from Poster B, which in turn contains a quote from Poster C and/or some external source (Figure 5).

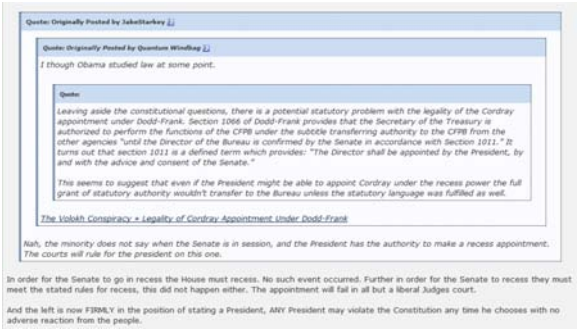

Figure 5: Multiple Embedded Quotes in a Post

Because of the importance of quotes for various parts of the BOLT data pipeline, it is highly desirable for the processed XML version of harvested threads to preserve markup for quoted text. Simply detecting the presence of quoted text in the original source data can be quite difficult given the wide range of HTML representations for quoted text, and there will be a certain number of cases in which the quote markup is missed. However, the majority of well-formed quote markups are preserved in the official XML format, including the possibility of embedded quotes-within-quotes.

3.2 Threading, Post Selection and Annotation

The threaded nature of discussion forums is of particular interest to BOLT, given the program's emphasis on informal and interactive discourse. The content of a forum thread covers multiple posters' perspectives on a topic, and individual posts are best understood in the context of the previous posts within the thread. At the same time, while the unit of collection is full threads, the unit of annotation is individual posts and sentences within those posts. This reality presents some difficulties for downstream annotation, particularly for co-reference.

The co-reference task identifies different mentions of the same entity (person, organization, etc.) within a post; this primarily consists of linking definite referring noun phrases and pronouns to their antecedents. In threaded messages, the pronoun "you" will often be used to refer to a previous poster, while that poster's name does not appear explicitly in the body text for any message. Moreover, in a long or complex thread it can be very difficult to tell which previous poster "you" refers to.

Co-reference annotation is made still more difficult by the BOLT practice of selecting individual posts rather than full threads for annotation. While post sub-selection is necessary given resource constraints and other factors, this does lead to cases where the co-reference chain is broken for a given entity. For instance, in Example 1 the second post would likely be labeled "Exclude" during triage due to the prevalence of quoted text (in italics), but ideally this post should be available for co-reference annotation since it is the only post in the thread where the entity's full name is stated.

Example 1

Post 1: OK guys, I have a new one for you: Billy H. was to Presidents as Pluto is to Planets. Discuss.

Post 2: *OK guys, I have a new one for you: Billy H. was to Presidents as Pluto is to Planets. Discuss.* William Henry Harrison is no longer considered a President?

Post 3: B-to-the-double-H was a small, meaningless President.

Post 4: I disagree. He ran the first modern campaign for president. He had tokens made and ribbons printed up and even slogans we still remember today. "Tippicanoe and Tyler Too" refered to the General winning a battle against the Indians at Tippicanoe and his V.P John Tyler. The log house and hard cider jug on his political tokens was a slap at opponents who tried to portray him as a hard drinker.

While triage annotators are encouraged to consider such issues during post selection, such problems may only be apparent after the downstream annotation tasks have begun. To overcome this challenge, annotators for all downstream tasks are given two versions of the BOLT data to work with: an official version of each file that contains just the selected posts, and a full thread version containing all posts. Annotators can make use of the full thread version for context, and in cases like Example 1 where unselected posts contain information that is crucial for annotation, posts can be provisionally annotated and flagged for later inclusion.

3.3 Non-Standard Language Usage

Discussion forum data is of interest to BOLT largely because of its highly informal nature. Posters do not aim to produce carefully edited prose with standard spelling and punctuation. Non-standard variants, slang and internet abbreviations are common, as are typographical

errors and misspellings. Some intentional misspellings have become part of standard internet language (examples from English include *kitteh* for *kitty* and *pwned* for *owned*). These non-standard uses of language present particular challenges for downstream annotation, in particular translation. Translators must preserve something of the stylistic flavor of the source text while creating a literal, meaning-accurate translation suitable for training MT systems. Other non-standard language features like special text formatting and emoticons have potential complications for other tasks including information retrieval. For example, a poster may follow a statement with a winking smiley emoticon to indicate a non-serious stance. Annotation guidelines for each BOLT task specify how such challenges are handled.

4. Language-Specific Challenges

Beyond the general challenges presented by discussion forums, a number of language-specific issues require special attention.

4.1 Egyptian Orthographic Variation

A general pattern of diglossia in Arabic leads to the use of MSA (Modern Standard Arabic) in formal settings and writing, while dialectal Arabic varieties are primarily used in informal or spoken interactions. But while colloquial varieties like Egyptian Arabic are prevalent in social media such as discussion forums, Twitter and text messaging, there is a lack of commonly accepted orthographic standards for dialectal varieties, and inconsistencies in the way people spell the same words or sounds are to be expected. An example of the orthographic variation in Egyptian Arabic is the frequent use of *alif maqsura* for *yaa* and *ta marbuta* for *haa*, which would both be considered typos or misspellings in MSA, as depicted in the boxed words in Example 2.

Example 2

عايز تناقش ايه و ازاى ... انا واحد من الناس اللى مانحبش تتغرب .. غصب عنى اضطريت اسيب مصر .. و الا حابقى باشتغل نصاب .. انا تخصص critical medicine عارف احنا كام واحد فى مصر ما نعديش 50 بأى شكل ..

We want to talk about what and why…. I am one of those who do not like to migrate… but I had to leave Egypt not by choice, otherwise I would continue to be a thief. I am specialized in critical medicine. Do you know how many of us are there in Egypt? We are 50 at the most.

Additionally, Egyptian Arabic is frequently written using a Romanized script, as in Example 3.

Example 3

ana s2alt 3an ezay w fen a2dar aktb so2aly w 2ab3ato le2ostaz mustafa w no one answer me untill now.rabena ysam7km.

I asked how and where I can write my question and send it to Mr. Mustafa, and no one answer me until now. May God forgive you.

This reality poses an additional challenge for consistency throughout the BOLT annotation pipeline. In order to avoid the likely scenario in which annotators at different phases of the pipeline make different decisions in dealing with nonstandard representations of the language, an additional level of semi-automated annotation to normalize the Egyptian data has been designed. During this optional normalization stage, Romanized text is converted to Arabic script and all text is normalized to a single, standardized representation that is propagated down through the rest of the annotation pipeline.

4.2 Codeswitching

Along with use of multiple orthographic representations of dialectal Arabic, an additional challenge is presented by the frequent use of foreign language(s) including English and other varieties of Arabic, especially Modern Standard Arabic. Codeswitching may occur in isolation, or more commonly, in combination with the orthographic variation described above. Figure 6 below shows a portion of a typical Egyptian Twitter feed, in which English, Romanized Egyptian Arabic, Egyptian written in Arabic script, and Modern Standard Arabic are freely utilized by a single author.

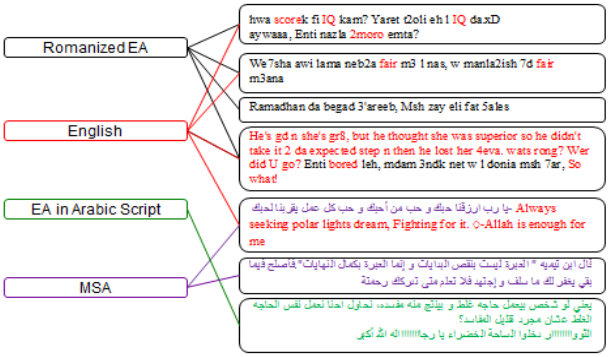

Figure 6: Variation in a Single Egyptian Twitter Feed

English content embedded in a post that is otherwise written in Arabic orthography is simple to detect and exclude from downstream annotation. However, many Egyptian Arabic posts are written using a Romanized script, making it considerably more difficult to distinguish real English borrowings from Arabic words whose transliteration is English-like. It can be even more difficult to clearly distinguish mixing among Egyptian Arabic and other dialects or MSA given lack of diacritics in written text.

4.3 Chinese Word Substitution

Orthographic variation in Chinese is also prevalent in discussion forums due to the informal nature of the data. Common uses of nonstandard orthography include number substitutions and homophones. Example 4 shows the use of a number substitution, which is prompted by the sound similarity between the pronunciation of the numbers and the pronunciation of the words of the

intended meaning. In this case, the pronunciation of 520 sounds like the Chinese for *I love you*, normally written as 我爱你.

Example 4

520，送给所有亲人，兄弟，朋友，想我的，我想的，还有我下一位女朋友！

I love you. My love goes to all my family, my brothers, friends, those missing me, those I miss and my next girlfriend!

In other cases the character for a commonly used homophonous word is substituted for the intended meaning. In Example 5, 萝卜丝 literally means radish slice, but in this context it is understood as a transliteration of Roberts.

Example 5

明明就是**萝卜丝**抓了刘翔的手、什么叫互相的拉拽？还你妹的拳击与动员、这个主持人，你是不是脑子有问题啊？

Obviously it is [**Roberts | radish slice**] who grasped Liu Xiang's hand. Where does the push and pull come from? And what is the nonsense of boxer about? Hey Anchor, are you out of your mind?

Sometimes such variations are induced by intentional substitutions of characters in order to circumvent censorship in the discussion forums. These often involve substitution via homophones for the controversial term, where the homophomes themselves have an innocuous meaning. In Example 6 below, the characters for *Li Yue Yue Niao* and *Wen the Best Actor award winner* are substituted for the potentially censorable *Li Peng* and *Wen Jiabao*, respectively.

Example 6

李月月鸟和温影帝比，谁家更有钱？？？

[Li Peng | Li Yue Yue Niao] and [Wen Jiabao | Wen the Best Actor award winner], whose family is richer???

These orthographic issues cannot be fully addressed by normalization, particularly because the current approach limits that annotation task to only a portion of the Egyptian Arabic data. Instead, annotation guidelines for each downstream task (translation, word alignment, Treebanking) provide explicit guidance on how such variants must be treated.

4.4 Topicalization in Threaded Posts

The practice of topicalization in Chinese allows the noun representing the topic or subject of a sentence to remain implicit once the topic has been established. Topicalization produces threads in which later posts may contain no explicit reference to the people, places, or events under discussion. In Example 7 below, the subject Wang Lijun is introduced in the first post; his name is not explicitly mentioned in subsequent posts. When another name, Bo, is introduced several posts later, that name also becomes implicit in following posts. In the final post in the thread, both individuals are understood to be participants but neither is mentioned explicitly. In this example, DROP-WL represents an implicit mention of Wang Lijun while DROP-BO represents an implicit mention of Bo.

Example 7

Post 1: @重庆市人民政府新闻办公室： 据悉，王立军副市长因长期超负荷工作，精神高度紧张，身体严重不适，经同意，现正在接受休假式的治疗。 转发(4776) | 评论(1429) 8 分钟前 来自新浪微博

It is reported that Deputy Mayor Wang Lijun has agreed to take vacation-style treatment due to unwellness from exhaustion and high pressure, after approval from DROP-WL.

Post 3: 软禁了哇。

DROP-WL imprisoned?

Post 7:他是薄的人？

Is he (Wang) in Bo's team?

Post 11: 铁杆头号手下啊！从东北带来的啊！

DROP-WL die-hard subordinate! DROP-WL accompanied DROP-BO from North East!

There are several annotation challenges associated with topicalization. For translation, the full thread context must be carefully reviewed in order to understand the implied topic/subject(s). Word alignment and co-reference annotation also must account for the empty subject on the source side and the explicitly stated subject on the translation side.

5. Conclusion

To support the BOLT Program's goal of improved machine translation and information retrieval technologies for informal genres, Linguistic Data Consortium is engaged in collection and annotation of discussion forums and other user-generated content in three languages. The BOLT corpora described here have been designed for variety, breadth and volume. The collection target is unconstrained, real-world data, reflecting the full spectrum of quality and content of such data on the web. The scale is very large, ultimately comprising over a billion words per language. These demands have required new approaches and new frameworks for both collection and annotation.

These resources described here will initially be distributed to BOLT performers as training, development and evaluation data. We will wherever possible distribute the data more broadly, for example to our members and licensees, through the usual mechanisms including publication in the LDC catalog.

6. Acknowledgements

This material is based upon work supported by Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. References

Olive, J.; Christianson, C. and McCary, J. (Eds.) (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer New York.