Revealing Contentious Concepts Across Social Groups

Zumrut Akcam[1], Ching-Sheng Lin[1], Samira Shaikh[1], Sharon Gower Small[2], Ken Stahl[1], Tomek Strzalkowski[1], Nick Webb[3]

[1]ILS Institute, University at Albany, State University of New York
[2] Siena College Institute for Artificial Intelligence, Siena College
[3]Department of Computer Science, Union College

{zakcam, clin3, sshaikh, kstahl, tomek}@albany.edu, ssmall@siena.edu, webbn@union.edu

Abstract

A computational model based on concept polarity is proposed to investigate the influence of communications across 'diacultural groups'. The hypothesis is that there are communities which can be characterized by a network of concepts and the corresponding valuations of those concepts that are agreed upon by the members of the community. We apply an existing research tool, ECO, to generate community specific Valuation Concept Networks (VCN). We then compare VCNs across communities, to attempt to find 'contentious concepts', which could subsequently be the focus of further exploration as points of contention between the two communities.

1. Introduction

Increasingly, groups of users congregate together around particular forums or blogs to express themselves in relation to new or recurring events. In this study, we aim to discover the contentious concepts between opposing diacultural groups to find those concepts which occur in both communities postings, but which are valued differently in each. At an elementary level, concepts can be though as singularities (e.g., persons, locations, organization) that are invoked by appropriate references in various forms of communication. We will identify these units, along with their valuation of a community, and compare them with similar concepts in a different, opposing community. The potential of this technique is in it's use for government or business units to identify and monitor different points of view, with respect to specific issues, or across specific groups.

We created a prototype, CPAM (Changing Positions, Altering Minds), as a proof of concept. In Section 2, we review related research approaches. In Section 3, we describe the components of the proposed technique and the way they are used to establish CPAM. In Section 4, we discuss empirical studies. In final section, we present conclusions and future work.

2. Related Work

The automatic detection of opinions and sentiment in text (cf. (Wiebe et al., 2005; Breck et al., 2007; Strapparava and Mihalcea 2008)) and speech (cf. (Vogt et al., 2008)) is a rapidly emerging area of research interest. In common with prior work, we depend on a subjectivity lexicon (derived from the MPQA corpus (Wiebe et al., 2005)) and opt for a mechanism that is of a deeper level of understanding than bag-of-words, and yet does not necessitate deeper syntactic relationships. For CPAM, initial concept identification and annotation is based on our prior work in ECO (Effective Communication Online) (Small et al., 2010). The purpose of ECO is to extract and model the valuation system of the community and compare whether the contents of a new message fits into the targeted community. Users are guided in ways to shape their communication such that it eliminates, or mitigates the number of conflicting concept valuations between the new message and target community. To achieve this, we first derive the salient concepts and corresponding polarities for the targeted community. We do this using a Transformation-Based Learning (TBL) approach, using lexical items, concepts, POS labels and the presence of polarity words in the input as learning features, and producing a Valuation System Vector (VSV). Accordingly, the same mechanism will be applied to the new message to obtain the Message System Vector (MSV). Finally, a comparison between two vectors is performed, identifying mismatches of concepts, and highlighting these to the user for possible amendment. An example of comparison of vectors is shown in Figure 1.

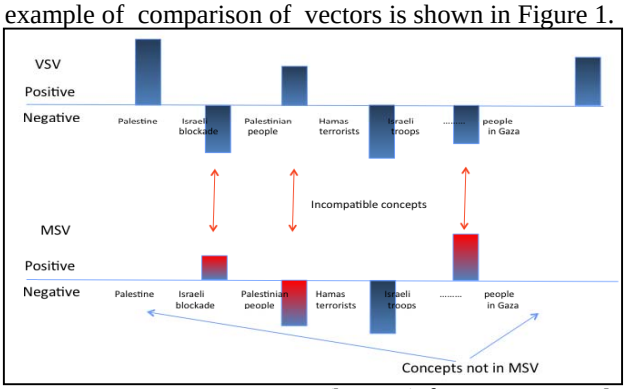

Figure 1. Comparing a MSV (bottom) from a pro-Israeli message to a VSV (top) from a pro-Palestinian blog. Incompatible concepts are highlighted in red.

3. CPAM Methodology

In this section, we describe the fundamental components for CPAM and explain how we use them to build the model for finding the contentious concepts.

3.1 Valuation Concept Network

Taking a Valuation System Vector from a particular community, we extend that vector by including inter-conceptual temporal relationships for each community. By temporal relationships, we mean 'temporal in text', where one concept occurs prior to another in the source material, rather than existing in any formal temporal relationship. We believe that concepts as they occur in text are not accidental, that much like newspaper text, the occurrence of concepts is deliberate, and that one may be able to infer loose, causal relationships between two valued concepts. Take the example from an Israeli blog, detailing the barricade erected around Israeli settlements: "the barriers positive effect is in the control of arms to Palestinian terrorists". Here 'barrier' is positively valued, and 'Palestinian terrorists' is negative, and we might infer that the positive value of the former is in controlling the negative concept that is the latter. For each community, we extract all the concepts and their valuations, then for each concept, we form a sequence of relationships by looking at the sequence of concepts in the text.

We use the open source graph software, Graphviz, to display the resulting network structure. Node size represents the overall frequency of the concept and colour represents the net cumulative count of negative and positive instances across all blog texts. Edge direction represents the ordering of concepts and edge thickness represents the frequency of the relationship. In Figure 2, the networks for blog data are shown.

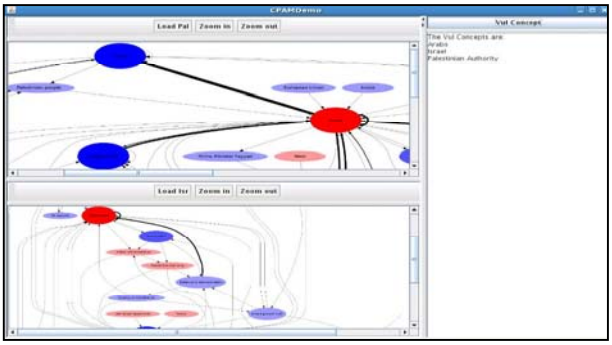

Figure 2. Left-top is the VCN for pro-Palestinian and left-bottom is for pro-Israeli.

3.2 Contentious Concepts

The goal of CPAM is to find and display potential contentious concepts. Our definition of a contentious concept is one that is shared between networks (therefore appears in blog postings of both communities) but the valuation of that concept is opposing between the communities. However, this by itself is insufficient. Ideally, we want to identify contentious concepts that share, across communities, some concept which can be used to negotiate the understanding of the target concept. Therefore, there also needs to be another shared concept, either preceding or following the contentious concept, where the valuation is also shared between the two networks. Once the networks of two communities are created, we identify contentious concepts by comparing all relationships between two communities. A contentious concept example is show in Figure 3.

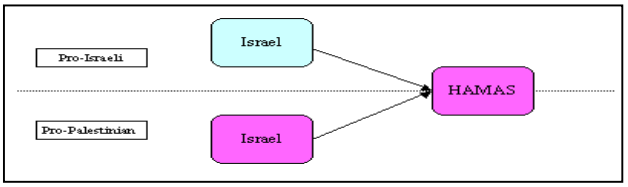

Figure 3. Two communities share a judgement on the concept "HAMAS" (which is negative). The two communities have different valuations on the concept "Israel".

4. Data Collection

In this section, we discuss the data we collect and annotate to form the basis of the CPAM system. The first step of data generation starts with manual annotation. Two blog communities, pro-Palestinian and pro-Israeli, are collected for evaluation. From those blog sites, we have collected 36000 words for pro-Palestinian data and 27000 words for pro-Israeli. Within these texts there are 577 and 362 concepts respectively that are assigned judgments by our multiple annotators. Of those, there are 335 and 213 unique concepts in the pro-Palestinian and pro-Israeli data sets. We then create the networks from these concepts as described in Section 3. These networks contain relationships of varying lengths, so to limit data sparsity, we use a manually selected threshold distance of 2 in our initial experiments. In the pro-Palestinian data there are 885 relationships between concepts within window length 2 and 552 in the pro-Israeli. Summary information of the initial data sets is in Table 1.

We then attempted to extract contentious concepts, as per our definition in Section 3. We found only 3 contentious concepts ("Israel", "Palestinian Authority" and "Arabs") that matched our definition. From table 1, there are 335 unique concepts in pro-Palestinian and 213 in pro-Israeli. However, we calculated the number of shared concepts between the two data sets consisted of only 34 concepts. It became clear that we needed to expand data set.

	Pro-Palestinian Data	Pro-Israeli Data
Words	36,000	27,000
Concepts	577	362
Unique Concepts	335	213
Relationships	885	552

Table 1: Data summary of our initial data sets.

We collected an additional 200,000 words of data from each set of blog sites. To annotate this data manually is infeasible, so we used our initial data to train our ECO annotation tool, based around a transformation-based learning approach. Applying this tagger to new data, we found around 2000 new concepts per community as predicted. Of those, there were around 775 unique concepts per data set. New data summary is in Table 2.

	Pro-Palestinian Data	Pro-Israeli Data
Words	200,000	227,000
Concepts	2177	2676
Unique Concepts	773	775
Relationships	3251	3610

Table 2: Data summary for expanded data set.

With the new data set, we see that there are 202 overlapping concepts that appear in both data sets providing a greater opportunity to find contentious concepts as per our criteria. Of those 202 concepts, we find 38 distinct contentious concepts, listed in Figure 4. Remember, these are concepts where the valuation of the concept is different between the two communities, yet are connected to concepts (not shown) which share valuations between communities. We see that we can conflate some concepts ("Israel", "state of Israel"), using word overlap or synonyms, and that there are errors in the automatic processing ("U" is cut off from "U S A"). It is instructive however to see known, highly contentious concept between the communities, such as "Goldstone Report" (one of the key phrase terms of a known contentious issue used to identify community specific blog sites).

MB, Obama, Egypt, Israel, Hamas, IDF, state of Israel, Israeli government, Middle East, West Bank, Goldstone, Gaza, Iran, Iraq, South Africa, Syria, U, US, PA, Gaza Strip, Jerusalem Post, Netanyahu, Palestinian Authority, Jerusalem, Lebanon, Mubarak, Geneva, America, Richard Goldstone, Tunisia, Jeffrey Goldberg, Europe, Goldstone Report, UN, United States, Mahmoud Abbas, Cairo, East Jerusalem

Figure 4: Concepts identified by our mechanism for detecting contentious concepts.

5. Conclusion

We described a computational model to discover and explore contentious concepts for two communities. The work is still in progress and we intend to enlarge the data set and try to determine intermediate roles played by concepts that are not explicitly judged in the source texts. Using these contentious concepts is another avenue of future work – first by attempting to identify inconsistencies within communities, to find those concepts where there is significant disagreement inside a single diacultural group.

References

Breck, E.; Choi, Y.; and Cardie, C. 2007. Identifying expressions of opinion in context. IJCAI.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21:543–565.

David, S., and Pinch, T. J. 2006. Six degrees of reputation: The use and abuse of online review and recommendation systems. First Monday. Special Issue on Commercial Applications of the Internet.

Hu, M., and Liu, B. 2004. Mining opinion features in customer reviews. In Proceedings of AAAI, 755–760.

Lager, T. 1999. The μ-TBL System: Logic Programming Tools for Transformation-Based Learning. CoNLL'99.

Small, S., Strzalkowski, T. and Webb, N. 2010. ECO: Effective Communication Online. Technical Report ILS-015, University at Albany.

Strapparava, C., and Mihalcea, R. 2008. Learning to Identify Emotions in Text. In SAC 2008.

Toutanova, K., and Manning, C. D. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of- Speech Tagger. EMNLP/VLC-2000, 63–70.

Vogt, T.; Andre´, E.; and Bee, N. 2008. Emovoice - a framework for online recognition of emotions from voice. In Proc. IEEE PIT 2008, volume 5078 of LNCS, pages 188–199. Springer

Webb, N.; Small, S.; and Shaikh, S. 2010. ECO Annotation Guidelines, Version 1.2. Technical Report ILS-012, University at Albany.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. Journal of Language Resources and Evaluation 39(2-3):165–210.