Calculus of Fuzzy Semantic Typing for Qualitative Analysis of Text

Pero Subasic
Claritech Corporation
5301 Fifth Ave
Pittsburgh PA 15232
Tel. +1-412-621-0570
pero@claritech.com

Alison Huettner
Claritech Corporation
5301 Fifth Ave
Pittsburgh PA 15232
Tel. +1-412-621-0570
ahuettner@claritech.com

ABSTRACT
Statistical approaches to text mining can be enhanced and improved through the qualitative representation of free text – ideally, a representation which accommodates ambiguity and imprecision. We introduce a specialized lexicon that assigns semantic categories to words, together with numeric values for centrality and intensity within each category. From this lexicon, we automatically generate an additional set of resources to implement some of the common operations of text mining – profiling, querying, and query/profile expansion and compression – in qualitative domains. We exploit the hierarchical structure of free text (i.e., sentence/ paragraph/ document) and develop a set of operators whose arguments are fuzzy representations ("profiles") of text at any hierarchical level. Various operators compute the centrality and intensity of categories within a profile, a profile's overall intensity, and the cardinality and fuzziness of a profile; others are used in profile merging, profile expansion or compression, and discovery of related categories from a profile. We address the meaning and modes of deployment of these operators using practical examples. Finally, we discuss the utility of fuzzy typing for various tasks, such as "qualitative browsing" and similarity estimates. We discuss how the existing approach can be enhanced using automatic lexicon expansion and information extraction techniques. We offer a practical software demonstration with several visualization examples, illustrating the power of the proposed operators in affect analysis of news reports and movie reviews.

Keywords
fuzzy logic, semantic typing, text management

1. INTRODUCTION
The affect domain is ambiguous and imprecise – first, because this is the nature of human emotion, and second, because it is characteristic of words in a natural language. Rather than attempting to constrain this ambiguity, we explicitly represent and process it using a fuzzy/possibilistic calculus. Specifically, we integrate techniques from fuzzy logic ([1], [8], [9]) with techniques from natural language processing (NLP). Since the central technique we use from NLP is semantic typing ([3], [4], [5], [6]), we refer to this approach as ***fuzzy semantic typing*** (FST-[7]). At the most basic level, it involves:

- Isolating a vocabulary of words belonging to a meta-linguistic domain (here, affect)
- Using multiple characterizations and scalar metrics to represent the meaning of each word in that domain
- Computing profiles for texts based on the categorizations and scores of their component parts
- Manipulating profiles to visualize the texts

2. CALCULUS OF FST
We propose the following set of operators, which apply to different resources and represent different aspects of underlying content:

- The *centrality computation operator* combines category centralities to yield a qualitative representation of content (e.g., to what degree is this sentence *aggressive*?)
- The *intensity computation operator* combines category intensities to represent the intensity of expressions (e.g. *how intense* is the aggression in this sentence?)
- The *cardinality* indicates the applicability of our lexicon and taggings to particular content, and the precision of the vocabulary representation that the lexicon handles
- The *fuzziness* shows the balance and precision of a particular distribution of centralities across categories

3. VISUALIZATIONS OF AFFECT SPACE

Some of the operations supported by our system are directly supported and visible through its graphical user interface:

- Browsing the document structure from a single affect through sentence, paragraph and document to corpus; and displaying affect profiles for each element of the document structure; display of movie profiles for action movies from the list given in the source view. This is one of five different movie categories we analyzed – action, comedy, family, romance, and science fiction – comprising about 100 movies; we also created personal movie profiles to find out about individuals' movie preferences. The profile for the movies (classified as action movies by reel.com) have a "negative" side and "positive" side. Placing opposite affect categories in opposite positions lets us see ratios between centralities for opposite pairs. For example, in action movies *horror* is more central than *humor*, *pain* than *pleasure*, and *fear* than *courage*, whereas other opposites are more balanced, like *justice* and *injustice*.
- Various display options. We can show categories arranged in a circle and sorted by centralities; opposite categories are across from each other. A bar chart represents the centralities of the affect categories in the selected paragraph.
- Discovery of affect category groups (similarity classes) through fuzzy thesaurus browsing. Thesaurus Browser displays relationships among affect categories from the present profile. Higher centralities are shown in the thesaurus with larger font and nodes. We can filter out relationships with degrees lower than a certain threshold, what allows us to display a set of closely related relationships. We can use the information on relationships to examine spurious categories in our profile, i.e., those which are not related to any other category, and which thus indicate a representation of the content that is erroneous.

4. CONCLUSION

We believe the system can be useful in text browsing, decision making, and other creative tasks that include text manipulation and mining, including support for creative writing. We see several directions for improving and expanding the ideas presented in this paper. First, information extraction techniques can be employed for structure-dependent typing. For example, modifiers can affect the centralities and/or intensities of the word they modify. We can also exclude certain phrases from affect analysis to avoid spurious taggings so that, e.g., *wild card* is distinguished from other uses of *wild*. Second, a weighting scheme can be introduced to emphasize elements of text that have higher importance, such as titles, summaries or sentences closer to the beginning of a document. This would make centralities of affects higher in these regions. We also plan to integrate our approach with existing statistical text management methods and to implement quasi-statistical approaches using affect categories. In a quasi-statistical approach, a sigma count of affect categories' centralities would be equivalent to term frequency count. Finally, we need to expand the affect lexicon to cover a broader affect vocabulary. We currently have about four thousand entries, and are considering automatic expansion using WordNet ([2]). Depending on the domain of utilization of our system, we may also shift our focus to other specialized vocabularies.

5. REFERENCES

[1] D. Dubois, H. Prade and C. Testemale, Weighted Fuzzy Pattern Matching, pp. 313-331, *Fuzzy Sets and Systems 28*, North-Holland, 1988.

[2] C. Fellbaum, Wordnet: An Electronic Lexical Database, MIT Press, 1998.

[3] C. Fillmore and B.T.S. Atkins, FrameNet and Lexicographic Relevance, pp. 417-420, *First International Conference on Language Resources & Evaluation: Proceedings*, 1998.

[4] T. Fontanelle, Semantic Tagging: A Survey, pp. 39-56, *Papers in Computational Lexicography, COMPLEX 99*, 1999.

[5] R. Krovetz and W.B. Croft, Lexical ambiguity and information retrieval, 10(2):115-141, *ACM Transactions on Information Systems*, 1992.

[6] G. Miller and C. Walter, Contextual correlates of semantic similarity, 6:1-28, *Language and Cognitive Processes*, 1991.

[7] P. Subasic, A. Huettner, Affect Analysis of Text Using Fuzzy Semantic Typing, *FUZZ-IEEE 2000, San Antonio*, May 2000.

[8] L. A. Zadeh, Similarity Relations and Fuzzy Orderings, pp.177-200, *Information Sciences 3*, Elsevier Science, 1977.

[9] R. Zwick, E. Carlstein and D. V. Budescu, Measures of Similarity Among Fuzzy Concepts: A Comparative Analysis, 1:221-242, *Int. Journal of Approximate Reasoning*, Elsevier Science, 1987.