Preface

PROCEEDINGS OF THE 1th WORKSHOP ON SEMANTIC SENTIMENT ANALYSIS

and

WORKSHOP ON SOCIAL MEDIA AND LINKED DATA FOR EMERGENCY RESPONSE

Proceedings edited by

- Aldo Gangemi
- Harith Alani
- Malvina Nissim
- Erik Cambria
- Diego Reforgiato Recupero
- Vitaveska Lanfranchi
- Tomi Kauppinen

The workshops were held in conjunction with the 11th European Semantic Web Conference (ESWC 2014), Crete, Greece, May 25th, 2014.

Preface

This volume collects the papers of two workshops held at the 11th European Semantic Web Conference (ESWC 2014) in Crete, Greece. For more detailed information about the 1th workshop on semantic sentiment analysis (SSA2014) and workshop on social media and linked data for emergency response (SMILE2014) please refer to the individual workshop prefaces.
The editors would like to thank the SSA2014 and SMILE2014 organizers, the members of the two scientific committees, the authors, and of course, the participants of the workshops.

January 2015
The editors

TABLE OF CONTENT

- *Preface and Table of Contents* Pag. 1-2
- *SSA2014 Preface* Pag. 3-4
- *Adapting Sentiment Lexicons using Contextual Semantics for Twitter Sentiment Analysis* Pag. 5-12
- *Usage of Abstract Features in Semantic Sentiment Analysis* Pag. 13-16
- *Sentiment Analysis for Dynamic User Preference Inference in Spoken Dialogue Systems* Pag. 17-24
- *Semantic Web-based Sentiment Analysis 25-28* Pag. 25-28
- *SMILE2014 Preface* Pag. 29-30
- *Towards emergency vehicle routing using Geolinked Open Data the case study of the Municipality of Catania* Pag. 31-42
- *Visualisation of User-Generated Event Information: Towards Geospatial Situation Awareness Using Hierarchical Granularity Levels* Pag. 43-54
- *An Ontology-Based Approach to Social Media Mining for Crisis Management* Pag. 55-66
- *Visual Analysis of Real-time Social Media for Emergency Response* Pag. 67-78

Preface

This edition of the Workshop on Semantic Sentiment Analysis (SSA2014) was the first on this topic in the semantic web community. It was organised as a half-day event, co-located with the 11th European Semantic Web Conference (ESWC2014). Our aim in launching this workshop was twofold: (i) providing a space for proposing and discussing new methods and applications of semantics in sentiment analysis; and (ii) broadening the sentiment analysis community towards the incorporation of semantic features. Semantics have shown to improve precision and accuracy of traditional sentiment analysis methods. The workshop was also linked to the Concept-Level Sentiment Analysis Challenge, where users submitted their practical systems that had to include semantics features to solve different tasks in the sentiment analysis domain.

The workshop included oral presentations, a panel session, and an invited talk by Mauro Dragoni, who participated in the semantic sentiment analysis challenge: *A Fuzzy System For Semantic Sentiment Analysis*, by Mauro Dragoni, Andrea Tettamanzi and Celia Da Costa Pereira.

Altogether, SSA2014 received six research paper submissions. From among these submissions, the Program Committee selected four research papers for presentation at the workshop. Additionally, the organisers of each ESWC Workshop were asked to indicate the workshop's "best paper". Authors were then invited to submit a revised and extended version of their work for inclusion in the main Conference proceedings. Based on the reviewers' scores, we selected the paper *Adapting Sentiment Lexicons using Contextual Semantics for Twitter Sentiment Analysis*, by Hassan Saif, Yulan He, Miriam Fernandez and Harith Alani.

We are grateful to the Programme Committee for their hard work, and to the authors for submitting their papers and for addressing the reviewers' comments. We would also like to thank the organisers of the 11th European Semantic Web Conference for hosting SSA2014 at ESWC2014 and for giving us the opportunity to organise this workshop in such an interesting and stimulating scientific environment. Further information about the Workshop on Semantic Sentiment Analysis can be found at:
`http://ontologydesignpatterns.org/wiki/SemanticSentimentAnalysis2014`

November 2014

SSA2014 Chairs

Aldo Gangemi
Harith Alani
Malvina Nissim
Erik Cambria
Diego Reforgiato Recupero

Adapting Sentiment Lexicons using Contextual Semantics for Sentiment Analysis of Twitter

Hassan Saif,[1] Yulan He,[2] Miriam Fernandez[1] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, m.fernandez, h.alani}@open.ac.uk
[2] School of Engineering and Applied Science, Aston University, United Kingdom
y.he@cantab.net

Abstract. Sentiment lexicons for sentiment analysis offer a simple, yet effective way to obtain the prior sentiment information of opinionated words in texts. However, words' sentiment orientations and strengths often change throughout various contexts in which the words appear. In this paper, we propose a lexicon adaptation approach that uses the contextual semantics of words to capture their contexts in tweet messages and update their prior sentiment orientations and/or strengths accordingly. We evaluate our approach on one state-of-the-art sentiment lexicon using three different Twitter datasets. Results show that the sentiment lexicons adapted by our approach outperform the original lexicon in accuracy and F-measure in two datasets, but give similar accuracy and slightly lower F-measure in one dataset.

Keywords: Sentiment Analysis, Semantics, Lexicon Adaptation, Twitter

1 Introduction

Sentiment analysis on Twitter has been attracting much attention recently due to the rapid growth in Twitter's popularity as a platform for people to express their opinions and attitudes towards a great variety of topics. Most existing approaches to Twitter sentiment analysis can be categorised into machine learning [7, 11, 13] and lexicon-based approaches [2, 8, 15, 6].

Lexicon-based approaches use lexicons of words weighted with their sentiment orientations to determine the overall sentiment in texts. These approaches have shown to be more applicable to Twitter data than machine learning approaches, since they do not require training from labelled data and therefore, they offer a domain-independent sentiment detection [15]. Nonetheless, lexicon-based approaches are limited by the sentiment lexicon used [21]. Firstly, because sentiment lexicons are composed by a generally static set of words that do not cover the wide variety of new terms that constantly emerge in the social web. Secondly, because words in the lexicons have fixed prior sentiment orientations, i.e. each term has always the same associated sentiment orientation independently of the context in which the term is used.

To overcome the above limitations, several lexicon bootstrapping and adaptation methods have been previously proposed. However, these methods are either supervised [16], i.e., they require training from human-coded corpora, or based on studying the statistical, syntactical or linguistic relations between words in general textual corpora (e.g., The Web) [17, 19] or in static lexical knowledge sources (e.g., WordNet) [5] ignoring, therefore, the specific textual context in which the words appear. In many cases, however, the sentiment of a word is implicitly associated with the semantics of its context [3].

In this paper we propose an unsupervised approach for adapting sentiment lexicons based on the contextual semantics of their words in a tweet corpus. In particular, our approach studies the co-occurrences between words to capture their contexts in tweets and update their prior sentiment orientations and/or sentiment strengths in a given lexicon accordingly.

As a case study we apply our approach on Thelwall-Lexicon [15], which, to our knowledge, is the state-of-the-art sentiment lexicon for social data. We evaluate the adapted lexicons by performing a lexicon-based polarity sentiment detection (positive vs. negative) on three Twitter datasets. Our results show that the adapted lexicons produce a significant improvement in the sentiment detection accuracy and F-measure in two datasets but gives a slightly lower F-measure in one dataset.

In the rest of this paper, related work is discussed in Section 2, and our approach is presented in Sections 3. Experiments and results are presented in Sections 4. Discussion and future work are covered in Section 5. Finally, we conclude our work in Section 6.

2 Related Work

Exiting approaches to bootstrapping and adapting sentiment lexicons can be categorised into dictionary and corpus-based approaches. The dictionary-based approach [5, 14] starts with a small set of general opinionated words (e.g., good, bad) and lexical knowledge base (e.g., WordNet). After that, the approach expands this set by searching the knowledge base for words that have lexical or linguistic relations to the opinionated words in the initial set (e.g., synonyms, glosses, etc).

Alternatively, the corpus-based approach measures the sentiment orientation of words automatically based on their association to other strongly opinionated words in a given corpus [17, 14, 19]. For example, Turney and Littman [17] used *Pointwise Mutual Information* (PMI) to measure the statistical correlation between a given word and a balanced set of 14 positive and negative paradigm words (e.g., good, nice, nasty, poor). Although this work does not require large lexical input knowledge, its identification speed is very limited [21] because it uses web search engines in order to retrieve the relative co-occurrences of words.

Following the aforementioned approaches, several lexicons such as MPQA [20] and SentiWordNet [1] have been induced and successfully used for sentiment analysis on conventional text (e.g., movie review data). However, on Twitter these lexicons are not as compatible due to their limited coverage of Twitter-specific expressions, such as abbreviations and colloquial words (e.g, "`looov`", "`luv`", "`gr8`") that are often found in tweets.

Quite few sentiment lexicons have been recently built to work specifically with social media data, such as Thelwall-Lexicon [16] and Nielsen-Lexicon [8]. These lexicons have proven to work effectively on Twitter data. Nevertheless, such lexicons are similar to other traditional ones, in the sense that they all offer fixed and context-insensitive word-sentiment orientations and strengths. Although a training algorithm has been proposed to update the sentiment of terms in Thelwall-Lexicon[16], it requires to be trained from human-coded corpora, which is labour-intensive to obtain.

Aiming at addressing the above limitations we have designed our lexicon-adaptation approach in away that allows to (i) work in unsupervised fashion, avoiding the need for labelled data, and (ii) exploit the contextual semantics of words. This allows capturing their contextual information in tweets and update their prior sentiment orientation and strength in a given sentiment lexicon accordingly.

3 A Contextual Semantic Approach to Lexicon Adaptation

The main principle behind our approach is that the sentiment of a term is not static, as found in general-purpose sentiment lexicons, but rather depends on the context in which the term is used, i.e., it depends on its contextual semantics.[3] Therefore, our approach functions in two main steps as shown in Figure 1. First, given a tweet collection and a sentiment lexicon, the approach builds a contextual semantic representation for each unique term in the tweet collection and subsequently uses it to derive the term's contextual sentiment orientation and strength. The SentiCircle representation model is used to this end [10]. Secondly, rule-based algorithm is applied to amend the prior sentiment of terms in the lexicon based on their corresponding contextual sentiment. Both steps are further detailed in the following subsections.

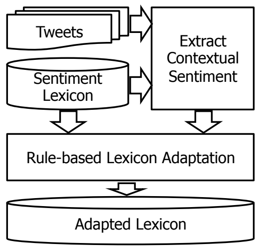

Fig. 1. The systematic workflow of our proposed lexicon adaptation approach.

3.1 Capturing Contextual Semantics and Sentiment

The first step in our pipeline is to capture the words contextual semantics and sentiment in tweets. To this end, we use our previously proposed semantic representation model, SentiCircle [10].

Following the distributional hypothesis that words that co-occur in similar contexts tend to have similar meaning [18], SentiCircle extracts the contextual semantics of a word from its co-occurrence patterns with other words in a given tweet collection. These patterns are then represented as a geometric circle, which is subsequently used to compute the contextual sentiment of the word by applying simple trigonometric identities on it. In particular, for each unique term m in a tweet collection, we build a two-dimensional geometric circle, where the term m is situated in the centre of the circle, and each point around it represents a context term c_i (i.e., a term that occurs with m in the same context). The position of c_i, as illustrated in Figure 2, is defined jointly by its Cartesian coordinates x_i, y_i as:

$$x_i = r_i \cos(\theta_i * \pi) \qquad\qquad y_i = r_i \sin(\theta_i * \pi)$$

Where θ_i is the polar angle of the context term c_i and its value equals to the prior sentiment of c_i in a sentiment lexicon before adaptation, r_i is the radius of c_i and its value represents the degree of correlation (tdoc) between c_i and m, and can be computed as:

$$r_i = tdoc(m, c_i) = f(c_i, m) \times \log \frac{N}{N_{c_i}}$$

[3] We define context as a textual corpus or a set of tweets.

where $f(c_i, m)$ is the number of times c_i occurs with m in tweets, N is the total number of terms, and N_{c_i} is the total number of terms that occur with c_i. Note that all terms' radii in the SentiCircle are normalised. Also, all angles' values are in radian. The trigonometric properties of the SentiCircle allows us to encode the contextual semantics of a term as sentiment orientation and sentiment strength. Y-axis defines the sentiment of the term, i.e., a positive y value denotes a positive sentiment and vice versa. The X-axis defines the sentiment strength of the term. The smaller the x value, the stronger the sentiment.[4] This, in turn, divides the circle into four sentiment quadrants. Terms in the two upper quadrants have a positive sentiment ($\sin\theta > 0$), with upper left quadrant representing stronger positive sentiment since it has larger angle values than those in the top right quadrant. Similarly, terms in the two lower quadrants have negative sentiment values ($\sin\theta < 0$). Moreover, a small region called the "*Neutral Region*" can be defined. This region, as shown in Figure 2, is located very close to X-axis in the "*Positive*" and the "*Negative*" quadrants only, where terms lie in this region have very weak sentiment (i.e, $|\theta| \approx 0$).

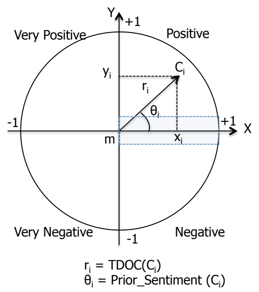

Fig. 2. SentiCircle of a term m. Neutral region is shaded in blue.

Calculating Contextual Sentiment In summary, the Senti-Circle of a term m is composed by the set of (x, y) Cartesian coordinates of all the context terms of m. An effective way to compute the overall sentiment of m is by calculating the geometric median of all the points in its SentiCircle. Formally, for a given set of n points $(p_1, p_2, ..., p_n)$ in a SentiCirlce Ω, the 2D geometric median g is defined as: $g = \arg\min_{g \in \mathbb{R}^2} \sum_{i=1}^{n} || |p_i - g||_2$. We call the geometric median g the **SentiMedian** as its position in the SentiCircle determines the final contextual-sentiment orientation and strength of m.

Note that the boundaries of the neutral region can be computed by measuring the density distribution of terms in the SentiCircle along the Y-axis. In this paper we use similar boundaries to the ones used in [10] since we use the same evaluation datasets.

3.2 Lexicon Adaptation

The second step in our approach is to update the sentiment lexicon with the terms' contextual sentiment information extracted in the previous step. As mentioned earlier, in this work we use Thelwall-Lexicon [16] as a case study. Therefore, in this section we first describe this lexicon and its properties, and then introduce our proposed adaptation method.

Thelwall-Lexicon consists of 2546 terms coupled with integer values between -5 (very negative) and +5 (very positive). Based on the terms' prior sentiment orientations and strengths (SOS), we group them into three subsets of 1919 negative terms (SOS∈[-2,-5]), 398 positive terms (SOS∈[2,5]) and 229 neutral terms (SOS∈{-1,1}).

The adaptation method uses a set of antecedent-consequent rules that decides how the prior sentiment of the terms in Thelwall-Lexicon should be updated according to the positions of their SentiMedians (i.e., their contextual sentiment). In particular, for a term m, the method checks (i) its prior SOS value in Thelwall-Lexicon and (ii) the SentiCircle

[4] This is because $\cos\theta < 0$ for large angles.

quadrant in which the SentiMedian of m resides. The method subsequently chooses the best-matching rule to update the term's prior sentiment and/or strength.

Table 1 shows the complete list of rules in the proposed method. As noted, these rules are divided into *updating rules*, i.e., rules for updating the existing terms in Thelwall-Lexicon, and *expanding rules*, i.e., rules for expanding the lexicon with new terms. The *updating rules* are further divided into rules that deal with terms that have similar prior and contextual sentiment orientations (i.e., both positive or negative), and rules that deal with terms that have different prior and contextual sentiment orientations (i.e., negative prior, positive contextual sentiment and vice versa).

Although they look complicated, the notion behind the proposed rules is rather simple: *Check how strong the contextual sentiment is and how weak the prior sentiment is* $\rightarrow$ *update the sentiment orientation and strength accordingly*. The strength of the contextual sentiment can be determined based on the sentiment quadrant of the SentiMedian of m, i.e., the contextual sentiment is strong if the SentiMedian resides in the "*Very Positive*" or "*Very Negative*" quadrants (See Figure 2). On the other hand, the prior sentiment of m (i.e., $prior_m$) in Thelwall-Lexicon is weak if $|prior_m| \leqslant 3$ and strong otherwise.

Id	Antecedents	Consequent
Updating Rules (Similar Sentiment Orientations)		
1	$(\|prior\| \leqslant 3) \wedge (SentiMedian \notin StrongQuadrant)$	$\|prior\| = \|prior\| + 1$
2	$(\|prior\| \leqslant 3) \wedge (SentiMedian \in StrongQuadrant)$	$\|prior\| = \|prior\| + 2$
3	$(\|prior\| > 3) \wedge (SentiMedian \notin StrongQuadrant)$	$\|prior\| = \|prior\| + 1$
4	$(\|prior\| > 3) \wedge (SentiMedian \in StrongQuadrant)$	$\|prior\| = \|prior\| + 1$
Updating Rules (Different Sentiment Orientations)		
5	$(\|prior\| \leqslant 3) \wedge (SentiMedian \notin StrongQuadrant)$	$\|prior\| = 1$
6	$(\|prior\| \leqslant 3) \wedge (SentiMedian \in StrongQuadrant)$	$prior = -prior$
7	$(\|prior\| > 3) \wedge (SentiMedian \notin StrongQuadrant)$	$\|prior\| = \|prior\| - 1$
8	$(\|prior\| > 3) \wedge (SentiMedian \in StrongQuadrant)$	$prior = -prior$
9	$(\|prior\| > 3) \wedge (SentiMedian \in NeutralRegion)$	$\|prior\| = \|prior\| - 1$
10	$(\|prior\| \leqslant 3) \wedge (SentiMedian \in NeutralRegion)$	$\|prior\| = 1$
Expanding Rules		
11	$SentiMedian \in NeutralRegion$	$(\|contextual\| = 1) \wedge AddTerm$
12	$SentiMedian \notin StrongQuadrant$	$(\|contextual\| = 3) \wedge AddTerm$
13	$SentiMedian \in StrongQuadrant$	$(\|contextual\| = 5) \wedge AddTerm$

Table 1. Adaptation rules for Thelwall-Lexicon, where `prior`: prior sentiment value, `StrongQuadrant`: very negative/positive quadrant in the SentiCircle, `Add`: add the term to Thelwall-Lexicon.

For example, the word "`revolution`" in Thelwall-Lexicon has a weak negative sentiment (*prior=-2*) while it has a neutral contextual sentiment since its SentiMedian resides in the neutral region ($SentiMedian \in NeutralRegion$). Therefore, rule number 10 is applied and the term's prior sentiment in Thelwall lexicon will be updated to neutral ($|prior| = 1$). In another example, the words "`Obama`" and "`Independence`" are not covered by the Thelwall-Lexicon, and therefore, they have no prior sentiment. However, their SentiMedians reside in the "*Positive*" quadrant in their SentiCircles, and therefore rule number 12 is applied and both terms will be assigned with a positive sentiment strength of 3 and added to the lexicon consequently.

4 Evaluation Results

We evaluate our approach on Thelwall-Lexicon using three adaptation settings: (i) the *update* setting where we update the prior sentiment of existing terms in the lexicon, (ii) The *expand* setting where we expand Thelwall-Lexicon with new opinionated terms, and (iii) the *update+expand* setting where we try both aforementioned settings together. To this end, we use three Twitter datasets OMD, HCR and STS-Gold. Numbers of positive and negative tweets within these datasets are summarised in Table 2, and detailed in the references added in the table. To evaluate the adapted lexicons under the above settings, we perform binary polarity classification on the three datasets. To this end, we use the sentiment detection method proposed with Thelwall-Lexicon [15]. According to this method a tweet is considered as positive if its aggregated positive sentiment strength is 1.5 times higher than the aggregated negative one, and negative vice versa.

Dataset	**Tweets**	**Positive**	**Negative**
Obama-McCain Debate (OMD)[4]	1081	393	688
Health Care Reform (HCR)[12]	1354	397	957
Standford Sentiment Gold Standard (STS-Gold)[9]	2034	632	1402

Table 2. Twitter datasets used for the evaluation

Applying our adaptation approach to Thelwall-Lexicon results in dramatic changes in it. Table 3 shows the percentage of words in the three datasets that were found in Thelwall-Lexicon with their sentiment changed after adaptation. One can notice that on average 9.61% of the words in our datasets were found in the lexicon. However, updating the lexicon with the contextual sentiment of words resulted in 33.82% of these words flipping their sentiment orientation and 62.94% changing their sentiment strength while keeping their prior sentiment orientation. Only 3.24% of the words in Thelwall-Lexicon remained untouched. Moreover, 21.37% of words previously unseen in the lexicon were assigned with contextual sentiment by our approach and added to Thelwall-Lexicon subsequently.

	OMD	HCR	STS-Gold	Average
Words found in the lexicon	12.43	8.33	8.09	9.61
Hidden words	87.57	91.67	91.91	90.39
Words flipped their sentiment orientation	35.02	35.61	30.83	33.82
Words changed their sentiment strength	61.83	61.95	65.05	62.94
Words remained unchanged	3.15	2.44	4.13	3.24
New opinionated words	23.94	14.30	25.87	21.37

Table 3. Average percentage of words in the three datasets that had their sentiment orientation or strength updated by our adaptation approach

Table 4 shows the average results of binary sentiment classification performed on our datasets using (i) the original Thelwall-Lexicon (*Original*), (ii) Thelwall-Lexicon induced under the *update* setting (*Updated*), and (iii) Thelwall-Lexicon induced under the *update+expand* setting.[5] The table reports the results in accuracy and three sets of precision (P), recall (R), and F-measure (F1), one for positive sentiment detection, one for negative, and one for the average of the two.

From these results in Table 4, we notice that the best classification performance in accuracy and F1 is obtained on the STS-Gold dataset regardless the lexicon being used. We also observe that the negative sentiment detection performance is always higher than the positive detection performance for all datasets and lexicons.

As for different lexicons, we notice that on OMD and STS-Gold the adapted lexicons outperform the original lexicon in both accuracy and F-measure. For example, on OMD the adapted lexicon shows an average improvement of 2.46% and 4.51% in accuracy and F1 respectively over the original lexicon. On STS-Gold the performance improvement is

[5] Note that in this work we do not report the results obtained under the expand setting since no improvement was observed comparing to the other two settings.

Datasets	Lexicons	Accuracy	Positive Sentiment			Negative Sentiment			Average		
			P	R	F1	P	R	F1	P	R	F1
OMD	Original	66.79	55.99	40.46	46.97	70.64	81.83	75.82	63.31	61.14	61.4
	Updated	**69.29**	58.89	51.4	54.89	74.12	79.51	76.72	66.51	65.45	65.8
	Updated+Expanded	69.2	58.38	53.18	55.66	74.55	78.34	76.4	66.47	65.76	**66.03**
HCR	Original	66.99	43.39	41.31	42.32	76.13	77.64	76.88	59.76	59.47	**59.6**
	Updated	**67.21**	42.9	35.77	39.01	75.07	80.25	77.58	58.99	58.01	58.29
	Updated+Expanded	66.99	42.56	36.02	39.02	75.05	79.83	77.37	58.8	57.93	58.19
STS-Gold	Original	81.32	68.75	73.1	70.86	87.52	85.02	86.25	78.13	79.06	78.56
	Updated	81.71	69.46	73.42	71.38	87.7	85.45	86.56	78.58	79.43	78.97
	Updated+Expanded	**82.3**	70.48	74.05	72.22	88.03	86.02	87.01	79.26	80.04	**79.62**

Table 4. Cross comparison results of original and the adapted lexicons

less significant than that on OMD, but we still observe 1% improvement in accuracy and F1 comparing to using the original lexicon. As for the HCR dataset, the adapted lexicon gives on average similar accuracy, but 1.36% lower F-measure. This performance drop can be attributable to the poor detection performance of positive tweets. Specifically, we notice from Table 4 a major loss in the recall on positive tweet detection using both adapted lexicons. One possible reason is the sentiment class distribution in our datasets. In particular, one may notice that HCR is the most imbalanced amongst the three datasets. Moreover, by examining the numbers in Table 3, we can see that HCR presents the lowest number of new opinionated words among the three datasets (i.e., 10.61% lower than the average) which could be another potential reason for not observing any performance improvement.

5 Discussion and Future Work

We demonstrated the value of using contextual semantics of words for adapting sentiment lexicons from tweets. Specifically, we used Thelwall-Lexicon as a case study and evaluated its adaptation to three datasets of different sizes. Although the potential is palpable, our results were not conclusive, where a performance drop was observed in the HCR dataset using our adapted lexicons. Our initial observations suggest that the quality of our approach might be dependent on the sentiment class distribution in the dataset. Therefore, a deeper investigation in this direction is required.

We used the SentiCircle approach to extract the contextual semantics of words from tweets. In future work we will try other contextual semantic approaches and study how the semantic extraction quality affects the adaptation performance.

Our adaptation rules in this paper are specific to Thelwall-Lexicon. These rules, however, can be generalized to other lexicons, which constitutes another future direction of this work.

All words which have contextual sentiment were used for adaptation. Nevertheless, the results conveyed that the prior sentiments in the lexicon might need to be unchanged for words of specific syntactical or linguistic properties in tweets. Part of our future work is to detect and filter those words that are more likely to have stable sentiment regardless the contexts in which they appear.

6 Conclusions

In this paper we proposed an unsupervised approach for sentiment lexicon adaptation from Twitter data. Our approach extracts the contextual semantics of words and uses them to update the words' prior sentiment orientations and/or strength in a given sentiment lexicon. The evaluation was done on Thelwall-Lexicon using three Twitter datasets. Results showed that lexicons adapted by our approach improved the sentiment classification performance in both accuracy and F1 in two out of three datasets.

Acknowledgment

This work was supported by the EU-FP7 project SENSE4US (grant no. 611242).

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May. Valletta, Malta (2010)
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
3. Cambria, E.: An introduction to concept-level sentiment analysis. In: Advances in Soft Computing and Its Applications, pp. 478–483. Springer (2013)
4. Diakopoulos, N., Shamma, D.: Characterizing debate performance via aggregated twitter sentiment. In: Proc. 28th Int. Conf. on Human factors in computing systems. ACM (2010)
5. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: EACL. vol. 6, p. 2006 (2006)
6. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd World Wide Web conf (2013)
7. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM. Barcelona, Spain (2011)
8. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
9. Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. In: Proceedings, 1st ESSEM Workshop. Turin, Italy (2013)
10. Saif, H., Fernandez, M., He, Y., Alani, H.: Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In: Proc. 11th Extended Semantic Web Conf. (ESWC). Crete, Greece (2014)
11. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proc. 11th Int. Semantic Web Conf. (ISWC). Boston, MA (2012)
12. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland (2011)
13. Suttles, J., Ide, N.: Distant supervision for emotion classification with discrete binary values. In: Computational Linguistics and Intelligent Text Processing, pp. 121–136. Springer (2013)
14. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of words using spin model. In: Proc. 43rd Annual Meeting on Association for Computational Linguistics (2005)
15. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. American Society for Information Science and Technology 63(1), 163–173 (2012)
16. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. American Society for Info. Science and Technology 61(12) (2010)
17. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 315–346 (2003)
18. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1), 141–188 (2010)
19. Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R.: The viability of web-derived polarity lexicons. In: Human Language Technologies: ACL (2010)
20. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proc. Empirical Methods in NLP Conf. (EMNLP). Vancouver, Canada (2005)
21. Xu, T., Peng, Q., Cheng, Y.: Identifying the semantic orientation of terms using s-hal for sentiment analysis. Knowledge-Based Systems 35, 279–289 (2012)

Usage of Abstract Features in Semantic Sentiment Analysis

Mohammed Almashraee[1], Kia Teymourian[1], Dagmar Monett-Diaz[2], and Adrian Paschke[1]

[1] Freie Universität Berlin, Corporate Semantic Web Working Group
[2] Department of Computer Science, Berlin School of Economics and Law
{almashraee,kia,paschke}@inf.fu-berlin.de
dagmar.monett-diaz@hwr-berlin.de

Abstract. Feature-based sentiment analysis can be realized on different types of object features. Some of these features might be about technical aspects of the objects and some others might be application-oriented features. The application-oriented features are more abstract features and can be of interest to the broad number of people than only the technical experts of specific products. In this paper, we propose an approach for extraction of abstract object features from a set of sub-features. In our approach we use a knowledge base about the application domain to extract related sub-features of high-level abstract features. On the basis of such related sub-features our approach performs the extraction of more abstract features that are only implicitly included in the analysis text.

1 Introduction

People express their opinions about certain objects using features. For example, in photography application domain, users of digital cameras express their opinions about features of cameras like *flash* or *lens*. Consumers make use of the opinions expressed to know about the quality of a product and its aspects so that they can make the right purchase decisions [1,2].

Sentiment analysis of technical-oriented features like *flash, lens, optical zoom, shutter, sensor quality* are interesting for professional photographers who are familiar with the technical details of cameras and know which one of the features are important for which kind of photography modes. Contrary to that, non-expert users are interested in features which are more abstract and are application-oriented. For example non-experts are interested to know if the camera can make good pictures of kids or if the camera can take pictures of landscape during their vacations. Such implicit high-level abstract features are mostly not explicitly mentioned in the review corpus or are only implicitly mentioned in some of the review items. The main difference between abstract features and sub-features is that abstract features are non-technical aspects which are rarely available among related reviews, while sub-features are technical aspects which can be found frequently in the review corpus.

In this paper, we propose an approach for the semantic sentiment analysis of abstract features based on the related sub-features. The abstract features can be derived from the explicitly mentioned sub-features that are related to the abstract feature. The system in general performs sentiment analysis on the review text on the basis of an extracted set of sub-features. Table 1 shows some examples of such abstract futures and the related sub-features.

Table 1. Examples of Abstract- and Sub-Features of Digital Cameras

Abstract-Features	**Related Sub-Features**
Night photo	Flash, Lens, Image Processor, Sensors
Portrait	Optical Zoom, Lens, Image Processor
Sports	Shutter, Image Processor, Flash, Sensors
Landscape	Optical Zoom, Shutter, Flash, Sensors
Kids Photography	Shutter, Image Processor, Sensors, Flash

As an example consider the following of review text[3]: *"It automatically selects the* ***best shooting settings*** *for optimal quality based on the environmental factors (****lightning*** *I guess) to provide point'n'* ***shoot simplicity****.* ***16.0 Megapixels****, with loads of resolution pictures are* ***still clear****.* ***High resolution*** *is also* ***good for producing biggest printouts****. 5x* ***Optical Zoom is sufficient*** *in most cases.* ***DIGIC 4 Image Processor*** *is not as fast as* ***DIGIC 5*** *though* ***fast and powerful enough*** *to give you advanced system options, provide* ***quick-shoot with reliable performance*** *and* ***low battery consumption****. As far as I know* ***DIGIC 4*** *is currently* ***Canon's most efficient processor*** *for* ***budget cameras****. BTW it has some* ***Eco mode****, that is said to be providing even* ***faster warm-up times*** *and* ***saves the standard battery****, but I haven't tested it yet.* ***Very lightweight****, just put it into your* ***pocket****, can take it everywhere. Like A2300 it* ***lacks optical image stabilization****, though it's got digital image stabilization.* ***1/2.3" sensor****, well,* ***entry level*** *CCD providing good pictures,* ***not of a DSLR quality****, that's all I can say."*

2 Abstract Features in Semantic Sentiment Analysis

Our approach consist of the following processing tasks:

1. *Feature Extraction:* In this task the related sub-features are identified.
2. *Knowledge-based Feature Annotation:* By using a knowledge-based annotator the sub-features can be annotated with their background knowledge resources.
3. *Feature Preparation:* The background knowledge for each annotated resource is retrieved from knowledge base and enriched to them.
4. *Sentiment Relation Calculation:* Based on the specified relations of sub-features to the abstract features in a background knowledge base the sentiment of abstract feature are calculated.

As a general conceptual solution, we propose to parse the text to collect features, names, name phrases and other parts which constitute the features. We split each review into sentences and then parse each sentence to extract the feature(s) it contains. For knowledge-based feature extraction we propose to

[3] Review Example from Amazon Online Store `http://www.amazon.com/Canon-PowerShot-A2500-Stabilized-2-7-Inch/dp/B00B5HE2UG/`

use a knowledge-based feature annotation that can recognize names of concepts or entities have been mentioned in the text. Using knowledge-based resource annotation systems like DBpedia Spotlight[4] or AlchemyAPI[5] it is possible to collect the target features from the review text. Such entity annotation system can be used with a knowledge base specially made for the application domain.

Knowledge-based feature annotation and feature preparation system can extract from the given example features like: *"best shooting settings", "lightning", "good result", "shoot simplicity", "16.0 Megapixels", "DIGIC 4 Image Processor" and "faster warm-up times".*

Annotation is a task of adding more information to an existing object like text, image and video. The major advantage of using semantic annotation is that we can relate the entities to their knowledge base resources so that we can extract background knowledge about them. As a general conceptual solution, the set of extracted features from the feature extraction task is enriched and extended using entity recognition and ontological reasoning. The feature enrichment process is realized using a knowledge-based annotator. The examples of such features and their knowledge base types are shown in Table 2.

Table 2. Examples of Features and their Knowledge Base Types

Enriched Features	**Knowledge Base Types**
Shoot Setting	camera-onto:Camera_Setting
Megapixels	camera-onto:Image_Quality
Eco mode	camera-onto:Camera_Shooting_Mode
DIGIC 4	camera-onto:Image_Processor
DIGIC 5	camera-onto:Image_Processor

We propose to start with a set of ontological relationships that can be used to extract further knowledge resources like equivalence, direct hypernyms and direct hyponyms. This list can be extended with additional relationships depending on the structure of the ontology in use and on its granularity. The sentiment value of each resource can be computed based on the sentiment of the related sub-feature. We propose the following correspondences for ontological relationships:

1. equivalence = the same sentiment value is given to the sub-feature
2. hyperonymy = a factor to be applied to the sentiment value of sub-feature
3. hyponymy = a factor to be applied to the sentiment value of sub-feature

These factors should be specified manually in the ontology by the domain experts who are familiar with the relations of sub-features to abstract features. The ontology should include the knowledge required about the application domain, e.g., in our example it should conceptualize the camera concept and photography world in general so that one can extract the related concepts, e.g., for the *"Night Photography"*.

As an example for abstract features sentiment calculation, we consider the calculation for the *night photography and kids photography.* By inferencing on an

[4] `http://spotlight.dbpedia.org./`
[5] `http://www.alchemyapi.com`

ontology about the relations of sub-features to each other and to abstract features, we can calculate different affecting factors that can be used for abstract features sentiment calculation. We extract sentiments of related sub-features in the whole product corpus. For example, we use the subsequent sentiment calculation of the abstract features *"Night Photography" and "Kids Photography"*.
$S_{NightPhotography} = 0.1{*}S_{Flash} + 0.3{*}S_{Lens} + 0.4{*}S_{ImageProcessor} + 0.2{*}S_{Sensors}$
$S_{KidsPhotography} = 0.1{*}S_{Shutter} + 0.5{*}S_{ImageProcessor} + 0.3{*}S_{Sensors} + 0.1{*}S_{Flash}$

In the above example the sentiment factors (e.g., 0.1, 0.5) of sub-features are extracted by using an ontology that include the relations between abstract features and sub-features. Our approach depends highly on the existence of an ontology that can describe relations between features and can be used for inferencing on feature relations.

3 Conclusion and Future Work

Our main research question in this research was *"To which extent is it possible to use ontological background knowledge to derive abstract upper-level features based on more technical sub-features"?*. To answer this question, we structured the solution into three main tasks and from each task we tackled a number of sub-tasks. The first task extracts features from reviews using natural language processing tools. The second task extends features collected based on entity recognition and ontological reasoning. The third task finds relations between features and maps sub-features into related abstract features.

We have been working on approaches for recognition of features relevant to application domain and extraction of relations between sub-features to their related abstract-features.

Our future work is to specify details of background knowledge usage in the process of feature extraction, e.g., the reasoning on background knowledge can help to understand about the features that are not explicitly connected to the abstract features in the ontology. We also need to find methods to relate and evaluate specific features to more abstract ones. Furthermore, we have to evaluate our approach on a larger corpus using a domain ontology.

References

1. S. Kim, J. Zhang, Z. Chen, A. H. Oh, and S. Liu, "A hierarchical aspect-sentiment model for online reviews," in *AAAI* (M. desJardins and M. L. Littman, eds.), AAAI Press, 2013.
2. Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, (New York, NY, USA), pp. 815–824, ACM, 2011.

Sentiment Analysis for Dynamic User Preference Inference in Spoken Dialogue Systems

Yves Vanrompay[1], Mario Cataldi[2], Marine Le Glouanec[1], Marie-Aude Aufaure[1], and Myriam Lamolle[2]

[1] MAS Laboratory, Ecole Centrale Paris
Grande Voie des Vignes,
F-92 295 Chatenay-Malabry Cedex, France
[2] Université Paris 8, Saint-Denis, France {yves.vanrompay,marine.le-glouanec, marie-aude.aufaure}@ecp.fr,{m.cataldi,m.lamolle}@iut.univ-paris8.fr

Abstract. Many current spoken dialogue systems for search are domain-specific and do not take into account the preferences of the user and his opinion about the proposed items. In order to provide a more personalized answer, tailored to the user needs, in this paper we propose a spoken dialogue system where user interests are expressed as scores in modular ontologies and his sentiment about the system propositions is considered. This approach allows for a dynamic and evolving representation of user interests. In fact, in order to improve the performance of the detection mechanism of users preferences, we propose a hybrid model which also makes use of a sentiment analysis module to detect the opinion of the user with respect to the proposition of the system. This allows the system to leverage the degree of user satisfaction and improve the overall recommendation mechanism being more precise about the expressed user interest. An evaluation on a representative set of dialogues is presented and highlights both the validity and the reliability of the proposed preference inference mechanism.

1 Introduction

The traditional goal of spoken dialogue system is to approach human performance in conversational interaction, specifically in terms of the interactional skills needed to do so. With this objective, as an attempt to enhance human-computer interaction, in the PARLANCE project[3], we build a system for interactive, *personalized*, *hyper-local* search for open domains such as restaurant search and tourist information. Current search engines work well only if the user has a single search goal and does not have multiple trade-offs to explore. For example, standard search works well if you want to know the phone number of a specific business but poorly if you are looking for a restaurant with several different search criteria of varying importance, e.g. food type versus location versus price etc. The latter requires the user to collaborate conversationally over several turns. In order to provide a personalized answer, tailored to the specific user, in

[3] https://sites.google.com/site/parlanceprojectofficial/

this work we focused on three levels: (1) user modelling in terms of preferences and interests inferred from past interactions with the system, (2) personalization of search approach and (3) the opinion of the user himself about the suggested products.

Personalization in the context of spoken dialogue system has slowly progressed compared to the field of non-natural language systems. Indeed, current spoken dialogue systems are mostly domain-specific, using rather static information from experts and knowledge bases. In PARLANCE, we choose to represent the dynamic domain knowledge through modular ontologies, where each ontology module represents a domain and can be dynamically loaded at run-time to meet the current needs of the user. In order to provide a semantic representation of the user model, the concepts and attributes in these ontology modules are annotated with scores representing the preferences and interests of the user. This allows us to learn the specificities of a user, and give responses that fit the user's profile. Additionally, user opinions can be detected where system recommendations are made. The sentiment analysis of the user answers permits to detect the degree of user satisfaction to systems replies and provide more tailored recommendations in the future.

This paper is organized as follows: Section 2 presents related work on preference systems. Section 3 motivates our research and provides an overview of the different components in the system. Section 4 introduces the representation of user interests while Section 5 describes how to detect and exploit user opinions about the recommended items. Then, we provide our experimental evaluation. Finally, section 7 concludes the paper.

2 Related Work

The automatic detection and categorization of sentiment expressed in natural language constitutes a remarkable research challenge, and has generated a nexus of techniques and approaches. Recently, [1] and [2] have conducted extensive surveys of the open challenges in this research area. Most of the works in this area focus on the categorization of *overall* sentiment, capturing a form of average polarity at the document level [3, 4].

A number of researchers have applied natural language processing (NLP) techniques to detect features in small chunks of text [5, 6]. Some methods make use of a lexical approach to focus on subjective words of the considered dialogue, namely adjectives and adverbs[7]. The study of linguistic complements, negations, syntactic dependencies, etc., can also be positively used to categorise the terms in the dialogue [8, 9].

Various API are available for sentiment analysis purposes. Most of them are built for study of specific environments, like social networks and/or blogs, and they try to focus on particular subjects, enterprises or events. Among all, SentiWordnet [10] analyses relationships like antonymy, hyponymy and/or hyperonymy to give a triplet value (objective, positive and negative) to each term of the considered phrase/document.

3 PARLANCE Architecture

This section starts with an example dialogue between the user and the system motivating the need for personalized interaction. Let us suppose that Paul is visiting San Francisco and he is looking for a good restaurant in the center of the city. An example dialogue between Paul (U) and the system (S) can be as follows:

1. U: I'm looking for a moderately priced restaurant near the centre of town.
2. S: I have found 2 Indian and 1 Chinese restaurant. Star India is a moderately priced restaurant in the centre of town. It has free Wifi.
3. U: Ok, what's the address?
4. S: The address is 3721 Geary Ave.
5. U: Oh, I don't want that one, it's hard to park on that street. Are there any other Indian restaurants?

In this conversation, not only information on user interests learned from past dialogues is exploited, but also new information on interests and preferences is learned. In item 2, the system has learned from previous conversations that the user is quite fond of Indian and Chinese restaurants. Therefore, it starts with proposing restaurants with these food types. The system also knows that the availability of Wifi is important to the user, so it proactively gives this information. In item 5, as the user asks for an alternative restaurant, user preferences are updated: the user does not like Geary Avenue.

Considering this scenario, the starting point for inferring user interests are abstract representations of the history of dialogues (dialogue act units) between the user and the system. A user model manager analyses the dialogue history of the user and derives interest scores associated to concepts, attribute types and attribute values in a weighted ontology module corresponding to a specific domain.

4 Evolving User Preferences as Weighted Modular Ontologies

User preferences regarding concepts and attributes are inferred from the user's dialogue history. All user and system utterances from past dialogues are saved in a so called dialogue act unit (DAU), which is the abstract representation of utterances in PARLANCE. For example, when the user asks for the price range of a restaurant, this is represented as the DAU *request(price)*. Interest scores are derived from logged traces of DAUs. We keep track of the positive and negative occurrences of attribute values and concrete instances. These frequencies allow us to rank the different elements. Based on this ranking it is decided which system response is best suited with regards to the user interests. If the user often queries for pricing information with given attribute value "cheap" in searching for a restaurant, the value of the *price* attribute will have a high frequency, and the system will proactively inform the user about the price in its answers, and lead the system to recommend restaurants from a cheaper price range.

Our mechanism for expressing user interests is integrated in the approach which represents information as (hierarchical) modular ontologies. Ontology modularization is defined as a way to structure ontologies, so that large domain ontologies will be the aggregation of self-contained, independent and reusable knowledge components (considered as Ontology module (OM)). An OM can be seen as an ontology fragment that has a meaning from the viewpoint of applications or users. Each ontology module corresponds to a particular domain and its size should be small for easy maintenance. Each OM is characterized by a basic concept, called the *pivotal concept*. A tourism ontology can contain several ontology modules like lodging, transportation and restaurant information. Taking the restaurant ontology module, this contains (amongst others) the restaurant concept with attribute types name, food type, dress code and location. To personalize the responses given to the user, our user model incorporates the interests and preferences of the user by assigning scores to elements in the appropriate ontology modules. These scores are updated according to what is being learned from the history of past dialogues, ensuring that the interests evolve as user preferences may change through time. The weights are useful in two different aspects. First, the scores are used to rank and recommend concrete instances that are of interest to users. Second, attribute value scores are used to generate a system response tailored to the user needs. For example, based on the interest scores, the system can decide to proactively inform the user on the food type of the restaurant, but not on the dress code.

4.1 Evolving update of the User Model

Calculation of scores happens offline based on all available dialogues in the dialogue history component. This means that the user model contains both the representation of user preferences (as weights in modular ontologies), as the mechanism to calculate the scores. To update scores based on recent dialogues of the user with the system, the user model manager aims to the recalculation of scores on a regular basis. The scores are thus **dynamic**, within the modular ontology structure which itself is relatively static. The scores of the attribute values are relative and sum up to 1. If an attribute type has m possible values, the initial score w_i for each value will be $\frac{1}{m}$. The score w_i is updated by counting how often attribute value $attval_i$ was selected and dividing this number by the total number of times a value for the corresponding attribute type was specified by the user. This means however that the past is as important as the present. In our context, user interests will typically evolve over time. So, if the dialogue history includes for instance the dialogues of the user during the last six months, it is reasonable to have more recent dialogues having relatively more influence on the user interest model than older ones. To this end, the scores for attribute values should be updated in such a way that recent dialogues have more relevance than older ones, which we do using an exponential smoothing method as follows: $w_i = \alpha \times x_j + (1 - \alpha)w'_i$ where w'_i represents the old score and $x_j \in \{0, 1\}$ is the value for the choice taken at moment j in the dialogue history. If $x_j = 1$ then $attval_i$ was specified by the user, if $x_j = 0$ it was not.

Using this method, the sum of all scores of the attribute values belonging to an attribute type remains 1, and the scores represent the relative importance of each attribute value. The *learning rate* $\alpha \in [0, 1]$ is a real number that controls how important recent observations are compared to older ones.

5 Mining User's Opinions wrt System Recommendations

This component is responsible for analyzing the positive, neutral, or negative opinions produced from user with respect to the propositions of the system. In order to tackle this issue, we make use of a novel feature-based polarity analysis technique[11], which combines statistical techniques with natural language processing. As in literature, we define a polarity as a real number that quantifies the user's positive, neutral, or negative opinion about a feature.

With this goal, for each dialog (intended as the complete set of system-user interaction), we model the user's sentiment with respect to the proposition of the user by estimating the degree of positivity/negativity with respect to the considered features. To extract such fine-grained sentiment information from raw text, we model each review as a set of sentences. A sentence is then formalized as a syntactic dependency graph, used to analyze the semantic and syntactic dependencies between its terms, and identify the terms referring to features. More formally, a sentence S can be formalize as an ordered vector of terms $S = \{w_0, w_1...w_m\}$, where the order represents the original position of each term within the sentence. The sentence s can be represented as a dependency graph G. The dependency graph is a labeled directed graph $G = (V, E, l)$, where V is the set of nodes representing the lexical elements w_i and E the set of edges (i.e. dependency relations) among the nodes.

The graph is obtained through a preliminary POS tagging phase, achieved by training a tagging model on the annotated corpus proposed by [12] and therefore by calculating the probability $p(t_j|w_i)$ of assigning a tag t_j to the term w_i using a maximum-likelihood estimation as in [13].

Subsequently, the dependency graphs are then utilized to detect the terms referring to a feature, which expresses some non-neutral opinion, including compound expressions, e.g. "the restaurant serves a very good pizza." In this phase, a SentiWordNet-like approach[10], which attributes polarity values to each WordNet synset, is used as a source of polarity values. In detail, using the synset graph proposed by WordNet, we calculate the polarities of each term by using a two-step algorithm. A first step is a semi-supervised learning step in which polarity values are assigned to two sets of *seed nodes*. This set consists of two subsets; one subset of "paradigmatically positive" synsets and another one consisting of "paradigmatically negative" synsets [14]. The polarities are then propagated automatically to other synsets of the WordNet graph by traversing selected semantic relations. This propagation is performed within the minimal radius that guarantees no conflicts among the relations, that is, until a node labeled as positive points to a node already linked to some negative seed, or vice-versa. In other words, we only propagate the polarities to the nodes that are univocally connected to a positive or a negative seed. Second, a random-walk step is executed on the whole WordNet graph starting from the seed nodes, and iteratively propagates the positive and negative polarity to all of the synsets. This approach preserves or inverts the polarity of each node based on the number of positive and negative relations that connect it to the seeds. The process ends when a convergence condition is reached. This condition is satisfied when all the nodes have maintained the same polarity sign (positive or negative) after two consecutive steps. Finally, the polarities of terms are aggregated into single values, each one referring to a specific feature.

6 Evaluation

The evaluation of our approach consists in two different analysis; from one side, capturing the evolution of the user preference scores in time, i.e. with respect to the size of the dialogue history or the number of interactions of the user with the system, and from the other side, study the user feedback expressed in the dialog with the system.

6.1 Analysing the evolution of the user preference scores

By using the fading factor α in our interest score update formula, we make sure that more recent dialogues provide more information on interests than the older ones, leading to an evolution of interests. To obtain real spoken dialogues, we used Amazon Mechanical Turk, which is a tool for crowd-sourcing where users can call a toll-free number, solve tasks assigned to them, and earn money. First, a number of well defined tasks, expressed in natural language, were constructed. As an example, we ask a user to find an Italian restaurant in the center of town. As a check for task success, the user has to give in the phone number of the restaurant he has found. This means the test users have every incentive to succeed, since they are only paid in case of task success. We varied the content of the tasks as to reflect changing user interests over time. The basic task was in each case to find a restaurant with variations in the attribute types *food*, *area* and *price*. The experiments are based on 60 dialogues, and we set α to be 0.1. In analyzing the evolution of the interest scores, we plot the maximum score P across attribute values for each attribute type as a function of the number of dialogues considered, as can be seen in Figure 1. Indeed, the score P serves as a basic metric for showing how outspoken the user interests are at a certain moment in time. A low P signifies that the user changes his interest rather quick, while a high P means that in recent dialogues he has shown a rather consistent and stable choice behavior.

Figure 1 shows the evolution of the values of P over the dialogues for different attribute types. It can be noticed that there is a correlation between the "peaks" in the graph for *food* and *area*, reflecting that when the user is in a stable period with respect to his interests, this holds for some attribute types. The evolution of the graph allows us to identify periods in which the user very dynamically changes his interests, and periods in which his choices remain merely stable. Also, by looking at the relative values of P for the different attribute types at a given moment, it is possible to make a ranking of those types where the user does not change his behavior a lot, compared to the ones that are more fluctuating.

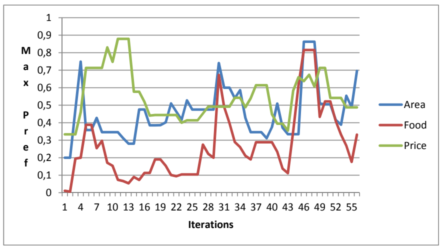

Fig. 1. Evolution of user interests in function of dialogue history

6.2 Dissecting user opinion with respect to System recommendations

In order to assess the proposed approach to sentiment analysis, we analysed the detected sentiments in the considered corpus of dialogs. The aim of this analysis is to study, from one side, how the users interact with the system and, on the second side, analyse the retrieved opinion with respect to the recommended items. For this, we considered three features: food type, area and price. Figure 2 shows the obtained polarities.

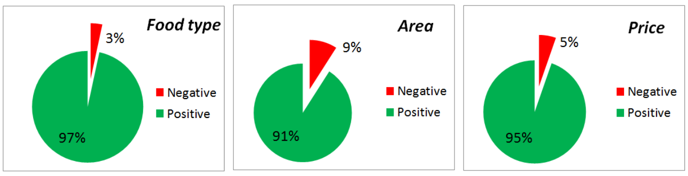

Fig. 2. Statistics about the performed feature-based polarity on user dialogs.

These results highlight, for all the considered features, that the user express more likely positive opinions rather than negative ones. In a sense, the users accept, in great majority of the cases (up to 97% of the cases regarding the food type) the recommendations of the system, explicitly proving the goodness of the whole recommender system and the user preference model. In fact, the previously collected preferences lead the system to recommend an item that perfectly match the real user preferences.

Notice that this feedback value can be positively use to dynamically tune the *learning rate* α (explained in Section 4.1). In fact, a negative opinion can suggest a change of the user preference wrt the attended one. Thus, we can dynamically set the learning rate, which express how important recent observations can be compared to older ones, proportionally to the negative sentiment expressed by the user in order to reflect a sort of user preference change. Following this idea, the more negative the sentiment expressed by the user, the higher the learning rate which will reflect the necessity of the system to quickly update the user preferences.

7 Conclusion

In this paper we described an approach for modeling the dynamics of user preferences in a spoken dialogue system for searching items of interest. A fading method allows for the interests to keep track of the evolution of user behavior. We then leverage the opinion of the user about the suggested items to improve the recommender system and tune the learning rate of the system. We evaluated our approach on a set of real dialogues and showed it can provide useful insights into changing interests. The ontology-based representation of interests lets us tailor recommendation of items to the recent preferences the user has exhibited, for each search domain involved.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2) (2008) 1–135
2. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1) (2012) 1–167
3. Beineke, P., Hastie, T., Manning, C., Vaithyanathan, S.: Exploring Sentiment Summarization. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004) 1–4
4. Hiroshi, K., Tetsuya, N., Hideo, W.: Deeper Sentiment Analysis Using Machine Translation Technology. In: Proceedings of COLING '04. (2004) 1–7
5. Holz, F., Teresniak, S.: Towards Automatic Detection and Tracking of Topic Change. In: Computational Linguistics and Intelligent Text Processing. Volume 6008. (2010) 327–339
6. Missen, M., Boughanem, M., Cabanac, G.: Opinion Mining: Reviewed from Word to Document Level. (2012) 1–19
7. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the ACL '02. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 417–424
8. Benamara, F., Chardon, B., Mathieu, Y.Y., Popescu, V.: Towards context-based subjectivity analysis. In: IJCNLP. (2011) 1180–1188
9. Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? finding strong and weak opinion clauses. In: Proceedings of the AAAI'04, AAAI Press (2004) 761–767
10. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of LREC'10. (2010) 2200–2204
11. Cataldi, M., Ballatore, A., Tiddi, I., Aufaure, M.A.: Good location, terrible food: detecting feature sentiment in user-generated reviews. Social Netw. Analys. Mining **3**(4) (2013) 1149–1163
12. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330
13. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. ACL '03 (2003) 423–430
14. Turney, P.D., Littman, M.L.: Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Trans. Inf. Syst. **21**(4) (October 2003) 315–346

Semantic Web-based Sentiment Analysis

Diego Reforgiato Recupero[1], Sergio Consoli[1], Aldo Gangemi[12], Andrea Giovanni Nuzzolese[13], Valentina Presutti[1], and Daria Spampinato[1]

[1] National Research Council (CNR), Institute of Cognitive Sciences and Technologies, Semantic Technology Laboratory, Via Gaifami 18, 95028 Catania (CT)
[2] LIPN, University Paris 13, Sorbone Cit'e, UMR CNRS, France
[3] Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, 40127 Bologna (BO), Italy
{diego.reforgiato,sergio.consoli,aldo.gangemi, andrea.nuzzolese,valentina.presutti,daria.spampinato}@istc.cnr.it

Abstract. [1] The introduction of semantics in Sentiment Analysis research has proved to bring several benefits for what performances are concerned and has allowed to identify new challenging tasks to be accomplished. Semantics helps structuring the plain natural language text with formal representation. The current system we are developing performs sentiment analysis by hybridizing natural language processing techniques with Semantic Web technologies. Our system, called Sentilo, is able to recognize the holder of an opinion, to detect the topics and sub-topics in its scope, and to measure the sentiment expressed by them. This information is formally represented by means of RDF graphs according to an OWL opinion ontology, while holders and topics identity is resolved on the Linked Open Data cloud.

Keywords: Sentiment Analysis, Sentic Computing, Semantic features

1 Introduction

Sentiment Analysis (SA) is one of the hottest problems currently studied in Natural Language Processing (NLP), and recently it has entered the Semantic Web world: [16] provides evidence that including semantic features to SA algorithms improves their performance. However, existing approaches at SA, even those that include semantic features, are basically supervised and rely on the availability of manually annotated samples, hence they are usually domain-dependent. Semantic sentiment analysis can take advantage from linked data, ontologies, controlled vocabularies, and lexical resources (e.g. DBpedia, YAGO, ConceptNet [13], SenticNet [4], Nell [11], OIE [7], etc.), which help aggregating the conceptual and affective information associated with natural language opinions.

Combining NLP and Semantic Web approaches could provide us with the flexibility of language processing techniques, as well as with the depth of semantic

[1] This work is supported by the project PRISMA SMART CITIES, funded by the Italian Ministry of Research and Education under the program Programma Operativo Nazionale

knowledge bases, through which also sentiments that are expressed in a subtle manner can be detected, as in the case of concepts that do not explicitly convey any emotion, but which are implicitly linked to other concepts that do so. What is challenging is the way those techniques can be used and combined to yield highly performant systems. With semantics, we can expand the current state of the art in sentiment analysis to track, correlate, and compare sentiment of specific entities or group of related entities over time and across different contexts.

Another common aspect of most existing SA methods is that they neglect the identification of holders and topics of an opinion as a task per se. In fact, they mainly focus on interpreting the tone of a sentence by identifying terms that carry a particular sentiment polarity; it has been demonstrated that including topic detection in models used by algorithms for SA improves their results [2, 12, 17]. However, in such approaches, the SA task melts with the topic detection task, which is never evaluated separately.

Sentilo[2] is a semantic SA system introduced in [10] that analyses the sentiment of a sentence: it identifies the holder of an opinion, the topics and sub-topics of that opinion, the sentiment expressed in each of them by the holder, as well as the sentiment of the overall sentence. Topics, holders, and sentiments are represented as RDF graphs compliant to an OWL ontology [15] described in [10], while topics and holders are resolved on the Linked Open Data cloud in order to aggregate sentiments expressed on the same topic in different contexts or from different sources.

2 Analyzing Opinions

Sentilo implementation is inspired by Davidsons view [6]: events and situations are primary objects for the representation of a domain. Based on this view of the world, sentences are represented as linked events or situations, with participating objects. We use DOLCE+DnS [8, 9] as a vocabulary for events and situations, and VerbNet [14] as reference for thematic roles of events. Based on this rationale, we distinguish main topics from sub-topics of an opinion. The distinction between topics and subtopics, as well as the event- and situation-based representation of opinions, impacts on the strategy used for computing the sentiment scores of individual topics and of the whole sentences. To compute sentiment scores we rely on two resources: Sentic.net [5, 3], a publicly available resource that provides polarized scores of concepts, and SentiWordNet [1], a lexical resource for opinion mining. Given an entity, identified as a topic of an opinion (either a main or sub-topic), we compute its sentiment score by combining the scores of its associated opinion features, which are extracted from the RDF graph representing the opinionated sentence. If the topic participates in an event or a situation occurrence, we say that such occurrence provides a context to it, and affects its sentiment score.

We also want to tackle issues contained in sentences like the following: *"John is happy because President Alvarez was arrested"*. For such a sentence, a common

[2] http://wit.istc.cnr.it/stlab-tools/sentilo/service

reader would understand a positive emotion for John as he is happy and a negative event (not opinion as that would depend on the context) for the President Alvarez as he was arrested. A careful reader however would also consider John as the holder of a negative opinion for the President Alvarez as John is having a positive reaction to a bad event happened to President Alvarez. To this aim we introduce the concepts of *Role sensitivity* and *Factual impact*. These concepts have been the basis for the design of a novel resource of annotated verbs, named *SentiloNet*. A role is sensitive with respect to an event (referred to by a verb) if it is indirectly affected by an opinion directly expressed on the event. As far as the annotation of a verb (frame) is concerned, the *sensitivity* is an attribute of its thematic roles. The value of the sensitivity attribute of a role with respect to a verb can be either *true* or *false*, meaning that the role is sensitive or is not, respectively. Factual impact indicates that an event (referred to by a verb) has either a positive or negative impact on a specific role. As far as the annotation of a verb is concerned, the *factual impact* is an attribute of its sensitive roles. The value of this attribute for a role can be *positive*, *negative*, meaning that the inherited opinion will keep its polarity or change it, respectively. The current version of SentiloNet includes 1,100 annotated verbs. Given the high number of different thematic roles of verbs, we have devised a heuristics that allowed us to manually annotate a good amount of verbs in a rather limited amount of time. SentiloNet indicates, for 1,100 verbs, the value of *sensitivity* and *factual impact* attributes for each class of roles.

Sentilo sentiment score $sc_{Sentilo}$ of a topic t can be defined as a function f taking the following arguments:

$$sc_{Sentilo}(t) = f(\sum_{i=0}^{n} sc(q_i(t)), \sum_{i=0}^{m} sc(type_i(t)), truth(t), sc(trig), sc(ctx(t)), mod(t))$$

- *sc(x)* is the score of x as provided by Sentic.net or SentiWordNet;
- q_i is the object value of a triple `t dul:hasQuality q`$_i$. Such triples represent the opinion features, i.e. adjectives and adverbs, associated with entities composing the opinion sentence;
- *type*$_i$*(t)* is the type of t expressed in the RDF graph by means of `rdf:type` triples;
- *truth(t)* is a truth value associated with an entity in the graph, typically an event or situation occurrence, or a quality. If its value is *false* it means that the entity is negated. E.g. in a sentence such as *"John is not a good guy"*, a RDF triple `situation_1 boxing:hasTruthValue fred:False` would be included in the graph, and its effect would be to change the sign of the sentiment score assigned to the feature *good*;
- *trig* is the opinion trigger verb;
- *ctx(t)* is the context of t, if any. It can be either a situation or an event to which t participates in;
- *mod(t)* is the modality of the verb t, if any. E.g. in a sentence such as *I would like a dog*, an RDF relationship `fred:like_1 boxing:hasModality boxing:Necessary` would be included.

3 Conclusions

In this paper we have given our view on SA and shown an example with Sentilo, a semantic SA system that we are currently developing. Sentilo is able to analyses the sentiment of a sentence, identify holders, topics and subtopics. As future direction we are designing a sentiment scoring algorithm that takes into account all the semantics information provided by Sentilo in order to correctly propagate the scores from topics/sub-topics to situations/events and viceversa.

References

1. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 2010.
2. K. Cai, W. S. Spangler, Y. Chen, and L. Z. 0007. Leveraging Sentiment Analysis for Topic Detection. *Web Intelligence and Agent Systems*, 8(3):291–302, 2010.
3. E. Cambria. Senticnet. `http://sentic.net`, Dec. 2013.
4. E. Cambria, C. Havasi, and A. Hussain. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In G. M. Youngblood and P. M. McCarthy, editors, *FLAIRS Conference*, pages 202–207. AAAI Press, 2012.
5. E. Cambria and A. Hussain. *Sentic computing: Techniques, tools, and applications*, volume 2. Springer, 2012.
6. D. Davidson. The Logical Form of Action Sentences. In N. Rescher, editor, *The Logic of Decision and Action*, pages 81–120. University of Pittsburgh Press, 1967.
7. O. I. Extraction. Artificial intelligence - university of washington. `http://ai.cs.washington.edu/projects/open-information-extraction`.
8. A. Gangemi. What's in a Schema? In C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prevot, editors, *Ontology and the Lexicon*, pages 144–182. Cambridge University Press, 2010.
9. A. Gangemi. Dolce Ultra Lite Ontology. `http://ontologydesignpatterns.org/ont/dul/DUL.owl`, Dec. 2013.
10. A. Gangemi, V. Presutti, and D. Reforgiato. Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Computational Intelligence Magazine*, 2014.
11. N.-E. L. Learning. Research project at carnegie mellon university. `http://rtw.ml.cmu.edu/rtw/`, 2010.
12. C. Lin, Y. He, R. Everson, and S. M. Rüger. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Trans. Knowl. Data Eng.*, 24(6):1134–1145, 2012.
13. H. Liu and P. Singh. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, 22(4):211–226, 2004.
14. M. Palmer. The VerbNet Project. `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`, Dec. 2013.
15. D. Reforgiato, A. G. Nuzzolese, S. Consoli, A. Gangemi, and V. Presutti. Sentilo. `http://wit.istc.cnr.it/stlab-tools/sentilo/service`, Dec. 2013.
16. H. Saif, Y. He, and H. Alani. Semantic Sentiment Analysis of Twitter. pages 508–524, Boston, UA, 2012. Springer.
17. I. Titov and R. McDonald. Modeling Online Reviews with Multi-grain Topic Models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA, 2008. ACM.

Preface

Emergencies require significant effort in order for emergency workers and the general public to respond effectively. Emergency Responders must rapidly gather information, determine where to deploy resources and make prioritization decisions regarding how best to deal with the emergency. Good situation awareness [1] is therefore paramount to ensure a timely and effective response. Thus, for an incident to be dealt with effectively, citizens and responders must be able to share reliable information and help build an understanding of the current local and global situation and how this may evolve over time [2]. Information available on Social Media is increasingly becoming a fundamental source for situation awareness. During a crisis, citizens share their own experiences, feelings and often, critical local knowledge. Integrating this information with Linked Data, such as geographic or demographic data, could greatly enrich its value to better prevent and respond to disasters and crises. Analysing, modelling and integrating social media content and Linked Data presents significant technical as well as social challenges. Social data is: (i) high in volume, rapidly changing and constantly increasing, (ii) often duplicated, incomplete, imprecise and potentially incorrect; (iii) textual content may be written in informal style (i.e., short, unedited and conversational), thus much less grammatically bounded and containing extensive use of shorthand, symbols (e.g., emoticons), misspellings etc.; (iv) generally concerning the short-term zeitgeist; and (v) covering every conceivable domain. For this the workshop on Social Media and Linked Data for Emergency Response (SMILE2014) called for papers on innovative approaches for exploitation of social media and Linked Data for emergency response and crisis management using semantic web technologies. To address this the proceedings include papers ranging from vehicle routing to crowd-sourcing, social media mining and visualization.

January 2015

SMILE2014 Chairs

Vitaveska Lanfranchi
Tomi Kauppinen

References

[1] Wong, W. and Blandford, A. Describing Situation Awareness at an Emergency Medical Dispatch Centre. In Proc. Human Factors and Ergonomics Societys 48th Annual Meeting. Santa Monica, CA: HFES. 285–289. (2004)

[2] Endsley, Mica R Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society 37.1, (1995). pp. 32–64.

Towards emergency vehicle routing using Geolinked Open Data: the case study of the Municipality of Catania*†

Sergio Consoli[1,2], Aldo Gangemi[1,3], Andrea Giovanni Nuzzolese[1,2], Silvio Peroni[1,2], Valentina Presutti[1], Diego Reforgiato Recupero[1], and Daria Spampinato[1]

[1] National Research Council (CNR), Institute of Cognitive Sciences and Technologies, Semantic Technology Laboratory, ITALY ✉ `{firstname.lastname}@istc.cnr.it`
[2] Department of Computer Science, University of Bologna, ITALY
✉ `{nuzzoles,speroni}@cs.unibo.it`
[3] LIPN, University Paris 13, Sorbone Cité, UMR CNRS, FRANCE
✉ `aldo.gangemi@lipn.univ-paris13.fr`

Abstract. Linked Open Data (LOD) has gained significant momentum over the past years as a best practice of promoting the sharing and publication of structured data on the semantic Web. Currently LOD is reaching significant adoption also in Public Administrations (PAs), where it is often required to be connected to existing platforms, such as GIS-based data management systems. Bearing on previous experience with the pioneering data.cnr.it, through Semantic Scout, as well as the Agency for Digital Italy recommendations for LOD in Italian PA, we are working on the extraction, publication, and exploitation of data from the Geographic Information System of the Municipality of Catania, referred to as SIT ("Sistema Informativo Territoriale"). The goal is to boost the metropolis towards the route of a modern Smart City by providing prototype integrated solutions supporting transport, public health, urban decor, and social services, to improve urban life. In particular a mobile application focused on real-time road traffic and public transport management is currently under development to support sustainable mobility and, especially, to aid the response to urban emergencies, from small accidents to more serious disasters. This paper describes the results and lessons learnt from the first work campaign, aiming at analyzing, reengineering, linking, and formalizing the Shape-based geo-data from the SIT.

Keywords: Geo-Linked Open Data applications; Linked eGovernment Data extraction and publication; sustainable mobility; emergency vehicle routing.

---

*This work has been supported by the PON R&C project PRISMA, "PiattafoRme cloud Interoperabili per SMArt-government", ref. PON04a2 A Smart Cities, under the National Operational Programme for Research and Competitiveness 2007-2013.

†Authors are listed in alphabetical order as their contributions are equally distributed.

1 Preliminary discussion

In a currently on-going project we are investigating the extraction, enrichment, publication and reuse of Linked Open Data (LOD) [1, 2] for the Municipality of Catania (MoC), Italy, by means of the application of latest semantic technologies and software components [3]. The main motivation of the work consists of experimenting social eGovernment systems aimed at optimizing the performance of the Public Administration (PA) of the MoC for the provision of intelligent ICT services to citizens and businesses, supporting the external evaluation of the PA by the detection of the community trust. The work falls within the spirit of the Smart Cities initiatives of the European Commission, which aims at bringing together cities, industry and citizens to improve urban life through more sustainable integrated solutions. Although the methodology has been designed for the case study of the city hall of Catania, the approach is completely generalizable and can be replicated to any PA worldwide. One of the main development objectives of the project consists in conceiving, designing and prototyping applications for the MoC related to certain areas of experimentation, such as online social services and health, traffic management and transport, and urban decor. With the aim of detecting and collecting the required data and processes for these applications, meetings with the Leadership of the Directorate of Information Systems Service of the MoC were carried out.

A particular field of experimentation is specially focused on the management of mobility, i.e. road traffic and public transport. Within this context, the scenario has identified the development of a prototype mobile application implementing a real-time system to inform on the state of roads in urban areas to support sustainable mobility and, in particular, to aid the response to urban emergencies, from small scale accidents to more serious disasters. The system aims at connecting drivers to one another, helping people create local driving communities that work together to improve the quality of everyone's daily driving. That might mean helping them avoid the frustration of sitting in traffic, advising them on unexpected accidents or other traps, or just shaving five minutes off of their regular commute by showing them new routes they never even knew about. But most importantly, the application may have any extremely important role on emergency logistics. Response to an emergency incident requires careful planning and professional execution of plans, when and if an emergency occurs [4]. During these events there is the need to find rapidly the nearest hospitals, or to obtain the best way outs from the emergency zones, or to produce the optimal path connecting two suburbs for redirecting the road traffic, etc. Technically, this system should be able to locate the best path between source and destination not only in a static environment, but particularly in a dynamic one. That is, the user feedback serves at placing in the map some obstacles, or inaccessible zones, coming from accidents or emergency events, and the system responds in real-time producing the optimal path without these forbidden zones. After typing in their destination address, users just drive with the application open on their phone to passively contribute traffic and other road data, but they can also take a more active role by sharing road reports on accidents, advising on unexpected traps, or any other hazards along the way, helping to give other users in the area real-time information about what's to come [5]. For the realization of the app for our case study, it is necessary to process the data and diagrams in the Geographic Information System of the MoC, referred to as SIT: "Sistema Informativo Territoriale" [6]. Therefore it was decided, by mutual agreement with the chief officers and experts of the city hall of Catania, to process the data in order to make them open, interoperable and compatible with the principles of Linked Open Data.

The paper is structured as follows. Background on the state of the art on the use of LOD for PA, often referred to as Linked eGovernment Data [7], is reported in Section 2. Techniques and tools used to deal with LOD for the MoC are introduced in Section 3, while the extracted ontology is described in Section 3.3, along with the means used to consume the accessible data. Section 4 ends the paper with conclusions and the future research where we are directed.

2 Linked eGovernment Data

LOD are currently bootstrapping the Web of Data by converting into RDF and publishing existing datasets available to the general public under open licenses [1, 2]. LOD offers the possibility of using data across domains or organisations for purposes like statistics, analysis, maps and publications. These major changes in technology and society are involving also the way of doing politics, administration and the relationship between politicians, public servants and citizens. Transparency, participation and collaboration are the main issues of the integration of citizens in the paradigm of Open Government [8]. Because PAs have large amounts of data which could be made accessible for the purpose of the LOD movement, research on the opening process, data reengineering, linking, formalisation and consumption is of primary interest [9].

The Digital Administration Code incorporates a wide range of best-practices in the usage of Linked eGovernment Data, which can be synthetized as: portals for the supply of the Linked eGovernment Data sets; portals providing raw data sets of LOD for PAs along with technical tools or developer kits for understanding, interpreting, or processing the provided data; existing portals acting as showrooms for best practices for Linked eGovernment Data; mobile apps for smartphones using LOD for PAs [7].

The main thrust on the publication of LOD for PA is coming from big initiatives in the United States (`data.gov`) [10, 11] and the United Kingdom (`data.gov.uk`) [12], both providing thousands of raw sets of LOD within their portals, but there are also some other experiences and notable initiatives that are in line with the international state of the art. In Germany, one of the first examples for a LOD portal is the one from the state of Baden-Württemberg (`opendata.service-bw.de`), divided into three main parts: LOD, applications, and tools. In addition to their potentials, Linked eGovernment Data can provide great benefits in the matter of accountability, as shown in the LOD portal example of Kenya (`opendata.go.ke`).

In addition, LOD have been published in Italy by the city hall of Florence[4], Agency for Digital Italy[5], from the Piedmont region[6], the Chamber of Deputies[7]. Beside these initiatives, another notable for the Italian PA is "`data.cnr.it`" [13, 14], the open data project of the National Research Council (CNR), designed and maintained by the Semantic Technology Laboratory of ISTC-CNR, and shared with the unit Information Systems Office of CNR.

3 Extraction of Linked eGovernment Data for the MoC

In this section we present the methodology used for the extraction and publication of LOD for the Municipality of Catania. The methods are based on the standards of the W3C[8], on good international practices, on the guidelines issued by the Agency for Digital Italy [15, 16] and those by the Italian Index of Public Administration[9], as well as on the in-depth experience of the research participants on this field, in particular related to the development of the "`data.cnr.it`" [13, 14] portal.

3.1 Scenario analysis

During the phase of selection of the source data, a thorough analysis of the reference domain was made. Thanks to the close interaction with the PA experts of the MoC, the Geographic Information System, SIT [6], was identified as the source dataset for the enrichment and publication of data. The SIT is a data warehouse used for reporting and data analysis, and consisting of databases, hardware, software, and technicians, which manages, develops and integrates information of the province of Catania based on a geographical space [6]. The various territorial levels (hydrography, topography, buildings, infrastructure, technological networks, administrative boundaries and land, ...) form the geo-localised common part of the information flow of the MoC, according to which all the constituent parts are related to each other.

The SIT is designed to contain all the available data of the PA in Catania for the purpose of in-depth knowledge of the local area. Basically tit contains three types of data: *register base*, *registry office*, and *toponymy*, provided in the form of Shape-based files [17] for each data record, i.e. files with extensions: .dbf, .shp, .shx, .sbn, .sbx, .xml. Through the consultation platform on the web it is possible to display the following information: basic cartography; ortho-photos; road graph; buildings with a breakdown by main body of some areas of the city; cadastral sections; data from the 1991 and 2001 census of the population; last Master Plan; gas network on-going works; resident population in selected

[4] Available at: `http://opendata.comune.fi.it/linked_data.html`
[5] Available at: `http://www.digitpa.gov.it`
[6] Available at: `http://www.dati.piemonte.it/rdf.html`
[7] Available at: `http://dati.camera.it`
[8] Available at: `http://www.w3.org/standards/semanticweb/`
[9] Available at: `http://spcdata.digitpa.gov.it/data.html`

areas (municipalities, entire street, polygonal, circular area); total population, distributed into bow street, house number, etc; breakdown of the population by municipality, blocks, nationality, gender, family components, age, marital status, etc; extraction and search of resident persons, and their location on the bow streets; competence areas of pharmacies; location and alphanumeric information of: municipality, hospitals, universities, schools, pharmacies, post offices, areas or emergency, public safety, fire departments, public green areas, public community centres, institutions for minors and orphanages. The SIT also includes maps containing geo-referenced information related to: sub-services (electricity-gas-water pipes); data on stoppage areas; occupation stalls; stalls for disadvantaged people; occupation of public land; public transport fleet; management and working state of the fleet; data on lines and stops of public transport; accident traffic data; road signs and markings; maintenance state of roads and sidewalks; management of roadway construction; data of the municipal police; the accounting of the Municipality. Note that the information contained in the SIT are in Italian language, therefore the produced Linked Open Data will be in Italian too (although the whole generation process is completely language-independent).

3.2 Geo-data modelling and reengineering

To reengineer the dataset according to the target conceptual model we used Tabels[10], a software tool developed by the research foundation CTIC, which, using the GeoTools libraries[11], is capable of transforming the information encoded in the shape files into RDF representations. From the shape files supplied for each data record (in particular, the files with extensions .dbf and .shp), Tabels encoded the shape files into RDF triples related to the designed ontology, that it will be described in more detail in Section 3.3. On the one hand the characteristics of the table are stored as RDF representation, and, on the other hand, the spatial geometry is modelled on the standard KML representation [18]. At this

Fig. 1: Example of a geo-localised entity of "pharmacies".

[10] Available at: `http://idi.fundacionctic.org/tabels/`
[11] Available at: `http://geotools.org`

stage we are mapping to existing vocabularies, in particular NeoGeo[12], suitable for geo-data. The geometric coordinates in KML are expressed according to the Geodetic reference system Gauss-Boaga (or Rome 40). By means of different conversion tools publicly available on-line (e.g. `http://www.ultrasoft3d.net/Conversione_Coordinate.aspx`), it is possible to produce the coordinates of latitude, longitude and altitude in meters using the Geodetic system WGS84 [19]. In particular, the application of Tabels to each pair of files, .dbf and .shp, of the data tables is able to produce a set of RDF triples stored in a repository with other geometric resources contained in a public server. For example, from the information stored in the database of the SIT representing an entity of "pharmacies" (Figure 1), Tabels produces the related RDF triples, shown in Figure 2, and the file with the geometric KML coordinates (Figure 3).

```
<http://www.essepuntato.it/2013/10/prisma/resource/farmacia/finocchiaro-giuseppa>
a <http://www.essepuntato.it/2013/10/prisma/CATANIA.SDO_Farmacie> ;
    <http://www.essepuntato.it/2013/10/prisma/CODICE-of-CATANIA.SDO_Farmacie>
      "10625" ;
    <http://www.essepuntato.it/2013/10/prisma/MUNI-of-CATANIA.SDO_Farmacie>
      "5" ;
    <http://www.essepuntato.it/2013/10/prisma/Municipali-of-CATANIA.SDO_Farmacie>
      5 ;
    <http://www.essepuntato.it/2013/10/prisma/NOME-of-CATANIA.SDO_Farmacie>
      "FINOCCHIARO GIUSEPPA" ;
    <http://www.essepuntato.it/2013/10/prisma/NUMERO-of-CATANIA.SDO_Farmacie>
      "60" ;
    <http://www.essepuntato.it/2013/10/prisma/OBJECTID-of-CATANIA.SDO_Farmacie>
      1 ;
    <http://www.essepuntato.it/2013/10/prisma/OBJECTID_1-of-CATANIA.SDO_Farmacie>
      1 ;
    <http://www.essepuntato.it/2013/10/prisma/PROPRIETA-of-CATANIA.SDO_Farmacie>
      "FINOCCHIARO GIUSEPPA" ;
    <http://www.essepuntato.it/2013/10/prisma/RECAPITO-of-CATANIA.SDO_Farmacie>
      "VIA SAN GIOVANNI BATTISTA 74" ;
    <http://www.essepuntato.it/2013/10/prisma/Shape-of-CATANIA.SDO_Farmacie>
      "http://www.w3.org/2003/01/geo/wgs84_pos#Point" ;
    <http://www.essepuntato.it/2013/10/prisma/kml-of-CATANIA.SDO_Farmacie>
      <http://www.essepuntato.it/2013/10/prisma/farmacia/kml/Farmacie.1.kml> .
```

Fig. 2: RDF triples produced by Tabels for the example of entity in "pharmacies".

```
<?xml version="1.0" encoding="UTF-8"?>
<kml:kml xmlns:kml="http://earth.google.com/kml/2.1">
    <kml:Document id="featureCollection">
        <kml:Placemark id="Farmacie.1">
            <kml:Point>
                <kml:coordinates>15.0520808419018,37.5490041443454</kml:coordinates>
            </kml:Point>
        </kml:Placemark>
    </kml:Document>
</kml:kml>
```

Fig. 3: KML coordinates produced by Tabels for the example of entity in "pharmacies".

[12] Available at: `http://geovocab.org/doc/neogeo.html`

Tabels is able to import common file formats, such as XLS or CSV, including shape files. Afterwards it generates automatically a transformation program from the input data files. The generated program is able to transform each row of the input data into a new instance of a RDF class ad-hoc. In addition, each value in the column of the input tables is converted into a new triple where the subject is the instance mentioned, the predicate is a property based on the name of the column header, and the object is the value of the column as a *rdfs:Literal*. It is worth noting that the transformation program automatically generated, is a SPARQL-based script completely customisable by the user. Thus it is possible to change classes, names and associated properties, and then to annotate them appropriately. Once the transformation program is defined, the execution of Tabels generates the corresponding RDF in output, which we make publicly available online through a dedicated SPARQL endpoint. In addition, information regarding each resource object of the ontology data can be obtained through negotiation mechanisms of the content (content negotiation) based on HTTP REST that make them accessible, for example, through a browser or as REST web service. Data consumption is described in more detail in Section 3.5.

3.3 Resulting ontology for the SIT

Starting from the definition of the tables of the SIT, a first version of OWL ontology was developed. This provides classes and properties representing the database entities of the SIT, and is publicly available at the following URI:

`http://ontologydesignpatterns.org/ont/prisma/ontology.owl`

having the namespace (i.e. the default address of the entities in the ontology):

`http://www.ontologydesignpatterns.org/ont/prisma/`.

The creation process of this ontology was divided into two main phases and has followed the good practice of formal representation, naming, and semantic assumptions in use in the domain of the Semantic Web and Linked Open Data [15, 16]. In the first phase, the entire structure of the tables was converted into a draft OWL ontology, where each table (i.e. each entity type described by the supplied data) is represented by a class and each field of the table by a data property. This translation was carried out in a fully automatic way from the sources provided in XML format (extension .shp.xml) by means of the use of an XSLT transformation. Note that fields with the same name but belonging to different tables have been provided with distinct properties. For example, the fields "Name" of the tables "Nursing Homes" ("Case Riposo") and "Pharmacies" ("Farmacie") have been translated with two different data properties, respectively "Name-of-CATANIA.SDO_NursingHomes" and "Name-of-CATANIA.SDO_Pharmacies".

From this interim draft ontology and from the available data, a first version of the ontology in OWL was produced. At this stage we have followed the suggestions of the W3C Organization Ontology[13], a set of guidelines for generating,

[13]Available at: `http://www.w3.org/TR/2014/REC-vocab-org-20140116/`

publishing and consuming LOD for organizational structures. In this respect we have named the graph nodes as URIs and pursued the following principles:

- The name of all the classes was taken to the singular (e.g., from "Pharmacies" to "Pharmacy");
- The names of the data properties were aligned when they were clearly showing the same semantics. For example, the properties "Name-of-CATANIA.SDO_NursingHomes" and "Name-of-CATANIA.SDO_Pharmacies" ended in the same property "name", assigned to "NursingHome" and "Pharmacy" as domain or entity class;
- The data properties that seemed to refer to individuals of other classes, probably having foreign key functions on the data base, were transformed into object properties. For example, the property "MUNI-of-CATANIA.SDO_NursingHomes" became "municipality" in order to connect individuals of class "Nursing Home' with individuals of class "Municipality";
- The data properties having values clearly assigned to some resources were transformed into object properties and their values were *reified* as individuals of specially created classes.

All changes made to the intermediate draft ontology for the implementation of the first version of the ontology have been documented in the form of SPARQL CONSTRUCT. This allowed us to create a simple script to convert the data extracted by Tabels in order to make them fully compliant with the final expected ontology, produced as output in RDF format.

3.4 Example of conversion from the geo-data to the final ontology

In this section we want to focus on the phase of transformation from shape files to the final RDF ontology by reporting an example. Consider as reference the data record "Traffic Lights" ("Semafori"). The SQL schema of this table includes the fields:

- *ObjectID* - unique number incremented sequentially;
- *Shape* - type Geometry that represents the coordinates defining the geometric characteristics of the entity;
- *Id* - Identification number of type Double;
- *name* - String type name of the entity;
- *Sde_SDE_se* - integer number;
- *Se_ANNO_CAD_DATA* - blob representing the date.

Passing the .shp and .dbf files to Tabels, this generates the transformation program, that is the SPARQL-based script used to import the data (see Figure 4). As already mentioned, it is possible to edit the script to suit custom requirements. Once any change in the transformation program is completed, it is possible to save and run it, which generates the RDF triples from the table data given as input. Figure 5(a) shows the RDF/Turtle produced by Tabels by using the methodology already described for a single 'Traffic Light" entity as example. Figure 5(b) shows the corresponding final ontology of this entity obtained

```
PREFIX project: <http://www.essepuntato.it/2013/10/prisma/semaforo/>
PREFIX my: <http://www.essepuntato.it/2013/10/prisma/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/#>

FOR ?rowId IN rows FILTER get-row(?rowId)
    MATCH [?oBJECTID,?iD,?nOME,?sDESDESE,?geometry,?kml] IN horizontal
    LET ?resource = resource(replace(replace(replace(lower-case(?nOME)," +","-"),"--+","-"),",|\.",""),
                             <http://www.essepuntato.it/2013/10/prisma/resource/semaforo/>)

CONSTRUCT {
    my:TabelsDataCatalog a dcat:Catalog .
    my:TabelsDataCatalog dct:title "Tabels AutoGenerated Catalog" .
    my:TabelsDataCatalog dct:description "Tabels AutoGenerated Catalog" .
    my:TabelsDataCatalog dct:publisher my:TabelsAutoGenerator .
    my:TabelsDataCatalog dcat:dataset my:DataSet
}

CONSTRUCT {
    ?resource a my:CATANIA.SDO_semafori .
    ?resource my:OBJECTID-of-CATANIA.SDO_semafori ?oBJECTID .
    ?resource my:Id-of-CATANIA.SDO_semafori ?iD .
    ?resource my:NOME-of-CATANIA.SDO_semafori ?nOME .
    ?resource my:sde_SDE_se-of-CATANIA.SDO_semafori ?sDESDESE .
    ?resource my:Shape-of-CATANIA.SDO_semafori ?geometry .
    ?resource my:kml-of-CATANIA.SDO_semafori ?kml
}

CONSTRUCT {
    my:CATANIA.SDO_semafori a rdfs:Class
}

CONSTRUCT {
    my:OBJECTID-of-CATANIA.SDO_semafori a rdf:Property .
    my:Id-of-CATANIA.SDO_semafori a rdf:Property .
    my:NOME-of-CATANIA.SDO_semafori a rdf:Property .
    my:sde_SDE_se-of-CATANIA.SDO_semafori a rdf:Property .
    my:Shape-of-CATANIA.SDO_semafori a rdf:Property .
    my:kml-of-CATANIA.SDO_semafori a rdf:Property
}
```

Fig. 4: A view on the transformation program used by Tabels to convert the shape files to RDF for the table "Traffic Lights" ("Semafori").

```
@prefix my: <http://www.essepuntato.it/2013/10/prisma/> .
@prefix : <http://www.essepuntato.it/2013/10/prisma/resource/semaforo/> .

:cso-italia-cso-provincie-vle-ionio a my:CATANIA.SDO_semafori ;
    my:Id-of-CATANIA.SDO_semafori 8 ;
    my:NOME-of-CATANIA.SDO_semafori "C.so Italia-C.so Provincie-V.le Ionio" ;
    my:OBJECTID-of-CATANIA.SDO_semafori 10 ;
    my:Shape-of-CATANIA.SDO_semafori "http://www.w3.org/2003/01/geo/wgs84_pos#Point" ;
    my:kml-of-CATANIA.SDO_semafori
      <http://www.essepuntato.it/2013/10/prisma/semaforo/kml/Semafori.10.kml> ;
    my:sde_SDE_se-of-CATANIA.SDO_semafori 1 .
```

(a)

```
@prefix ont: <http://ontologydesignpatterns.org/ont/prisma/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .

<http://ontologydesignpatterns.org/ont/prisma/semaforo/cso-italia-cso-provincie-vle-ionio>
  a ont:Semaforo ;
  ont:forma geo:Point ;
  ont:identificativoOggetto 10 ;
  ont:nome "C.so Italia-C.so Provincie-V.le Ionio" ;
  ont:sde 1 .
```

(b)

Fig. 5: Top panel (a): RDF/Turtle produced by the transformation program of Tabels for a single entity of the table "Traffic Lights" ("Semafori"). Bottom panel (b): Corresponding final RDF/Turtle ontology obtained through SPARQL CONSTRUCT conversion to fully match the designed ontology.

by conversion through SPARQL CONSTRUCT of the related data extracted by Tabels, in order to fully match the designed ontology.

This example further shows the ability and simplicity of the proposed methodology to gather the complex structure of a non-structured database, allowing a rapid analysis, retrieval, and conversion of the data into a structured RDF format, and the publication in the form of Linked Open Data.

3.5 Data consumption

The produced ontology consists of 854,221 triples and can be publicly queried by selecting the RDF graph called <*prisma*> on the dedicated SPARQL endpoint accessible at `http://wit.istc.cnr.it:8894/sparql`. Queries can be made by editing the text area available into the interface for the SPARQL query. The SPARQL endpoint is also accessible as a REST web service, whose synopsis is:

- URL ⇒ `http://wit.istc.cnr.it:8894/sparql`
- Method ⇒ GET
- Parameters ⇒ query (mandatory)
- MIME type supported output ⇒ *text/html*; *text/rdf+n3*; *application/xml*; *application/json*; *application/rdf+xml*.

Data are also accessible through content negotiation. The reference namespace for the ontology is `http://www.ontologydesignpatterns.org/ont/prisma/` which is identified by the prefix *prisma-ont*. The namespace associated with the data is, instead `http://www.ontologydesignpatterns.org/data/prisma/` which is identified by the prefix *prisma*. These two namespaces allow content negotiation related to the ontology and the associated data. The negotiation can be done either via a web browser (in this case the MIME type of the output is always *text/html*), or by making HTTP REST requests to one of the two namespaces. The synopsis of the REST requests to the web service associated with the namespace identified by the prefix *prisma-ont* is the following:

- URL ⇒ `http://www.ontologydesignpatterns.org/ont/prisma/`
- Method ⇒ GET
- Parameters ⇒ ID of the ontology object (mandatory the PATH parameter)
- MIME type supported output ⇒*text/html*; *text/rdf+n3*; *text/turtle*; *text/owl-functional*; *text/owl-manchester*; *application/owl+xml*; *application/rdf+xml*; *application /rdf+json*.

Instead, the synopsis of the REST requests to the web service associated with the namespace identified by the prefix *prisma* is the following:

- URL ⇒ `http://www.ontologydesignpatterns.org/data/prisma/`
- Method ⇒ GET
- Parameters ⇒ ID of the ontology object (mandatory the PATH parameter)
- MIME type supported output ⇒*text/html*; *text/rdf+n3*; *text/turtle*; *text/owl-functional*; *text/owl-manchester*; *application/owl+xml*; *application/rdf+xml*; *application /rdf+json*.

4 Conclusion

This paper presents an application of Linked Open Data for PA. The used methodology was implemented by following the standards of the W3C, the good international practices, the guidelines issued by the Agency for Digital Italy and the Italian Index of Public Administration, as well as by the in-depth experience of the research participants in the field. The method was applied to the case study of the PA of the MoC, in particular from their data stored in the Geographic Information System, SIT. By using tools and technologies for the extraction and publication of data, it was possible to produce an ontology of the SIT according to the paradigm of Linked Open Data. The data are publicly accessible to users through queries to a dedicated SPARQL endpoint, or alternatively through calls to dedicate REST web services.

In currently on-going work a mobile application based on this LOD and related to sustainable mobility and emergency vehicle routing is under development and will be released soon. This will support the real-time management of road traffic and public transport, informing citizens on the state of roads in urban areas, in particular during urban emergencies, from small accidents to more serious disasters, and redirecting the road traffic by providing best alternatives routes to find way outs, the nearest hospitals or other locations of interest. The user will be able to contribute traffic and other road data, sharing road reports on accidents, advising on unexpected obstacles or inaccessible zones, or any other hazards along the way, helping to give other users in the area real-time information about what is currently happening. Soon, when the mobile app based on these LOD will be launched, user-centric tests and an experimental evaluation will be object of investigation. Our work is a concrete step supporting the Municipality of Catania to move into the paradigm of Open Government and Linked Data, boosting the metropolis towards the route of a modern Smart City.

References

[1] Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D.: Tabulator: Exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop, SWUI 2006, Athens, USA (2006)

[2] Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22

[3] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Boston (2006)

[4] Diaz, R., Behr, J., Toba, A., Giles, B., Manwoo, N., Longo, F., Nicoletti, L.: Humanitarian/emergency logistics models: A state of the art overview. Volume 45. (2013) 261–268

[5] Liu, B.: Route finding by using knowledge about the road network. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans. **27**(4) (1997) 436–448

[6] Municipality of Catania: Il Sistema Informativo Territoriale. [Online] `http://www.sitr.provincia.catania.it:81/il-sit` (last accessed: January 2014)

[7] Geiger, C.P., von Lucke, J.: Open Government and (Linked) (Open) (Government) (Data). JeDEM - eJournal of eDemocracy and Open Government **4**(2) (2012)

[8] Geiger, C.P., von Lucke, J.: Open Government Data. In Parycek, P., Kripp, J.M.., Edelmann, N., eds.: CeDEM11. Conference for E-Democracy and Open Government. Volume 6317 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2011) 183–194

[9] Alani, H., Dupplaw, D., Sheridan, J., O'Hara, K., Darlington, J., Shadbolt, N., Tullo, C.: Unlocking the potential of public sector information with Semantic Web technology. In: Proceedings of the 6th International Semantic Web Conference (ISWC 07). Volume 4825 of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Berlin Heidelberg (2007) 708–721

[10] Ding, L., Difranzo, D., Graves, A., Michaelis, J., Li, X., McGuinness, D., Hendler, J.: Data-gov Wiki: Towards Linking Government Data. In: Proceedings of the AAAI 2010 Spring Symposium on Linked Data Meets Artificial Intelligence, Palo Alto, CA. Volume SS-10-07. AAAI Press (2010) 38–43

[11] Ding, L., Lebo, T., Erickson, J.S., DiFranzo, D., Williams, G.T., Li, X., Michaelis, J., Graves, A., Zheng, J.G., Shangguan, Z., Flores, J., McGuinness, D.L., Hendler, J.A.: TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web **9**(3) (2011) 325 – 333

[12] Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., Schraefel, M.: Linked Open Government Data: Lessons from data.gov.uk. Intelligent Systems, IEEE **27**(3) (2012) 16–24

[13] Baldassarre, C., Daga, E., Gangemi, A., Gliozzo, A., Salvati, A., Troiani, G.: Semantic scout: Making sense of organizational knowledge. In Cimiano, P., Pinto, H., eds.: Knowledge Engineering and Management by the Masses. Volume 6317 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 272–286

[14] Gangemi, A., Daga, E., Salvati, A., Troiani, G., Baldassarre, C.: Linked Open Data for the Italian PA: the CNR Experience. Informatica e Diritto **1**(2) (2011)

[15] Agency for a Digital Italy: Linee guida per i siti web delle PA. Art. 4 della Direttiva n. 8/2009 del Ministro per la pubblica amministrazione e l'innovazione [Online] `http://www.digitpa.gov.it/sites/default/files/linee_guida_siti_web_delle_pa_2011.pdf` (2011)

[16] Agency for a Digital Italy: Linee guida per l'interoperabilità semantica attraverso Linked Open Data. Commissione di coordinamento SPC [Online] `http://www.digitpa.gov.it/sites/default/files/allegati_tec/CdC-SPC-GdL6-InteroperabilitaSemOpenData_v2.0_0.pdf` (2012)

[17] Lamb, A., Johnson, L.: Virtual expeditions: Google Earth, GIS, and geovisualization technologies in teaching and learning. Teacher Librarian **37**(3) (2010) 81–85

[18] Dodsworth, E., Nicholson, A.: Academic uses of Google Earth and Google Maps in a library setting. Information Technology and Libraries **31**(2) (2012) 81–85

[19] EUROCONTROL: WGS 84 implementation manual. Institute of Geodesy and Navigation (IfEN), University FAF Munich, Germany (1998)

Visualisation of User-Generated Event Information: Towards Geospatial Situation Awareness Using Hierarchical Granularity Levels

Heidelinde Hobel[1], Lisa Madlberger[2], Andreas Thöni[3], and Stefan Fenz[4]

[1] Vienna University of Technology,
Doctoral College Environmental Informatics, Austria
[2] Vienna University of Technology,
Vienna Phd School of Informatics, Austria
[3] Vienna University of Technology,
Institute of Software Technology and Interactive Systems, Austria
[4] SBA Research, Vienna, Austria
hobel@geoinfo.tuwien.ac.at, lisa.madlberger@tuwien.ac.at,
andreas.thoeni@tuwien.ac.at, sfenz@sba-research.org

Abstract. In recent years, enterprises and emergency response teams have started to use user-generated content to monitor crises, events and trends. Especially in critical situations, decision makers must, above all, quickly assess huge amounts of data. Effective geographical visualization and aggregation of collected data is an important prerequisite to enable decision makers to infer the impact of a detected event on, for example, their supply chains and other physical establishments. However, in existing literature the aspect of geographical visualization of automatically analysed events is hardly addressed. In this paper, we propose to introduce hierarchical levels of detail, a concept from Geographic Information Systems, for the visualization of user-generated data describing a local event. We developed a tool which can improve the assessment of regional impacts by offering the possibility to browse and visualize results on layers aggregating data along individually defined hierarchical dimensions, e.g. geographical or political districts.

1 Introduction

Social networks, RSS feeds, and platforms for microblogging are becoming more and more important for users to share feelings, experiences and to report about recent events at any time. As a result, huge amounts of data are created, which are often publicly available. This raw information can be used by enterprises as well as governments to rapidly learn about the latest events and to monitor the public opinion about certain topics. However, processing these huge amounts of information is a challenging task and, in case of a crisis, the accessible information must be quickly assessed to be useful for decision-making. Globalization led to an increase in the multinational footprint of enterprises and their supply chains. Consequently, the critical infrastructure is spread over large and disconnected territories. Knowing geographical references is a key input for decision-making processes in such an environment.

Therefore, the assessment of the collected information needs to be linked with geospatial information about regions and points of interest. Microblogs and news feeds are often enriched with temporal and geospatial data, either explicitly provided by tags in the meta-data or implicitly in the messages' content itself. However, this information inherits the intrinsic properties of user-generated data and is therefore likely to be incomplete, incorrect, and imprecise. Furthermore, it is a non-trival task to monitor large regions of interest while still quickly assessing the impact of a detected event with globally dispersed points of interest. Therefore, knowing the geographical reference area of a feed can be an important starting point.

The goal of our research is, therefore, to improve geospatial assessment of events reported in microblogs by using hierarchical levels for the analysis of important events threatening the infrastructure of enterprises and visualization of results by browsing through these layers. While the detection of the geographic origin of an incident is often determined by the measurement of bursts, e.g., [7], we focus on monitoring regions of interest, assessing the impact of detected events, and providing better user-support for decision-making. To this end, we developed a tool for the assessment of collected microblogs at hierarchical geospatial levels while considering relevance factors that are assigned to individual microblog messages. The contributions of this paper are as follows:

- We developed a tool aimed at supporting enterprises and emergency response workers in geospatial assessment of incidents by using hierarchical levels.
- We discuss architectural decisions, implementation details, and our semantic model for the analysis of microblogs.

It is important to note that any monitoring-approach relying on user-generated web data is restricted to situations where both users as well as decision makers are still able to access communication infrastructure. Moreover, it is limited by the extent of data being augmented with geospatial information, either by explicit tags (e.g. GPS-tags) or implicitly in the content. Considering for example the microblog platform Twitter[5], around 2 percent of microblogs had been GPS-tagged in 2012. Given a baseline of around 400 million entries per day, the amount tagged was still significant. A fulltext geocoder including additional fields such as the location could even reference around 28% of the entries [6].

Throughout the paper we are considering exemplary User Generated Text Content (UGTC), which includes microblogs, RSS feeds and content from social media platforms. We focus on the generic aspects when using UGTCs in critical response systems - an implementation in a certain domain using a specific provider always requires the consideration and adherence to the specific applicable data protection laws, privacy terms, and terms of use of the provider.

The remainder of this paper is structured as follows: In Section 2, we present related work on the use of microblogs for event-detection and in Section 3, we

[5] http://www.twitter.com

discuss the opportunities to retrieve geospatial information from microblogs and RSS feeds. We present the visualization approach in Section 4, and describe our architecture and implementation in Section 5. In Section 6, we conclude our work and offer an outlook on future work.

2 Related Work

The huge amounts of publicly available user-generated data motivated research in various areas. Several studies focus on detecting events in microblog data - one of the earliest of this kind was developed by Sakaki et al. [11], who developed a system to reliably detect earthquakes in Japan exclusively based on Twitter data. They do not particularly address the aspect of visualization in their study, however in a provided screenshot they use coloured pins to depict individual tweets on related to earthquakes on a map. The problem with pins is that multiple instances at the same location cannot be visually distinguished from single occurrences. The same applies for Sadilek et al. [10] who use microblog data to predict disease transmission and used pins to visualize geographic locations of users, but again, visualization was not a core aspect of their work. However in both cases, the possibility to aggregate and view results on higher hierarchical levels, e.g. for each district might enable a better overview and lead to additional insights. Chunara et al. [3] use news and microblog data to visualize disease outbreaks on a health map. They present alerts, derived from Tweets on a heatmap, which indicates high and low-density of relevant messages both on a detailed level and aggregated on up to two hierarchical levels according to administrative districts. While this provides a good example for the use of geospatial hierarchies for data-visualization, our approach aims for a general solution allowing for hierarchical aggregations along multiple dimensions, e.g. administrative but also according to geographical or political attributes.

Additionally, several efforts have been made to detect events independently of a specified domain [1, 2, 5, 7]. These methods typically extract events based on the detection of high occurrences of words. While in most of these studies temporal and geospatial properties of the detected events are extracted, only little attention has been paid yet to the geographic representation of events.

One of the few studies explicitly devoting attention to visualization aspects was conducted in [7], where the authors validated their map-based visualization approach in a user study. As one of their results they found, that for an intuitive user experience the additional possibility to zoom in and out of visualized data as well as the aggregation of mapped results would be required. Rosi et al. [9] also point out the need for better visualization techniques and tools to view and understand data at multiple levels of granularity. Pouliquen et al. [8] geocoded news items and experimented with different visualization options. They suggested representing news stories as points on a map leveraging WorldKit[6] or used placeholders in GoogleEarth[7] with icons representing the frequency of news

[6] brainoff.com/worldkit/
[7] earth.google.com

items found referencing a specific place. In the later, they relied on the zooming features naturally provided by GoogleEarth. Furthermore, they experimented with Scalable Vector Graphics (SVGs), but relied on only one country level.

In our study, we want to address this gap by proposing a method to implement hierarchical geospatial layers, which allow for different aggregation levels during event-detection, visualization and assessment. We use UGTCs such as microblogs and RSS feeds to illustrate our approach.

3 Inferring Geospatial Attributes from UGTC

Multiple ways to infer geospatial information are applicable to Microblogs as well as to RSS feeds. By using microblogging platforms users can often decide whether the exact location (identified by GPS-information), the place (such as the city or neighbourhood) or no location information is attached to a message. Additionally to these location-tagged microblogs, geographic information could be obtained from the user's profile if available and accessible at a platform. The user's profile may include a location field, the time zone and may include further location information in the profile description or on a linked personal website. Eventually, information can also be extracted from the message text, which might relate to a certain event or directly to an area, territory, or jurisdiction. Using profile, user location and geo tags as input and again taking Twitter as an example for a microblogging platform, Leetaru et al. [6] reported a share of 34% of microblogs mappable at a correlation level of 72% against a baseline.

Considering standard RSS feeds, location information can be obtained analogously from the content or the author's information. Since RSS feeds are linked to more comprehensive blogs or news articles, more information about where an event occurred could be provided. Moreover, the W3C GeoRSS standard is designed to explicitly provide information about the geographic location a post relates to in form of geographical points, lines and polygons, which can be automatically processed by geographic software. In this case, the location information is certainly more accurate since it is explicitly annotated and aims at describing the geospatial features of a report. However, even when having only full-text with geographic information available, accuracy rates of 77% have been achieved [8].

Locations inferred from user-generated data can be obtained by different methods. However, it must be distinguished between reports about events or incidents from the place where the event or incident occurred and reports in which the event or incident are discussed. While certainly both of categories are important, we require schemas to reflect these geospatial dependencies. Furthermore, location information is often ambiguous or may be incorrect. As a consequence, we have to account for the possibility to allow users to correct errors and infer the geographic dependencies between analysed topics.

4 Geospatial Visualization of UGTCs

The goal of our tool is to improve geospatial assessment of events reported in UGTCs by using hierarchical levels for the analysis of important events threatening the infrastructure of enterprises and the visualization of results by browsing through these layers. These layers are diversely designed according to the needs of the domain, which is encompasses the enterprises' requirements and the (global) dependencies of its supply chain infrastructure. For example, comparing the states of the US with the countries of the European Economic Area requires to set these areas on the same hierarchical level. This might be interesting when enterprises are planning or evaluating establishments of their infrastructure in these regions of interest. Furthermore, the user has to decide which regions and layers are of importance in the analysis. For example, in an industrial use case, an enterprise might concentrate on geographic regions where critical infrastructure or suppliers are located. Measuring the influence of earthquakes might motivate the user to define the center of the earthquake as central point and then to define concentric circles as hierarchical levels around the earthquake's epicentre. For our implementation, we created a hierarchical structure according to the United Nations Statistics Division[8]. According to this website, the geographical regions and compositions are structured in the following hierarchy: World $\mapsto$ continental regions $\mapsto$ geographic sub-regions (e.g., Eastern Africa) $\mapsto$ Countries. We extended the structure with Countries $\mapsto$ country-specific Regions.

In the following two scenarios, we will motivate the usage of hierarchical layers which facilitate the browsing of data in two conceptually contrary models: top-down and bottom-up.

Top-down assessment. In the top-down approach, users are monitoring the highest level of interest. This visualization approach is aimed at providing a holistic overview of gathered information and allowing to zoom into the defined levels of interest, where each area reveals its own and from sub-areas inherited UGTCs. By using the top-down approach, the user can compare regions of interest globally (i.e., at the highest abstraction level) with other areas at this level. Our tool then provides the possibility to seamlessly start zooming in to explore more specific areas in more detail. For instance, this visualization scenario is suitable when an enterprise is planning new establishments for critical infrastructure or to monitor geographically large areas. In case of an emergency or crisis this approach allows to zoom into the relevant local areas, determine the geospatial impact and to plan appropriate measures.

Bottom-up assessment. The bottom-up approach is useful for monitoring specific geographic locations and to access the impact as as soon as an event is detected. Users can also zoom out of the region in order to browse the UGTCs in the history of upper-levels. The predefined hierarchical levels allow to systematically analyze the global impact of an event. For instance, in case of a detected earthquake an approximation of the epicenter is calculated and the user's definition

[8] `https://unstats.un.org/unsd/methods/m49/m49regin.htm`

of the radii of the concentric circles around the epicenter are used to infer the impact of the earthquake.

The following sections explain the features of our tool based on the Visual Analytics Mantra “Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand” [4]. We shortly describe how the visualization tool can be used to browse through the analyzed data (Show the Important – Zoom Filter and Analyse Further), and how details of analyzed UGTCs could be accessed. The setup of the system and the processing of UGTCs is explained in Section 5.

4.1 Interactive Maps for Visualization

We use Leaflet[9], which is an open source javascript library for interactive maps, for the visualization of results using the defined hierarchical layers. Figure 1 shows the main interface for the analysis of collected UGTCs.

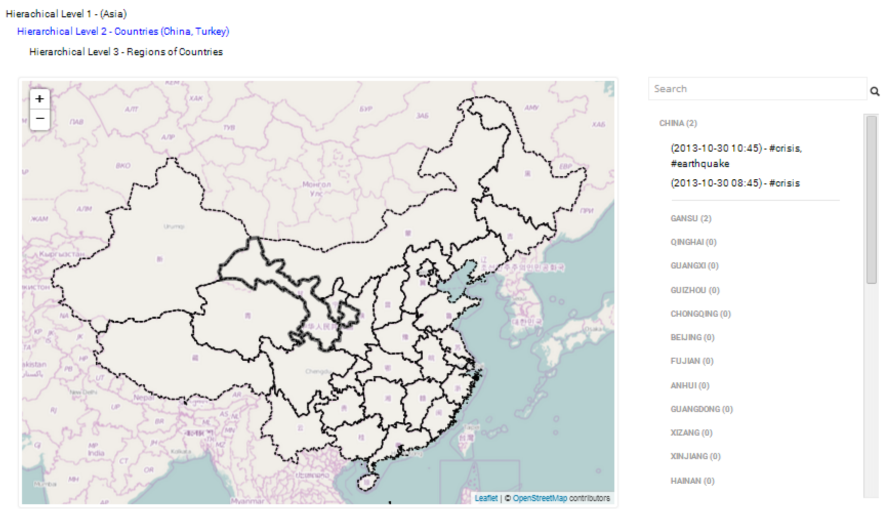

Fig. 1. Using Leaflet and a hierarchical tree to visualize the geographic correlation of UGTCs on interactive maps. (See for tools and data: Leaflet, OpenStreetMap, and NaturalEarth)

On the top of the tool, the user can choose the appropriate level for the analysis. In the example presented in Figure 1, the user is analyzing the secondary level, which comprises the countries China and Turkey. Furthermore, the user has opened the category China in the tree on the right side, which centers China’s geometry on the map. China’s category shows two identified

[9] `http://leafletjs.com`

UGTCs, which where classified into one of China's provinces, i.e., Gansu. How many UGTCs where detected in an area is displayed in brackets beneath after the name of the region of interest. Zooming into one of the provinces of China could be done by either opening the respective category or by double clicking onto the specific area on the map. In the case that a high number of UGTCs regarding defined topics, e.g. crisis, bomb, etc., for an area is detected, the specific area and every superordinate area is colored red. To specify "high", users may define a threshold value of number of UGTCs detected. Moreover, in order to minimize the effort to assess single UGTCs, only the most relevant UGTCs (see Section 5) are presented to the user. The time-frame, which could be used to retrieve relevant UGTCs can be manually adjusted.

When zooming into an area of the lowest level of the defined layers, the vector overlay will be transparent and UGTCs that have a location tag are shown on the map as markers, as shown in Figure 2.

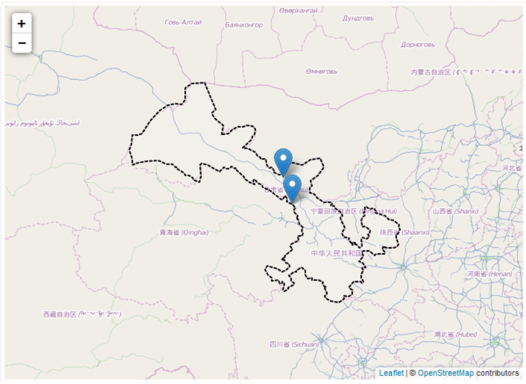

Fig. 2. Highlighting of detected UGTCs with location tags in the lowest hierarchical level. (See for tools and data: Leaflet, OpenStreetMap, and NaturalEarth)

4.2 Details on Demand

In our first prototype, we use timestamps of retrieval and the actual timestamp of exemplary messages, the content of exemplary messages, topical and geospatial tags, as well as information about the preprocessed information of the UGTC. Our system pre-classifies UGTCs based on a keyword analysis and assigns an indicator which is reflecting the relevance of each UGTC, as explained in Section 5. However, the classification of text content is a non-trivial task and a complete accuracy is almost not possible, therefore users should be able to manually assess detected incidents and correct possible misclassification. Hence, our interactive interface allows to open detail-sites when clicking onto a link of a incident that is displayed in the tree structure on the right side, in which the user can quickly assess the relevance of a message, edit its tags, and link further online resources. Semantic annotations allow to explore further details by following the links.

5 Implementation and Architecture

In this section, we present our framework for geospatial assessment of UGTCs. For our first prototype, we implemented a Web application for the processing of collected microblogs and visualization of results, its architecture is illustrated in Figure 3.

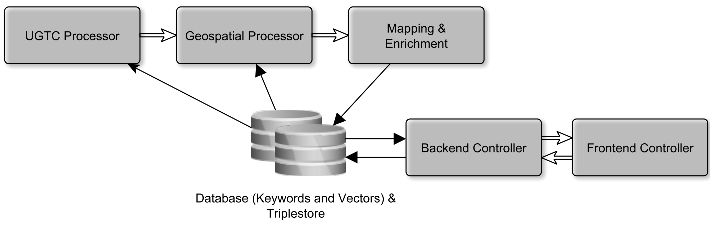

Fig. 3. Architecture

In the following, we will showcase the process using an exemplary fictive microblog message: "The earth is shaking - earthquake in Gansu".

UGTC Processor Our system is designed that it can be connected to various data sources, including microblogs and RSS feeds. The UGTC processor is designed to collect UGTCs from online sources or import datasets to search based on topical and geospatial keywords for relevant UGTCs. At the time this paper was written, we have tested our system based on an imported historic dataset. The keywords should be initially defined according to the domain of the system in order to restrict the input of UGTCs to the system. Enterprises with globally dispersed critical infrastructure could define keywords for the names of establishments and critical facilities (i.e. topical keywords) as well as all related names of areas corresponding to the facilities (i.e. geospatial keywords). In more generic use cases, the users should define keywords such as crisis, earthquake and thunderstorm. In our example message, we detected the following keywords: `earthquake` and `Gansu`.

Geospatial Processor. Geospatial keywords and hierarchical dependencies can be inferred from the GeoNames API[10]. GeoNames provides a huge amount of geospatial features, however, the location of many areas is often just provided as a single point, since the boundaries are not yet available for every record.

[10] `http://www.geonames.org/export/ws-overview.html`

Furthermore, the hierarchies of queried areas are not customizable and must be mapped to self defined hierarchies to allow customized comparisons.

The hierarchical geospatial structure for the prototype is based on continental and administrative boundaries, where we used the following taxonomy as mentioned in Section 4: World $\mapsto$ continental regions $\mapsto$ geographic sub-regions (e.g., Eastern Africa) $\mapsto$ Countries $\mapsto$ country-specific regions. We identified the datasets provided by NaturalEarth[11] as the most appropriate dataset for the visualization, since the data layers are preprocessed and provide consistent geographic shapes which lineworks are independent from other shapes, e.g. countries that share one line as a boundary. For countries and their regions we used the large scale data (1:10m) and directly exported the vector data in GeoJSON format by using the open source tool Quantum GIS[12]. For the world-wide and continental layers, and geographic sub-regions, we started from the dataset for administrative level one and merged the country-specific features into the appropriate structure. For each of the extracted areas, we added GeoJSON properties to designate the hierarchical type of the layer and the part-of relation of the respective area. Each layer and feature is enriched with the geospatial features, names and alternative names retrieved from GeoNames. The hierarchical geospatial database for our first prototype implementation is based upon a NoSQL database to store vectorial features in GeoJSON[13] format.

Each UGTC that is passed on from the UGTC Processor is processed according to the hierarchical information stored in the databases. Our first prototype supports geospatial keyword analysis for the content of a UGTC as well as geospatial queries such as "is this point in this polygon" for possible location metadata information of the UGTC. If the UGTC is classified based on the location information, then it is assigned to the feature of the lowest level of the hierarchical layers. If the UGTC contains a geographical keyword of the layers, then it is assigned to the specific feature. Since the fictional microblog message encompasses no explicit location tag, it is assigned to the feature that relates to "Gansu".

Mapping and Enrichment of Microblogs. Once a UGTC is detected from the geospatial processor and preclassified according to the identified feature, it is mapped to our RDF schema and stored in a triple store. To allow queries and aggregation functions on the stored set of UGTCs, we map each UGTC into a semantic model. We link the UGTC to the geospatial feature of interest by using a feature tag. The term location tag is used to refer to explicit GPS information in the meta data of the UGTC if available. Identified topical and geospatial keywords are annotated as tags, as well as location and temporal tags of UGTC are added as annotations. For the annotation we used the geonames[14] and dcterms[15] vocabularies. The following triples show an excerpt of the tags

[11] http://www.naturalearthdata.com/
[12] http://www.qgis.org/en/site/
[13] http://geojson.org/
[14] http://www.geonames.org/ontology/documentation.html
[15] http://dublincore.org/documents/dcmi-terms

used to annotate our exemplary microblog (we simplified it for presentation by linking geonames' RDF resource for Gansu and one entry for dcterms).

```
@prefix geovis: <http://environmental.tuwien.ac.at/geovis/1.0/> .
<http://environmental.tuwien.ac.at/geovis/1.0/id12938572552353>
    <http://purl.org/dc/terms/created> "2014-04-23 2:11:02" ;
    geovis:relatesToGeoNames <http://sws.geonames.org/1810676/about.rdf ;
    geovis:feature "feature8984762834456" ;
    geovis:tag "earthquake" ;
    geovis:geo-tag "Gansu" ;
    geovis:relevance 0.4 .
```

Moreover, we added a relevance tag which is calculated as follows: Each topical tag is assigned a relevance factor between 0 and 1. Assume that a UGTC has the topical tags: $tkw_1, \ldots, tkw_n$. Each topical tag can be mapped to a relevance factor ($rel(tkw_i)$), defined according of the domain of the enterprise. The maximum of these relevance factors is then multiplied by a location factor (f_{loc}), as shown in equation (1). We explain this process below.

$$UGTC_{relevance} = f_{loc} \times \max\{rel(tkw_0), \ldots, rel(tkw_n)\} \tag{1}$$

First we calculate the maximum of the relevance factors of the respective tags. In addition, χ_{locTag} denotes the boolean decision whether there is a location tag in this specific UGTC. In the second step, it is analyzed where the UGTC originates. If the UGTC was annotated with a location tag that is within one of the the lowest level of the regions of interest, then its values obtained by the previous maximum calculation is multiplied by 1. When the UGTC contains geospatial keywords referring to the region of interest, however the location tag indicates that it originates from another location, it is multiplied by 0.3. If the UGTC does not contain a location tag, it is multiplied by 0.5, since it is not clear whether the UGTC originates from one of the regions of interest. In our example microblog, the relevance would be calculated as follows: $max(rel(\text{“}earthquake\text{”})) = 0.8$ and $f_{loc} = 0.5$, which results in the relevance indicator 0.4. The described semantic model is still in an early stage, however it provides the basic links to the fundamental information for the assessment. Future extensions and links between the UGTCs could provide useful information in the assessment of the UGTC and aggregated results.

We believe that the usage of the maximum function is justified, because it allows the direct translation of the modeled relevance factors as the most important contributing element, as we assume that the topical keywords with the highest relevance has the highest impact. This is in accordance with the informal understanding, that the most severe element is the major contributor to its impact. We would like to note that the assessment of these values is out of the scope of this paper, since we are focussing on presenting a coherent visualization tool.

Controller. The controller is responsible for the processing of queries. The system allows to request the defined hierarchical layers and their assigned features. The request for a specific layer of interest is accomplished by the following steps:

First, the features of the layers are loaded that are displayed as overlays on the map. Second, for each feature of the layer all UGTCs are retrieved that were assigned as relevant to the feature. Subsequently, all UGTCs of all lower levels are retrieved. To limit the amount of results, the user has the option to specify the timeframe, the minimum of the calculated relevance, and specific topical keywords. The user can further explore the details of a UGTC by requesting the stored information and the enriched tags as well as he has the option to update the information and link the UGTC to further online resources.

6 Conclusion and Future Work

Enterprises and emergency response teams can make use of user-generated data to monitor events impacting critical infrastructures and, in case of crises, quickly assess resulting impacts to be able to decide about appropriate measures. In existing research, little focus has been onto the visualization and user-interaction in such systems for event-detection and the assessment of impact. We developed a tool for the visualization of geospatial impacts, which allows to analyze and assess impacts on different abstraction layers using geospatial hierarchical layers and enables therefore more intuitive and more user-specific representations.

This work presents a first approach of mapping UGTCs onto different hierarchical visualization layers. For future work we want to extend our prototype to a completely generic framework with comprehensive vocabulary for the hierarchical layer for visualization as well as the annotation of collected data. This would allow to set up such a system in a generic way and share and reuse information processed from the system. Moreover, we identified the following major aspects that allow further improvements. First, during the manual assessment of UGTCs supported by the hierarchical visualization layers, possible links between identified topics and contents could be identified and integrated into the data set. This semantically enriched data could be shared between domains and significantly improve the analysis and future decision-making in critical situations. Second, the recognition of geographical references could be improved by applying more sophisticated approaches in information retrieval. This could, among others, include the exploration of machine learning or the usage of further geographical attributes provided. Third, the presented visualization could be advanced by including additional attributes of analyzed feeds or by seamlessly adjusting polygons to different projections. Finally, the relevance calculation could be derived in a more sophisticated way including spatial features. All these adjustments could enhance the approach presented to a more sophisticated tool that would be applicable in multiple domains.

Acknowledgements This research was partially funded by the Vienna University of Technology through the Doctoral College Environmental Informatics and the Vienna PhD School of Informatics as well as by COMET K1, FFG – Austrian Research Promotion Agency.

References

1. L. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. *Multimedia, IEEE Transactions on*, 15(6):1268–1282, Oct 2013.
2. A. Boettcher and D. Lee. Eventradar: A real-time local event detection scheme using twitter stream. In *GreenCom*, pages 358–367, 2012.
3. R. Chunara, J. R. Andrews, and J. S. Brownstein. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.
4. D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining*, pages 76–90. 2008.
5. M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler. Getting There First: Real-Time Detection of Real-World Incidents on Twitter. In *Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics "Task-Driven Analysis of Social Media" as part of the IEEE VisWeek 2012, October 15th, 2012, Seattle, Washington, USA*, 2012.
6. K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.
7. A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, R. Miller, and O. Arenz. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 227–236. ACM, 2011.
8. B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, and A.-C. Forslund. Geocoding multilingual texts: Recognition, disambiguation and visualisation. `http://publications.jrc.ec.europa.eu/repository/handle/111111111/12654`, 2006.
9. A. Rosi, M. Mamei, F. Zambonelli, S. Dobson, G. Stevenson, and J. Ye. Social sensors and pervasive services: Approaches and perspectives. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 525–530, March 2011.
10. A. Sadilek, H. A. Kautz, and V. Silenzio. Modeling Spread of Disease from Social Interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, pages 322–329, 2012.
11. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *In Proceedings of the Nineteenth International WWW Conference (WWW2010). ACM*, 2010.

An Ontology-Based Approach to Social Media Mining for Crisis Management

Vanni Zavarella, Hristo Tanev, Ralf Steinberger, and Erik Van der Goot

Joint Research Center - European Commission

Abstract. We describe an existing multilingual information extraction system that automatically detects event information on disasters, conflicts and health threats in near-real time from a continuous flow of on-line news articles. We illustrate a number of strategies for customizing the system to process social media texts such as Twitter messages, which are currently seen as a crucial source of information for crisis management applications. On one hand, we explore the mapping of our domain model with standard ontologies in the field. On the other hand, we show how the language resources of such a system can be built, starting from an existing domain ontology and a text corpus, by deploying a semi-supervised method for ontology lexicalization. As a result, event detection is turned up into an ontology population process, where crowdsourced content is automatically augmented with a shared structured representation, in accordance with the Linked Open Data principles.

1 Introduction

Monitoring of open source data, such as news media, blogs or micro-blogging services, is being considered worldwide by various security agencies and humanitarian organizations as an effective contribution to early detection and situation tracking of crisis and mass emergencies. In particular, several techniques from text mining, machine learning and computational linguistics are applied to user-generated ("crowdsourced") content, to help intelligence experts to manage the overflow of information transmitted through the Internet, extract valuable, structured and actionable knowledge and increase their situation awareness in disaster management[3].

A massive amount of messages on social media platforms such as Twitter, Facebook, Ushahidi etc. are generated immediately after and during large disasters for exchange of real-time information about situation developments, by people on the affected areas.The main added value of this content is that: a. it is nearly real-time, and typically faster than mainstream news; b. it potentially contains fine-grained, factoid information on the situation on the field, far more densely distributed geographically than official channels from crisis management organizations. Nonetheless, there are several problems that hinder social media content to unfold its full potential for real-world applications.First, content is massive, containing a high rate of information duplication, partly due to several people reporting about the same fact, partly due to platform-specific content-linking practices, such as re-tweeting. Secondly, while social media messages can be generated with a certain amount of platform-specific metadata (such as time, geocoding and hashtags) a crucial part of their content is still encoded in natural language text[1].

Both issues raise from the same general problem that crowdsourced content is unstructured and lacks interpretation with respect to a shared conceptualization of the domain, so that it cannot be integrated with knowledge bases from humanitarian agency information systems and thus be converted into actionable knowledge. This issue clearly emerged in a survey with disaster management experts active on the field right after the 2010 Haiti earthquake, where Twitter users produced a significant amount of reports, which were though underexploited because of the lack of semantic integration with existing information systems [3].One solution that has been explored consists in providing the users with tools for structuring their content "on the fly", for example by engaging a layer of domain experts in encoding unstructured observations into RDF triples by using the Ushahidi platform ([27]). We propose instead a general architecture for augmenting unstructured user contributions with structured data that are automatically extracted from the same user-generated content. The overall method is top-down.We then apply a semi-supervised method for the lexicalization of the target ontology classes and properties from text [1]. The method learns a mapping from classes of linguistic constructions, such as noun and verb phrases, to semantic classes and event patterns, respectively. Then, with a relatively limited human intervention, such constructions can be linearly combined into finite-state grammars for detection of event reports. Finally, we run the output grammar on crowdsourced content and populate the target ontology with structured information from that content, by deploying an instance of the event detection engine tuned to the social media streams.

As the ontology lexicalization method is language and domain independent, the proposed architecture is highly portable across languages, including for instance the il-formed ones used in social media platforms.

Next section is devoted to related work. Section 3 outlines the architecture of event extraction from news text and describe its implicit domain model. Section 3 proceeds by describing the ontology lexicalization algorithms. Section 5 shows how the resulting ontologies can be populated from user-generated content by information extraction techniques. We conclude by exploring open issues and future developments of the proposed architecture.

2 Related Work

Event detection and tracking in social media has gained increased attention in last years and many approaches have been proposed, to overcome the information overload in emergency management for large scale events, and to increase sensitivity of small case incident monitoring ([20], [17],[29]). Different approaches vary with respect to the level of analysis they perform on crowdsourced content and the amount of structure they extract. This ranges from taking tweets as simple "sensor reading", associated with a

[1] The equally important issue of the confidence and trust of user-generated content, and the specific techniques for content validation which can be deployed to cope with it, are out of the scope of this paper.

time and location, of a target earthquake ([18]), to deriving RDF triples posts for content merging ([19]).

Many studies have explored the use of ontologies for enhancing semantic interoperability in various domains of interest for the crisis management professionals (e.g. spatial data and GIS, [21]). Nonetheless, a comprehensive standard ontology encompassing all those subject areas does not exist, while different sub-ontologies cover critical subject areas (such as Resource, Damage, Disaster) (see [6] for a survey). This poses an issue of ontology mapping and integration for a crisis response information system which would like to make use of Linked Open Data for its data sharing.

Methods for ontology learning and population have been explored in recent years both within the Natural Language Processing community and in the Knowledge Representation field (see [22] and [23] for an overview). Strongly related to our work is the concept learning algorithm described in [24], which finds concepts as sets of semantically similar words by using distributional clustering.

Finally, there are many approaches for text mining from Twitter data ([25]). Relevant to our approach are the methods for automatic event detection from Twitter like the one described in [26]. However, in contrast to the already existing approaches, we perform structured event extraction from the tweets rather than only event detection-based tweet classification.

3 Event Extraction Architecture

We have created a multilingual event extraction engine NEXUS for global news monitoring of security threats, mass emergencies and disease outbreaks [9]. The system is part of the Europe Media Monitor family of applications (EMM) [8], a multilingual news gathering and analysis system which gathers an average of 175,000 online news articles per day in up to 75 languages, taken by a manually selected set of mainstream newspapers and news portals. NEXUS builds upon the output of EMM modules and identifies violent events, man-made and natural disasters and humanitarian crises from news reports. It then fills an event type-specific template like the one depicted in Figure 1, including fields for number and description of dead, injured, kidnapped and displaced people, descriptions of event perpetrators, event time and location, used weapons, etc.

Nexus can work in two alternative modes, cluster-based and on single articles. In the latter the full article text is analysed. In the former, we process only the title and the first three sentences, where the main facts are typically summarized in simple syntax according to the so-called “inverted pyramid” style of news reports. [2].

Figure 2 sketches the entire event extraction processing chain, in cluster-based mode.

First, news article are scanned by EMM modules in order to identify and enrich the text representation with meta-data such as entities and locations, that are typically

[2] This is feasible without significant loss in coverage as the news clusters contain reports from different news sources about the same fact, so that this redundancy mitigates the impact on system performance of linguistic phenomena which are hard to tackle, such as anaphora, ellipsis and long distance dependencies.

$$\begin{bmatrix} type & \\ subtype & \\ date & \\ location & \\ dead & \begin{bmatrix} number \\ description \end{bmatrix} \\ injured & \begin{bmatrix} number \\ description \end{bmatrix} \\ kidnapped & \begin{bmatrix} number \\ description \end{bmatrix} \\ displaced & \begin{bmatrix} number \\ description \end{bmatrix} \\ arrested & \begin{bmatrix} number \\ description \end{bmatrix} \\ perpetrators & \begin{bmatrix} description \end{bmatrix} \\ weapons & \end{bmatrix}$$

Fig. 1. The output structure of the event extraction system

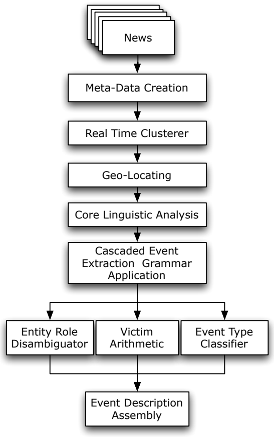

Fig. 2. Event extraction processing chain

separate from the ones deployed in the event extraction process proper. Next the articles are clustered and then geo-located according to extracted meta-data. Each article in the cluster is then linguistically preprocessed by performing fine-grained tokenization, sentence splitting, domain-specific dictionary look-up (i.e. matching of key terms indicating numbers, quantifiers, person titles, unnamed person groups like "civilians", "policemen" and "Shiite"), and finally morphological analysis, simply consisting of lexicon look-up on large domain-independent morphological dictionaries. The aforementioned tasks are accomplished by CORLEONE (Core Linguistic Entity Online Extraction), our in-house core linguistic engine.

Subsequently, a multi-layer cascade of finite-state extraction grammars in the ExPRESS formalism [12] is applied on such more abstract representation of the article text, in order to: a) identify entity referring phrases, such as *persons*, *person groups*, *organizations*, *weapons* etc. b) assign them to event specific roles by linear combination with event triggering surface patterns. For example, in the text "Iraqi policemen shot dead an alleged suicide bomber" the grammar should extract the phrase "Iraqi policemen" and assign to it the semantic role *Perpetrator*, while the phrase "alleged suicide bomber" should be extracted as *Dead*.

We use a lexicon of 1/2-slot surface patterns of the form:

```
<DEAD[Per]> was shot by <PERP>
<KIDNAP[Per]> has been taken hostage
```

where each slot position is assigned an event-specific semantic role and includes a type restriction (e.g. *Person*) on the entity which may fill the slot. ExPRESS grammar rules are pattern-action rules where the left-hand side (LHS) is a regular expression over Flat

Feature Structure (FFS) and the right-hand side (RHS) consists of a list of FFS, which is returned in case the LHS pattern is matched. See Section 5 for a sample rule.

The systems includes an event type classification module consisting of a blend of keyword matching, event role detection and a set of rules controlling their interaction. First, for each event type, we deploy: a) a list of weighted regular expression keyword patterns, allowing multiple tokens and wild cards b) a set of boolean pattern combinations: `OR` pattern lists are combined by the `AND` operator, each pattern is a restricted regular expression and conjunctions are restricted by proximity constraints.

As contradictory information on the same event may occur at the cluster level, a last processing step in the cluster-based mode consists of cross-article cluster-level information fusion: that is, the system aggregates and validate information extracted locally from each single article in the same cluster, such as entity role assignment, victim counts and event type.

The system architecture is highly customizable across languages and domains, by making use of simple local parsing grammars for semantic entities, backed by a number of lexical resources learned by semi-supervised machine learning methods [11]. Currently instances are in place for English, French, Italian, Spanish, Portuguese, Romanian, Bulgarian, Czeck, Turkish, Russian and Arabic language. The live event extraction results are freely accessible online for anybody to use, both in text form (`http://emm.newsbrief.eu/NewsBrief/eventedition/all/latest.html`) and displayed on a map (`http://emm.newsbrief.eu/geo?format=html&type=event&language=all`).

While extracted data are not delivered by the system in any Linked Data standard, a fine-grained categorization of domain entities and relations is implicitly used throughout the extraction process. An ontology representation of the domain model shared by the different instances of Nexus engine is illustrated in Figure 3.

As it can be seen, currently the system is profiled to model crisis event occurrence, rather than emergency tracking and relief operations. However, we experimented with a statistical method for building semi-automatically the core resources of an event extraction engine for a given language, starting from a pre-existing ontology and a text corpus.

4 Ontology Lexicalization Method

The general schema of our method is outlined here (see [1] for more details). Given a set of classes and properties from an event ontology:

1. we learn terms which refer to ontology classes by using a multilingual semantic class learning algorithm. These classes typically represent event-related entities, such as *Building* and *EmergencyCrew*.
2. we learn pre-and post-modifiers for the event-related entities in order to recognize entire phrases describing these entities. We do not give details on this part here (see [1]).
3. we learn surface event-triggering patterns for the properties relating ontology classes. As an example, the pattern *[BUILDING] was destroyed* instantiate the property *involvesBuilding* relating the event class *BuildingDamage* to a *Building* entity.

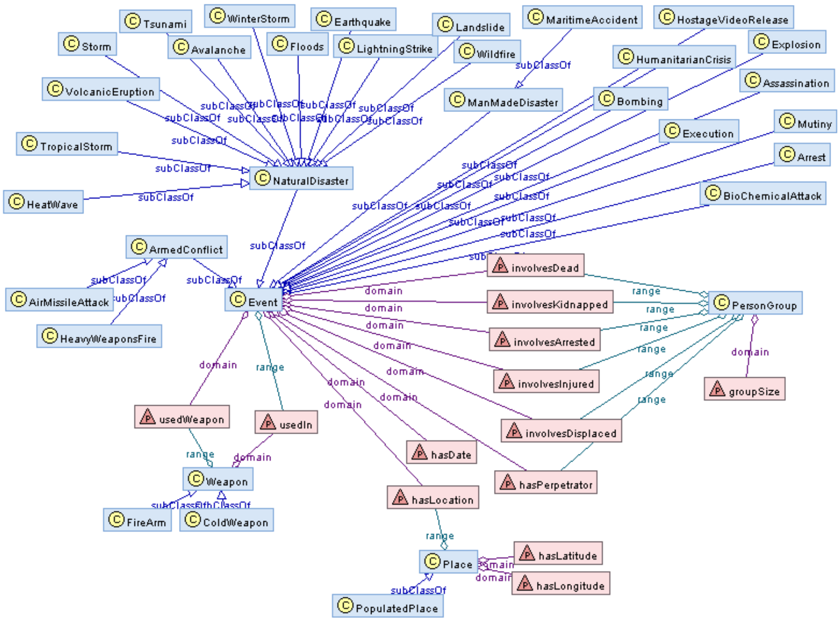

Fig. 3. Ontology representation of the Event Extraction system domain model.

4. Using the lexical classes and patterns learned in steps 1, 2 and 3, we create a finite-state grammar to detect event reports and extract the entities participating in the reported events, using the text processing modules and grammar engine described in Section 3.

4.1 Semantic Class Learning Algorithm

The algorithm we describe here accepts as input a list of small seed sets of terms, one for each semantic class under consideration in addition to an unannotated text corpus. Then, it learns additional terms that are likely to belong to each of the input semantic classes. As an example for English language, starting with two semantic classes *Building* and *Vehicle*, and providing the seed term set *home*, *house*, *houses* and *shop*, and the seed term set *bus*, *train* and *truck*, respectively, the algorithm will return extended classes which contain additional terms like *cottage, mosque, property*, for *Building* and *taxi,* *lorry, minibus, boat* for *Vehicle*. The semantic class learning algorithm looks like the following:

input : OriginalSeedSets- a set of seed sets of words for each considered class; Corpus- non-annotated text corpus; NumberIterations- number of the bootstrapping iterations
output : Expanded semantic classes

```
CurrentSets ← OriginalSeedSets;
for i ← 1 to NumberIterations do
    CurrentSets ← SeedSetExpansion(CurrentSets, Corpus);
    CurrentSets ← ClusterBasedTermSelection(CurrentSets,
    OriginalSeedSets, Corpus);
end
return CurrentSets
```

It makes use of two sub-algorithms for: (a) Seed set expansion; (b) Cluster-based term selection.

The seed set expansion learns new terms which have similar distributional features to the words in the seed set. It consists of two steps: (a)Finding contextual features where, for each semantic class c_i, we consider as a *contextual feature* each uni-gram or bi-gram n that co-occurs at least 3 times in the corpus with any of its seed terms $seed(c_i)$ and that is not composed only of stop words (we have co-occurrence only when n is adjacent to a seed term on the left or on the right); (b) Extracting new terms which co-occur with the extracted contextual features. In other terms, we take for each category the top scored features and merge them into a contextual-feature pool, which constitutes a semantic space where the categories are represented, and represent each term t as a vector $v^{context}(t)$ in the space of contextual features. Then we score the relevance of a candidate term t for a category c by using the projection of the term vector on the category vector.

The presented algorithm is semi-supervised ([16]) and in order to improve its precision we introduce a second term selection procedure: the learned terms and the seed terms are clustered, based on their distributional similarity; then, we consider only the terms which appear in a cluster, where at least one seed term is present. We call these clusters *good clusters*. The increased precision introduced by the cluster-based term selection allows for introducing bootstrapping in our process, which otherwise is typically affected by the problem of error propagation across iterations (so called “semantic drifting”).

4.2 Learning of Patterns

The pattern learning algorithm acquires in a weakly supervised manner a list of patterns which describe certain actions or situations. We use these patterns to detect event reports.

Each pattern looks like the ones shown in Section 3. For example, the pattern *damaged a [BUILDING]* will match phrases like *damaged a house* and *damaged a primary school*.

The algorithm accepts as its input: (a) A list of action words, e.g. *damaged*, *damaging*, etc. (b) a representation of the semantic category for the slot as a term list, e.g. *house*, *town hall*, etc. (c) an unannotated text corpus. It then returns a list of additional patterns like *[BUILDING] was destroyed*.

The main idea of the algorithm is to find patterns which are semantically related to the action, specified through the input set of action words and at the same time will co-occur with words which belong to the semantic class of the slot. It consists of three steps (see [1]):

1. find terms similar to the list of action words, e.g. *destroyed*, *inflicted damage*, using the semantic class expansion algorithm above;
2. learn pattern candidates which co-occur with the slot semantic category (e.g. *Building*), using the contextual feature extraction sub-algorithm. Each contextual feature of the slot class is considered a candidate pattern.
3. select only candidate patterns which contain terms similar to the action words (discovered in the first step). In this way, only contextual patterns like *inflicted damage on a [BUILDING]* will be left.

The algorithms presented here make use of no language analysis or domain knowledge, consequently they are applicable across languages and domains. In particular, this makes our method potentially applicable to the ill-formed language used in Twitter and other social media, provided that a relatively large text corpus is available.

5 Populating Disaster Management Ontologies

Once the lexicalization of an ontology has been carried out in a target language, we can apply part of the event extraction infrastructure described in Section 3 to populate that ontology from text streams. In [1] we report about an evaluation on populating a micro-ontology for disaster management with instances of events extracted from a stream of tweets published during several big tropical storms, in English and Spanish language. The ontology structure is shown in Figure 4

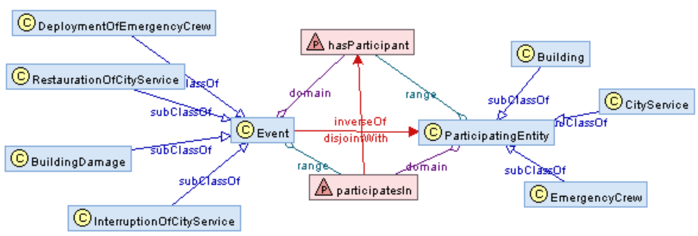

Fig. 4. The micro ontology used for the experiment

Each object from the sub-types of *Event*, namely *BuildingDamage*, *InterruptionOfCityService*, *RestaurationOfCityService* and *DeploymentOfEmergencyCrew*, is related to exactly one object from a sub-class of *ParticipatingEntity* via *has-a-participant* relation. Consequently, the detection of an event report can be done by detecting an action or situation with one participant. Event detection was performed by devising a simple two-level grammar cascade, whose rules combine semantic classes, patterns and modifiers learned with the previously described algorithms. Namely, it detects at the first level event participant entities: then, it combines them with event patterns to detect event-specific actions or situation, by rules like the one in Figure 5: where the boolean

```
event_rule :> ( leftPattern & [CLASS:"A"]
      participatingEntity & [CLASS:"C", SURFACE:#surf] )
-> event & [CLASS:"A", PARTICIPANT:#surf ]
& PossibleSlotFor(A,C).
```

Fig. 5. Rule schema for single participant event detection.

operator *PossibleSlotFor*(A) returns a list of all possible sub-classes of *ParticipatingEntity* which may fill the slots of a pattern of class *A*. As an example, when matching an *BuildingDamage* pattern followed by a *Building* entity expression on a tweet like

```
@beerman1991 yeah...i mean, there was thatb  neighborhood
in queens where like, 70 houses burned down during #sandy
```

this rule will populate the micro-ontology with an instance of the *BuildingDamage* event class and an instance of *Building* class, and relate them through an instance of *has-a-participant* property.

On an evaluation performed on a mixed-language English and Spanish corpus of 270,000 tweets, this method proved accurate in extracting events (**87%** and **95%** for English and Spanish, respectively), while recall on a small test set turned up to be very low (**23%** and **15%**). This seems to be due to missing action/situation patterns and to the finite-state grammars not being able to parse pattern-entity combination.

We plan to tackle the current limitation by deploying the ontology-derived grammar within the full-fledged event extraction infrastructure outlined in Section 3, where bag-of-words methods would be put into place, parallel to strict grammar matching, for the instantiation of event type entities and the extraction of participant entities.

To this end, we sketch a general strategy to map the current event extraction engine to the information structure of existing ontologies for Crisis Management. First, we manually map the implicit data model outlined in Section 3 with domain ontologies such as MOAC (`http://observedchange.com/moac/ns#`) or HXL (`http://hxl.humanitarianresponse.info/ns/`) and IDEA (Integrated Data for Events Analysis, see `http://vranet.com/IDEA.aspx`). We believe that, by doing this, we would increase the usefulness of our system output, both for human users working in the field of crisis management or conflict resolution and for services making use of Linked Open Data resources. In some cases, a one-to-one mapping already exists, as for instance between the hxl:PopulationGroup and nexus:PersonGroup classes, or between moac:Floods and nexus:Floods. In other cases, our ontology event type entities constitute an application of the more generic hxl:Incident entity.

Then, we add the missing, emergency-related sub-ontologies, encompassing classes like moac:InfrastructureDamage, moac:ContaminatedWater and moac:PowerOutag and we customize the system to detect events from them, via the method described in the previous section. Figure 6 illustrates the whole process:

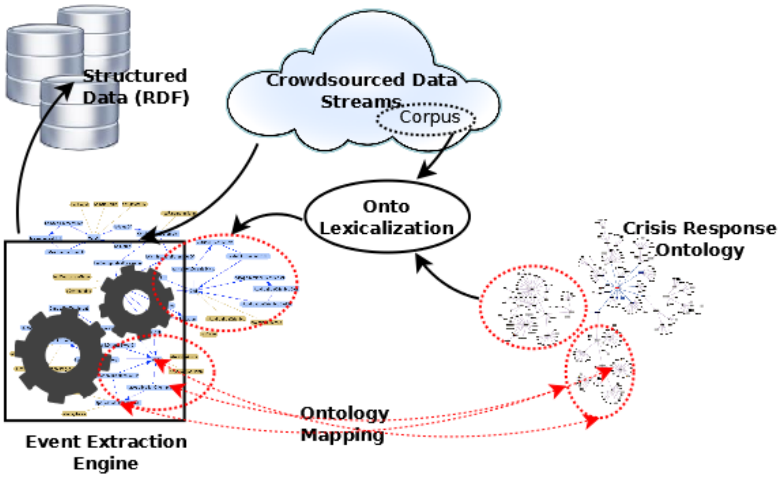

Fig. 6. A general schema for customizing the ontology-based event extraction engine.

6 Summary and Future Work

We outlined a procedure for mapping the domain model of an existing event extraction engine to larger scope ontologies for Crisis Management, in order to perform ontology-based event detection on social media streams. While the core of the procedure is a language-independent ontology lexicalization method that proved promising in supporting text mining from social media, processing of such sources remains challenging because: (a) potentially relevant messages have to be filtered from the large amount of user messages and (b) the usage of ill-formed text in social media messages (lower-casing names, omitting diacritics, doubling letters for stress etc.) reduces the information extraction recall and accuracy of automatic systems. We partially address (a) by generating a keyword-based query to retrieve event-related tweets starting from the text of a news cluster about that event ([28]). This will only be applicable though to large events which make to news clusters, while for the small scale incidents typically targeted for emergency tracking some methods for message duplication detection, clustering and merging appear necessary. (b) is currently addressed by deploying a preprocessing layer for tweet language normalisation ([13]) but work is still to be done in this direction.

Finally, as a prospective evaluation exercise, we plan to use a test corpus of around 18000 SMS messages (manually translated into English) collected by the Ushahidi platform, in the context of the Mission 4636 relief project, soon after the Haiti 2010 earthquake. We are currently searching for some validation event data from that context for measuring the extractive performance of our system.

References

1. Tanev, H. and Zavarella, V. (under publication) 'Multilingual Lexicalisation and Population of Event Ontologies. A Case Study for Social Media'.
2. Ashish, N., Appelt, D., Freitag, D. and Zelenko, D. (2006) 'Proceedings of the workshop on Event Extraction and Synthesis', held in conjunction with the AAAI 2006 conference, Menlo Park, California, USA.
3. Ortmann, J., Limbu, M., Wang, D., Kauppinen, T. (2011) 'Crowdsourcing Linked Open Data for Disaster Management', Proceedings of the 10th International Semantic Web Conference, Germany, 11-22.
4. Fan, Z., Zlatanova, S. (2011) 'Exploring Ontologies for Semantic Interoperability of Data in Emergency Response', Appl Geomat, 3, 2, 109-122.
5. Gruber, T. (2008) 'Collective knowledge systems: Where the Social Web meets the Semantic Web', Web Semantics: Science, Services and Agents on the World Wide Web, Volume 6 Number 1, pp. 4-13.
6. Shuangyan Liu and Duncan Shaw and Christopher Brewster (2013) 'Ontologies for Crisis Management: A Review of State of the Art in Ontology Design and Usability', Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM 2013 12-15 May, 2013)
7. Han, Bo, Cook, P. and Baldwin, T. (2012)'Automatically Constructing a Normalisation Dictionary for Microblogs', Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, Jeju Island, Korea, pp. 421–432.
8. Steinberger, R., B. Pouliquen, and E. van der Goot. (2009). 'An introduction to the europe media monitor family of applications'. In Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop, Boston, USA.
9. Piskorski, J. and Tanev, H. and Atkinson, M., Van der Goot, E. and Zavarella, V. (2011) 'Online News Event Extraction for Global Crisis Surveillance', Lecture Notes in Computer Science, volume 6910, pp. 182-212.
10. Tanev, H., Piskorski, J., Atkinson, M. (2008) 'Real-Time News Event Extraction for Global Crisis Monitoring', NLDB '08 Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, pp. 207 - 218.
11. Tanev, H., Zavarella, V. Linge, J. Kabadjov, M., Piskorski, J. Atkinson, M. and Steinberger, R.. (2009). 'Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish', Linguamatica, 1(2):5566.
12. Piskorski, J. (2007). 'ExPRESSextraction pattern recognition engine and specification suite'. In Proceedings of the International Workshop Finite-State Methods and Natural language Processing.
13. Küçük D., Guillaume J. and Steinberger, R. (2014) 'Named Entity Recognition on Turkish Tweets', Proceedings of the Conference Language Resources and Evaluation' (LREC2014).

14. Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. (2012). 'Joint Inference of Named Entity Recognition and Normalization for Tweets'. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 526535.
15. OBrien Sean P. (2002). 'Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability'. Journal of Conflict Resolution 46:791.
16. Tanev,H. and Magnini,B. (2008). 'Weakly Supervised Approaches for Ontology Population'. Ontology learning and population. Bridging the gap between text and knowledge, Berlin: Springer Verlag.
17. Schulz, A. Ristoski, P., Paulheim, H. (2013) 'I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs' Lecture Notes in Computer Science Volume 7955, 2013, pp 22-33.
18. Sakaki, Takeshi and Okazaki, Makoto and Matsuo, Yutaka (2010) 'Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors', Proceedings of the 19th International Conference on World Wide Web, ACM, pp 851–860.
19. Schulz, A. and Ortmann, J. and Probst, F. (2012) 'Getting User-Generated Content Structured: Overcoming Information Overload in Emergency Management', Global Humanitarian Technology Conference (GHTC), 2012 IEEE, pp 143-148.
20. Vieweg, S. and Hughes, A.L. and Starbird, K. and Palen, L., (2010) 'Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness' Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 1079-1088.
21. Guarino, N. (1998) 'Formal Ontology and Information Systems' in FOIS 2004: Proceedings of the Third International Conference. IOS Press pp 315.
22. Drumond,L. and Girardi,G. (2008) 'A survey of ontology learning procedures'. In The 3rd Workshop on Ontologies and their Applications, pp.1325, Salvador, Brazil.
23. Buitelaar, P., and Cimiano, P. (eds.) (2008) 'Ontology learning and population. Bridging the gap between text and knowledge', Berlin: Springer Verlag.
24. Pantel,P. and Lin, D. (2002) 'Discovering Word Senses from Text'. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 613619. Edmonton.
25. Breslin, J., Ellison, N., Shanahan, J., and Tufekci, Z. (Eds.) (2012) 'Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM - 12)', Dublin: AAAI Press.
26. Reuter,T. and Cimiano, P. (2002) 'Event-based classification of social media streams'. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong.
27. Okolloh, O. (2009) 'Ushahidi, or testimony: Web 2.0 tools for crowdsourcing crisis information', Participatory Learning and Action, vol. 59, no. 1, pp. 6570.
28. Tanev H., Ehrmann M. Piskorski, J. and Zavarella, V. (2012) 'Enhancing Event Descriptions through Twitter Mining'. In: AAAI Publications, Sixth International AAAI Conference on Weblogs and Social Media, pp 587-590.
29. Ritter, Alan and Mausam and Etzioni, Oren and Clark, Sam (2012) 'Open Domain Event Extraction from Twitter'. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,KDD '12, Beijing, China, pp. 1104–1112.

Visual Analysis of Real-time Social Media for Emergency Response

Suvodeep Mazumdar, Vitaveska Lanfranchi, Neil Ireson, and Fabio Ciravegna

Organisations, Information and Knowledge Group,
Department of Computer Science,
University of Sheffield, UK
{s.mazumdar, v.lanfranchi, n.ireson, f.ciravegna}@dcs.shef.ac.uk

Abstract. The prevalence of Social Media in sharing day to day information regarding all aspects of our life is ever increasing. More so, with access to cheap Internet-enabled devices and proliferation of Social Media applications. Among the variety of information shared, the most relevant, in the context of this paper is how individuals assess their surroundings and how they or their loved ones are affected by adverse events, disasters and crises. Traditional channels of communication often fall behind in providing timely information for emergency responders to formulate an accurate picture of the situation on the ground. The role of Social Media in complimenting such sources of information is thus invaluable and Social Media has been recognised as a key element of assessing evolving situations. Timely, accurate and efficient means to explore and query Social Media is essential for an effective response during emergencies, and hence this gives rise to a Knowledge Management issue. Our paper presents our approach to analysing Real-Time Social Media data streams using Visual Analytic techniques. We discuss the highly visual and interactive approach we employ to provide emergency responders means to access data of interest, supporting different information seeking paradigms.

Keywords: Social Media, Emergency Response, Visual Analytics

1 Introduction

The challenge of gathering a good understanding of large volumes of Social Media data streams is a significant one. The nature of Social Media itself, being highly dynamic, multi-lingual, geographically distributed, and highly relevant to the short-term zeitgeist poses enormous challenges. Additionally, the highly repetitive and noisy nature of Social Media also adds significant challenges to analysis efforts. In spite of these challenges, the potential of Social Media is immense and has been recognised as significant in Emergency Response by various organisations. Social Media empowers analysts with better means to understand public perception of events and situations on the ground as well as facilitating decision making processes for planning rescue operations[1,2] [20, 31]. The Emergency Response domain, owing to the need for the efficient delivery of critical information for decision makers, requires the assimilation, analysis and visualisation of Real-Time information. How such information can be presented to end users is a significant research challenge, as evolving situations require means to ensure the dynamicity of information is communicated. Social Media is multi-dimensional and hence, the information delivered to users must be communicated in a multi-faceted paradigm.

The goal of our research is to facilitate exploration of large volumes of Social Media information for gathering a very quick understanding of a situation at hand. It is important to note that this paper discusses events and situations from a broader perspective – events refer to anything that occurs[3] that is of interest to an analyst, while situations relate to the conditions and state of affairs (including events that occur to shape situations). To this end, we have developed a system, TUI (Tracking User Intelligence) [13] that exploits various visualisations and employs different interaction paradigms and approaches to help users improve their Situation Awareness [11, 32] by exploring various facets of Social Media. The context of our paper is all types of emergencies (major and minor) and events, but the effectiveness of our solution depends on the volume of Social Media generated. For example, low scale events such as small accidents in less populated areas would generate less interest, and hence, minimal Social Media posts. Our system also explores how users can have access to information based on temporal windows and can observe the evolution of events.

This paper is structured as follows: the next section describes related work. Section 3 presents an overview of the TUI interface. Section 4 describes how various types of information needs are addressed by elements within the framework, and how they are relevant to Emergency Response. Section 5 briefly discusses several field studies we conducted in real world examples and we conclude the paper with lessons learned and a discussion of future work.

2 Related Work

We present related work in two areas: Situational awareness and Visual Analytics. Situational Awareness in emergencies is paramount to deliver a timely and effective response [9]. To achieve effective Situational Awareness, emergency services must collate information from multiple sources and use it to build an understanding of the current situation and how this may evolve over time [11]. Leveraging data from citizens to build a form of collective intelligence [26], during emergencies or for security purposes, is becoming a vital resource for Situation Awareness [22]. During the 2007 southern California wildfires, two bulletin boards were set up to facilitate the exchange of information between citizens

[1] http://www.unocha.org/top-stories/all-stories/disaster-relief-20-future-information-sharing-humanitarian-emergencies
[2] http://www.unocha.org/hina
[3] http://www.complexevents.com/2011/08/23/event-processing-glossary-version-2-0/

and authorities [24]. A later analysis of Twitter postings during the 2009 Red River flooding [27] indicated that the service was being used by citizens and communities to collate and propagate information in a concise and responsive manner. Several systems have been developed to support citizen participation during emergencies that either directly foster data from citizens through custom apps [19] or analyse public data stream to extract real-time knowledge [2,29,30]. Existing techniques for searching Social Media involve exploiting entity-based semantic features [28]; entity mentions, hashtags, URLs and metadata [17]; and entity annotations coupled with user models for personalised searches [1]. Recommendations and filtering systems are used to help users reduce information overload, i.e. recommending links that users may find interesting; using dynamic semantic models of user interests [2]; recommending posts and friends based on categories [10, 21].

Visual Analytic techniques have been proposed to represent and filter Social Media at different levels of specificity [15] [4] and to convey information evolution in the crisis management domain [23]. When visualising large scale Social Media data, Visual Analytics is mainly used to provide high level overviews. [14] explores information regarding Social Media campaigns, [23] uses Twitter to understand the progression of earthquakes and [33] explores trends in emergency medicine. While these systems manage to efficiently display the chosen information, they are limited in the amount of data displayed. Systems with a broader focus try to capture the properties of generic data, allowing users to filter the data to items of interest. [15] for example, is a system for visualising and summarising events on Twitter. [8] allows users to explore Real-Time data streams relating to a given keyword. [3] is a system that improves Situation Awareness during small-scale crisis response, such as factory fires or music festivals by focussing on geotagged tweets and employing classification algorithms to identify messages relevant to specific events. Whilst most Social Media visualisation approaches rely on geographical and temporal features such as [18], some systems are starting to exploit the semantic of the data to enhance the visualisations. Examples of such systems are [6, 12, 15]. [15] uses features such as sentiment and link popularity to geographically plot the data. [12] uses features such as sentiment to create news flow diagrams that analyses the evolution of keywords and sentiments over time. [5] also focuses on interactive colour-coded timeline displays. [6] cluster groups of users and their evolution over time for a particular topic.

3 Tracking User Intelligence – An Overview

We designed our solution centering to the different modes of Information Seeking proposed by Bates [7]. Figure 2 presents the four modes: Searching, Monitoring, Browsing and Being Aware.

TUI is a multi visualisation platform that consists of multiple views over different facets of Social Media data. Users can interact with the system via each visualisation widget or a set of generic query elements to define filters. The user interface is separated from the backend processing, in order to deliver a more responsive performance, and hence our solution consists of two independent solutions. Several interaction paradigms govern how users can interact with the system, and therefore, provides multiple ways of engaging with the data. While being founded on the four information seeking models of Bates, TUI supports Schneiderman's well-known information seeking paradigm "overview first, zoom and filter, and details on demand" [25].

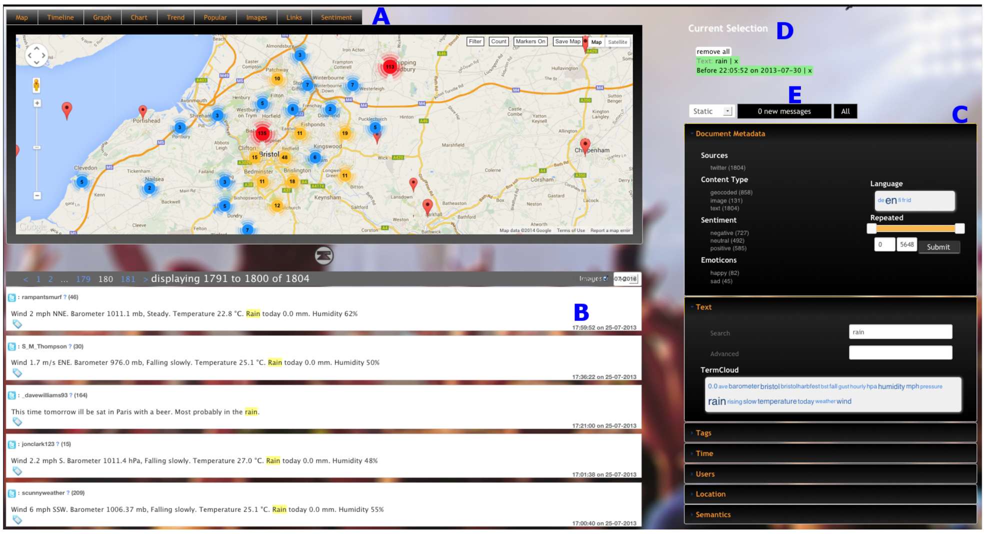

Fig. 1. A Screenshot of the developed system, Tracking User Intelligence (TUI). Five main sections can be observed (A: Contextual Information Visualisation; B: Content Presentation; C: Filters; D: Current Selection/filters; and E: Real Time Updates)

Figure 1 shows the four main components of the TUI interface. The interface is designed to enable users simultaneously access the most important information that is relevant to their present session. We enable this by providing a set of tabs (Section A: Contextual Information Visualisation), each presenting a piece of contextual information depending on the user's information need. The tabs (most relevant ones are discussed in the paper) are as follows: Map View ('Map'), Timeline View ('Timeline'), Graph view ('Graph'), Chart View ('Chart'), Trend View ('Trend'), Freq View ('Popular'). Map view provides contextual spatial information on a geographical map, while timeline provides visual summaries on a temporal scale. Graph view provides topical information, presented as a graph of topical relation among Social Media messages, where each topic is visually encoded based on the number of occurrence, associated collective sentiment or any other parameter that is decided as appropriate. The type of graph that is deployed depends on the particular use case, and TUI can select from a set of candidate graphs based on the number of topics. For example, a typical node-link graph can present a lower number of topics, but can provide an easier way of communicating relational information between co-occurring topics. The Context and Hierarchy chain [16], on the other hand presents a larger number of topics along with their hierarchical relations but requires a greater amount of interactivity to understand co-occurring topics. Chart view provides an overview of the entire dataset, presented as a series of pie charts. Trend view and Freq view provide advanced scatterplot visualisations to display which topics are trending and the most shared Social Media posts. The last two views are hidden from a typical Emergency Responder's profile, as they are presently under development and more investigation is needed into the best possible visualisations. These contextual tabs address the first part of Schneiderman's paradigm 'overview first'.

The Social Media posts section (Section B: Content Presentation) shows the data instances, where each Social Media post is presented as a snippet, chronologically listed. The display of Social Media posts is designed to address the last part of Schneiderman's 'details-on-demand' paradigm, where the session of user exploration dictates the relevant content to be displayed. The central part of Schneiderman's paradigm, 'zoom and filter' is the final filtering section (Section C: Filters). The filtering interface provides two functions – visual communication and overviews, and filtering. This section presents tag clouds of several facets of the data such as authors, places, type of post (photo, video, audio, link etc.), keyword or hashtag. Additionally, other interaction mechanisms such as text entry boxes, sliders, calendar widgets and buttons are presented to users for entering specific filters (or queries). Overall, interactions within Section A and C result in defining the context of the user's information needs and eventually retrieve the content for Section B. The Real-Time Updates section (Section E) provides a real-time reflection of the background data collection: this informs the user that in the time that has passed since the last time the visualisations have been refreshed, there has been a number of new messages that have been collected. The user has three choices: ignore the message, view only the new messages, or view all the data that has been collected. The final component presents the filters and selections that are active on the present exploration session of the user. These filters are generated by the user by interacting with visualisations and interface elements from the Sections A and C. Users can disable each filter by clicking on the (x) button if required.

4 Supporting Multiple Modes of Information Seeking

Within the scope of Emergency Response, Searching involves directed and active information seeking activities, where the user has a very specific information need, and knows the information exists. For example, an emergency responder is investigating a situation at hand (such as a flood) and is looking for more information related to the state of rivers overflowing. Monitoring involves the user having a specific information need, but is unsure if the information exists and hence monitors for relevant information. For example, an emergency responder is aware that flooding is a regular occurrence in a city, and is monitoring the condition of the city with a higher than usual predicted rainfall. Browsing, on the other hand involves the user actively looking for any relevant information, but in an undirected manner. An example of browsing would be an emergency responder looking for any occurrence of floods within a large geographical area. Being aware involves the user being passively looking for relevant information in an undirected manner. For example, an emergency responder looking for anything which can indicate an incident occurring that might be of interest to be further investigated.

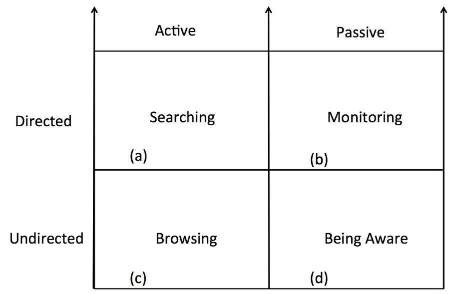

Fig. 2. Four modes of Information Seeking, proposed by [7]

As explained previously, Schneiderman's information seeking paradigm was central to the design of the interface. Bate's information seeking model provided a basis for understanding how users may need to search/explore/monitor existing situations using the system. During initial design phases, Bate's model along with interviews and focus groups with emergency responders and practitioners provided realistic scenarios that can explain what are the information needs of users at different stages of an emergency situation. This section presents four main scenarios that relate to the different elements of Bate's model, and how TUI can be typically used to identify data of interest. It is to be noted that while all of the aspects of the model requires active searching, the manner in which the user can trigger the searches and the follow up tasks play a role in addressing different types of information needs.

The filters section (Section C) is key to this task, where the user can drill-down to a set of data instances of interest, based on the information need of the user. The selection can be done using different interaction techniques and via a variety of facets. The variety of facets can be observed on the accordion menu on the right – three levels of filters are provided to users. Figure 3 describes how these filters are organised - Content filters, Immediate Context filters and Wider Context filters. Information that can be easily extracted from the content of the Social Media posts are referred to Content Filters. The content filters are the features that are aligned the closest to the Social Media post itself. A minimal analysis of the Social Media posts can provide a few more features, which have been organised into Immediate Context. While some features like Language or Type can be easily retrieved from the data provider, they may require some basic analysis to interpret content and infer. The Wider context filters indicate the features that require a further level of analysis/interpretation/query. Features such as identifying locations may be easily provided by geolocated posts, but the process may be considerably more difficult if there is a need to analyse text to interpret locations. Information surrounding a post such as the relevant users (who posted, who replied or who were mentioned) can be retrived by further queries to the data providers and may require further processing.

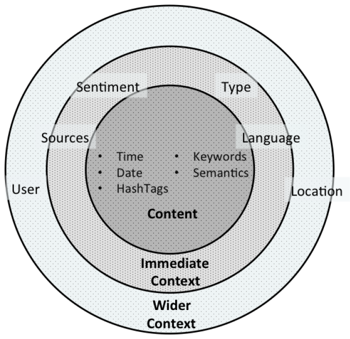

Fig. 3. Three levels of filters: Content filters indicate elements that can be quickly retrieved from the content of the Social Media posts; Immediate Context filters indicate information that can be inferred from the Social Media posts with minimal analysis; and Wider Context filters refer to the information that can be derived from the user who are relevant to the post, or the locations related to the posts.

A combination of filters can be used to identify an initial subset of the data, and then users can progressively apply further filters to reach the data instances that are of interest. For example, if the user is interested in the negative/positive messages that have had a high impact among the users, he/she can make a combination of selections of emotions/sentiments as well as selection of a high number of repeated messages. Observing the hashtags or keywords can then provide a rough summary of the topics of discussions and provide a greater overview of the situation.

4.1 Searching – Active and Directed

The first, and possibly most relevant to most Semantic Web applications is the ability to search for existing information, when there is a highly specific information need. Within the context of Emergency Response, the need to search for information, plays a significant role once an incident has been identified. Another role searching plays is post-event analysis where all data is searched to understand how an event evolved from different perspectives. Effectively, searching provides answers to questions that are already known by analysts and to facilitate follow-on analyses. In order to perform searching tasks, users simply enter (or select) an appropriate set of filters and reach the data of interest. Section C (filtering section) is the most appropriate in this context, as it provides direct means for users to search. Interactions within Section A (visualisations) also generate filter queries but such features are more pertinent to browsing activities. The search terms that are involved in this process are highly dependent on the evolution of the event as well as the event itself. Hashtags and keywords vary on how they have propagated within Social Media, and therefore, it is difficult to predict which terms would best suit the scenario being investigated. This calls for the need to monitor situations based on an initial 'guess' to encapsulate possible terms, and then fine-tuning queries to capture and retrieve more relevant information.

4.2 Monitoring – Passive and Directed

Perhaps most relevant to Emergency Response is the notion of monitoring, where the analyst needs to monitor an evolving situation, and looks out for information that may be relevant. This is achievable by setting queries and filters, and during standard data exploration via visualisations (Section A) and reading Social Media content (Section B), users can be updated when new relevant content is made available after harvesting recent Social Media posts. Once new content is available that is relevant to the present search criteria, users are communicated by a clickable label which states 'X new messages available'. The analyst is faced with one of three options in such situation: the first being, continue exploration and ignore the update. This would have no effect on the present exploration, and the user can proceed with finishing his/her analysis. With more information being made available, the notification is updated. This is enabled/disabled via the dropdown (session type) selection option 'Static' (the other options are 'dynamic' and 'batch-update'). The second option the user has is to proceed with analysing only the new content that has been retrieved. This notes the time when the previous analysis had started (the time when the last query was triggered). Selecting 'batch-update' from the dropdown enables this option and the user can then click on the button to view the relevant data instances that have been recently added. The third option the user has is to proceed analysing all of the data, with the new instances added to the analysis. This is performed by selecting 'static' and clicking on the click-able label.

The last option is a dynamic monitoring option, which regularly updates with new posts continually appearing during exploration sessions. This options is the least used, and hence is presently not encouraged to be used – more work is needed to understand how continuously evolving data can be presented to users without causing confusion and loss of analysis effort.

4.3 Browsing – Active and Undirected

Browsing involves users looking for information as and when their interests evolve. This is a mode which is also highly relevant to Emergency Response. Monitoring continuously evolving data, and searching existing data can indicate interesting events, keywords, or hashtags that might be relevant to improve the understanding of an evolving situation. In our approach, interactive Visualisations are key to this: while exploring a relevant dataset, users have the possibility to add new filtering terms by clicking elements from tag clouds, charts, timelines or maps. Clicking on sensitive areas of the visualisations immediately triggers queries, which result in a new filter being added and the system to drill-down into a more relevant subset of the data being explored. The user can remove the recently added filter by deselecting the filter from Section D (Current Selection).

4.4 Being Aware – Passive and Undirected

The most complex of information seeking modes is the state of being aware. TUI facilitates this by easily combining some of the approaches that are employed in the other three modes. The state of being aware in Emergency Response implies that an analyst is aware of the wider context of his/her analytic activity, but unaware of what event/situation may arise. This is often perceived to be the precursor to some of the modes and is conducted as a part of an initial or continued survey of the broader area (geographic, temporal or topical). In a typical ER (Emergency Response) activity, it is expected that an instance of the interface is continuously dynamically updated to provide overviews, based on query terms that are most likely to retrieve interesting events – generic terms such as 'flood', 'accident', '911' or 'fire' could form a good candidate set of keywords to look for[4]. Identifying any relevant interesting event would then initiate monitoring, browsing and searching activities on other instances of TUI.

5 Architecture and Implementation

The TUI interface is the second part of two systems - the first being a Social Media harvesting system which enables gathering of Social Media posts, and the second being a system which interfaces with the harvesting system and a local data store. While the first system (harvester) involves querying multiple sources of data continuously for new information and storing the results in local datastores following several iterations of backend processing, the second (TUI) is the interactive interface that users can use to access the data being stored/analysed/harvested. TUI is implemented as a standard web application, and written completely in HTML and Javascript. Several toolkits have been used

[4] It is to be noted that several of such 'generic' terms can also occur in posts that are highly irrelevant such as song lyrics, quotes from speeches etc. The decision to select relevant terms is mostly left with the analysts, based on their interpretation of the content being currently monitored.

within TUI to facilitate interactions such as jQuery[5], and provide customised look and feel such as jQueryUI[6], Less[7] and BlockUI[8]. Visualisations are provided by D3.js[9], Highcharts[10], Google Maps[11] and Javascript Infovis toolkit[12]

6 Discussions and Continuing Work

Over the summer of 2013, as a part of multiple projects, TUI was used to monitor several large events across the UK. Several Emergency Response organisations, Police, City Councils, event organisers and authorities were involved in the events. During the events the different types of information seeking modes were employed, and the system was used to monitor by several analysts at the same time. One of the key findings was the need for supporting multiple types of information seeking at the same time. Analysts need to explore, query, browse and be aware of situations at the same time, and hence, several instances of TUI are necessary to be active for improving Situation Awareness.

The TUI system is currently being redesigned to provide support for multiple tasks within one interface. Several layouts are presently being evaluated to understand which layout from a set of candidate layouts would be the most effective. The system is also planned to be evaluated with emergency responders during planned and unplanned events. Several techniques such as focus group, contextual inquiry and shadowing are planned to be used in the evaluations to understand how the system compliments traditional techniques for Emergency Response.

References

1. F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. *The Semantic Web–ISWC 2011*, pages 1–17, 2011.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web. In *Proceedings of ACM WebSci '11, 3rd International Conference on Web Science, Koblenz, Germany*. ACM, June 2011.
3. F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, editors, *WWW (Companion Volume)*, pages 305–308. ACM, 2012.

[5] http://jquery.com/
[6] http://jqueryui.com/
[7] http://lesscss.org/
[8] http://malsup.com/jquery/block/
[9] http://d3js.org/
[10] http://www.highcharts.com/
[11] https://developers.google.com/maps/
[12] http://philogb.github.io/jit/

4. A. M. Abhishek Pandya and S. T. Teoh. Enhancing online timeline visualizations with events and images. In *Electronic Imaging, Visualization and Data Analysis (VDA)*, 2011.
5. B. Adams, D. Phung, and S. Venkatesh. Eventscapes: visualizing events over time with emotive facets. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1477–1480, New York, NY, USA, 2011. ACM.
6. D. Archambault, D. Greene, P. Cunningham, and N. Hurley. Themecrowds: multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 77–84, New York, NY, USA, 2011. ACM.
7. M. J. Bates. Toward an integrated model of information seeking and searching. *The New Review of Information Behaviour Research*, 3:1–15, 2002.
8. M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.
9. A. Blandford and B. W. Wong. Situation awareness in emergency medical dispatch. *International Journal of Human-Computer Studies*, 61(4):421 – 452, 2004.
10. J. Chen, R. Nairn, and E. Chi. Speak little and well: recommending conversations in online social streams. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 217–226. ACM, 2011.
11. M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.
12. A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In *2nd Workshop on Making Sense of Microposts (MSM-2012*, 2012.
13. V. Lanfranchi, R. F. Carvalho, A. L. Gentile, S. Mazumdar, and F. Ciravegna. A semantic knowledge management framework for informal communication exchanges. In *Demo Track, 10th International Semantic Web Conference (ISWC 2011)*.
14. K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui. Content-driven detection of campaigns in social media. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 551–556, New York, NY, USA, 2011. ACM.
15. A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.
16. S. Mazumdar, F. Ciravegna, A. L. Gentile, and V. Lanfranchi. Visualising context and hierarchy in social media. In *International Workshop on Intelligent Exploration of Semantic Data (IESD'2012) at EKAW 2012 2012,Galway City,Ireland. October 9 2012*, 2012.
17. P. N. Mendes, A. Passant, and P. Kapanipathi. Twarql: tapping into the wisdom of the crowd. In *Proceedings of the 6th International Conference on Semantic Systems*, page 45. ACM, 2010.
18. B. Meyer, K. Bryan, Y. Santos, and B. Kim. Twitterreporter: Breaking news detection and visualization through the geo-tagged twitter network. In W. Li, editor, *CATA*, pages 84–89. ISCA, 2011.
19. O. Okolloh. Ushahidi, or'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1):65–70, 2009.

20. J. Ortmann, M. Limbu, D. Wang, and T. Kauppinen. Crowdsourcing linked open data for disaster management. In *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web in conjunction with the ISWC*, pages 11–22, 2011.
21. A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010, 2010.
22. L. Palen and S. B. Liu. Citizen communications in crisis: anticipating a future of ict-supported public participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 727–736, New York, NY, USA, 2007. ACM.
23. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.
24. I. Shklovski, L. Palen, and J. Sutton. Finding community through information and communication technology in disaster response. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 127–136, New York, NY, USA, 2008. ACM.
25. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.
26. V. Solachidis, P. Mylonas, A. Geyer-schulz, B. Hoser, S. Chapman, F. Ciravegna, V. Lanfranchi, A. Scherp, C. Contopoulos, I. Gkika, and et al. Advances in data analysis, data handling and business intelligence. *Knowledge Creation Diffusion Utilization*, pages 765–774, 2010.
27. K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In K. I. Quinn, C. Gutwin, and J. C. Tang, editors, *CSCW*, pages 241–250. ACM, 2010.
28. K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant for a topic? *Making Sense of Microposts (# MSM2012)*, page 9, 2012.
29. T. Terpstra, R. Stronkman, A. de Vries, and G. Paradies. Towards a realtime twitter analysis during crises for operational crisis management. In *Proceedings of the 9th International ISCRAM Conference. Vancouver, Canada*, 2012.
30. L. V. Realtime situation awareness via social media streams. In *Demo Session at ISCRAM 2012, Vancouver, Canada*.
31. P. Vinck, editor. *World Disasters Report*. International Federation of Red Cross and Red Crescent Societies, 2013.
32. C. D. Wickens. Situation awareness: Review of mica endsley's 1995 articles on situation awareness theory and measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):397–403, 2008.
33. K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1747–1756, New York, NY, USA, 2011. ACM.