The LivingKnowledge Project: Exploring the Spectrum of Opinions over Time

Richard Johansson and Alessandro Moschitti

DISI, University of Trento
Trento, Italy
{johansson, moschitti}@disi.unitn.it

Abstract. The last two decades of research in Information Retrieval have shown that *bag-of-words* models are sufficient for the design of document search and categorization systems. This has made useless, at least for document retrieval purposes, the development of semantic models more advanced than the simple *bag-of-words*. In contrast, recent research in sentiment classification has shown that, when the required semantic information is not limited to query-document relatedness, e.g. opinion mining, advanced semantic processing is crucial.

In this perspective, the LivingKnowledge (LK) project, funded by the seventh EU framework program, aims at studying and developing a technology based on semantic processing, which can be exploited to solve complex semantic tasks such as opinion extraction, opinion analysis in terms of diversity and their evolution over time.

The role of LK's work is twofold: (a) it is fundamental for the design of innovative future digital libraries since different opinions can be other dimensions for searching or categorization of the digital content and (b) the evolution of opinions also refers to the study of the evolution of knowledge and categorization schemes during time. This document gives an overview of the two main objectives of the project.

1   Introduction

LivingKnowledge (LK) is a project on future emerging technology, funded by the seventh EU framework program. Among other its research subjects, e.g. design of automatic tools for social science analysis, LK aims at studying and developing semantic processing models for opinion extraction, opinion analysis and knowledge evolution over time.

The first two aspects are rather interesting for digital library research since the automatic extracted metadata like for example: `opinionated` or `pos./neg. opinion`, allows for searching or categorizing documents according to standard topics as well as this new semantic dimension. For example, we can categorize films based on genres: *adventure*, *dramatic*, *horror* and so on along with the polarity of opinions[1]. The latter can be also characterized with a finite interval of values, e.g. from 1 to 5.

---

[1] Another interesting and related task is the product review mining [1]

Work on sentiment classification [2] has shown that, in contrast to standard text categorization [3, 4], syntactic/semantic processing is required to boost the performance of the *bag-of-words* models. In this perspective, LK will explore the most advanced technology for encoding syntax and semantics, i.e. support vector machines [5] based on structured kernels, e.g. [6], for encoding syntactic parse tree information along with predicate–argument structures [7–9] (semantic structures) in the automatic opinion analyzers.

Regarding knowledge evolution, for which opinion dynamics is just an instance of knowledge, the project is studying: (a) ways to adapt categorization systems to the evolution of document content over time such data they maintain a satisfactory accuracy; and (b) approaches to the scheme evolution so that categories are automatically created, deleted or merged.

In the remainder of this paper for sake of space we will only illustrate the opinion mining aspects along with our preliminary results on simple models and the planned advanced approaches.

2 Automatic Retrieval of Opinionated Pieces of Text

Automatic retrieval of opinionated pieces of text may be carried out on a number of different levels. On the most coarse level, *documents* are categorized as opinionated or factual; for instance, this may be used to distinguish editorials from news [10].

At the other end of the spectrum, methods have been proposed to carry out fine-grained subjectivity analysis on the level of linguistic expressions [2].

In the ongoing project, we currently focus mainly on the automatic classification of individual *sentences* as opinionated or not. This will later pave the way for a more fine-grained analysis that can support a detailed exploration over time of the opinions held by various groups of people.

2.1 Preliminary Experiments in Sentence-level Opinion Classification

As a first step, we formalized the problem of detecting opinionated sentences as a binary text categorization problem. The problem could then be approached using classical statistical text categorization techniques. We thus represented a sentence as a vector in a high-dimensional space using a bag-of-words representation and trained a binary statistical classifier to distinguish the two types of sentences (subjective or objective). The classifiers were linear support vector machines, which have previously been shown to be effective in text categorization problems [5]. To train and evaluate the classifier, we used the MPQA corpus, a collection of 692 documents in English (containing 15,768 sentences) manually annotated with information about expressions of subjectivity [2].

In addition to the classifier based on a pure bag-of-words representation, we implemented a classifier that also used a subjectivity lexicon to determine the presence of strongly or weakly subjective words in the sentence (such as
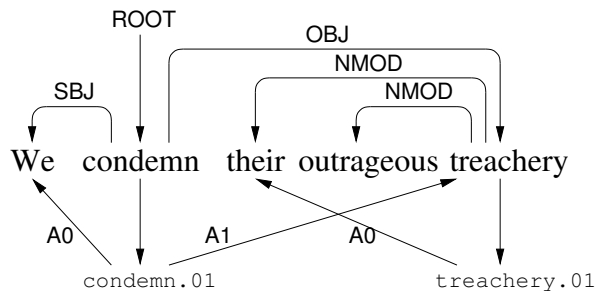
Fig. 1. Example of a syntactic–semantic dependency graph.

wonderful, condemn). We evaluated the classifiers for subjective sentences on a subset of 2,185 sentences of the MPQA corpus by obtaining a precision, a recall, and an F1-measure of 0.79, 0.76 and 0.78, respectively. When we used the lexicon the figures improved to 0.82, 0.79 and 0.81, respectively.

2.2 Future Work on Linguistic Structure for Opinion Extraction

It is still an open question which linguistic information is useful for the automatic retrieval of opinionated sentences. Previous methods for finding opinions have relied either on on simple cues or keyword spotting [11] or on bag-of-words methods [12], as described in the previous section.

We hypothesize that deeper linguistic structures may be useful for opinion retrieval, and we will explore various linguistic representations as a part of this research. As a start, we will see whether it is possible to use automatic syntactic and role-semantic analysis of sentences to improve the classifiers similarly to the approach followed for Question/Answer classification [6].

As an example, Figure 1 shows the analysis of the sentence *We condemn their outrageous treachery.* The sentence was automatically analyzed by the LTH parser [13]. In the figure, the syntactic representation is shown above the sentence and the semantic representation below. For instance, the syntactic graph shows that *We* is a syntactic subject of *condemn*, and the semantic graph shows that *their* has the A0 (BETRAYER) semantic role in the event represented by the word *treachery.* Such structure can easily be encoded in SVMs by means of structured kernels, [6], which represent it in terms of all its substructures (i.e. each feature is a portion of the graph). In addition to the syntactic and role-semantic graphs, we plan to explore other types of linguistic representation such as discourse graphs [14].

To conclude, we believe that the LK research will help to advance the research on digital libraries on three different lines: (i) the opinion classification will provide automatic metadata, which can refine the categorization schema and the access methods, (ii) advanced syntactic/semantic representation for opinions will provide theory and methods for the representation of other digital content,

e.g. definitions, explanations, and (iii) the study of the evolving categorization schema is directly related to the evolution and management of future digital libraries.

References

1. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Edmonton, Canada (2002) 341–349
2. Wilson, T.A.: Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. PhD thesis, University of Pittsburgh, Pittsburgh, United States (2008)
3. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In McDonald, S., Tait, J., eds.: Advances in Information Retrieval – ECIR, Sunderland, UK. (2004)
4. Basili, R., Moschitti, A., Pazienza, M.T.: A text classifier based on linguistic processing. In: Proceedings of IJCAI 99, Machine Learning for Information Filtering. (1999)
5. Joachims, T.: Learning to Classify Text using Support Vector Machines. PhD thesis, University of Dortmund, Dortmund, Germany (2002)
6. Moschitti, A.: Kernel methods, syntax and semantics for relational text categorization. In: Proceeding of CIKM '08, NY, USA (2008)
7. Giuglea, A.M., Moschitti, A.: Knowledge Discovery using Framenet, Verbnet and Propbank. In Meyers, A., ed.: Workshop on Ontology and Knowledge Discovering at ECML 2004, Pisa, Italy (2004)
8. Moschitti, A., Pighin, D., Basili, R.: Tree kernels for semantic role labeling. Computational Linguistics **34**(2) (2008) 193–224
9. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J.: The CoNLL–2008 shared task on joint parsing of syntactic and semantic dependencies. In: CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, Manchester, United Kingdom (2008) 159–177
10. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), Sapporo, Japan (2003) 129–136
11. Wiebe, J., Bruce, R., O'Hara, T.: Development and use of a gold standard data set for subjectivity classifications. In: Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics. (1999)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP. (2002)
13. Johansson, R., Nugues, P.: Dependency-based syntactic–semantic analysis with PropBank and NomBank. In: CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, Manchester, United Kingdom (2008) 183–187
14. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse Treebank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). (2008)