HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

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

Meaning-based machine learning for information assurance

*Courtney Falk**, *Lauren Stuart*

Purdue University, West Lafayette, IN, United States

HIGHLIGHTS

- Describes the combination of semantic knowledge bases with machine learning.
- Natural language processing application for phishing detection.
- Semantic machine learning improves on existing approaches.

ARTICLE INFO

Article history:
Published online 23 November 2016

Keywords:
Natural language processing
Machine learning
Information security

ABSTRACT

This paper presents meaning-based machine learning, the use of semantically meaningful input data into machine learning systems in order to produce output that is meaningful to a human user where the semantic input comes from the Ontological Semantics Technology theory of natural language processing. How to bridge from knowledge-based natural language processing architectures to traditional machine learning systems is described to include high-level descriptions of the steps taken. These meaning-based machine learning systems are then applied to problems in information assurance and security that remain unsolved and feature large amounts of natural language text.

1. Introduction

This paper outlines a research program called meaning-based machine learning (MBML). MBML combines the meaningful input provided by ontological semantics with the pattern searching abilities of established machine learning.

First, the paper explains the novelty of MBML and establishes how it interconnects with different fields.

Second, the end-to-end data flow of an MBML system is described. Special attention is paid to leveraged established formalisms from ontological semantics.

Finally, there is a discussion of how this general MBML approach is applicable to problems of information assurance. The problems of phishing detection and stylometry are addressed in-depth.

1.1. *Machine learning*

Machine learning (ML), particularly statistical ML, has matured and grown in popularity over the past decade for natural language processing (NLP) applications. Some, but not necessarily all, of the most popular ML approaches

Peer review under responsibility of Qassim University.

* *Corresponding author.*
E-mail address: courtney.falk@gmail.com (C. Falk).

http://dx.doi.org/10.1016/j.jides.2016.10.007

center around statistical techniques [1]. Performance of these statistical methods improve with larger amounts of well-annotated data.

Different ML approaches attempt delve below surface language features such as word frequency and syntactic structure into semantic meaning with varying levels of success. Whether or not statistical approaches can identify semantic information remains an open question that is outside the scope of this paper. Instead, the MBML approach described in detail later on will start from the position of using semantically meaningful data derived from an ontological semantics system. It is the position of the authors that only by beginning with semantic data as the input will the output resemble anything approaching what humans understand to be semantically meaningful.

It is always worth noting that the sense in which the aforementioned statistical ML systems use the word "semantics" differs from the "semantics" of ontological semantics. In the former sense "semantics" describes a structure that is sufficiently complex to example the observed data while in the latter sense "semantics" describes the philosophical, linguistic, and cognitive models of meaning.

1.2. *Ontological semantics technology*

Ontological Semantics Technology (OST) is a development of the theory of Ontological Semantics [2]. Ontological Semantics first began to be formalized as a comprehensive system with the Mikrokosmos project [3] before it was described in detail in the text of the same name [4].

At its core, ontological semantics is a frame-based system [5] where language-dependent lexicons define syntactic behavior and extend the semantic concepts stored in the language-independent ontology. The development of these resources (the lexicons, other language-specific knowledge repositories or tools, the ontology, and other language-independent knowledge repositories or tools) is named acquisition; its practitioners are acquirers [4].

The process of acquisition involves the careful description of linguistic-semantic behaviors and distinctions, as observed or theorized in human use of language, via the OST framework. The two basic resources, the lexicon and the ontology, are the two we will discuss in depth here because the details of their specification and intended use most impact the array of features we wish to introduce. Other elements in the ecology of OST are described elsewhere.

The ontology is a large, dense graph of nodes, called concepts, connected by relations. A concept represents a separable, cohesive meaning unit, such as *automobile*, *travel*, *rice*, or *freedom*. Relations provide relative information for concepts; they have a domain (originating concept), and range (target concept, literal, or scalar) by which additional information is encoded. The strength of an ontology is in its dense connections between concepts: the use of a *automobile* for a *human* in an instance of *travel* is modeled by appropriately-restricted (loose enough to make semantic distinctions where actual text does, but tight enough to reduce sense-making where actual text would not) relations (where *human* is the *AGENT* of *travel*), along which some very basic reasoning can be performed. The methods and directions of such reasoning become application-specific (for instance, in detecting and flagging possible instances of insider threat) but OST assumes a reusable kernel of these, that we also assume here to be in any OST implementation regardless of application.

The lexicon provides the first mapping from word (or other separable part of a text or utterance) to concept, relation, attribute, or graph of these. A lexicon entry gives, for each sense of a word, the base lexeme, morphological rules, syntactic and grammar rules and representation, and semantic representation. This semantic representation specifies the ontological concepts, relations, or literals that express the meaning of the lexeme. In text processing, each word (or phrasal set of words, in the case of common multiple-word expressions with non-compositional semantics) is queried in the lexicon, which gives one or several sets of morphological, syntactic, and semantic dependencies to be resolved in assembling the semantic map of the text's meaning. (Some special cases may be handled instead by other lookup-type elements of OST; for example, proper names are stored in a separate resource, the onomasticon, and have some other considerations for how they show up in the map.)

OST processes a text into TMRs, text meaning representations. A TMR constitutes a modified subgraph of the ontology, encoding information that has been explicitly or implicitly called out in the text. The granularity is application-determined: some applications may find that a one-to-one sentence-to-TMR transformation is all that is needed or can be done with what is available, and some may operate on a whole text and produce one large and complicated TMR. It is this graph of concepts, relations, and literals that we use as the input for MBML.

1.3. *Information assurance and security*

Information assurance and security (IAS) are ripe fields for NLP applications [6–8]. Because natural language remains an unsolved problem and yet is central to how humans use technology it is an important research area for IAS.

Semantically meaningful results in NLP can offer new insight into text-heavy domains such as social network analysis, business intelligence, and social engineering detection. As in [7], we use our Section 3 to explore a few problem areas in information assurance and security in which we have noted a need.

1.4. *What is meaning-based machine learning?*

MBML bridges disciplines. It begins in the realm of ontological semantics and uses techniques popularized by machine learning (ML) to find patterns in meaningful data. For an MBML system that relies on OST the meaning is represented in the TMRs. ML techniques examining these meaningful TMRs will in turn derive meaningful results from the TMRs.

The kinds of patterns in TMRs varies. Different linguistic phenomena aren't necessarily represented solely in the text itself. Novelty of information and referencing information across documents assume a certain level of background knowledge. It is in areas such as these that ML algorithms, operating on the TMR structures generated by OST, that ML might add new layers of meaning by building on the existing meaning described by OST.