Harvesting Knowledge from Social Networks: Extracting Typed Relationships among Entities

Andrea Caielli, Marco Brambilla, Stefano Ceri, and Florian Daniel

Politecnico di Milano, DEIB
Via Ponzio 34/5, I-20133, Milano, Italy
andrea.caielli@mail.polimi.it, {name.surname}@polimi.it

Abstract. Knowledge bases like DBpedia, Yago or Google's Knowledge Graph contain huge amounts of ontological knowledge harvested from (semi-)structured, curated data sources, such as relational databases or XML and HTML documents. Yet, the Web is full of knowledge that is not curated and/or structured and, hence, not easily indexed, for example social data. Most work so far in this context has been dedicated to the extraction of entities, i.e., people, things or concepts. This paper describes our work toward the extraction of relationships among entities. The objective is reconstructing a typed graph of entities and relationships to represent the knowledge contained in social data, without the need for a-priori domain knowledge. The experiments with real datasets show promising performance across a variety of domains.

Keywords: Social Networks, Relationship Extraction, Domain Graph

1 Introduction

In [3], we outlined a roadmap of work toward the identification and capturing of knowledge that is not yet contained in any well formalized knowledge base but only emerges from the observation of *social data* (data collected from social networks, such as Facebook, Twitter, Instagram). The problem is relevant, as understanding large volumes of social data is complex, and tools able to aid this understanding are still missing. The problem is timely, as it is no longer enough to describe a document only by the sentiment it expresses; it is important to also put that sentiment into context and to move toward comprehensive Social Media Analytics [8]. Finally, the problem is hard, as data in social networks is unstructured, ephemeral, and constantly changing.

In [4], we concentrated on the first building block, i.e., the semi-supervised extraction of *entities* from social data. In this paper we complement that work and report on our first experience with the extraction of *relationships*, able to put entities into context and to give meaning to the co-occurrence of entities inside a document. For instance, if we analyze the tweet in Figure 1, we are able to identify two entities and one relationship that allow us to draw a typed triple. If we do so for a set of documents, we are able to draw a complete *domain graph*, producing the desired output.

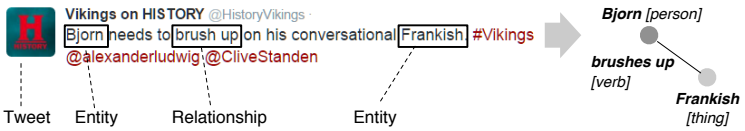

Fig. 1. Analysis of a tweet on the Vikings TV series

With this paper, we contribute to the state of the art with (i) an integrated social data processing pipeline able to extract typed entities and relationships from Facebook posts and tweets, and (ii) a set of experiments with real datasets that demonstrate the practical viability of the approach. The key distinguishing feature of the work is its focus on highly unstructured social data (tweets and Facebook posts) without reliable grammar structures. Traditional relation extraction approaches – supervised [7], semi-supervised [1] or unsupervised – [6], commonly assume the availability of grammatically correct language corpora.

2 Extraction of Relationships

We approach the extraction of relationships as follows: Social data is extracted from social networks using their APIs or scraping content from their HTML pages (in this work, we specifically concentrate on Twitter and Facebook). Collected data is analyzed for entities and for relationships using different techniques in parallel to increase quality. Once entities and relationships are available, they are consolidated so as to eliminate duplicates and errors and to form correct tuples with consistent entity and relationship types. After analyzing all documents, a dedicated graph viewer enables the user to interactively inspect the obtained graph and to drill down into details.

Before executing this process, we prepare the content so that it becomes most similar to correct natural language, by substituting special symbols with natural language tokens. Social data heavily leverage on # *hashtags* (for topics) and @ *handles* (for identities); we substitute them with their corresponding entities. We also drop *URLs* from the documents, as we don't analyse them (although they are heavily used in social media). Known *acronyms* are written in their full texts and *author names* are added.

Extracting Entities. The extraction of entities leverages on Dandelion (`https://dandelion.eu`) and the Named Entity Recognizer (NER) of the Stanford coreNLP library (`http://nlp.stanford.edu/software/CRF-NER.shtml`). The former is based on DBpedia and enables the identification of entities contained in DBpedia. The latter is able to identify entities by analyzing the grammar structure of sentences. Both instruments are fed with the pre-processed data, and outputs are consolidated into one set of entities. After integration, entities are identified with good precision (see below).

Extracting Relationships. The extraction of relationships leverages on coreNLP OpenIE (`http://nlp.stanford.edu/software/openie.html`) and a purpose-

fully designed extension (heuristic) inspired by the work of Bird, Klein and Loper [2]: subjects and objects in subject–relationship–object triples identified by OpenIE are associated with an abstract "thing" type if OpenIE fails to identify a proper type. This enables identifying triples for cases where OpenIE would fail and deciding which triple is best if OpenIE extracts multiple conflicting triples for a given document. In addition, using the linguistic tokenization of coreNLP we extract noun-predicate-noun triples by applying pre-defined templates. The relationships identified by the two methods are again combined to avoid repetitions. The integration is based on relationship similarity and containment and analyzes the verbs, subjects and objects, giving preference to the most expressive relationships (containing the others).

Integrating and Typing Relationships. A good domain graph requires typed relationships. This is achieved by means of two complementary techniques: First, all identified verbs are clustered into synonym classes using wordnet-magic (`https://www.npmjs.com/package/wordnet-magic`), a node.js module for WordNet (`https://wordnet.princeton.edu`). Second, verbs are categorized linguistically using VerbNet [5] and by looking for the membership in classes of the verb describing a relationship. We use both techniques and consolidate identified types.

3 Evaluation and Lessons

We ran the described relationship extraction process on five different datasets with documents retrieved from Facebook and Twitter (see the used Twitter/Facebook handles and hashtags between parentheses): *Black Sails* (#BlackSails, @BlkSails_STARZ, @blacksails.starz), *Teen Wolf* (# TeenWolf, @MTVteenwolf, @TeenWolf), *Vikings* (#Vikings, @Vikings, @Vikings), the *Milan Fashion Week 2016* (#MFW, Twitter only), and *Rugby* (#AsOne, #RBS6nations, @rbs_6 _nations, @rbs6nations). Table 1 reports some statistics about the datasets: *# docs* tells the number of documents extracted per domain, *# entities* and *# rels* the number of entities and relationships extracted, and *# verb rels* the number of relationships for which meaningful descriptive verbs could be identified (types of relationships). Figure 2 exemplifies the types of relationships extracted for the Rugby dataset using synonym classes: besides the predominance of "be" and "have," the identified verbs provide effective insight into the typical terminology of the domain.

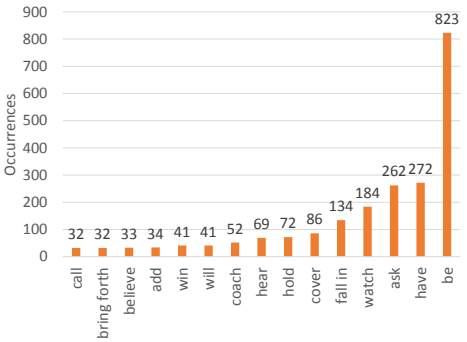

Fig. 2. Relationship types identified for the Rugby dataset

In order to assess the precision ($P = \frac{TP}{TP+FP}$) and recall ($R = \frac{TP}{TP+FN}$) of the extraction process, we randomly picked 100 samples from each domain,

Table 1. Description of datasets with P/R of algorithm

	Dataset				
	Black Sails	**Teen Wolf**	**Vikings**	**Milan Fashion Week**	**Rugby**
# docs	2495	2346	1969	1136	1796
# entities	1243	1045	978	1157	1558
# rels	2025	1549	1378	2311	5356
# verb rels	66	81	146	288	437
P(entities)	83.7%	76.0%	78.8%	67.3%	73.1%
R(entities)	79.9%	72.9%	76.3%	59.4%	74.1%
P(rel)	73.8%	75.8%	71.4%	54.7%	71.4%
R(rel)	92.1%	90.2%	82.1%	82.7%	95.2%

manually created a ground truth of relationships, and manually labeled the automatically extracted relations as true positive (TP), false positive (FP) or false negative (FN). The second part of Table 1 plots the results for the five datasets, distinguishing P and R for entities only and for complete relationships.

The results show that the joint use of syntactic techniques (that identify subject-predicate-object triples) and semantic techniques (that also require that entities and relationships be typed) produces good precision and recall. Precision is above 70% in all the use cases except the "Milano Fashion Week" – where however we miss social data from Facebook. Recall is in all cases above 80% and goes up to 95% in the case of Rugby; therefore, in our method false negatives are very few: all tokens which are labeled as relationships indeed correspond to relationships. Note that we obtain higher recall on relationships that on entities (therefore, we may miss some entities, but once they are understood then relationships are normally understood).

If we look at the quality of extracted relations, we found stronger results in the case of Rugby, due to the higher quality of tweets and posts, which commented aspects of the game and made statements about reality; in the case of TV series, with comparatively lower precision and recall, many tweets and posts expressed just comments or sentiments unrelated to the actual content of the series.

References

1. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.
2. S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. "O'Reilly Media, Inc.", 2009.
3. M. Brambilla, S. Ceri, F. Daniel, and E. D. Valle. On the quest for changing knowledge. pages 3:1–3:5, 2016.
4. M. Brambilla, S. Ceri, E. Della Valle, R. Volonterio, and F. Acero Salazar. Extracting Emerging Knowledge from Social Media. In *WWW 2017*, 2017, in print.
5. C. Gardent. Doing things with meanings. 2005.
6. H. Poon and P. Domingos. Unsupervised ontology induction from text. In *ACL 2010*, pages 296–305, 2010.
7. S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.
8. S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger. Social media analytics. *Business & Information Systems Engineering*, 6(2):89–96, 2014.