Towards Building Knowledge Resources from Social Media Using Semantic Roles

Diana Trandabăț(✉)

University “Al. I. Cuza” of Iasi, Iași, Romania
dtrandabat@info.uaic.ro

Abstract. Text semantics is a well-hidden treasure, whose deciphering requires deep understanding. Artificial Intelligence enhances computers with human-like judgments, thus decoding the covered message and sharing it between machines is one of the main challenges that the computational linguistics domain faces nowadays. In an attempt to learn how humans communicate, computers use language models derived from human knowledge. While still far from completely understanding insinuated messages in political discourses, computer scientists and linguists have joined efforts in modeling a human-like linguistic behavior. This paper aims to introduce the *VoxPopuli* platform, an instrument to collect user generated content, to analyze it and to generate a map of semantically-related concepts to capturing crowd intelligence.

Keywords: Semantic roles · Knowledge resources · Social media

1 Introduction

Language technology is generally acknowledged today as one of the key growth areas in information technology. The META-NET White Paper series “Europe’s Languages in the Digital Age” [15], warned that languages may find it difficult to survive in the digital age, if the support for language technologies will not receive a boost. Building machine-readable knowledge bases takes a huge amount of time and resources, both financial and human (trained experts). Since today we found ourselves in an era in which software learns from its users and all of the users are connected, this paper proposes a natural language processing application which explores the social web in a new and innovative way, based on semantic frames, in order to extract the wisdom of crowds captured within.

With such knowledge bases, 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.

Social media refers in fact to Web 2.0 applications which support user content-creation and collaboration. People seek and share ideas, information, experiences, expertise, opinions, and emotion with both acquaintance and strangers on the Internet, based on the effect of the Wisdom of Crowds [13]. Over the last few years, the use of Social Media

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 585–591, 2017.
DOI: 10.1007/978-3-319-67008-9_50

has increased tremendously all over the world. Through *VoxPopuli*, people's contribution can reach a much wider audience than their small group of friends, by contributing to a "universal" 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.

The paper is structured as follow: Sect. 2 gives a short overview of the current state of the art in analyzing user generated content and semantic roles, while Sect. 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

Since the emerging of user generated contents (UGC), researchers have tried to automatically understand the opinions and sentiments that people are communicating [10, 11]. However, most analyses over social media were so far limited to identify user profiles or group behavior, extract sentiments expressed in specific posts, or identify topics in order to adapt recommendation systems. This paper is a position paper proposing the extraction of structured knowledge from social media using semantic frames, a direction yet unexplored.

Semantic roles [4] allow to identify when, where, why or how an event takes place, by clarifying the context of a sentence in terms of relations between the predicational word [3] and its semantic roles. The SRL system we propose: (a) is adapted for UGC (as opposed to existing systems, trained on news date); (b) incorporates, besides syntactic information, named entity recognition and topic information.

For extracting events and relations from texts, worth noticing is the work done by the Watson group at IBM on relationship extraction and snippets evaluation with applications to question answering [12], but also the work in [2], where semantic roles and event extraction are considered structurally identical tasks. Our approach extracts relations between concepts using semantic roles, similar up to some extend to the work in [2], but tailored for UGC.

3 Methodology

The architecture of the *VoxPopuli* platform involves 4 distinct modules (see Fig. 1), each of them specialized on a specific task and corresponding to a different objective.

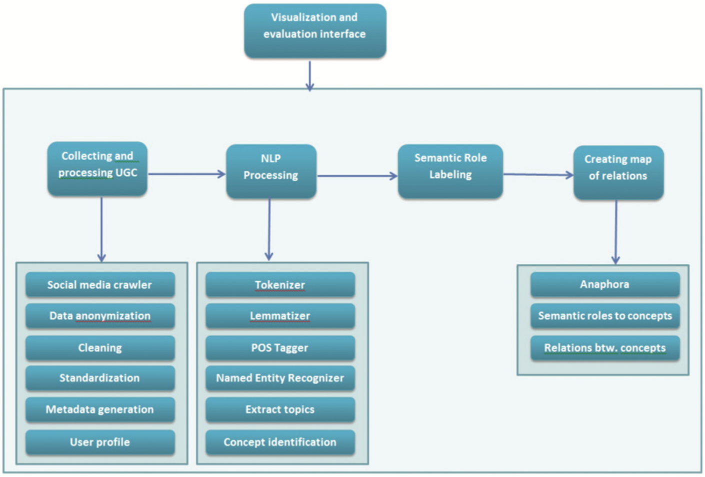

Fig. 1. Architecture of the *VoxPopuli* platform

3.1 Collecting and Pre-processing UGC

User-generated content (UGC) is defined 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*" [1].

The *VoxPopuli* platform collects textual UGC for analysis of individual and collective behavior, without accessing any personal data of users[1]. Privacy and copyright in using social media data is an open issue. Hoser and Nitschke [8] discuss the ethics of mining social networks, suggesting that researchers should not access personal data that users did not share for research purpose, 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 exist, allowing collection of social media data.

This module performs the following consecutive tasks:

- Identify User Generated Content (UGC) sources;
- Apply a social web crawler;
- Anonymyzing, cleaning and standardization;
- Metadata generation;
- Identify user profiles and classify user generated content types.

[1] According to the Directive 95/46/EC of the European Parliament and of the Council, personal data is defined as: "'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".

After identifying UGC sources (social media site, folksonomies, blogs and review sites, etc.), a collective 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.

VoxPopuli platform ensures no relation to a natural person is made from collected data, no personal data are stored or used, and that all texts are properly shuffled and anonymized, before being cleaned and standardized.

The standardization step is focuses 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 [7], coupled with location-based linguistic variations in shortening styles in microtexts [6].

The last step identifies of user types, depending on the most frequent concepts in a user's content and its writing style, based on the ontology proposed in [9].

3.2 NLP Processing

Once the collection of texts from UGC is created, data is explored using a series of NLP processes: tokenization, lemmatization, part-of-speech tagging, named entity recognition, sentiment analysis, topics and concept identification.

We tested our platform for the Romanian language, therefore specific NLP tools were applied, but the Platform is built to be language independent, allowing the inclusion of different language specific tools.

For sentiment analysis, *VoxPopuli* uses regular expressions to: convert the texts to lowercase; discard words shorter than two characters; remove special diacritic signs, URLs, as well as unsupported symbols (such as "?" or "@"); remove duplicated vowels in the middle of the words (e.g. cooooool). Subsequently, the sentiment is extracted by combining two methods: (1) a Naïve Bayes classifier, trained on Semeval 2016 data, using the following features: tokenized unigrams, emoticons, hash tags, similar to [5] and (2) the AlchemyAPI, applies to tweets to extract the expressed opinion.

The Topic Extraction module identifies different types of significant events using named entities, concepts extracted from the Romanian WordNet using hypernyms and hash tags. After the identification of topics, they are classified using a hybrid text classification model, combining statistical classification with rule-based filtering, identifying the thematic area of the message (transport, economy, daily life), alert situations (road accidents, fires, street violence), specific locations (building, means of transport) or events to which the text refers (cultural or sport events).

3.3 Semantic Role Labeling

The next module of *VoxPopuli* applies a semantic role labeling system. We adapted the SRL parser [14] developed for news texts, in order to cope with social media input, due to a set of challenging social media characteristics, syntactically and semantically different than the ones of the texts the role labeler is trained on: short messages, noisy content, temporal and social context, multilingual.

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.

3.4 Creating a Semantically-Related Map of Concepts

The most challenging module of *VoxPopuli* extracts individual and collective intelligence from UGC based on the semantic annotation. This modules populates a knowledge resource in three steps: (1) first, it extracts from each UGC the predicational words, for which semantic roles are annotated; (2) semantic frame analysis is used to extract relations between concepts (semantic roles) collocated with the specific predicate; and (3) the concepts found in the UGC in relation to a specific predicational word are mapped to the ones already introduced in the knowledge base (if any) using anaphora resolution and/or the WordNet (5) hyponymy hierarchy.

The concepts and their references are linked using a simple anaphora resolution method, based on a set of reference rules.

The created knowledge base, stored in RDF format, can be validated through a specialized interface. At this stage, we are still fine-tuning *VoxPopuli* platform, so we only validated a small number of relations (2000 relations).

4 Evaluation

For the evaluation of *VoxPopuli* platform, we analyses a set of 2000 relations extracted from user-generated content using semantic roles. The distribution of the total number of annotated roles per sentence is: 14% sentences with 3 annotated roles; 62% sentences with 2 roles; 24% sentences with 1 role. For the evaluation we only considered sentences having at most 3 annotated semantic roles, a limitation imposed in the testing version of our system which can be removed latter.

In this first stage of analyzing our platform, we focused on six semantic roles: Entity, Item, Manner Duration, Place and Time.

Figure 2 presents a distribution of the types of semantic roles. The overall accuracy for the identified relations is over 86%. 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 was selected; (3) errors in generalization using WordNet, e.g. the pronoun he is generalized as helium.

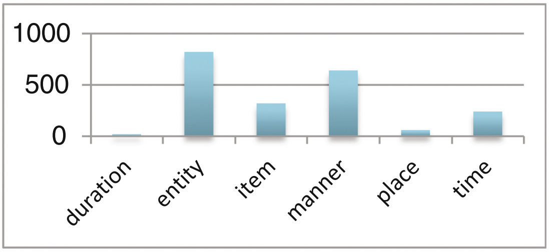

Fig. 2. Distribution of semantic roles

5 Discussion

The major contribution of this paper is a method for building a knowledge base from user generated content. Our results 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, including different types of semantic roles.

The next obvious stage is to merge our resource to existing linked open data repositories. We also intend to expand the *VoxPopuli* platform in order to extract similar map of relations from scientific literature in order to predict future "hot" research topics.

References

1. Chua, T.-S., Li, J., Moens, M.-F.: Mining User Generated Content. Chapman and Hall/CRC, Boca Raton (2014)
2. Chen, C.-M., Chen, L.-H.: A novel approach for semantic event extraction from sports webcast text. Multimed. Tools Appl. **71**(3), 1937–1952 (2014)
3. Curteanu, N.: Contrastive meanings of the terms "predicative" and "predicational" in various linguistic theories (i, ii). Comput. Sci. J. Moldova **11**(4), 2003 (2003)
4. Daniel, G., Jurafsky, D.: Automatic labeling of semantic roles. Comput. Linguist. **28**(3), 245–288 (2002)
5. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision, Technical report (2009)
6. Gouws, S., Metzler, D., Cai, C., Hovy, E.: Contextual bearing on linguistic variation in social media. In: Proceedings of Workshop on Languages in Social Media, LSM-2011, pp. 20–29 (2011)
7. Han, B., Baldwin, T.: Lexical normalisation of short text messages: makn sens a #twitter. In: Proceedings of the 49th ACL-HLT 2011, pp. 368–378 (2011)
8. Hoser, B., Nitschke, T.: Questions on ethics for research in the virtually connected world. Soc. Netw. **32**(3), 180–186 (2010). doi:10.1016/j.socnet.2009.11.003
9. Macovei, A., Gagea, O., Trandabăţ, D.: Towards creating an ontology of social media texts. In: Trandabăţ, D., Gîfu, D. (eds.) RUMOUR 2015. CCIS, vol. 588, pp. 18–31. Springer, Cham (2016). doi:10.1007/978-3-319-32942-0_2
10. Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V., Sebastiani, F.: SemEval-2016 task 4: sentiment analysis in Twitter. In: Proceedings of SemEval 2016 (2016)

11. Russell, M.A.: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More (2013)
12. Schlaefer, N., Chu-Carroll, J., Nyberg, E., Fan, J., Zadrozny, W., Ferrucci, D.: Statistical source expansion for question answering. In: Proceedings of CIKM (2011)
13. James, S.: The wisdom of crowds. Doubleday (ed.) (2005). ISBN: 0-385-50386-5
14. Diana, T.: Mining Romanian texts for semantic knowledge. In: Proceedings of ISDA 2011, Cordoba, Spain, pp. 1062–1066 (2011)
15. Trandabăţ, D., Irimia, E., Barbu, M.V., Cristea, D., Tufis, D.: The Romanian language in the digital age. In: White Paper Series, p. 87. Springer (2012). ISBN: 978-3-642-30702-7