Towards a Cognitive Approach for the Automated Detection of Connotative Meaning

Jaime Snyder*, Michael A. D'Eredita, Ozgur Yilmazel, and Elizabeth D. Liddy

School of Information Studies, Syracuse University, Syracuse, NY, USA
* Corresponding author, jasnyd01@syr.edu

1 Introduction

The goal of the research described here is to automate the recognition of connotative meaning in text using a range of linguistic and non-linguistic features. Pilot results are used to illustrate the potential of an integrated multidisciplinary approach to semantic text analysis that combines cognitive-oriented human subject experimentation with Machine Learning (ML) based Natural Language Processing (NLP). The research presented here was funded through the Advanced Question and Answering for Intelligence (AQUAINT) Project of the U.S. federal government's Intelligence Advanced Research Projects Activity (IARPA) Office. Funded as an exploratory "Blue Sky" project, this award enabled us to develop an extensible experimental setup and to make progress towards training a machine learning system.

Automated understanding of connotative meaning of text requires an understanding of both the mechanics of text and the human behaviors involved in the disambiguation of text. We glean more from text than what can be explicitly parsed from parts of speech or named entities. There are other aspects of meaning that humans take away from text, such as a sincere apology, an urgent request for help, a serious warning, or a perception of personal threat. Merging cognitive and social cognitive psychology research with sophisticated machine learning could extend current NLP systems to account for these aspects. Building on current natural language processing research [**?**], this pilot project encapsulates an end-to-end research methodology that begins by 1) establishing a human-understanding baseline for the distinction between connotative and denotative meaning, 2) then extends the analysis of the mechanics of literal versus non-literal meaning by

Proceedings of the 8th International Conference on Computational Semantics, pages 326–332,
Tilburg, January 2009.

applying NLP tools to the human-annotated text, and 3) uses these cumulative results to feed a machine learning system that will be taught to recognize the potential for connotative meaning at the sentence level, across a much broader corpus. This paper describes the preliminary iteration of this methodology and suggests ways that this approach could be improved for future applications.

2 Analytic framework: A cognitive approach

We view an excerpt of text to be a stimulus, albeit much more complex than most stimuli used in typical psychological experiments. The meaning of any excerpt of text is tieds to a constructive cognitive process that is heavily influenced by previous experience and cues, or features, embedded within the text. Our goal is to gain a better understanding of (1) what features are attended to when the text is being interpreted, (2) which of these features are most salient and (3) how these features affect connotative meaning.

One's ability to derive connotative meaning from text is behavior that is learned, becoming intuitive in much the same way an individual learns any skill or behavior. When this process of attending and learning is repeated across instances, specific skills become more automatic, or reliable [**?**, **?**]. This process is considered to be constructive and episodic in nature, yet heavily dependent upon "cues" that work to draw or focus one's attention [**?**]. Further, research on communities suggests that the meaning of an artifact (e.g., a specific excerpt of text) is heavily influenced by how it is used in practice [**?**] The meaning of text is constructed in a similar manner. Members of a speech community tend to make similar assumptions, or inferences. The mechanics of making such inferences are scaled to the amount of contextual information provided. Our preliminary research suggests that when presented with a sentence that is out of context an individual seemingly makes assumptions about one or all of the following: who created the text, the context from which it was pulled and the intended meaning given the features of the text.

3 Methods

3.1 Data

Blog text was used as the corpus for this research. Sentences were deemed the most practical and fruitful unit of analysis because words were consid-

ered too restrictive and pieces of text spanning more than one sentence too unwieldy. A single sentence presented enough context while still allowing for a wide range of interpretation. Sentences were randomly selected from a pool of texts automatically extracted from blogs, using a crawler set with keywords such as "oil", "Middle East" or "Iraq." Topics were selected with the intention of narrowing the range of vocabulary used in order to aid the machine learning experiments.

3.2 Preliminary phase

To start, we conducted a series of eight semi-structured, face-to-face interviews. Individuals were presented with 20 sentences selected to include some texts that were expected to be perceived as highly connotative as well as some expected to be perceived as highly denotative. Each interviewee was asked to exhaustively share all possible meanings they could derive from the stimulus text, while also pinpointing what it was about the text that led them to make their conclusions. Based on these interviews, we modified our probes slightly and moved the human evaluation process to an open-ended, on-line instrument in order to increase the number of responses. We presented a series of 20 sentences to participants (N=193) and, for each stimulus text, asked: 1) "What does this sentence suggest?" & "What makes you think this?"; and 2) "What else does this sentence suggest?" & "What makes you think this?" Upon analysis of the responses, we found that while interpretations of the text were relatively idiosyncratic, how people allocated their attention was more consistent. Most people tended to be making assumptions about the (1) author (addressing who created the artifact), (2) context (addressing from where the sentence was taken) and/or (3) intended meaning of the words. We interpreted this to mean that these three areas were potentially important for identifying inferred meanings of texts.

3.3 Design of pilot experiment

Next, our efforts focused on designing a reusable and scalable online evaluation tool that would allow us to systematically gather multiple judgments for each sentence using a much larger pool of stimulus text. Scaling up the human evaluations also allowed us to decipher between responses that were either systematically patterned or more idiosyncratic (or random). According to our forced-choice design, each online participant was presented with a series of 32 pairs of sentences, one pair at a time, and asked to identify the sentence that provided more of an opportunity to read between the lines.

Half the participants were presented with a positive prompt (which sentence provides the most opportunity) and half were presented with a negative prompt (which sentence provides the least opportunity). Positive/negative assignment was determined randomly. The 16 sentences selected during the first round were re-paired in a second round. This continued until 4 sentences remained, representing sentences that were more strongly connotative or denotative, depending on the prompt. Final sentence scores were averaged across all evaluations received.

The forced choice scenario requires a sample of only 13 participants to evaluate 832 sentences. This was a significant improvement over previous methods, increasing the number of sentences and the number of evaluations per sentence and therefore increasing the reliability of our findings. For example, using this scalable setup on a set of 832 sentences we need only 26 participants to generate two evaluations per sentence in the set, 39 participants to yield three evaluations per sentence, etc. We ran the system with a randomly selected sample of both sentences and participants with the intent to eventually make direct comparison among more controlled samples of sentences and participants. This has direct implication for the evaluation phase of our pilot. Because sentences were selected at random, without guarantee of a certain number of each type of sentence, our goal was to achieve results on a par with chance. Anything else would reveal systematic bias in the experiment design or implementation. This also provides us with a baseline for future investigations where the stimulus text would be more wilfully controlled.

4 Results

4.1 Evaluation of text by human subjects

In the first iteration of the pilot setup, each of 832 sentences were viewed by six different participants, three assigned to a positive group and three to a negative group, as described above. The denotative condition ranged in ratings from 0 to -3 while the connotative condition ranged in rating from 0 to 3. These were then averaged to achieve an overall score for each sentence. Because they were randomly selected, each sentence had predictable chance of ultimately being identified as connotative or denotative. In other words, each sentence had an equal chance of being identified as connotative.

Having established a baseline based on chance, we can next control for various features and evaluate the relative impact as systematic differences from the baseline. We will be able to say with a relatively high degree of

certainty that "x," "y" or "z" feature, sentence structure, behavior, etc. was responsible for skewing the odds in a reliable manner because we will be able to control for these variables across various experimental scenarios. This, combined with improved validity resulting from an increased number of human judgments and an increased number of sentences viewed, marks the strength of this methodology.

Additionally, we will be able to compare sentences within each scenario even when an overall chance outcome occurs. For example, in the initial run of our sentences, we achieved an overall chance outcome. However, "anomalies" emerged, sentences that were strongly skewed towards being assigned a neutral evaluation score or towards an extreme score (either distinctly connotative or distinctly denotative). This allowed us to gather a reliable and valid subset of data that can be utilized in ML experiments. See below for a very short list of sample sentences grouped according to the overall scores they received determine by the six human reviewers:

Denotative examples-

- The equipment was a radar system.
- Kosovo has been part of modern day Serbia since 1912.
- The projected figure for 2007 is about $ 3100.

Connotative examples-

- In fact, do what you bloody well like .
- But it's pretty interesting , in a depressing sort of way .
- It's no more a language than American English or Quebecois French

4.2 Experimental Machine Learning system

Our preliminary analysis suggests that humans are consistent in recognizing the extremes of connotative and denotative sentences and an automatic recognition system could be built to identify when a text is likely to convey connotative meaning. Machine Learning (ML) techniques could be used to enable a system to first classify a text according to whether it conveys a connotative or denotative level of meaning, and eventually, identify specific connotations. ML techniques usually assume a feature space within which the system learns the relative importance of features to use in classification. Since humans process language at various levels (morphological, lexical, syntactic, semantic, discourse and pragmatic), some multi-level combination of features is helping them reach consistent conclusions. Hence, the initial machine learning classification decision will be made based on a class of critical

features, as cognitive and social-cognitive theory suggests happens in human interpretation of text.

TextTagger, an Information Extraction System developed at Syracuse University's Center for Natural Language Processing, currently can identify sentence boundaries, part-of-speech tag words, stem and lemmatize words, identify various types of phrases, categorize named entities and common nouns, recognize relations, and resolve co-references in text. We are in the process of designing a ML framework that utilizes these tags and can learn from a few examples provided by the human subject experiments described above, then train on other sets of similar data marked by analysts as possessing the features illustrated by the sentences consistently identified as conveying connotative meaning.

For preliminary ML-based analysis, the data collection included 266 sentences (from the original 832 used in human subject experiments), 145 tagged as strongly connotative and 121 tagged as strongly denotative by subjects. Fifty sentences from each set became a test collection and the remaining 95 connotative and 71 denotative sentences were used for training. Our baseline results (without TextTagger annotations) were: Precision: 44.77 ; Recall: 60; F: 51.28. After tagging, when we only use proper names and common nouns the results improved: Precision: 51.61 Recall: 92; F: 67.13. Although these results are not as high as some categorization results reported in the literature for simpler categorization tasks such as document labeling or spam identification, we believe that using higher level linguistic features extracted by our NLP technology will significantly improve them. More sophisticated analysis will be conducted during future applications of this methodology.

5 Discussion and Future Work

By allowing the ML system to do time- and labor-intensive analysis, and exploiting a natural human ability to "know it when they see it" (in this case "it" referring to connotative meaning), we feel that this pilot methodology has great potential to deliver robust results. In addition to the significant contribution this research will make in the area of natural language processing, it will also provide a model for future work that seeks to create similar bridges between psychological investigation and system building. Preliminary results suggest that our approach is viable and that a system composed of multiple layers of analysis-with each level geared towards reducing the variability of the next-holds promise.

Future work will concentrate efforts in two areas. First, the notion of speech communities will be addressed. The pilot study looked at a very generalized speech community, expecting to achieve equally generalized results. While this has merit, there is much to be learned by implementing this approach using a more targeted community. Second, the protocol used in this pilot study was run using a relatively modest number of human evaluators and a relatively small set of data. With the experience gained during the pilot, the reliability of the data used to train the ML system can be easily improved by increasing the size of both human subject samples and data sets. With a more robust set of initial data, ML experiments can progress beyond the basic proof-of-concept results reported here and produce actionable feature sets tuned to specific speech communities.

References

[1] M. A. D'Eredita and C. Barreto. How does tacit knowledge proliferate? *Organization Studies*, 27(12):1821, 2006.

[2] E.D. Liddy, E. Hovy, J. Lin, J. Prager, D. Radev, L. Vanderwende, and R. Weischedel. Natural Language Processing. *Encyclopedia of Library and Information Science*, pages 2126–2136, 2003.

[3] G. D. Logan. Toward an instance theory of automatization. *Psychological Review*, 95(4):492–527, 1988.

[4] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1999.

[5] R.S. Wyer and J.A. Bargh. *The Automaticity of Everyday Life*. Lawrence Erlbaum Associates, 1997.