Building Tagged Linguistic Unit Databases for Sentiment Detection

Arno Scharl
(MODUL University Vienna, Austria
arno.scharl@modul.ac.at)

Albert Weichselbraun
(Vienna University of Economics and Business Administration, Austria
albert.weichselbraun@wu-wien.ac.at)

Stefan Gindl
(MODUL University Vienna, Austria
stefan.gindl@modul.ac.at)

Abstract: Despite the obvious business value of visualizing *similarities* between elements of evolving information spaces and mapping these similarities e.g. onto geospatial reference systems, analysts are often more interested in how the *semantic orientation (sentiment)* towards an organization, a product or a particular technology is changing over time. Unfortunately, popular methods that process unstructured textual material to detect semantic orientation automatically based on tagged dictionaries [Scharl et al. 2003] are not capable of fulfilling this task, even when coupled with part-of-speech tagging, a standard component of most text processing toolkits that distinguishes grammatical categories such as article (AT), noun (NN), verb (VB), and adverb (RB). Small corpus size, ambiguity and subtle incremental change of tonal expressions between different versions of a document complicate the detection of semantic orientation and often prevent promising algorithms from being incorporated into commercial applications. Parsing grammatical structures, by contrast, outperforms dictionary-based approaches in terms of reliability, but usually suffers from poor scalability due to their computational complexity. *This paper* addresses this predicament by presenting an alternative approach based on automatically building *Tagged Linguistic Unit (TLU)* databases to overcome the restrictions of dictionaries with a limited set of tagged tokens.
Key Words: sentiment detection, semantic orientation, tagged linguistic unit
Category: I.2.7, I.2.7, E.m

1 Introduction

The field of *Sentiment Detection* is an intriguing one, since the revelation of the semantic content of a written text shows the real opinion and meaning of the writer. This information can be used for several reasons: enterprises can research on the question, why a certain product failed a success in the market, in politics this knowledge can be used to predict the electoral behavior to adapt the political course or search engines can profit from these methods to augment search results.

The detection of the semantic content of writings has fascinated many researchers, leading to a vast amount of different approaches to absolve this task. Some of these only use binary decisions (a positive or negative sentiment), others use more sophisticated classifications.

Pang et al. [Pang et al. 2002] present an approach using movie reviews as text corpus. The basis for the sentiment detection builds a manually annotated lexicon. The authors use machine learning methods, i.e. Naïve Bayes, maximum entropy and support vector machines (SVMs), and the SVM performed best. The approach is refined in the their subsequent work [Pang and Lee 2004]. Kushal et al. [Kushal et al. 2003] apply three machine learning methods to product reviews, comparing their results to a baseline algorithm. In the baseline algorithm, the score of a term is decided by the number the term occurs in a class divided by all terms in this class.

The polarity of the context, in which a sentiment term occurs, influences the significance of this term [Wilson et al. 2005] - e.g., in '...the president of the National Environment Trust...', with 'trust' a large enterprise is meant and not the synonym for 'confidence'. They use a set of 28 features, such as modifiers or adjacent terms, to determine the context's polarity.

Subasic and Huettner [Subasic and Huettner 2001] use a lexicon to ascertain the semantic orientation of a document. They discriminate a word's meaning by focusing on three features: the *affect* class a word belongs to is decided via part-of-speech tagging (e.g., 'alert' as an adjective targets *intelligence*, whilst as a verb it targets *warning*). The feature *centrality* shows how strongly a word belongs to the affect class, whereas the *intensity* represents the strength of a word (e.g., 'abhor' is stronger than 'displeasure', but both belong to the affect 'repulsion'). The words are aggregated in a fuzzy thesaurus (using max-min combination) and the overall sentiment is calculated via the weighted average of the occurring terms of each class and their intensities. Mullen and Collier [Mullen and Collier 2004] work on SVMs, processing music reviews. The semantic wordlist contains three dimensions: *potency* (representing the strength of a term), *activity* (active or passive) and *evaluation* (positive or negative).

Lexical entries can also be distinguished from each other by using so called 'appraisal taxonomies' [Whitelaw et al. 2005]. These contain information on the attitude of a word (e.g., 'appreciation' or 'affect'), the 'orientation' (positive vs. negative), the 'force' (can be increased by modifiers like 'very'), or the 'polarity' (a binary decision depending on the existence of a negation trigger).

2 Lexical Approaches versus Shallow Parsing

Capturing the evolution of information spaces calls for a new generation of robust, language-independent and distributed natural language processing tech-

niques optimized for throughput and scalability. From a stakeholder perspective, the semantic orientation expressed in textual material (e.g., media coverage) is a particularly interesting aspect when identifying semantic relations [Scharl and Weichselbraun 2006]. Automated methods to compute semantic orientation, however, usually belong to one of the following two categories: (i) full parsing of grammatical structures implicates in good results, but suffers from poor scalability; (ii) simple, scalable methods that focus on the lexis of text but, compared to the first category, fall short in terms of reliability and validity. This paper presents an alternative approach based on automatically generated databases of *tagged linguistic units* with a focus on heterogeneous data in terms of sample composition and entity type (e.g., content versus social).

Past research often preferred to gather a large corpus of text, compiled from many sources and typically sampled in regular intervals - the *US Election 2004/2008 Web Monitor* (www.ecoresearch.net/election2008), for example, or the *Media Watch on Climate Change* (www.ecoresearch.net/climate). Using part-of-speech tagged and partially parsed corpora to identify relevant sketches (= co-occurrence lists for grammatical patterns provided by a grammar rule engine) improves the performance of existing techniques for computing semantic orientation [Kilgarriff et al. 2003, Kilgarriff et al. 2004], but processing arbitrarily long blocks of text still requires a fundamentally new strategy. The ability to maximize the algorithm's validity even when working with very short textual segments is paramount when trying to analyze the *evolution* of knowledge reflected in corpora. Longitudinal studies of specific topics or events often yield few additional occurrences of a term in a given interval, as incremental changes to existing documents are common. This complicates the analysis, since the validity of many text processing methods depends on corpus size and frequency of target terms.

3 Tagged Linguistic Units

Generic methods for computing semantic orientation, i.e. those that do not use machine learning algorithms on a narrowly defined domain, rely on a tagged dictionary that distinguishes between positive and negative sentiment words. The semantic orientation towards a target term is then calculatedby measuring the distance (in words) between the target term and these sentiment words [Scharl et al. 2003].

Tagged dictionaries typically contain a few thousand words, annotated with positive or negative charges. They can be subjected to a reverse lemmatization procedure (word stemming), adding inflections to the initial list of sentiment words. But even assuming such an extended dictionary, dictionary-based approaches do not suffice for robust, scalable components to be embedded in corporate knowledge architectures.

This paper addresses this shortcoming by developing a hybrid method based on spreading activation networks coupled with machine learning algorithms for assigning sentiment charges to encountered linguistic units. For this purpose, the following linguistic units for computing the semantic orientation should be distinguished: *tokens* (single words), *terms* (multiple-word units of meaning), and *concepts* (units of meaning not tied to a particular lexical form and represented via rules or regular expressions).

A sentiment value and a context (for instance part-of-speech tag, geo location, or named entity) is assigned to each linguistic unit. The matrix of sentiment values is constantly being updated based on new data from the knowledge acquisition services, and can be customized for specific domains, applications or users. The outlined approach represents a significant improvement in terms of scalability and reliability. Generating and using a *Tagged Linguistic Unit (TLU)* database instead of a *tagged dictionary* that only contains words and binary classifications. This allows a fine-grained differentiation between the sentiment values associated with morphologically similar but semantically different linguistic units such as *cell*, *fuel cell* and *prison cell* and the consideration of tags like part-of-speech tags, named entity tags, and geo tags.

Work by Scharl et al. [Scharl et al. 2008] has demonstrated the usefulness of assigning sentiment values to geographic locations, but also shows how heavily these values depend on the text's context. Future approaches therefore will address these dependencies by combining tags with more sophisticated context information as for instance hierarchical classifications [Weichselbraun 2004] or topic tags. This approach is (i) *language-independent* in the sense that only a small set of seed terms (e.g., 100 positive and 100 negative words) and grammar patterns are required to initialize the machine learning algorithm and fine-tune the semantic values to any language that is decomposable into concepts, terms and tokens, (ii) not restricted to the categories of 'positive' and 'negative', but supporting an arbitrary number of linguistic categorizations such as weak $\longleftrightarrow$ strong, passive $\longleftrightarrow$ active, etc., (iii) ensures that *every* sentence or document can be annotated; traditional approaches often encounter sentences that do not contain any of the words listed in the tagged dictionary. Annotating the context unit (= sentence, paragraph or document) based on the *average sentiment vector* for all linguistic units encountered in the context unit, instead of only considering a few sentiment words.

Figure 1 illustrates sentiment scoring based on linguistic units. A tagging engine identifies part-of-speech tags, named entities, and geo locations facilitating the identification of linguist units by the phrase engine. The sentiment engine processes linguistic units and associated tags based on the data in the tagged linguistic units database, computing a sentiment value for the given text. Tagging provides important background information for these tasks - in the easiest case

the sentiment of linguistic units as for instance the word `like` depend on the assigned part-of-speech tag (`like/VB` versus `like/IN`), in more complex cases named entity tags or even geo tags might be necessary to correctly identify the TLU's sentiment value (e.g., in the case of `National Environment Trust`).

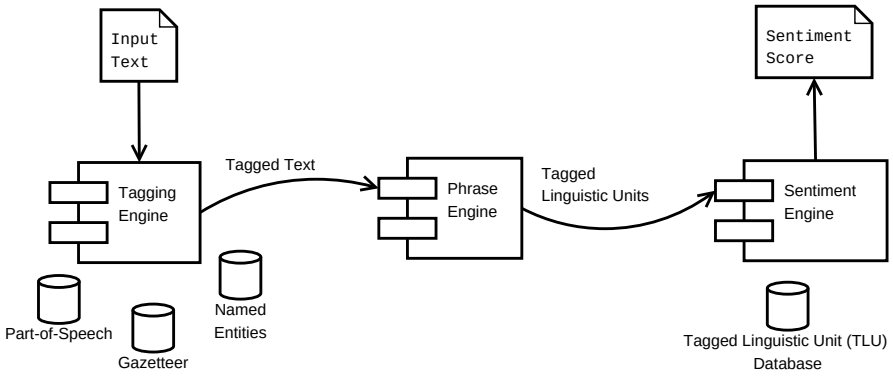

Figure 1: Sentiment Scoring based on Linguistic Units

4 Iterative Extension and Optimization

As outlined in the previous section Tagged Linguistic Unit databases can be easily customized to specific domains and use cases. A domain specific corpus, language specific grammar rules and a set of seed terms with "known" sentiment values (as for instance provided by the conventional tagged dictionaries as the General Inquirer project) initialize the TLU database, the architecture identifies unknown linguistic units in the corpus and determines their sentiment value as illustrated in Figure 2.

The tagging component marks sentences with part-of-speech tags and identifies named entities such as people, organizations, and geographic locations. Combining co-occurrence analysis with a grammar rule engine yields candidate terms for extending the TLU database. Annotating these terms with named-entity tags and encoding characteristic grammatical patterns and known phrases creates a complex semantic network, which describes the relations between the identified linguistic units. Liu et al. [Liu et al. 2005] demonstrated how decomposing and translating semantic networks based on heuristic rules yields a spreading activation network facilitating the domain specific extension of domain ontologies.

Applying this approach for dynamically identifying and tracking tagged linguistic units builds a spreading activation network used to distribute the semantic charges between the units based on the features and annotations generated during the annotation step. Activation of concepts with known semantic charges in accordance to sign and strength of the charge leads to the propagation of energy pulses through the network, eventually distributing charges to all linguistic

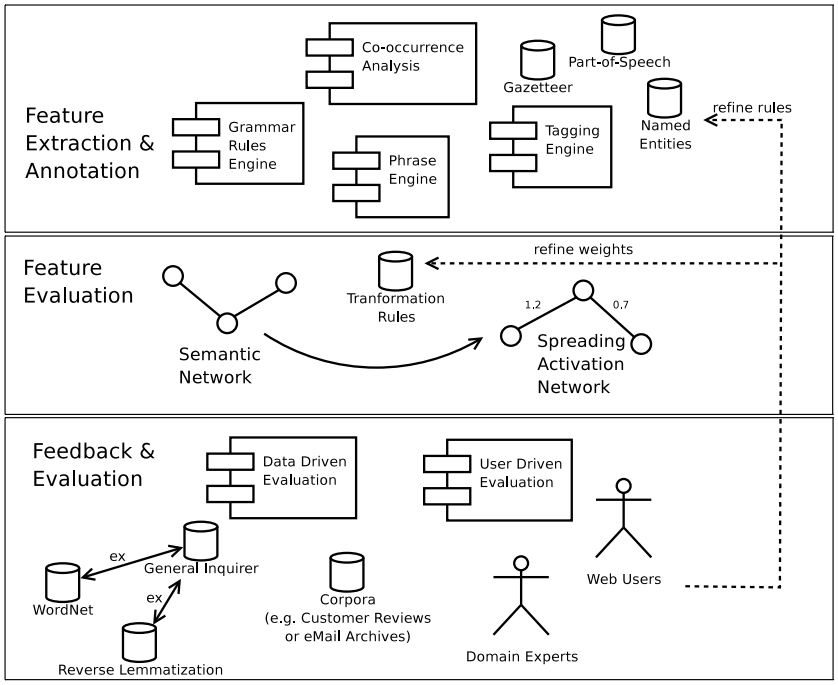

Figure 2: Iterative fine-tuning of the Tagged Linguistic Unit (TLU) Database

units. Analyzing the sentiment values' variance allows estimating confidence levels and facilitates the identifying of synonym $\leftrightarrow$ antonym relationships.

Feedback gathered in the evaluation step adjusts and optimizes the transformation rules for the given domain and corpus, improving the quality of the TLU database with every subsequent step.

Automatic data-driven evaluation on a TLU level will help assess overall performance. Facilitating publicly available corpora like product (amazon.com), movie (www.imdb.com) and music reviews (www.metacritic.com, mp3.com) test cases for sentiment scoring will be developed and applied to the TLU database. Automated processes will be complemented by user-driven evaluations from domain experts and Web users. The feedback gathered by the data- and user-driven evaluations will be utilized to refine the transformation rules of the feature evaluation, and to identify candidate patterns for the inclusion into the databases of the grammar rule engine and the phrase engine.

Automatically generating TLU databases faces the problem of determining the correct charge (+0.4 vs. -0.4, for example) of the sentiment value to be assigned to the linguistics unit. The problem arises from the fact that synonyms and antonyms have very similar (co-)occurrence patterns in a given corpus. Advanced relation discovery techniques developed within the AVALON project

[Weichselbraun et al. 2007] will help overcome this challenge and facilitate the automation of this classification process. The machine learning algorithms can be trained and evaluated on augmented tagged dictionaries (created through reverse lemmatization and adding WordNet synonym and antonym pairs), as well as on publicly available tagged corpora that can serve as the 'gold standard'.

5 Conclusion and Outlook

Simple approaches to sentiment detection based on co-occurrence patterns with terms from a tagged dictionaries scale well, but provide inferior results when comparing their output to complex methods that require a full parsing of sentence structures. The sheer volume of textual data, however, frequently rules out the most sophisticated approaches. Continuously updated databases of tagged linguistic units aim to balance accuracy and throughput. They represent a radical improvement over static sentiment scoring approaches based on tagged dictionaries, which still tend to be compiled manually.

Preliminary results from the described approach are promising. Following a formal evaluation of different approaches to sentiment detection, recall and precision were significantly improved by adding WordNet synonyms and antonyms to the tagged dictionary (only considering synsets with high frequencies to exclude rare and uncommon expressions) [Gindl and Liegl 2008]. Currently we are extracting terms from media corpora as candidates for assigning polarity values via co-occurrence analysis, which will further extend the tagged dictionary.

The increased effectiveness of sentiment detection algorithms will pave the way for a more widespread use in both academic and commercial applications. Refined versions of the sentiment detection method presented in this paper will generate a richer set of context information (e.g., ontology concepts or explicit references to other types of structured knowledge), and consider this information in the scoring process.

Acknowledgment

The authors wish to thank Johannes Liegl for his feedback and suggestions. This work has been developed as part of RAVEN (Relation Analysis and Visualization for Evolving Networks) research project funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.fit-it.at).

References

[Gindl and Liegl 2008] Gindl, S. and Liegl, J.: "Evaluation of Different Sentiment Detection Methods for Polarity Classification on Web-Based Reviews". In Methods for

Polarity Classification on Web-Based Reviews. International Workshop on Computational Aspects of Affectual and Emotional Interaction, 18th European Conference on Artificial Intelligence. Patras, Greece. (2008).

[Kilgarriff et al. 2003] Kilgarriff, A., Evans, R., Koeling, R., Rundell, M., and Tugwell, D.: "WASPBENCH: A Lexicographer's Workbench Supporting State-of-the-Art Word Sense Disambiguation". In 10th Conference on European Chapter of the Association For Computational Linguistics, Morristown, USA. Association for Computational Linguistics (2003).

[Kilgarriff et al. 2004] Kilgarriff, A., Rychl, P., Smrz, P., and Tugwell, D.: "The Sketch Engine". In 11th Euralex international Congress, Lorient, France (2004).

[Kushal et al. 2003] Kushal, D., Lawrence, S., and Pennock, D. M.: "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". In WWW '03: Proceedings of the twelfth international conference on World Wide Web. ACM Press, (2003), 519–528.

[Liu et al. 2005] Liu, W., Weichselbraun, A., Scharl, A., and Chang, E.: "Semi-Automatic Ontology Extension Using Spreading Activation", Journal of Universal Knowledge Management, 0, 1 (2005), 50–58.

[Mullen and Collier 2004] Mullen, T. and Collier, N.:. "Sentiment Analysis Using Support Vector Machines with Diverse Information Sources" (2004).

[Pang and Lee 2004] Pang, B. and Lee, L.: "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". In Proceedings of the 42nd ACL, (2004), 271–274.

[Pang et al. 2002] Pang, B., Lee, L., and Vaithyanathan, S.: "Thumbs up? Sentiment Classification using Machine Learning Techniques". In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2002).

[Scharl et al. 2008] Scharl, A., Dickinger, A., and Weichselbraun, A.: "Analyzing News Media Coverage to Acquire and Structure Tourism Knowledge", Information Technology and Tourism, 10, 1 (2008), 3–17.

[Scharl et al. 2003] Scharl, A., Pollach, I., and Bauer, C.: "Determining the Semantic Orientation of Web-based Corpora". In Liu, J., Cheung, Y., and Yin, H., editors, Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003, Hong Kong (Lecture Notes in Computer Science, Vol. 2690), Berlin. Springer, (2003), 840–849.

[Scharl and Weichselbraun 2006] Scharl, A. and Weichselbraun, A.: "Web Coverage of the 2004 US Presidential Election". In 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy. Association for Computational Linguistics, (2006), 35–42.

[Subasic and Huettner 2001] Subasic, P. and Huettner, A.: "Affect Analysis of Text Using Fuzzy Semantic Typing", IEEE Transaction on Fuzzy Systems, 9, 4 (2001), 483–496.

[Weichselbraun 2004] Weichselbraun, A.:. "Ontologiebasierende Textklassifikation mittels mathematischer Verfahren". PhD thesis, Vienna University of Economics and Business Administration (2004).

[Weichselbraun et al. 2007] Weichselbraun, A., Wohlgenannt, G., Scharl, A., Granitzer, M., Neidhart, T., and Juffinger, A.: "Applying Vector Space Models to Ontology Link Type Suggestion". In 4th International Conference on Innovations in Information Technology, Dubai, United Arab Emirates. IEEE Press, (2007), 566–570.

[Whitelaw et al. 2005] Whitelaw, C., Garg, N., and Argamon, S.: "Using Appraisal Taxonomies for Sentiment Analysis". In Proceedings of MCLC-05, the 2nd Midwest Computational Linguistic Colloquium, Columbus, US (2005).

[Wilson et al. 2005] Wilson, T., Wiebe, J., and Hoffmann, P.: "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis". In Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA (2005).