T

Text Mining for the Semantic Web

Marko Grobelnik[1], Dunja Mladenić[2], and Michael Witbrock[3]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Department for Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia
[3]Cycorp Inc, Austin, TX, USA

Definition

▸ Text Mining methods allow for the incorporation of textual data within applications of semantic technologies on the Web. Application of these techniques is appropriate when some of the data needed for a Semantic Web use scenario are in textual form. The techniques range from simple processing of text to reducing vocabulary size, through applying shallow natural language processing to constructing new semantic features or applying information retrieval to selecting relevant texts for analysis, through complex methods involving integrated visualization of semantic information, semantic search, semiautomatic ontology construction, and large-scale reasoning.

Motivation and Background

Semantic Web applications usually involve deep structured knowledge integrated by means of some kind of ontology. Text mining methods, on the other hand, support the discovery of structure in data and effectively support semantic technologies on data-driven tasks such as (semi)automatic ontology acquisition, extension, and mapping. Fully automatic text mining approaches are not always the most appropriate for combination with Semantic Web content, because often it is too difficult or too costly to fully integrate the available background domain knowledge into a suitable representation. For such cases, semiautomatic methods, such as ▸ Active Learning and ▸ Semisupervised Text Processing (see ▸ Semisupervised Learning), can be applied to make use of small pieces of human knowledge to provide guidance toward the desired ontology or other models. Application of these semiautomated techniques can reduce the amount of human effort required to produce training data by an order of magnitude while preserving the quality of results.

To date, Semantic Web applications have typically been associated with data, such as text documents, and corresponding metadata that have been designed to be relatively easily manageable by humans. Humans are, for example, very good at reading and understanding text and tables. General semantic technologies, on the other hand, aim more broadly at handling data modalities including multimedia, signals from emplaced or remote sensors, and the structure and content of communication and transportation graphs and networks. In handling such multimodal data, much of which is not readily comprehensible by unaugmented humans, there must be signifi-

C. Sammut, G.I. Webb (eds.), *Encyclopedia of Machine Learning and Data Mining*,
DOI 10.1007/978-1-4899-7502-7_835-1

cant emphasis on fully or semiautomatic methods offered by knowledge discovery technologies whose application is not limited to a specific data representation (Grobelnik and Mladenic 2005).

Data and the corresponding semantic structures change over time, and semantic technologies also aim at adapting the ontologies that model the data accordingly. For most such scenarios, extensive human involvement in building models and adapting them according to the data is too costly, too inaccurate, and too slow. Stream mining (Gaber et al. 2005) techniques (Data Streams: Clustering) allow text mining of dynamic data (e.g., notably in handling a stream of news or of public commentary).

Ontology is a fundamental method for organizing knowledge in a structured way and is applied, along with formalized reasoning, in areas from philosophy to scientific discovery to knowledge management and the Semantic Web. In computer science, an ontology generally refers to a graph or network structure consisting of a set of concepts (vertices in a graph), a set of relationships connecting those concepts (directed edges in a graph), and, possibly, a set of distinguished instance concepts assigned to particular class concepts (data records assigned to vertices in a graph). Although much useful knowledge can be represented by the ground binary relations most conveniently encoded as graphs, more complex relationships involving more than two entities are needed, and the graph metaphor is more remote. In many cases, knowledge is structured in one of these ways to allow for automated inference based on a logical formalism such as the predicate calculus (Barwise and Etchemendy 2002); for these applications, an ontology often further comprises a set of rules or produces new knowledge within the representation from existing knowledge. An ontology containing both instance data and rules for its application is often referred to as a knowledge base (KB) (e.g., Lenat 1995).

Machine learning practitioners refer to the task of automatically constructing these ontologies as ontology learning. From this point of view, an ontology is seen as a class of models – somewhat more complex than most used in machine learning – which need to be expressed in some ▸ Hypothesis Language. This definition of ontology learning (from Grobelnik and Mladenic 2005) enables a decomposition into several machine learning tasks, including learning concepts, identifying relationships between existing concepts, populating an existing ontology/structure with instances, identifying change in dynamic ontologies, and inducing rules over concepts, background knowledge, and instances.

Following this scheme, text mining methods have been applied to extending existing ontologies based on Web documents, learning semantic relations from text based on collocations, semiautomatic data-driven ontology construction based on document clustering and classification, extracting semantic graphs from text, transforming text into RDF triples (a commonly used Semantic Web data representation), and mapping triplets to semantic classes using several kinds of lexical and ontological background knowledge. Text mining is also intensively used in the effort to produce a Semantic Web for annotation of text with concepts from ontology. For instance, a text document is split into sentences, each sentence is represented as a word vector, sentences are clustered, and each cluster is labeled by the most characteristic words from its sentences and mapped upon the concepts of a general ontology. Several approaches that integrate ontology management, knowledge discovery, and human language technologies are described in Davies et al. (2009).

Extending the text mining paradigm, efforts are aimed at giving machines an approximation of the full human ability to acquire knowledge from text. Some of the systems (Curtis et al. 2009; Mitchell 2005; Rusu 2014) actively use background knowledge in the extraction process for disambiguation or knowledge structuring. Machine reading aims at full text understanding by integrating knowledge-based construction and use into syntactically sophisticated natural language analysis, leading to systems that autonomously improve their ability to extract further knowledge from text (e.g., Curtis et al. 2009; Etzioni et al. 2007; Mitchell 2005; Starc and Fortuna 2012; Starc and Mladenic 2013).

Biomedical Text Mining

Because of the development and widespread use of high-quality biomedical knowledge bases, such as the Gene Ontology (Ashburner et al. 2000), Cell Ontology (Bard et al. 2005), and Linked Neuron Data (Zeng et al. 2015), and the overwhelming volume of the relevant literature (24 million biomedicine citations in PubMed), biomedical knowledge extraction is subject to a great deal of research. Relevant shared evaluation tasks include BioCreative (Hirschman et al. 2005) and BioNLP (Cohen et al. 2014). Although much of the work on biological fact extraction still relies on supervised training with closely annotated training data, with the risk of over-constraining the mapping of semantics to particular text substrings, volume of high-quality Semantic Web fact bases has enabled more natural training methods, such as distant supervision (Augenstein et al. 2014).

Cross-References

- Active Learning
- Classification
- Document Clustering
- Semisupervised Learning
- Semisupervised Text Processing
- Text Mining
- Text Visualization

Recommended Reading

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29

Augenstein I, Maynard D, Ciravegna F (2014) Relation extraction from the web using distant supervision. In: Janowicz K et al (eds) EKAW 2014. LNAI 8876. Springer, pp 26–41

Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biol 6(2):R21

Barwise J, Etchemendy J (2002) Language proof and logic. Center for the study of language and information. ISBN:157586374X

Buitelaar P, Cimiano P, Magnini B (2005) Ontology learning from text: methods, applications and evaluation, frontiers in artificial intelligence and applications. IOS Press, Amsterdam

Cohen K, Demner-Fushman D, Ananiadou S, Tsujii J-i (2014) Proceedings of BioNLP 2014, Baltimore. Association for Computational Linguistics

Curtis J, Baxter D, Wagner P, Cabral J, Schneider D, Witbrock M (2009) Methods of rule acquisition in the TextLearner system. In: Proceedings of the 2009 AAAI spring symposium on learning by reading and learning to read. AAAI Press, Palo Alto, pp 22–28

Davies J, Grobelnik M, Mladenić D (2009) Semantic knowledge management. Springer, Berlin

Etzioni O, Banko M, Cafarella MJ (2007) Machine reading. In: Proceedings of the 2007 AAAI spring symposium on machine reading

Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Mining data streams: a review. ACM SIGMOD Rec 34(1):18–26. ISSN:0163-580

Grobelnik M, Mladenic D (2005) Automated knowledge discovery in advanced knowledge management. J Knowl Manag 9:132–149

Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinform 6(Suppl 1):S1

Lenat DB (1995) Cyc: a large-scale investment in knowledge infrastructure. Commun ACM 38(11):33–38

Mitchell T (2005) Reading the web: a breakthrough goal for AI. Celebrating twenty-five years of AAAI: notes from the AAAI-05 and IAAI-05 conferences. AI Mag 26(3):12–16

Rusu D (2014) Text annotation using background knowledge. Doctoral Dissertation, Jozef Stefan International Postgraduate School, Ljubljana

Starc J, Fortuna B (2012) Identifying good patterns for relation extraction. In: Proceedings of the 15th international multiconference information society – IS 2012. Institut Jožef Stefan, Ljubljana, pp 205–208

Starc J, Mladenic D (2013) Semi-automatic construction of pattern rules for translation of natural language into semantic representation. In: Proceedings of the 5th Jožef Stefan International Postgraduate School Students Conference, Jožefa Stefana International Postgraduate School, pp 199–208

Zeng Y, Wang D, Zhang T Linked brain data. Web http://www.linked-neuron-data.org/. Retrieved 11 Jan 2015