Neural Networks 58 (2014) 82–97

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2014 Special Issue

Sentiments analysis at conceptual level making use of the Narrative Knowledge Representation Language

Gian Piero Zarri *

Sorbonne University – STIH Laboratory, 1, rue Victor Cousin – 75005 Paris, France

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Behaviors
Representation of deep sentiments
Narratives
Standard ontologies
Ontologies of events
Querying and inference techniques

ABSTRACT

This paper illustrates some of the knowledge representation structures and inference procedures proper to a high-level, fully implemented conceptual language, NKRL (Narrative Knowledge Representation Language). The aim is to show how these tools can be used to deal, in a sentiment analysis/opinion mining context, with some common types of human (and non-human) "behaviors". These behaviors correspond, in particular, to the concrete, mutual relationships among human and non-human characters that can be expressed under the form of non-fictional and real-time "narratives" (i.e., as logically and temporally structured sequences of "elementary events").

1. Introduction

"Sentiment analysis" (or "opinion mining") concerns all the possible computer-based applications that try to identify and extract "subjective information" (opinions, beliefs, emotional states and views about specific entities) from source materials, usually represented in textual form (Feldman, 2013; Westerski, 2007). Related disciplines are "behavior computing" – or "behavioral informatics" (Cao and Yu, 2012) – and "affective computing" (Ahn, 2010). Most common, practical sentiments analysis applications are in the area of reviews of consumer products and services.

The research tools used in the sentiment analysis domain consist normally of computational linguistics and text mining techniques that perform some sort of "surface" analysis of the original sources in order, e.g., to determine the "negative/positive polarity" of words or sentences, recognizing the presence of words or expressions within specific sentiment lexica, detecting sentences that contain comparative opinions, etc. In this paper, we suggest that these surface techniques, often strongly statistically-oriented, could be usefully complemented by "deep" conceptual analysis tools aiming at describing, in sufficient detail, the behaviors (according to the most general meaning of this term) and the mutual relationships of the (human and non-human) characters that appear in the original natural language documents. To this end, this paper focuses on the conceptual representation tools proper to a (wholly implemented) knowledge representation language and computer system environment, NKRL, the Narrative Knowledge Representation Language (Zarri, 2009).

In a nutshell, the term *narrative* denotes a general unifying framework used for relating real-life or fictional stories (novels, tales…) involving the *common relationships* between concrete or imaginary characters. Narratives deal then, among other things, with those opinions, beliefs, emotional states and viewpoints about specific entities that, as already stated, represent the *basic, raw material* used to perform the sentiment analysis operations. Narratives are normally conveyed by NL supports as, in a *non-fictional* context, news stories, corporate memory documents (memos, reports, minutes…), normative and legal texts, medical records, etc. However, they can also be represented by multimedia documents like audio records, surveillance videos, actuality photos for newspapers and magazines, etc. A photo representing President Obama addressing the Congress, or a short video showing three nice girls chattering on a beach, must be considered as "narrative" documents even if they are not, of course, NL documents. A narrative is a *highly-dynamic entity*, since it can be synthetically defined as a *sequence of logically structured and temporally and spatially bounded "elementary events" (a non-linear 'stream' of elementary events)*. An "elementary event" corresponds in turn to the conceptual representation of the bundle of mutual relationships among characters associated with a *single* "generalized predicate" isolated within the natural language formulation of the whole stream. Generalized predicates correspond not only to the usual tensed/untensed "verbs", but also to "adjectives" ("…worth several dollars…", "…a dormant volcano…"), nouns ("…Jane's amble along the park…", "…a possible attack…), etc., when they have a predicative function.

* Tel.: +33 1 40463288.
E-mail addresses: zarri@noos.fr, gian-piero.zarri@u-pec.fr.

http://dx.doi.org/10.1016/j.neunet.2014.05.010
0893-6080/

To justify the use within the sentiment analysis/opinion mining domain of formal tools created for the analysis of "narrative" documents, let us examine briefly other "*conceptual*" – as opposed to pure statistical – approaches used in this domain. For example, the so-called "*sentiment (or opinion) lexica*" can be defined in general as lists of *words and expressions* used to denote people's subjective feelings and sentiments/opinions ("negative" or "positive" *prior polarities*). The term "expressions" is used here to denote not just individual words, but also phrases and idioms. These lexica can be built up according to three main ways, a manual approach (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), a corpus-based approach that relies on the detection of syntactic patterns in large corpora (Ding, Liu, & Yu, 2008; Kaji & Kitsuregawa, 2007) and a dictionary-based approach. Lexica pertaining to this last category are often developed by making use of WordNet's synsets and hierarchies to acquire opinion words, see in this context, e.g., WordNet-Affect (Strapparava & Valitutti, 2004) and SentiWordNet (Esuli & Sebastiani, 2006). WordNet-Affect was developed through the selection and labeling, using the terms included in a specific hierarchy of "affective domain labels", of the WordNet synsets representing affective concepts. SentiWordNet is a version of WordNet where the independent values "positive", "negative", and "objective" are associated with 117,660 WordNet's synsets. Each of the three values ranges from 0.0 to 1.0, and their sum is 1.0 for every synset.

In a "sentiment lexica" context, one of the most well-known and advanced approaches is represented by SenticNet. This system exists in three versions of increasing complexity, SenticNet 1 (Cambria, Speer, Havasi, & Hussain, 2010), SenticNet 2 (Cambria, Havasi, & Hussain, 2012) and SenticNet 3 (Cambria, Olsher, & Rajagopal, 2014). Partially inspired from WordNet-Affect and SentiWordNet, SenticNet makes use of the so-called "*sentic computing*" *approach*. This is a new paradigm that exploits both AI and Semantic Web techniques to recognize, interpret, and process natural language opinions going beyond a simple "syntactic" strategy. In its version 2 for example, it provides the semantics and the "*sentic information*" – i.e., the *cognitive* and *affective* information – that concern over 14,000 concepts. Unlike SentiWordNet, SenticNet discards concepts with neutral or almost neutral polarity, i.e., concepts with polarity magnitude close to zero. Moreover, while SentiWordNet stores three values for each synset, SenticNet associates each concept $\boldsymbol{c}$ with just one value $\boldsymbol{p_c}$, i.e., a float in the range $[-1, 1]$ representing its polarity. This choice allows SenticNet to avoid redundancy and facilitates its representation as a (ConceptNet, see below) semantic network. In SenticNet, eventually, concepts like **make good impression**, **look attractive**, **show appreciation** or **good deal** are likely to have a $\boldsymbol{p_c}$ very close to 1 while concepts such as **being fired**, **leave behind** or **lose control** are likely to have $\boldsymbol{p_c} \approx -1$ (Cambria et al., 2010: 16). An important, common characteristic of the three SenticNet versions concerns the fact that their 'basic knowledge' derives from ConceptNet (Liu & Singh, 2004; Speer & Havasi, 2012), a semantic network built up from *nodes* representing *concepts* in the form of *words or short phrases in natural language* and from *labeled relationships between them*. The relationships (21, including the standard **IsA**) are in the form of, e.g., **CreatedBy**, **PartOf**, **UsedFor**, **PrerequisiteOf**, **DefinedAs**, **LocatedNear**. Thus, ConceptNet knowledge is mainly associated with *general compound concepts* instead of single words/concepts. The compound concepts are represented in semi-structured English by composing, using the labeled relationships, a verb/concept with a noun phrase/concept or a prepositional phrase/concept. (Recursive) compound concepts can then be, e.g., "**[*wake up in the morning*] PrerequisiteOf [*eat breakfast*]**", "**[*kitchen table*] UsedFor [*eat breakfast*]**", "**[*chair*] LocatedNear [*kitchen table*]**", etc. (Liu & Singh, 2004: 213).

Independently from the formal semantic/syntactic details, the *knowledge* included in all the systems mentioned above has in common the fact of being, basically, a sort of *terminological/definitional knowledge*. It denotes, then, some *stable, self-contained, a priori and basic notions/concepts* that can be considered, *at least in the short term*, as '*a-temporal*' (or '*static*') and '*universal*'. This means that their definitions *are not subject to change*, at least within the framework of a given application, even if they can evolve *in the long term* as a consequence, e.g., of the progress of our knowledge or of criticisms/comparisons with different approaches. These static notions can be very general, see concepts like **human being**, **color** or **chair** that are proper to several application domains, or linked to more specific contexts as **make person happy**, **feel guilty** or **shed tears** in a sentiment analysis environment.[1]

The self-contained and stable character of this terminological/definitional knowledge (where, as stated above, the temporal phenomena can be ignored) justifies the use of a relatively simple *formal model* for its conceptual representation/definition. This can be limited to the description of some main properties—sometimes, only the use of the genus/species **IsA** relationships is actually required. This formal model can then correspond to the usual binary one, where properties are simply expressed as a *binary* (i.e., accepting only two arguments) relationship linking two individuals or an individual and a value. And this independently from the fact that these binary relationships are organized into, e.g., frame format as in the original Protégé software (Noy, Fergerson, & Musen, 2000) or take the form of a set of "property" statements used to define a "class" (a "concept") in some W3C language. In a sentiment analysis/opinion mining framework we can note that, accordingly, WordNet 3 is now represented in (*binary*) RDF/W3C format; RDF is also used in a ConceptNet 5 environment and to encode the nodes of the SenticNet network.

In the context of the "*narrative information*" analysis evoked above and of similar applications, the *main* knowledge to be dealt with corresponds, on the contrary, to a sort of particularly *complex and "structured"* information. This type of knowledge denotes, in fact, the *dynamic, interpersonal, often accidental and unpredictable, spatio-temporal characterized behavior* proper to *specific subsets* of the *terminological/definitional entities* examined above. Examples of this sort of dynamic/structured knowledge that can be of interest in a sentiment analysis/opinion mining environment correspond, e.g., to the description of "*elementary events*" in the style of "On November 17, 2003, in an unspecified location in Afghanistan, an armed group of people shot a woman dead", "Yesterday, John gave a book to Mary for her birthday", "Peter has recently bought his first iPhone in the Carrousel Apple Store of Paris", "On November 20, 1999, in Sulu province, the family of the kidnapped journalist was asked for a ransom", "On August 8, 2012, at Beta Bank's premises, Mary Collins fired John Smith", "Tom returned his new Ultrabook yesterday", etc. In a "structured/dynamic" context, then, some *static, terminological/definitional entities* ("John", "Mary", "woman", "present"',

[1] We can note that this terminological/definitional knowledge coincides largely with the "*common knowledge*" as defined, e.g., in Cambria, Olsher, et al. (2014) and Cambria and White (2014). More precisely, Cambria and his colleagues make a distinction between "*common knowledge*" and "*common-sense knowledge*". The first corresponds to general knowledge about the world, e.g., "a chair is a type of furniture". On the other hand, common-sense knowledge denotes "…accepted things that people normally know about the world but which are usually left unstated in discourse, e.g., that *things fall downwards* (*and not upwards*) and *people smile when they are happy*" (Cambria & White, 2014: 51). The two types can be both classified as static, a priori, a-temporal knowledge as the terminological/definitional knowledge introduced above. In an NKRL context we prefer, however, to think about the common-sense knowledge as that "*operational knowledge*" definitely needed for setting up useful *inference rules*, see Section 4 below.