Mining Economic Sentiment using Argumentation Structures

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska de Jong

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, The Netherlands

{hogenboom, fhogenboom, kaymak, wouters, fdejong}@ese.eur.nl

Abstract. The recent turmoil in the financial markets has demonstrated the growing need for automated information monitoring tools that can help to identify the issues and patterns that matter and that can track and predict emerging events in business and economic processes. One of the techniques that can address this need is sentiment mining. Existing approaches enable the analysis of a large number of text documents, mainly based on their statistical properties and possibly combined with numeric data. Most approaches are limited to simple word counts and largely ignore semantic and structural aspects of content. Yet, argumentation plays an important role in expressing and promoting an opinion. Therefore, we propose a framework that allows the incorporation of information on argumentation structure in the models for economic sentiment discovery in text.

1 Introduction

Today's economic systems are complex with interactions amongst ever more actors and with increasing dynamics. Tracking and monitoring is important in any dynamic system in order to be able to exercise control over it, and is essential in complex systems like economic systems. As our ability to collect and process information increases, actors in economic systems (e.g., businesses) feel a growing need for automated information monitoring tools that can help to identify issues and patterns that matter and that track and predict emerging events.

A key element for decision makers to track is stakeholders' sentiment. The relevance of insight in sentiment has been studied in various contexts. For instance, recent research demonstrates that the detection of occupational fraud – a 652 billion dollar problem – can be supported by the automated detection of employee disgruntlement in a vast amount of archived e-mails [1]. In the context of organizational change processes, Hartelius and Browning [2] argue that managers' most important actions are persuasive actions. Furthermore, recent research demonstrates the influence of investor sentiment on financial markets through the impact of news messages [3].

The recent turmoil in the financial markets has illustrated the need for advanced monitoring and tracking tools that enable timely intervention. The key conceptualization of economic sentiment considered here is consumer confidence, which is the degree of optimism that consumers have about the future of the economy and their own financial situation. Consumer spending tends to vary with the consumer confidence [4]. Since consumer spending is an important element of economic growth, consumer confidence can be considered to be an important indicator for economic expansion. As such, the formation of expectations regarding future developments in the economy significantly influences future states of the economy, such as a recession [5] or economic recovery [6]. Hence, economic analysts and policy makers must keep track of economic sentiment in order to anticipate the future state of the economy.

Back in 1975, Katona [7] argued that economic sentiment may represent a subjective state of mind of actors within an economic system. Economic sentiment has commonly been characterized as a latent variable, correlated with traditional macro-economic indicators, e.g., employment conditions [8]. More recent studies however consider additional macro-economic indicators to capture economic sentiment, e.g., the University of Michigan Consumer Sentiment Index (CSI) or the Consumer Confidence Index (CCI) [4]. Traditional indicators have been operationalized using publicly available macro-economic data, whereas the CSI and CCI have been based on regular, allegedly representative surveys. Conversely, Bovi [9] points out that people's expectation formation is thwarted by structural psychologically driven distortions. The structural difference between surveyed ex ante expectations and subsequent realizations may be caused by respondents considering questions to be vague or hard to assess, which may trigger them to provide heuristic, biased answers [10]. Moreover, Oest and Franses [11] stress that over time, the small survey panels encompass different respondent samples. This complicates generalizability of survey findings, as observed sentiment shifts may be largely driven by differences in respondent samples.

In a recent analysis, Vuchelen [6] argues that the broader view on economic sentiment pioneered by Katona may complement the more restrictive view based on macro-economic indicators. In this light, we envisage a more deliberate conceptualization of economic sentiment when common macro-economic indicators are complemented with a general mood, which is typically represented using an indicator of polarity (possibly assessed on multiple features). In their communication, people reveal their mood to a certain extent. With the advent of the Internet, traces of human activity and communication have become ubiquitous, partly in the form of written text. An overwhelming amount of textual publications (e.g., scientific publications, blogs, and news messages) is available at any given moment. Analyzing free-text information can enable us to extract the information tailored to the needs of decision makers. The amount of data available to decision makers is overwhelming, whereas decision makers need a complete overview of their environment in order to enable sufficient tracking and monitoring of business and economic processes, which in turn can facilitate effective, well-informed decision making.

The abundance of digitally stored text opens possibilities for large-scale (semi-)automatic text analysis, focused on uncovering interesting patterns: text mining. Text mining may lead to valuable insights, but raw textual data does not necessarily explicitly reveal the writer's sentiment. Existing sentiment mining approaches enable quantitative analysis of texts, mainly based on their statistical properties, possibly combined with numeric data. Most approaches are limited to word counts and largely ignore semantic and structural aspects of content. We hypothesize that argumentation structure analysis can support economic sentiment mining, as argumentation structures play an important role in expressing and promoting opinions. Moreover, not all parts of a text may contribute equally to expressing or revealing the underlying sentiment. The relative contribution of a certain linguistic element to the overall sentiment may depend on its position within the overall structure of the text and argumentation. For instance, a conclusion may contribute more than a refuted argument.

In this paper, we propose a framework combining knowledge from the areas of text mining – and more specifically sentiment mining – and argumentation discovery. This framework is inspired by a review of the state-of-the-art in these areas. Not only will this research contribute to the existing body of knowledge on sentiment mining by bridging the theoretical gap between qualitative text analyses and quantitative statistical approaches for sentiment mining, but the envisaged link between argumentation structures and associated sentiment may also enable decision makers and researchers to obtain insight in *why* things are happening in their markets, rather than just *what* is happening.

The remainder of this paper is organized as follows. First, the interrelated concepts of text mining and sentiment mining are presented in Sect. 2. Then, Sect. 3 shifts focus to discovery of argumentation structures. Subsequently, we propose a framework in which the knowledge from the disparate fields of sentiment mining and argumentation discovery is combined. We conclude in Sect. 5.

2 Text Mining

Much linguistic information is available in textual format. Text is a direct carrier of linguistic information as opposed to, e.g., audio data in which the waveforms need to be processed before the linguistic content can be identified. Therefore, textual data is a convenient mode for representing or processing linguistic data. This section hence continues with an introduction to textual data, before proceeding to a discussion on extracting information and sentiment from these data.

2.1 Textual Data

Text is typically considered to be unstructured data. Yet, text has a kind of structure that arbitrary collections of words or sentences generally lack. From a linguistic perspective, text documents typically have some implicit notion of structure, constituted by semantic or syntactical structure, as well as typographical elements, lay-out, and word sequence [12].

The analysis and interpretation of linguistic information conveyed by text or natural language in general is a complex process. Different levels of abstraction disclose different aspects of information, which all contribute to the overall meaning of the text. Liddy [13] distinguishes seven levels of analysis. Some levels deal with individual elements in texts (words or sounds), whereas others deal with larger units (e.g., sentences). Finally, some levels deal with texts as a whole.

The levels of language considered by Liddy are synchronic and can interact in a variety of ways. Liddy claims that meaning is conveyed by each and every level of language and that humans have been shown to use all levels of language to gain understanding. In contrast, some scholars appear to question the assumed interactions and interdependencies between levels of language. For instance, Kracht [14] breaks a lance for a definition of meaning that is independent of syntactic structure. Kracht decouples the levels of language by arguing that it is not the task of semantics to state in which way things are syntactically put together. This position goes against the Principle of Compositionality that is being adhered to in some Natural Language Processing communities and that goes back to work of Gotltob Frege (1848-1925). Yet, such a nuance only demonstrates the complexity of natural language as a carrier of information; interpreting and understanding text is an intricate, non-trivial matter.

2.2 Extracting Knowledge from Textual Data

In the last couple of decades, a substantial amount of research has been focusing on automated ways of gaining understanding from text by means of text mining. Text mining is a broad term that encapsulates many definitions and operationalizations, which appear to be distributed in a continuum between two extremes. On one hand, text mining refers to retrieving information that already is in the text (typically using predefined patterns). On the other hand, text mining could refer to a more inductive approach, where patterns are to be discovered in textual data. Theory (i.e., the model) follows the data.

In an attempt to provide an initial sketch of the continuum in which the various notions of text mining are distributed, Hearst [15] distinguishes these notions based on the extent to which they add new information. Whereas Hearst more or less focuses on approaches ranging from information retrieval to knowledge discovery, Kroeze et al. [16] emphasize that the field of text mining is slowly evolving towards a deliberate process of creating new knowledge that did not exist before and cannot simply be retrieved or discovered by accessing existing records of knowledge. Correspondingly, they add a concept to Hearst's high-level overview of the text mining landscape: intelligent text mining. Such perspectives on text mining involve the creation of new, important knowledge about the world outside of the data collection itself. Intelligent text mining involves deriving implications of found patterns and trends, e.g., implied business decisions or the social impact suggested by the found linkages may be assessed.

Many definitions of text mining exist, yet the common denominator is that text mining seeks to extract high-quality information from unstructured data which is textual in nature, where quality is often conceptualized as a measure of interestingness or relevance. The dispersion of conceptualizations of text mining is reflected in the terminology used to refer to text mining. For instance, text mining approaches fitting in the frameworks of Hearst and Kroeze et al. are occasionally addressed as text analytics, intelligent text analysis, knowledge discovery in texts, and text data mining. The latter term indicates a connection between data mining and text mining. Data mining is used to find patterns and subtle relationships in structured data, and rules that allow prediction of future results, whereas text mining focuses on finding patterns and relations in unstructured, textual data. Feldman and Sanger [17] however argue that from a linguistic perspective, text is typically not completely unstructured. A text document can already be referred to as weakly structured or free-format when it has some indicators to denote linguistic structure (e.g., key terms related to argumentation, headers, or templates adhered to in scientific research papers and news stories). Furthermore, Feldman and Sanger distinguish semi-structured documents which contain extensive and consistent format elements, such as HTML documents.

With respect to text mining in its broadest sense, literature exhibits a rough distinction between three stages: preprocessing, processing, and presentation. Feldman and Sanger [17] provide an extensive overview of preprocessing routines, pattern-discovery algorithms, and presentation-layer elements. Most text mining tools utilize their own framework for processing texts with the purpose of extracting information. However, GATE [18], a freely available text processing framework, has become increasingly popular due to its flexibility and extensibility. Amongst supported linguistic analyses are tokenization, Part-Of-Speech (POS) tagging, and semantic analysis. Tools like GATE could prove useful in a setting in which economic discourse is to be analyzed for interesting patterns. Yet nowadays, patterns in raw text are not enough anymore; insight in (patterns of) associated sentiment is crucial for decision makers.

2.3 Sentiment Mining

The field of sentiment mining is relatively young. The discovery of sentiment is usually focused on reviews of products, movies, etcetera. The focus of work on analyzing online discussions and blogs [19, 20] is more on distinguishing opinions from facts than on extracting and summarizing opinions. Existing toolkits are limited to simple word counts and relevant linguistic resources are absent or do not always fit into the applied framework. Today's text analytical tools are ill-equipped to deal with highly dynamic domains, because they have been developed without adaptation in mind [21, 22] and until recently largely ignore structural aspects of content [23, 24].

Early attempts to incorporate structural aspects of texts have been made by Pang et al. [25], who stress that for instance a review with a predominant number of negative sentences may actually have a positive conclusion and thus have an overall positive sentiment. Therefore, Pang et al. include location information of tokens for sentiment in their analysis. Devitt and Ahmad [26] use theories of lexical cohesion for sentiment polarity classification of financial news. Mao and Lebanon [27] model the sentiment of a document as a flow of local sentiments, which are simply related to position in the text. Yet so far, no attempts have been made for utilizing information encompassed in argumentation structures, whereas argumentation structures are closely related to the sentiment of the message they convey.

3 Discovering Argumentation Structures

By using argumentation structure and elements such as specific metaphors, analogies, vocabularies, or supportive non-textual data, a specific mood or opinion can be expressed and promoted. For example, the use of analogies or vocabularies invoking negative associations in means of communication concerning change processes may lead people to have negative expectations. Our framework starts from the hypothesis that sentiment mining in economic texts can thus be improved if the information in the structural elements of a text can be harvested.

3.1 Argumentation

Argumentation is central in any discourse. Humans discuss and argue by exchanging information in natural language. In all societies, there is a tendency for idle, free-flowing exchange of ideas and thoughts, which is called *conversation* [28]. In economics literature, conversation is often seen as *cheap talk* in which the act of conducting a conversation does not influence the payoffs in a game-theoretic setting [29]. Here, conversation is considered only to convey direct information, either in the form of imperatives (e.g., issuing orders) or in the form of information that is actionable (e.g., by revealing private information). Although classical economic theory posits that all information is incorporated in a market-based pricing system, the importance of private information and asymmetric distribution of information has been subject to many economic studies. Conversation provides a mechanism to diffuse asymmetric information.

In addition to the direct information content, argumentation and persuasion are important aspects of linguistic communication. People exchange ideas with a goal. Argumentation is incorporated to convince the listener of the validity of the reasoning. Anyone engaged in argumentation selects and presents information in a particular way that enhances the acceptance of the argument. Hence, rhetoric, argumentation structures, and presentation styles are very important since they facilitate persuasion, as acknowledged by various economists. McCloskey and Klamer [30] estimate that a significant part of national income can be attributed to persuasion. Cosgel [31] models consumption from a rhetorical perspective and shows how subjective information such as tastes can be understood form a different perspective than the more common choice framework.

3.2 Argumentation Mining

The above studies demonstrate that an analysis of discourse in which structural and semantic elements are incorporated can provide information that is otherwise not available. Qualitative text analyses, possibly guided by the Textual

Entailment (TE) framework [32] or the Rhetorical Structure Theory (RST) presented by Taboada and Mann [33], can enable the discovery of such information. In recent years, computational models of linguistic processing, text mining and argumentation discovery have been developed, especially in the fields of computer science and computational linguistics.

A pioneer in this area has been Teufel. In her early work [34], she relies on statistical classifiers to identify and classify sections on scientific documents as so-called argumentative zones. Early research, e.g., the work of Marcu [35], typically exploited keywords taken to be signaling a discursive relation, yet more recently, researchers like Webber et al. [36] argue that the true structure of discourse in a text is not necessarily formed by the actual textual units and their connecting keywords; they appear to employ a more high-level conceptualization of argumentation structures, which can however be linked to the relational meaning invoked by the keywords. Another perspective on argumentation discovery is advocated by Vargas-Vera. Initially focused on mere extraction of patterns from text – e.g., news articles [37] – Vargas-Vera focuses in more recent work on discovering argumentation structures in texts by representing these texts as networks of cross-referring claims [38], similarly to Buckingham Shum et al. [39]. Such efforts as described here are promising first steps towards principal ways of automatically detecting argumentation structures.

4 Argumentation-Based Economic Sentiment Mining

In order to be able to extract economic sentiment from text sources, we need an information system capable of inferring specific information on economic sentiment from natural language texts. The purpose of such a system is to analyze a given text collection and determine the sentiment in the texts. Inspired by our findings presented in Sect. 2.3, our envisaged system is to take into account argumentation structures, which can be detected automatically (see Sect. 3.2). We propose an Information Extraction pipeline, which divides specific roles and tasks amongst different components that are interconnected by their inputs and outputs. Such a pipeline facilitates stepwise abstraction from raw text to useable, formalized chunks of linguistic data and enables effective text processing, as each component can be optimized for a specific task.

In our framework, depicted in Fig. 1, we propose to employ the general purpose GATE framework, which allows for easy usage, extension, and creation of individual components. For initial lexico-syntactic analysis of input text (i.e., operations not specific to our envisaged sentiment mining approach), we propose to use several existing components from GATE's default pipeline, A Nearly New Information Extraction System (ANNIE). First of all, we clear documents from unwanted artifacts such as tags, by means of a *Document Reset* component. Subsequently, we employ an *English Tokenizer*, which splits text into separate tokens (e.g., words). Then, a *Sentence Splitter* is used, which splits the input text into sentences, after which a *POS Tagger* component is utilized in order to determine the part-of-speech of words within a text collection.

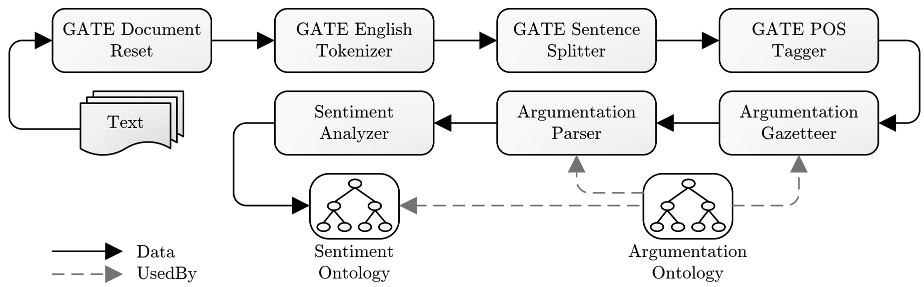

Fig. 1. Conceptual outline of the envisaged information processing pipeline.

After these basic syntactic operations, semantic analysis is to be performed by several novel components. Firstly, we employ an *Argumentation Gazetteer* for identifying argumentation markers, i.e., key terms related to argumentation. For this, we propose to employ a populated argumentation ontology that contains definitions of these argumentation markers and their relations to argumentative text elements (e.g., arguments, supports, conclusions), which are also defined in this ontology. Guided by the annotated argumentation key terms found by the gazetteer, the *Argumentation Parser* subsequently identifies text segments and determines their role in a document's argumentation structure, hereby utilizing the argumentation ontology. Finally, the *Sentiment Analyzer* identifies the sentiment in the individual text segments and connects the sentiment of these segments to the associated argumentation structure. Based on their role in the argumentation structure, text segments are assigned different weights in their contribution to the overall sentiment. The output of this process is an ontology that is populated on the fly in order for it to represent knowledge on the current economic sentiment in the text collection. This sentiment ontology in turn utilizes the argumentation ontology in order to enable a connection between argumentation and sentiment, hereby facilitating insight in opinion genesis. New knowledge on economic sentiment is stored in the ontology, thus enabling reasoning and inference of knowledge in order to support decision making processes.

5 Conclusions and Future Work

The disparate fields of text mining and sentiment mining on the one hand, and argumentation discovery on the other hand, offer a wide range of possibilities in order to advance economic discourse analysis. Firstly, text mining techniques, and more specifically sentiment mining techniques, can help researchers and decision makers to track important trends in their markets. Secondly, argumentation discovery techniques can facilitate insight in the reasoning utilized in economic discourse. Hence, we have proposed an information extraction framework that combines insights from these disparate fields by linking argumentation structures in economic discourse to the associated sentiment, which could offer researchers and decision makers a new perspective on the origins of economic sentiment.

As future work, we plan to further elaborate on this framework and to investigate principal ways of combining argumentation structures with sentiment analysis and subsequently representing economic sentiment in insightful ways. Special attention will be paid to the level of analysis; different types of text may require different levels of granularity due to their distinct characteristics with respect to, e.g., structure or content. Furthermore, we plan to implement the proposed pipeline and to perform analyses to assess the quality of its outputs on corpora of, e.g., news articles, scientific papers, or blogs, the sentiment of which is to be annotated by human experts in order to obtain a golden standard.

References

[1] Holton, C.: Identifying Disgruntled Employee Systems Fraud Risk Through Text Mining: A Simple Solution for a Multi-Billion Dollar Problem. Decision Support Systems **46**(4) (2009) 853–846

[2] Hartelius, J.E., Browning, L.D.: The Application of Rhetorical Theory in Managerial Research: A Literature Review. Management Communication Quarterly **22**(1) (2008) 13–39

[3] Arnold, I.J.M., Vrugt, E.B.: Fundamental Uncertainty and Stock Market Volatility. Applied Financial Economics **18**(17) (2008) 1425–1440

[4] Ludvigson, S.C.: Consumer Confidence and Consumer Spending. The Journal of Economic Perspectives **18**(2) (2004) 29–50

[5] Howrey, E.P.: The Predictive Power of the Index of Consumer Sentiment. Brookings Papers on Economic Activity **32**(1) (2001) 176–216

[6] Vuchelen, J.: Consumer Sentiment and Macroeconomic Forecasts. Journal of Economic Psychology **25**(4) (2004) 493–506

[7] Katona, G.: Psychological Economics. Elsevier (1975)

[8] Adams, F.G., Green, E.W.: Explaining and Predicting Aggregate Consumer Attitudes. International Economic Review **6**(3) (1965) 275–293

[9] Bovi, M.: Economic versus Psychological Forecasting. Evidence from Consumer Confidence Surveys. Journal of Economic Psychology **30**(4) (2009) 563–574

[10] Tversky, A., Kahneman, D.: Judgment under Uncertainty: Heuristics and Biases. Science **185**(4157) (1974) 1124–1131

[11] van Oest, R., Franses, P.H.: Measuring Changes in Consumer Confidence. Journal of Economic Psychology **29**(3) (2008) 255–275

[12] Freitag, D.: Machine Learning for Information Extraction in Informal Domains. Machine Learning **39**(2) (2000) 169–202

[13] Liddy, E.D.: Natural Language Processing. In: Encyclopedia of Library and Information Science. 2nd edn. Marcel Decker, Inc. (2003) 2126–2136

[14] Kracht, M.: The Emergence of Syntactic Structure. Linguistics and Philosophy **30**(1) (2007) 47–95

[15] Hearst, M.: Untangling Text Data Mining. In: 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999). (1999) 3–10

[16] Kroeze, J., Matthee, M., Bothma, T.: Differentiating Data- and Text-Mining Terminology. In: 3rd Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2003). (2003) 93–101

[17] Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2006)

[18] Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities **36**(2) (2002) 223–254
[19] Hu, M., Sun, A., Lim, E.P.: Comments-Oriented Blog Summarization by Sentence Extraction. In: 16th ACM SIGIR Conference on Information and Knowledge Management (CIKM 2007). (2007) 901–904
[20] Zhou, L., Hovy, E.: On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs. In: AAAI Symposium on Computational Approaches to Analysing Weblogs. (2006) 237–242
[21] Blitzer, J., McDonald, R., Pereira, F.: Domain Adaptation with Structural Correspondence Learning. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2006). (2006) 120–128
[22] Turmo, J., Ageno, A., Catala, N.: Adaptive Information Extraction. ACM Computing Surveys **38**(2) (2006)
[23] Daelemans, W., van den Bosch, A.: Memory-Based Language Processing. Cambridge University Press (2005)
[24] Shanahan, J.G., Qu, Y., Wiebe, J.M.: Computing Attitude and Affect in Text: Theory and Applications. Springer (2006)
[25] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up? Sentiment Classification using Machine Learning-Techniques. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). (2002) 79–86
[26] Devitt, A., Ahmad, K.: Sentiment Analysis in Financial News: A Cohesion-Based Approach. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007). (2007) 984–991
[27] Mao, Y., Lebanon, G.: Sequential Models for Sentiment Prediction. In: ICML Workshop on Learning in Structured Output Spaces. (2006)
[28] Shiller, R.J.: Conversation, Information, and Herd Behaviour. American Economic Review **85**(2) (1995) 181–185
[29] Farrell, J.: Talk is Cheap. The American Economic Review **85**(2) (1995) 186–190
[30] McCloskey, D., Klamer, A.: One Quarter of GDP is Persuasion. American Economic Review **85**(2) (1995) 191–195
[31] Cosgel, M.M.: Rhetoric in the Economy: Consumption and Audience. Journal of Socio-Economics **21**(4) (1992) 363–377
[32] Herrera, J., Penas, A., Verdejo, F.: Techniques for Recognizing Textual Entailment and Semantic Equivalence. Lecture Notes in Artificial Intelligence **4177** (2006) 419–428
[33] Taboada, M., Mann, W.C.: Rhetorical Structure Theory: Looking Back and Moving Ahead. Discourse Studies **8**(3) (2006) 423–459
[34] Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, University of Edinburgh (1999)
[35] Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. Computational Linguistics **26**(3) (2000) 395–448
[36] Webber, B., Stone, M., Joshi, A., Knott, A.: Anaphora and Discourse Structure. Computational Linguistics **29**(4) (2003) 545–587
[37] Vargas-Vera, M., Celjuska, D.: Event Recognition on News Stories and Semi-Automatic Population of an Ontology. In: 3rd International Conference on Web Intelligence (WI 2004). (2004) 615–618
[38] Vargas-Vera, M., Moreale, E.: Automated Extraction of Knowledge from Student Essays. International Journal of Knowledge and Learning **1**(4) (2005) 318–331
[39] Buckingham Shum, S.J., Uren, V., Li, G., Domingue, J., Motta, E.: Visualizing Internetworked Argumentation. In: Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Springer-Verlag (2002) 185–204