Computational Metaphor Identification in Communities of Blogs

Eric Baumer and Bill Tomlinson

Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440
{ebaumer, wmt} @ ics.uci.edu

Abstract

This poster presents a computational analysis of conceptual metaphors in a community of political blogs. Like sentiment analysis or opinion extraction, computational metaphor identification can provide an understanding of the framings or conceptualizations used in a community. This poster includes an implementation overview and results summary.

Introduction

As blogs become a more important part of internet culture, news media, and people's daily lives, there is an increasing interest in methods of making sense of this vast and growing body of content, e.g., automatically determining topics of conversation (Baumer and Fisher 2008), or extracting opinions expressed (Yang et al. 2007). Another potential perspective of a community can be provided by examining the conceptual metaphors that it uses. Consider, for example, the language used to describe having an argument. "I *attacked* his *position*." "Your claims are *indefensible*." "She *obliterated* her *opponent*." These words evince images of war. Lakoff and Johnson (1980) describe this language as evidence for the conceptual metaphor that argument is war. This poster explores the extent to which the set of metaphors used by a community of blogs can be informative about that community.

Previous approaches, e.g., (Martin 1990), have focused mostly on discerning metaphorical language from literal, and then performing additional processing on the metaphorical language in order to compute its literal, "true" meaning. In the computational metaphor identification (CMI) approached used here, the goal is not to view individual phrases as metaphorical or literal, but to identify conceptual metaphors that underly a body of text.

Implementation Overview

The techniques used in this implementation extend those from CorMet (Mason 2004). The main difference is that CorMet was designed to extract known metaphorical mappings between corpora of pre-determined domains, whereas the techniques presented here are used to identify potential conceptual metaphors in novel target corpora.

This CMI implementation hinges largely on selectional preference learning (Resnik 1993). For example, the English verbs "eat" or "drink" tend to have a *human*[1] or *animal* as the subject and *food* or *potable liquid* as the direction object, respectively. The selectional preference strength for a given verb-case slot pair is calculated by taking the relative entropy of the prior distribution over word classes and the posterior distribution conditioned on a given verb-case slot:

$$S_R(v) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}$$

A typed dependency parser (de Marneffe, MacCartney, and Manning 2006) generates grammatical relations for calculating selectional preferences. Selectional preference learning requires classes of words, but the corpus contains word tokens. WordNet synsets are used for word classes, where a single word token counts as a partial observation of any of the synsets it might represent. Selectional preference is the overall choosiness of a verb-case slot, while selectional association is the strength with which a verb-case slot select for some synset (Resnik 1993). Selectional associations are calculated for those verbs in a corpus with the highest frequency relative to general English, derived from the British National Corpus (BNC).

The synsets for which those relatively frequent verbs select are then clustered. Each synset is represented as a vector of selectional associations, where the n^{th} element of the vector is the degree to which the n^{th} verb-case slot selects for that synset. The result is a set of conceptually coherent clusters of synsets. For example, in the domain of a chemistry or biology LAB, the verbs "pour," "flow," "freeze," and "evaporate" all select for *liquids* and *fluids*, so the corpus of documents from this domain would result in a cluster of *fluids* and *liquids*.

The mapping process involves a corpus of documents in a specified source domain and a target corpus in which to identify metaphors. Source corpora are drawn from Wikipedia. Every Wikipedia article belongs to at least one category, and categories are organized into a directed

[1]ALL CAPS are domains, Capital first letters are Wikipedia categories, SMALL CAPS are concepts in a metaphor, and *italics* are word classes, and "quotes" are specific words that are part of an instance of a metaphor.

Metaphor	Target	Source	verb-case slot	Conf	Example Sentence
TERRORISM is an EVIL ATTACK	terrorism	evil, Satan	fight - prep_against	0.104	"The USA Today report... sparked a renewed debate over government intrusion into Americans' civil liberties in the *fight against terrorism*."
		hit, collision	protect - prep_against	0.036	
		war, hostility, campaign	violate - nsubj	0.002	
		battle, struggle, conflict	violate - nsubj	0.001	
TERRORISM is an EVIL ATTACKER	terrorism	evil, Satan	fight - prep_against	0.104	"US... efforts to *combat terrorism*."
		bandit, brigand	combat - dobj	0.007	
		intruder, stranger	violate - nsubj	0.007	
RELIGION is an ATTACK	religion, faith	intruder, stranger	violate - nsubj	0.018	"His '*faith* based' initiatives *violate* the first [amendment]."
		war, hostility, campaign	violate - nsubj	0.013	
		battle, struggle, conflict	violate - nsubj	0.012	

Table 1 *– Metaphors and supporting mappings in the Think Progress community. The first column is the manually assigned metaphor. The next four columns describe the computationally identified mappings supporting a metaphor: salient term(s) from the target and source clusters, mediating verb-case slot, and confidence score. The last column is an example sentence from the corpus (emphases added).*

acyclic graph. The corpus for a domain is all the Wikipedia pages in a given category, such as Military, and all the pages in all its subcategories, such as Military history.

Mappings are sought using selectional associations for clusters across different corpora. The selectional association of a verb-case slot for a cluster is the average of that verb-case slot's selectional association for each member of the cluster. The polarity of a mapping from cluster x in domain A to cluster y in domain B is:

$$\sum_{\alpha} \Lambda(\alpha,y,B)*0.75+\Lambda(\alpha,x,A)*0.25$$

where α ranges over the verb-case slots that select for x in A, and Λ(α, x, A) is the selectional association strength of the verb-case slot α for cluster x in domain A. The overall confidence for a mapping M is:

$$\frac{|vcs(M)|}{tot\_vcs}*0.25+\frac{pol(M)}{max\_pol}*0.75$$

where *vcs*(*M*) is the number of verb-case slots that mediate a given mapping, *max_vcs* is the total number of verb-case slots selecting for the two clusters being mapped, *pol(M)* is the polarity of the given mapping, and *max_pol* is the maximum polarity of all mappings between the two domains in question.

Test Evaluation

This CMI technique was applied to a community of blogs from the 2007 ICWSM dataset (http://www.icwsm.org/data.html). Starting with Think Progress (http://thinkprogress.org), a community of blogs was identified using (Bulters and de Rijke 2007). The resulting community from this data set contained 23 blogs with 1950 posts, yielding a corpus of 645,421 words. Mapping were then sought from the MILITARY domain, a corpus of 1867 articles containing 507,589 words.

The first two metaphors in table 1, framing TERRORISM as an EVIL ATTACK or an EVIL ATTACKER, are not entirely surprising, given the "war on terror" rhetoric in US politics of the time. The metaphor RELIGION is an ATTACK also makes sense, given that Think Progress is a liberal blog, and most US liberals negatively view the incorporation of religion in politics. For someone not familiar with the Think Progress community, these metaphors provide an introduction to the conceptual framings used in the community's discussions. The results demonstrate that CMI can effectively identify meaningful and informative conceptual metaphors.

References

Baumer, E. and Fisher, D. 2008. Smarter Blogroll: An Exploration of Social Topic Extraction for Manageable Blogrolls. In *Hawai'i Int'l Conf on System Sciences*, IEEE Computer Society.

Bulters, J. and de Rijke, M. 2007. Discovering Weblog Communities: A Content- and Topology-Based Approach. In *Int'l Conf on Weblogs and Social Media*, AAAI Press.

de Marneffe, M., MacCartney, B., and Manning, C. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In Language Resources and Evaluation.

Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By.* Chicago, IL: University of Chicago Press.

Martin, J. 1990. *A Computational Model of Metaphor Interpretation.* San Diego, CA: Academic Press.

Mason, Z. 2004. CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics* 30(1): 23-44.

Resnik, P. 1993. Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. diss, Dept of Comp and Info Sci, Univ of Penn., Philadelphia, PA.

Yang, K., et al. 2007. Fusion Approach to Finding Opinions in Blogosphere. In *Int'l Conf on Weblogs and Social Media*, AAAI Press.