Fuzzy Typing for Document Management

Alison HUETTNER
Clairvoyance Corporation
5301 Fifth Avenue
Pittsburgh, PA 15232 USA
a.huettner@clairvoyancecorp.com

Pero SUBASIC
Clairvoyance Corporation
5301 Fifth Avenue
Pittsburgh, PA 15232 USA
p.subasic@clairvoyancecorp.com

This prototype system demonstrates a novel method of document analysis and management, based on a combination of techniques from NLP and fuzzy logic. Since the central technique we use from NLP is semantic typing, we refer to this approach as *fuzzy typing for document management*.

The fuzzy typing approach is general in scope and can be applied to many different kinds of analysis. In this prototype, we illustrate its use in analyzing affect. At a basic level, it involves:

- Isolating a vocabulary of words belonging to a metalinguistic domain (here, *affect* or emotion)
- Using multiple categorizations 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 domain words
- Manipulating the profiles to categorize, differentiate, cluster, match, or visualize the texts

Our *affect lexicon* was generated from a list of roughly 4,000 English words denoting or connoting affect – from *abhor* and *abject* to *yen* and *yucky*. We created a small set of semantic categories, representing basic, core emotions (e.g., ***anger, happiness, fear***) as well as some recurring abstract themes (e.g., ***superiority, violence, death***). Each word in the lexicon is assigned to as many categories as necessary to capture all aspects of its meaning. Ambiguous words are not disambiguated, but simply assigned to all categories relevant to any of their meanings. Each association between a domain word and a category is assigned a numerical *centrality* ranging from 0 to 1, representing the degree of relatedness between the word and the category. It is also assigned a numerical *intensity*, representing the strength of the affect conveyed by the word. A lexical entry thus has the following fields:

```
<domain_word> <part_of_speech>
<semantic_category> <centrality>
<intensity>
```

As a concrete example, the lexical entries for the domain word *gleeful* look like this:

```
"gleeful" adj happiness 0.7 0.6
"gleeful" adj excitement 0.3 0.6
```

This representation captures the fact that *gleeful* is an adjective primarily expressing a kind of happiness (centrality 0.7 on the ***happiness*** scale), but including also a less prominent element of excitement (centrality 0.3 for ***excitement***). Its intensity is mid-range (0.6) with respect to both categories.

Centrality and intensity ratings are very subjective. They are typically arrived at by reviewing a large number of words assigned to a particular semantic category and manually ranking them from most to least central, then from most to least intense, with respect to that category. At centrality 0.7, *gleeful* is less central with respect to ***happiness*** than *contentment*, but more central than *dreamy*. At centrality 0.3, it is less central with respect to ***excitement*** than *agitation*, but more central than *cry*. The intensity of 0.6 indicates that *gleeful* denotes a ***happiness*** level less intense than *joy*, but more intense than *glad*. The other intensity score, also 0.6, conveys an ***excitement*** level below *hysterical* but above *disturbance*. In future, we plan to explore automating this ranking process, e.g., by using distance metrics on a thesaurus.

A central construct in our affect analysis is the *affect set*. An affect set comprises the set of unique affect categories from a given text, with attached centralities and intensities. It is generated by parsing the text to generate word/part-of-speech pairs; normalizing the words in accordance with morphological rules; and looking up the normalized pairs in the affect

lexicon. If such a pair has a lexical entry, we retrieve all of its affect categories with their associated centrality and intensity scores. We create the initial affect set for the text by merging all instances of the same category that were retrieved from the lexicon for the text as a whole. The centrality score for a merged category is computed as the fuzzy union of the centralities of all instances; it represents the centrality of that category for the text as a whole. The intensity score for a merged category is computed as a simple average of the intensities of its instances, and represents the intensity of that category for the text as a whole. An affect set can be computed for a text of any length, from a single word to a whole corpus.

A *query* is a specialized affect set for the purpose of retrieval. It can be created directly, by specifying a set of affect categories with attached centralities and intensities, or indirectly (automatically), from examples of desirable documents. Given an affect lexicon, we can also create a *fuzzy thesaurus*, which establishes relationships between pairs of affect categories based on the centralities of words assigned to both categories in the lexicon. The fuzzy thesaurus can then be used for expanding queries, in a process similar to thesaurus extraction in information retrieval. For example, if an affect set consists of ***love/***0.7, the user can expand it using the fuzzy thesaurus, so that related categories like ***attraction*** will be added to the set automatically.

Our demonstration prototype shows some interesting visualizations of affect sets, which can be understood as affective fingerprints for texts. The prototype supports:

- hierarchical browsing in text (affect word/sentence/paragraph/document), with simultaneous visualization of affect sets
- several visualization modes for affect sets, including ordered and opposite modes for centralities or intensities
- convenient exploration of category clusters
- quick search on single affect categories and on logical combinations of categories
- quick search on affect category clusters
- purely visual composition of complex affect profiles and queries, including individual centralities/intensities
- thresholds on overall intensity, fuzziness, and search targets

Higher-level tasks of direct interest for qualitative/semantic analysis of text can be achieved by combining these basic operations. For example, a user might compare affect profiles for multiple texts, query by example, or discover spurious affect categories. Finally, the model explored here can be adjusted to support summarization of affect profiles, as well as affect-based user profiling, classification, and clustering at different text levels.

Fuzzy typing represents an innovative way to capture metalinguistic facts about a text while accommodating linguistic ambiguity and vagueness. Computations and interface are independent of lexicon content, making it easy to change domains. The approach is useful in an indefinite number of areas, and lends itself to customization for a particular user or task.

Acknowledgements

Many thanks to Dr. Mark Kantrowitz for the use of his initial affect wordlist.

References

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

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

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

R. Krovetz and W.B. Croft, Lexical ambiguity and information retrieval, *ACM Transactions on Information Systems* 10(2), 1992, pp. 115—141.

G. Miller and C. Walter, Contextual correlates of semantic similarity, *Language and Cognitive Processes* 6, 1991, pp. 1—28.

R. W. Picard, *Affective Computing*, MIT Press, 1997.

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

A. Tversky, Features of Similarity, *Psychological Review* 84, 1977, pp. 327—352.

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

R. Zwick, E. Carlstein and D. V. Budescu, Measures of Similarity Among Fuzzy Concepts: A Comparative Analysis, *International Journal of Approximate Reasoning* 1, Elsevier Science, 1987, pp. 221—242.