Knowledge-Based Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Editorial

Q1

New avenues in knowledge bases for natural language processing

Between the birth of the Internet and 2003, year of birth of social networks such as MySpace, Delicious, LinkedIn, and Facebook, there were just a few dozen exabytes of information on the Web. Today, that same amount of information is created weekly. The advent of the Social Web has provided people with new content-sharing services that allow them to create and share their own contents, ideas, and opinions, in a time- and cost-efficient way, with virtually millions of other people connected to the World Wide Web. This huge amount of information, however, is mainly unstructured (because it is specifically produced for human consumption) and hence not directly machine-processable. The automatic analysis of text involves a deep understanding of natural language by machines, a reality from which we are still very far off.

Hitherto, online information retrieval, aggregation, and processing have mainly been based on algorithms relying on the textual representation of webpages. Such algorithms are very good at retrieving texts, splitting them into parts, checking the spelling and counting the number of words. When it comes to interpreting sentences and extracting meaningful information, however, their capabilities are known to be very limited, as most of the existing approaches are still based on the syntactic representation of text, a method that relies mainly on word co-occurrence frequencies. Such algorithms are limited by the fact that they can process only the information that they can 'see'. As human text processors, we do not have such limitations as every word we see activates a cascade of semantically related concepts, relevant episodes, and sensory experiences, all of which enable the completion of complex natural language processing (NLP) tasks – such as word-sense disambiguation, textual entailment, and semantic role labeling – in a quick and effortless way.

Knowledge-based NLP focuses on the intrinsic meaning associated with natural language text. Rather than simply processing documents at syntax-level, knowledge-based approaches rely on implicit denotative features associated with natural language text, hence stepping away from the blind usage of word co-occurrence count. Unlike purely syntactical techniques, knowledge-based approaches are also able to detect semantics that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

This special issue aimed at bringing together contributions from both academics and practitioners in the context of knowledge-based NLP in order to address the wide spectrum of issues related to NLP research and, hence, better grasp the current limitations and opportunities related to this fast-evolving branch of artificial intelligence. Out of the 54 submissions received for this special issue, 17 were accepted. Two of the accepted papers underwent four rounds of revisions, five papers underwent three, and the rest were revised twice.

The article *"Using Neural Word Embeddings to Model User Behavior and Detect User Segments"* by Ludovico Boratto, Salvatore Carta, Gianni Fenu, and Roberto Saia proposes to model user behavior for detecting segments of users to target for advertising. Various sources of data are mined and modeled in order to detect these segments, such as the queries issued by the users. Authors first show the need for a user segmentation system to employ reliable user preferences, since nearly half of the times users reformulate their queries in order to satisfy their information need. Then, they propose a method that analyzes the description of the items positively evaluated by the users and extracts a vector representation of the words in these descriptions (word embeddings). Since it is widely known that users tend to choose items of the same categories, the proposed approach is designed to avoid the so-called preference stability, which would associate the users to trivial segments. Authors performed different sets of experiments on a large real-world dataset, which validated the proposed approach and showed its capability to produce effective segments.

In *"Bilingual Recursive Neural Network Based Data Selection for Statistical Machine Translation"*, Derek Wong, Yi Lu, and Lidia Chao address the problem of data selection as an effective solution to domain adaptation in statistical machine translation (SMT). The dominant methods are perplexity-based ones, which do not consider the mutual translations of sentence pairs and tend to select short sentences. Authors propose bilingual semi-supervised recursive neural network data selection methods to differentiate domain-relevant data from out-domain data. The proposed methods are evaluated in the task of building domain-adapted SMT systems. Authors present extensive comparisons and show that the proposed methods outperform the state-of-the-art data selection approaches.

Next, the article *"Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering"* by Tiago Almeida, Tiago Silva, Igor Santos, and José Gómez Hidalgo proposes and then evaluates a method to normalize and expand online Instant Messaging and SMS text in order to acquire better attributes and enhance the classification performance. The proposed text processing approach is based on lexicographic and semantic dictionaries along with state-of-the-art techniques for semantic analysis and context detection. This technique is used to normalize terms and create new attributes in order to change and expand original

http://dx.doi.org/10.1016/j.knosys.2016.07.025
0950-7051/

text samples aiming to alleviate factors that can degrade the algorithms performance, such as redundancies and inconsistencies. Authors have evaluated the proposed approach with a public, real and non-encoded dataset along with several established machine learning methods.

The article *"Identifying Motifs for Evaluating Open Knowledge Extraction on the Web"* by Aldo Gangemi, Diego Reforgiato Recupero, Misael Mongiovì, Andrea Nuzzolese, and Valentina Presutti is in the context of Open Knowledge Extraction (OKE), the process of extracting knowledge from text and representing it in formalized machine readable format, by means of unsupervised, open-domain and abstractive techniques. Despite the growing presence of tools for reusing NLP results as linked data (LD), there is still lack of established practices and benchmarks for the evaluation of OKE results tailored to LD. In this paper, authors propose to address this issue by constructing RDF graph banks, based on the definition of logical patterns called OKE Motifs. They demonstrate the usage and extraction techniques of motifs using a broad-coverage OKE tool for the Semantic Web called FRED. Finally, authors use identified motifs as empirical data for assessing the quality of OKE results, and show how they can be extended through a use case represented by an application within the Semantic Sentiment Analysis domain.

Following, *"Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network"* – paper handled independently during review process – is elaborated upon by Soujanya Poria, Erik Cambria, and Alexander Gelbukh who present the first deep learning approach to aspect extraction in opinion mining. Aspect extraction is a subtask of opinion mining consisting in identifying the concepts about which the opinion is expressed in an opinionated text. Authors used a 7-layer Deep Convolutional Neural Network (CNN) to tag each word in the sentence as an aspect or non-aspect word. In addition to the CNN classifier, they developed a set of linguistic patterns useful for the same purpose and combined them with the CNN classifier. With this ensemble classifier, authors obtained significantly better accuracy than the state-of-the-art methods. Finally, they trained a word embeddings model specifically for sentiment analysis and opinion mining tasks, and made it publicly available.

"A New Hybrid Semi-supervised Algorithm for Text Classification with Class-based Semantics" is presented by Berna Altinel and Murat Can Ganiz who propose novel semantic smoothing kernels based on class specific transformations to represent certain aspects of natural language semantics. These kernels use class-term matrices, which can be considered as a new type of Vector Space Models (VSM). By using the class as the context, these matrices can extract class specific semantics by making use of word distributions both in documents and in different classes. The classification algorithms which are built on kernels like Support Vector Machines (SVM) can make use of these strictly supervised semantic kernels to achieve higher accuracy compared to traditional VSM based classifiers for text classification. The proposed algorithm uses Helmholtz principle based calculation of term meanings for initial classification and a class-based term weighting based semantic kernel with SVM for the final classification model. Term meaning calculations depend on the Helmholtz principle from the Gestalt theory and calculated in the context of classes. Authors perform various experiments on popular benchmark textual datasets and report the results with respect to wide range of experimental conditions in order to evaluate the proposed approach.

The possibility of *"Building a Twitter Opinion Lexicon from Automatically-annotated Tweets"* is analyzed by Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer who present a method that combines information from automatically annotated tweets and existing hand-made opinion lexicons to expand an opinion lexicon in a supervised fashion. The expanded lexicon contains part-of-speech (POS) disambiguated entries with a probability distribution for positive, negative, and neutral polarity classes. To obtain this distribution using machine learning, authors propose word-level attributes based on the syntactic information conveyed by POS tags and associations between words and the sentiment expressed in the tweets in which they occur. They consider tweets with both hard and soft sentiment labels. The sentiment associations are modeled in two different ways: using semantic orientation, which is based on mutual information, and using stochastic gradient descent, which learns a linear relationship between words and sentiment. The training dataset is labeled by a seed lexicon built from the combination of multiple hand-annotated lexicons. Experimental results show that the proposed method outperforms the three-dimensional word-level polarity classification performance obtained by using semantic orientation alone, a state-of-the-art measure for establishing world-level sentiment.

"Knowledge Base Population using Semantic Label Propagation' is subsequently suggested by Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder who study how the amount of manual labeling necessary for knowledge base population can be significantly reduced by applying distant supervision, which generates training data by aligning large text corpora with existing knowledge bases. Authors propose to combine distant supervision with minimal human supervision by annotating features (in particular shortest dependency paths) rather than complete relation instances. Such feature labeling eliminates noise from the initial training set, resulting in a significant increase of precision at the expense of recall. Authors further improve on this approach by introducing the Semantic Label Propagation (SLP) method, which uses the similarity between low-dimensional representations of candidate training instances to again extend the (filtered) training set in order to increase recall while maintaining high precision. The proposed strategy is evaluated on an established test collection designed for knowledge base population. The experimental results show that SLP leads to substantial performance gains when compared to existing approaches while requiring an almost negligible human annotation effort.

The contribution *"Contextual Sentiment Analysis for Social Media Genres"* by Aminu Muhammad, Nirmalie Wiratunga, and Robert Lothian introduce SmartSA, a lexicon-based sentiment classification system for social media genres which integrates strategies to capture contextual polarity from two perspectives: the interaction of terms with their textual neighborhood (local context) and text genre (global context). The lexicon-based approaches to opinion mining involve the extraction of term polarities from sentiment lexicons and the aggregation of such scores to predict the overall sentiment of a piece of text. It is typically preferred where sentiment labeled data is difficult to obtain or algorithm robustness across different domains is essential. A major challenge for this approach is accounting for the semantic gap between prior polarities of terms captured by a lexicon and the terms – polarities in a specific context (contextual polarity). This is further exacerbated by the fact that a term's contextual polarity also depends on domains or genres in which it appears. To this end, authors introduce an approach to hybridize a general purpose lexicon, SentiWordNet, with genre-specific vocabulary and sentiment. Evaluation results from diverse social media show that the proposed strategies to account for local and global contexts significantly improve sentiment classification, and are complementary in combination. the proposed system also performed significantly better than a state-of-the-art sentiment classification system for social media, SentiStrength.

In *"Leveraging Multimodal Information for Event Summarization and Concept-level Sentiment Analysis"*, Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Shaikh, and Roger Zimmermann discuss the rapid growth of online user-generated content (UGCs) and the need for social media companies to automatically