Mining social media to extract structured knowledge through semantic roles

Diana Trandabăț *dtrandabat@info.uaic.ro*
University Alexandru Ioan Cuza of Iasi
Romania

Daniela Gîfu *daniela.gifu@info.uaic.ro*
University Alexandru Ioan Cuza of Iasi
Romania

Dan Cristea *dcristea@info.uaic.ro*
University Alexandru Ioan Cuza of Iasi
Romania

Abstract

We use semantics in our daily communications without giving it too much attention. However, things are not so trivial when computers try to incorporate semantic knowledge. In an attempt to enhance machines with human-like behavior and understanding, computer scientists and linguists have joined efforts in making the language easier to be understood. Language models need to be derived from large knowledge bases, hence this paper presents a platform able to extract user generated content for social media websites, analyze it and generate a structured knowledge base, in an attempt to discover the crowd intelligence hidden within.

Keywords: semantic roles, text mining, knowledge resources, social media.

1. Introduction

One of the key concerns in natural language processing is storing human knowledge and making it accessible to computers. Huge human and financial resources are usually involved in developing knowledge resources; therefore we propose a language processing application to assist humans in this endeavor, by exploring the social web in a new and innovative way, based on semantic frames. If having such knowledge resources, easily and dynamically created for different users, contexts or time frames, a gap will be filled between where we are now and where we could be in artificial intelligence: computers could be engaged in "intellectual" cooperation (with humans, or even more futuristic, with each other) in order to foster creativity, innovation and inventiveness.

The main research question this paper intends to answer is how can user generated content from social web be used to build a structured knowledge base. Our research hypothesis proposes the use of semantic relations for solving this challenge.

Panini's theory, presented in [17], led linguists to consider that semantic relations may have been analysed since thousands of years ago, by enhancing morphology with semantic features. Despite their long history, semantic roles have not yet reached a commonly agreed classification, different variants existing, from more particular or verb-oriented to rather general, most of them with proven efficiency in various practical implementations.

The importance of this theme comes from the quantity of data involved in the social web. Social media are web applications allowing user to generate individually or collaboratively content. It is a way to communicate information, daily experiences, lifelong expertise, opin-

ions and emotion about any possible topic, with both acquaintance and strangers over the Internet, creating the effect of Wisdom of Crowds [25]. A society is, in essence, a large group of rational and adaptive individuals, taking decisions in a highly interconnected complex and dynamic environment. An emergent collective behavior emerges from such scenario without the necessity to provide specific goals to the users that belong to the group, community or any other kind of social based structure.

The use of Social Media has tremendously increased worldwide over the last few years,. Using the proposed platform, people's individual contribution can reach a much wider audience than their small group of friends, by contributing to a "universal" social knowledge base. The huge popularity of social networks provides an ideal environment for scientists to test and simulate new models, algorithms and methods to process knowledge and VoxPopuli provides a platform to do precisely this job. Structured social knowledge can be used by different actors (companies, public institutions, researchers and scholars interested in formal and empirical analysis of social trends) to understand the behavior of users or groups.

The paper is structured as follow: Section 2 gives a short overview of the current state of the art in analyzing user generated content and semantic roles, while Section 3 discusses the proposed methodology. Section 4 briefly discusses the evaluation of our platform before drawing some conclusions in the last section.

2. State-of-the-art

Social media websites have not seduced only their users, but also researchers trying to analyse human behavior. The first and most commonly used research over social media involves the manifestation of opinions and sentiments transmitted, directly or indirectly, by users [13, 14], [19], [21]. However, most analyses over social media were so far limited to identify communities, user profiles or group behavior and to identify topics of interest in order to fine tune recommendation systems. Social context is crucial for the correct interpretation of social media content. Semantic-based methods need to make use of social context (e.g. who is the user connected to, how frequently they interact), in order to automatically derive semantic models of social networks, measure user authority, cluster similar users into groups, as well as model trust and strength of connection. A different approach is the ontology of social media writing styles, discussed in [16]. This paper is a position paper, proposing the extraction of structured knowledge from social media using semantic frames, a direction yet unexplored. This paper's research area is a very innovative, where models and techniques are only at their beginning, and could go beyond current Information Science and Engineering approaches with contributions from Social Sciences.

In-depth semantic analysis for practical natural language processing (NLP) tasks starts receiving more attention every day. NLP systems gradually stopped relying so much on word-based techniques and started to exploit semantics, as discussed in [1]. Applying the theory of semantic frames comes in line with actual trends in the field. Semantic roles allow answering questions related to the place entities have in various contexts and could be considered as small atoms used to compose the meaning of a sentence. Semantic roles express the context of a sentence in terms of the relations between concepts; they can define who the doer of the action is, for whom is the action performed, through which means, at what time and with which goal. Semantic roles are annotated around a predicational word.

Predicationality is a lexical feature, equally identified in nouns, verbs and even adjectives, whose meaning evoke an event or process, corresponding to what in the literature is called the deverbal property, or the deverbality of these categories [5].

A word is considered to bear the predicationality feature if it demands a semantic role structure in order to reveal its meaning. While most verbs are predicational, there exists a set of state, auxiliary or support verbs which do not express a semantic role structure, such as the ones in square brackets in the example below.

```
I [shall] go.
I [used] to like you.
```

On the other side, several nouns can have a predicational behavior, demanding a role structure similar to the one of the corresponding verb, such as the predicative nominals `explanation, decision, receiving` etc.

```
This is his [decision] on your request.
```

To exemplify semantic relations, let's consider the scenario of an arrest: an *authority* charges a *suspect* for an *offense*. In this scenario, a specific *time*, *place*, *purpose* and probably also some *means* can also be identified. Among the list of predicates linked to this scenario, we can find the verbs *arrest*, *cop*, *bust*, *apprehend*, each of them evoking the same scenario.

Manually identifying semantic roles in texts takes time and needs trained experts. A solution is developing automatic role labelling systems through accurate and reliable methods. Automatic Labelling of Semantic Roles is defined as the task of finding semantic elements in a sentence and classifying them with a correct semantic role, using as input a sentence and the selected target word [7]. In other words, Semantic Role Labelling (SRL) tries to determine a label, from the predicate *p*'s semantic frame, for linguistic constituents of the sentence *s*.

The work on semantic role labeling (SRL) has included a broad spectrum of probabilistic and machine-learning approaches, from probability estimation [6], through decision trees [24] and support vector machines [9], to memory-based learning [18]. Most studies largely converge on a common set of syntactic information (path from predicate to constituent, phrasal type of constituent) and lexical information (head word of the constituent, predicate). The SRL system we consider is based on a previously developed system [26], which is adapted for social media and incorporates, besides syntactic information, named entity and concept information.

The relations and events identification task has been substantially researched by the Watson group from IBM, and their discoveries on snippets evaluation and relationship extraction had direct applicability to question answering [22]. The identification of events and semantic roles are presented as structurally similar problems in [4]. Our approach extracts relations between concepts using semantic roles, similar up to some extent to the work in [4], but tailored for social media.

Our approach differs from other existing platforms for annotation of texts, either manual or automated, such as GATE[1] or BRAT[2], in two directions: (1) it specializes in social media texts and semantic role labeling and (2) it has the additional feature of generating a knowledge base of related concepts.

3. Methodology

The system mines social media content in 3 different phases: 1) user generated content is extracted from the social web through the UGC Extractor; 2) extracted segments are analyses by the UGC Analysis module, 3) before relations are extracted in the last step. Figure XXX presents the succession of these phases.

3.1. UGC Extractor

The definition of user-generated content (UGC) in [3] considers content created by a user as *"any form of content (…) of media that was created by users of an online system or service, often made available via social media websites"*.

Our intention is to gather textual UGC towards assessing collective behavior, having a particular care not to touch on any personal data of users. The European Parliament and the Council's Directive 95/46/EC defines personal data as follow:

[1] https://gate.ac.uk/ - An open source platform able to perform automatic pipelined human language processing.

[2] http://brat.nlplab.org/configuration.html - an online environment for collaborative manual text annotation.

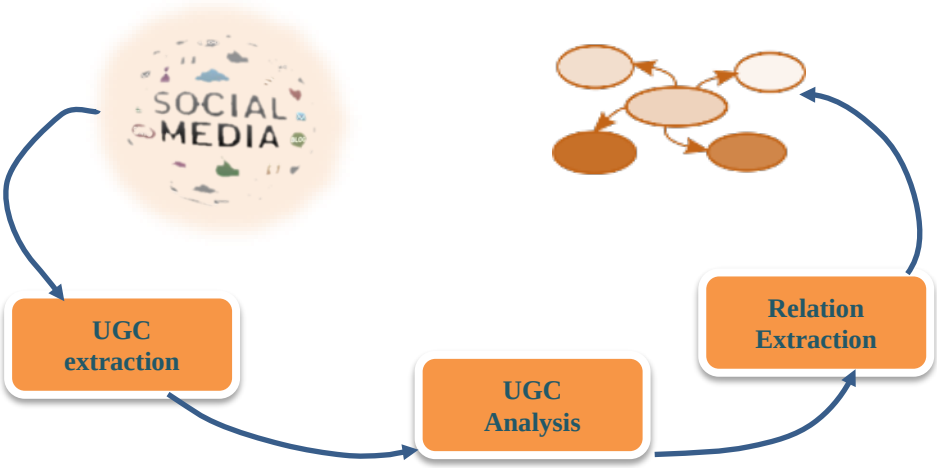

Fig 1. Architecture of the proposed platform

"'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, economic, cultural or social identity".

Privacy and copyright are still open issues when dealing with social media data. Hoser and Nitschke [11] discuss the ethics of mining social networks, and suggest that researchers should not access personal data that users did not specifically share for researchers, even when they are publicly available. On the other side, from a pure technical point of view, if for using the private data on social networks the user's agreement is needed, public postings, such as Facebook walls, Tweets, YouTube or Flickr comments, blogs and wikis count as public behavior. Furthermore, specialized APIs already exist for some of them, allowing collection of social media data.

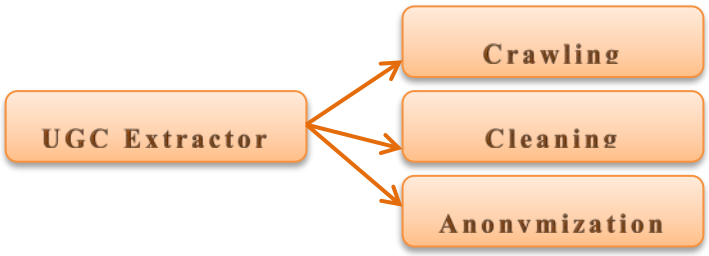

Fig 2. Extracting user-generated content

A set of sources have been used to mine for user-generated content, consisting mostly of social media, blogs and review websites. A data crawler uses a set of concurrent processes to query the social web using specialized search or streaming APIs, such as Archivist; YouTube Developer Page or Flickr API Gardens. The module offers the possibility for a filtered query by using a list of keywords to track (which may be expressions or entity names) and/or a set of geographical bounding boxes. The data to be process is cleaned to include only text, with no additional information, such as user, location, embedded media, or other similar data. In order to ensure that no relation to a natural person is made from the stored data, no personal data are stored or used, since all texts are properly shuffled and anonymized.

Another cleaning step involves a standardization focused on noisy content: social media content often has unusual spelling (e.g. 2moro), irregular capitalization (e.g. all capital or all lowercase letters), emoticons (e.g. :-P), and idiosyncratic abbreviations (e.g. ROFL, ZOMG). Spelling and capitalization normalization methods have been developed [10], coupled with studies of location-based linguistic variations in shortening styles in microtexts [8].

3.2. UCG Analysis

The analysis of the extracted content is performed in two major steps, at a superficial and deeper level, respectively. The system was evaluated using the NLP tools of our research group, specially designed for the Romanian language, but we consider extending it to allow for the inclusion of processing pipelines for different languages.

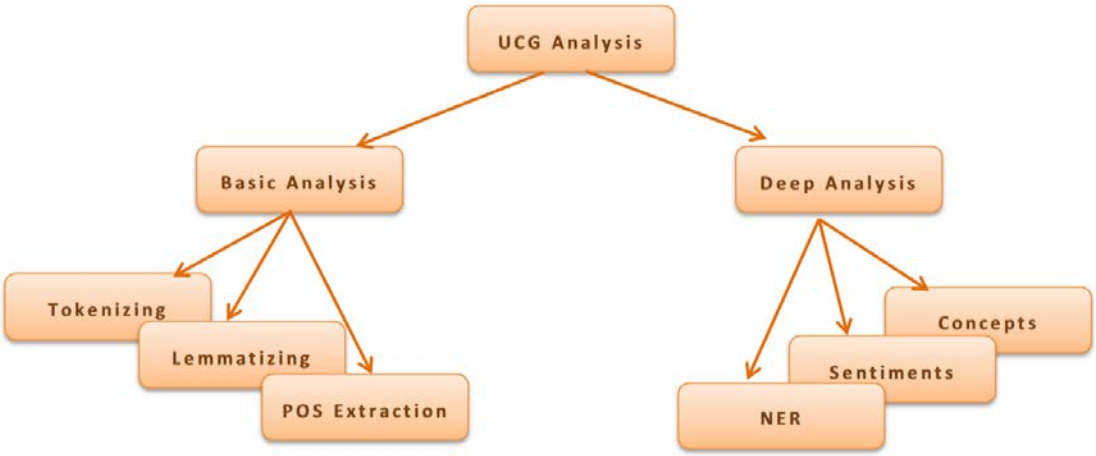

Fig 3. Analyzing user-generated content

At the basic level, while the tokenizer is a simple NLP tool, for the part of speech tagger we used an instrument developed by [23], which also performs lemmatization. For each UGC, the noun phrases were used to extract relevant concepts.

The deeper analysis part of this module performs named entity recognition, sentiment extraction and concept identification. Named entities are being extracted and classified using the parser developed in [12]. For extracting sentiments, a pre-processing phase made use of regular expressions to convert the texts to lowercase, discard words shorter than two characters, remove URLs or duplicated vowels in the middle of the words (e.g. cooooool). Then a module was used in order to attach polarity score to each word in the UGC. This simple sentiment analyzer used a manually acquired dictionary of about 2500 lemmas annotated with a sentiment score ranging from -5 (corresponding to the extreme negative sentiment) to +5 (the extreme positive one). The words not included in this list were considered neutral and received the polarity_score of 0. Furthermore, bigrams were extracted from the UGC, and a weighted score was computed for each bigram. The main idea behind this approach is that contrastive bigrams are more frequently indicating humor, such as “black milk”, and should receive a higher positive score. The sentiment was extracted by combining the obtained score with a Naïve Bayes classifier, trained on Semeval 2016 data, using features inspired by [7] and [20]: tokenized unigrams, emoticons and hash tags, and with the result of the AlchemyAPI for tweets.

The final step is the identification of concepts. In order to assure a high level of generalization, noun phrases are searched for in Romanian WordNet’s [27] network of hypernyms. Additionally, we attempt to unify instances by computing the similarity of two noun phrases from different UGCs using their synsets.

3.3. Relation Extraction

The next module applies a series of specialized language analyses in order to extract semantic relations between the concepts in the UGCs: a semantic role labeling platform, a pattern matching module and a generalization step.

The first step is applying a **semantic role** labeling system [26]. The challenge here is to deal with the social media input, syntactically and semantically different than the news one

used to train the parser. Twitter and most Facebook messages are very short (140 characters for tweets).

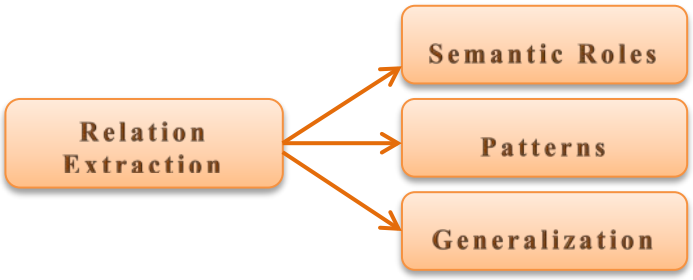

Fig 4. Relation extraction module

Many semantic-based methods supplement these with extra information and context coming from embedded URLs and hash tags. For semantic role labeling, we found that hash tags do more harm than good, therefore we eliminated them. Since in our tests we found very multilingual versions of UGC, we intend to extend this approach to also include an automatic language identification module [2].

Since it is time-consuming to annotate UGC with semantic roles in a large enough corpus to be used for training a classifier, our technique was to alter the training set, by including broken language, typing errors, limiting the number of words/characters in sentences, etc.) and run the machine learning algorithms again. The major shortcoming of this method is that it is not based on a real, naturally occurring language. Therefore, we decided to also use the initial SRL parser, improved with a set of post-processing patterns. The two methods are combined in a voting algorithm, which decides statistically on the semantic roles to apply for the user generated content.

For this study, semantic relations have been used in binary pairs of target (predicational) word plus different semantic roles, one by one, which we called **patterns**. For example, for the sentence *John obtained his diploma through hard work*, the semantic roles identified are presented in figure 5.

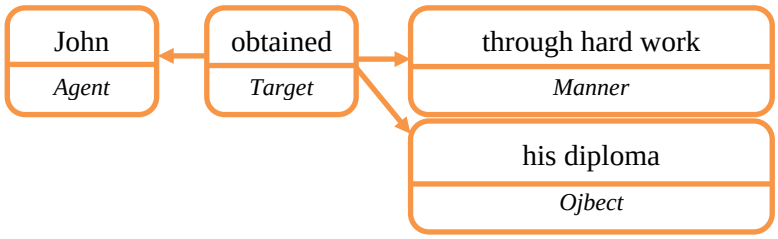

Fig 5. Example of semantic role annotation

The retained patterns are:

```
<John, Agent> <obtained, Target>
< through hard work, Manner> <obtained, Target>
< his diploma, Object> <obtained, Target>
```

The last step is the **generalization step**, which uses a simple anaphora resolution system and the WordNet hyponymy hierarchy to generalize over all obtained patterns.

The entity and its references are combined using a simplified version of an anaphora resolver, based on a couple of referential rules, focusing on anaphoric relations for named entities and discovered concepts. Thus, the rule-based system subsequently applies the following rules:

- Using a gazetteer (the most common method for identifying and classifying named entities). We extracted a gazetteer of named entities from Wikipedia list of names. As an example, `USA` and `United States of America` are co-references.
- Unify a part of a named entity with its full name, if both sequences can be extracted from closely located texts (at a distance of at most 2 sentences one another). To exemplify, `Caesar` is unified with `Julius Caesar` provided that both entities can be identified. Equally, the expression `the Minister of Defence` and `the Minister` refer to the same entity, if they co-occur in a narrow window of the text.
- Investigate different addressing techniques in order to match the ones that seem similar. For instance, `Mr. Smith` is a reference for `John Smith` if they appear in a narrow window, just as `The Smiths`, or `The Smith Family` refer most probably to `Mary and John Smith`.
- pronominal anaphora are solved in a similar manner. Thus, once a pronoun is found in the text, the previous sentence is scanned for an entity. If an entity is found, a link is established between the pronoun and the entity, taking into account the gender of the pronoun/entity.

For the considered example, the generalized patterns are:

```
<Person, Agent> <obtained, Target>
<Work, Manner> <obtained, Target>
<Certificate, Object> <obtained, Target>
```

When multiple UGC are annotated for the same target verb, they can be grouped together through these concepts. Thus, for instance, if having another input sentence: *Rehearsal allows you to obtain your price*, the patterns will be:

```
<you, Agent> <obtained, Target>
<rehearsal, Manner> <obtained, Target>
<your price, Object> <obtained, Target>
```

And their generalized versions:

```
<Person, Agent> <obtained, Target>
<Work, Manner> <obtained, Target>
<Gift, Object> <obtained, Target>
```

One can easily observe that there are matches between the two sentences in tern of generalized patterns. This approach will finally lead to generating a structured knowledge base, stored in RDF format, to be validated through a specialized visualization interface.

4. Evaluation

Since we are still in the developing phase of our platform, we only managed to validate the small number of 2000 relations, extracted from about 700 sentences. We limited the total number of annotated roles per sentence to three, because this was the average number of annotated roles in the Romanian FrameNet [26], but these limitations can be removed. We focused on the 6 semantic roles which are most representative in the Romanian FrameNet: Agent, Object, Duration, Place, Time and Manner.

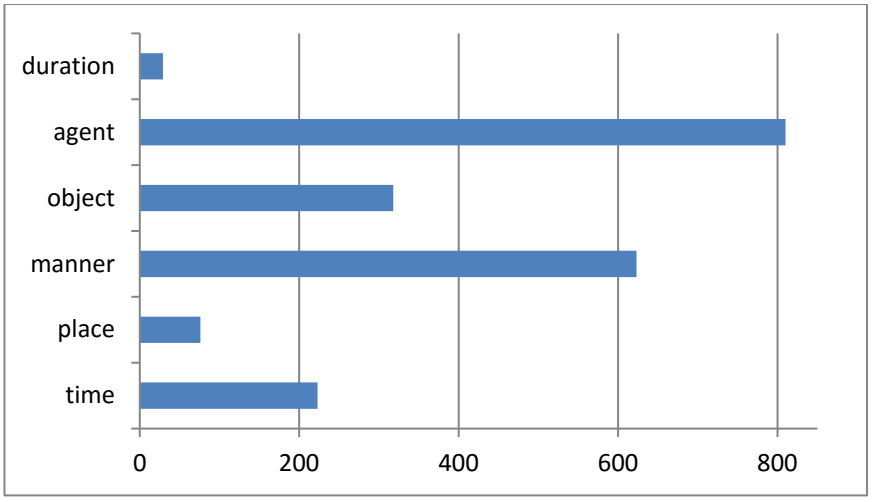

Fig.6. Distribution of semantic roles

Out of the total 2000 validated relations, we obtained an overall accuracy of over 86%. As expected, there were semantic roles which appeared in almost every sentence, sometimes even twice, if there were multiple predicational words in one sentence. One such example is the semantic role of `Agent`, roughly corresponding to the subject of the sentence. At the same time, other roles were only occasionally present, such as the duration role. Figure 6 presents the distribution of the semantic roles.

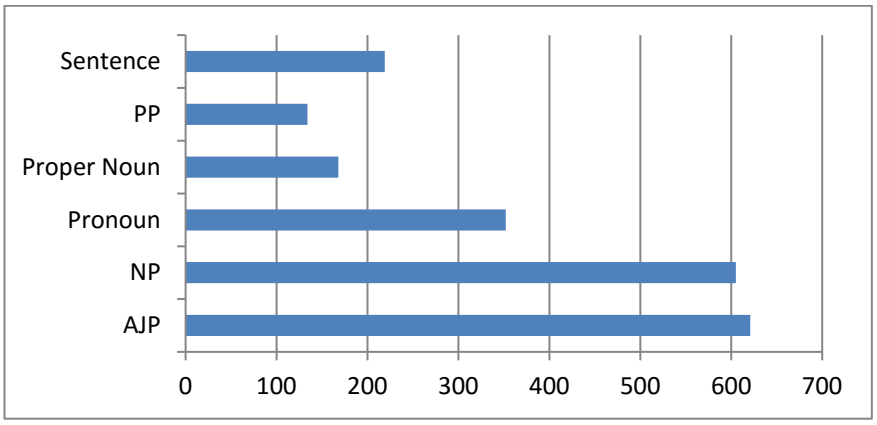

Fig. 7 Distribution of phrase types

Another factor important for the correct recognition of semantic relation was the type of phrase through which the role was expressed. Figure 7 presents the most common phrase types (PP – prepositional phrase, NP – Nouns Phrase, AJP – Adjectival Phrase, and Sentence – a subordinate clause). Thus, we observed that noun phrase were the most correctly identified types of phrases, while the roles expressed through Sentences were among the lowest recognized ones.

Most error cases were introduced by:

(1) incorrect mapping of semantic roles to their predicational word, in cases when more than one word appeared in the sentence;

(2) partial annotation of the semantic role, i.e. only the head of the constituent, not the whole constituent is selected;

(3) errors in generalization using WordNet, e.g. the pronoun `he` is generalized as `helium`.

Although it is expected (and true) that the frequency of the semantic role influence its correct recognition, another major factor was the length of the semantic role. Thus, the `duration` and `place` semantic roles, usually expressed through longer semantic roles, expressed as Sentences of prepositional phrases, have the lower recognition rate (see the table below).

Table 1. Average lengths of different phrase types

Phrase type	**Average no. of words)**	**Successful**
AJP	4,57	75.00
NP	6,27	68.27
Pronoun	4,00	94.44
Proper Name	2,50	92.50
PP	8,00	54.54
Sentence	12,57	9.09
AJP	4,57	75.00
NP	6,27	68.27
Pronoun	4,00	94.44

For instance, Manner semantic roles usually contain nominal or adverbial phrases, as well as long relative clauses. Therefore, the probability of generalizing them by finding a relevant hypernym in Wordnet decreased as the size of the semantic role increased.

5. Conclusion

This paper proposed a method for building a structured knowledge resource from user generated content. Our initial tests suggest that semantic role information can be used to automatically generate a knowledge resource. This pilot study needs to be extended to a larger scale, considering also other types of semantic roles.

The next obvious stage is to merge our resource to existing linked open data repositories.

As recent advances in information and communication technologies continue to reshape the relationship between governments and citizens, opportunities emerge at both ends. Citizens route their voices through new electronic channels, hoping to have their opinions heard at any time from any place. Thus, for content related to politics, we intend to add to VoxPopuli platform an application which makes use of technologies to allow governments and citizens alike to make the most of this explosion of user-generated content, by monitoring the social web for prediction of future “hot” topics.

Acknowledgements

The authors would like to thank <removed for reviewing> project for the financial support received while developing the platform proposed in this paper.

References

1. Cambria E. and B. White. 2014. Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence 9(2), 48-57
2. S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation Journal, Forthcoming.
3. Chua, Tat-Seng; Juanzi, Li; Moens, Marie-Francine (2014). Mining user generated content. Chapman and Hall/CRC. 2014.
4. Chun-Min Chen and Ling-Hwei Chen. A novel approach for semantic event extraction from sports webcast text. Multimedia Tools and Applications, vol. 71/3, pp 1937-1952, 2014.

5. Curteanu N. 2003. Contrastive meanings of the terms ”predicative” and predicational” in various linguistic theories (i, ii). Computer Science Journal of Moldova (R. Moldova), 11(4), 2003
6. Gildea Daniel and Daniel Jurafsky. “Automatic labeling of semantic roles”. Computational Linguistics, 28(3):245-288, 2002.
7. Go A., Bhayani R., Huang. L. (2009) Twitter Sentiment Classification using Distant Supervision, Technical Report.
8. S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In Proceedings of the Workshop on Languages in Social Media, LSM ’11, pages 20–29.
9. Hacioglu Kadri and Wayne Ward. “Target word detection and semantic role chunking using support vector machines”. In Proc. of HLT/NAACL-03, 2003
10. B. Han and T. Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT ’11, pages368–378, 2011.
11. Hoser B.and T. Nitschke. (2010) Questions on ethics for research in the virtually connected world. Social Net-works, 32(3):180–186, July 2010.DOI 10.1016/j.socnet.2009.11.003.
12. Iftene, A., Trandabăţ, D., Toader M., Corîci, M. 2011. Named Entity Recognition for Romanian in Studia Universitatis, Volume LVI, Number 2, pp. 19-24.
13. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Soc. for Inf. Science and Technology 60(11):2169-2188.
14. Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of ICWSM. 2011
15. Levin B. and M. Rappaport Hovav. Argument Realization. Research Surveys in Linguistics Series. Cambridge University Press, Cambridge, UK, 2005.
16. Andreea Macovei, Oana-Maria Gagea and Diana Trandabăţ (2016) Towards creating an ontology of social media texts, in Proceedings of RUMOUR2015, Springer CCIS/LNCS.
17. Misra Vidya Niwas. 1966. The Descriptive Technique of Panini. Mouton, The Hague.
18. Morante Roser, Walter Daelemans, and Vincent Van Asch. “A combined memory-based semantic role labeler of English”. In Proceedings of CoNLL, pp 208-212, 2008.
19. Nakov P., Ritter A., Rosenthal S., Stoyanov V., Sebastiani F.(2016) SemEval-2016 Task 4: Sentiment Analysis in Twitter, Proc. of SemEval '16.
20. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. (2002) "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of EMNLP, pp. 79-86.
21. Russell MA (2013) Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.
22. Schlaefer,N., J Chu-Carroll, E Nyberg, J Fan, W Zadrozny, D Ferrucci, "Statistical source expansion for question answering", Proceedings of CIKM 2011.
23. Radu Simionescu. 2011. Hybrid POS Tagger. In Proceedings of “Language Resources and Tools with Industrial Applications” Workshop (Eurolan 2011 summerschool).
24. Surdeanu M., S. Harabagiu, J. Williams, and P. Aarseth. “Using predicate-argument structures for information extraction”. In Proceedings of ACL2003, pp 8-15, Tokyo, 2003.
25. Surowiecki James (2005) The wisdom of crowds, published by Doubleday, ISBN: 0-385-50386-5
26. Trandabăţ Diana (2011) Mining Romanian texts for semantic knowledge, in Proceedings of ISDA2011, Cordoba, Spain, pp. 1062-1066.
27. Tufiș D, Radu Ion, Bozianu L, Ceaușu A., Ștefănescu D.. Romanian Wordnet: Current State, New Applications and Prospects. In Proceedings of Global WordNet Conference 2008, pp. 441-452, 2008.