Language Use: What can it Tell us?

[name]
[address1]
[address2]
[address3]
[email]

[name]
[address1]
[address2]
[address3]
[email]

[name]
[address1]
[address2]
[address3]
[email]

Abstract

For 20 years, information extraction has focused on facts expressed in text. In contrast, this paper is a snapshot of research in progress on inferring properties and relationships among participants in dialogs, even though these properties/relationships need not be expressed as facts. For instance, can a machine detect that someone is attempting to persuade another to action or to change beliefs or is asserting their credibility? We report results on both English and Arabic discussion forums.

1 Introduction

Extracting explicitly stated information has been tested in MUC[1] and ACE[2] evaluations. For example, for the text *Mushaima'a, head of the opposition Haq movement*, an ACE system extracts the relation *LeaderOf(Mushaima'a, HaqMovement).* In TREC QA[3] systems answered questions, e.g. '*When was Mozart born?*', for which the answer is contained in one or a few extracted text phrases.

Sentiment analysis uses implicit meaning of text, but has focused primarily on text known to be rich in opinions (product reviews, editorials) and delves into only one aspect of implicit meaning.

Our long-term goal is to predict social roles in informal group discussion from language uses (LU), even if those roles are not explicitly stated; for example, using the communication during a meeting, identify the leader of a group. This paper provides a snapshot of preliminary, ongoing research in predicting two classes of *language use*: *Establish-Credibility* and *Attempt-To-Persuade*. Technical challenges include dealing with the facts that those LUs are rare and subjective and that human judgments have low agreement.

Our hybrid statistical & rule-based approach detects those two LUs in English and Arabic. Our results are that (1) annotation at the message (turn) level provides training data useful for predicting rare phenomena at the discussion level while reducing the requirement for turn-level predictions to be accurate; (2)weighing subjective judgments overcomes the need for high annotator consistency. Because the phenomena are rare, always predicting the absence of a LU is a very high baseline. For English, the system beats those baselines. For Arabic, more work is required, since only 10-20% of the amount of training data exists so far.

2 Language Uses (LUs)

A language use refers to an aspect of the social intention of how a communicator uses language. The information that supports a decision about an implicit social action or role is likely to be distributed over more than one turn in a dialog; therefore, a language use is defined, annotated, and predicted across a thread in the dialog. Because our current work uses discussion forums, threads provide a natural, explicit unit of analysis. Our current work studies two language uses.

An *Attempt-to-Persuade* occurs when a poster tries to convince other participants to change their beliefs or actions over the course of a thread. Typically, there is at least some resistance on the part of the posters being persuaded. To distinguish between actual persuasion and discussions that involve differing opinions, a poster needs to engage

[1] http://www-nlpir.nist.gov/related_projects/muc/
[2] http://www.nist.gov/speech/tests/ace/
[3] http://trec.nist.gov/data/qa.html

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 341–345,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

in multiple persuasion posts (turns) to be considered exhibiting the LU.

Establish-Credibility occurs when a poster attempts to increase their standing within the group. This can be evidenced with any of several moves, e.g., explicit statements of authority, demonstration expertise through knowledge, providing verifiable information (e.g., from a trusted source or citing confirmable facts), or providing a justified opinion (e.g., a logical argument or personal experience).

3 Challenges

There were two significant challenges: (a) sparsity of the LUs, and (b) inter-annotator agreement. To address the sparsity of data, we tried to automatically select data that was likely to contain content of interest. Data selection focused on the number of messages and posters in a thread, as well as the frequency of known indicators like quotations. (withheld). Despite these efforts, the LUs of interest were rare, especially in Arabic.

Annotation was developed using cycles of guideline development, annotation, evaluation of agreement, and revision of guidelines. Elsewhere, similar, iterative annotation processes have yielded significant improvements in agreement for word sense and coreference (Hovy et al., 2006). While LUs were annotated for a poster over the full thread, annotators also marked specific messages in the thread for presence of evidence of the language use. Table 1 includes annotator consistency at both the evidence (message) and LU level.

	English				**Arabic**			
	Msg		**LU**		**Msg**		**LU**	
	Agr	**#**	**Agr**	**#**	**Agr**	**#**	**Agr**	**#**
Per.	0.68	4722	0.75	2151	0.57	652	0.49	360
Cred.	0.66	3594	0.68	1609	0.35	652	0.45	360

Table 1: Number of Annotated Data Units and Annotator Agreement (measured as F)

The consistency numbers for this task were significantly lower than we have seen in other language processing tasks. Discussions suggested that disagreement did not come from a misunderstanding of the task but was the result of differing intuitions about difficult-to-define labels. In the following two sections, we describe how the evaluation framework and system development proceeded despite low levels of consistency.

4 Evaluation Framework

Task. The task is to predict for every participant in a given thread, whether the participant exhibits Attempt-to-Persuade and/or Establish-Credibility. If there is insufficient evidence of an LU for a participant, then the LU value for that poster is negative. The external evaluation measured LU predictions. Internally we measured predictions of message-level evidence as well.

Corpora. For English, 139 threads from Google Groups and LiveJournal have been annotated for Attempt-to-Persuade, and 103 threads for Attempt-to-Establish-Credibility. For Arabic, threads were collected from al-handasa.net.[4] 31 threads were annotated for both tasks. Counts of annotated messages appear in Table 1.

Measures. Due to low annotator agreement, attempting to resolve annotation disagreement by the standard adjudication process was too time-consuming. Instead, the evaluation scheme, similar to the pyramid scheme used for summarization evaluation, assigns scores to each example based on its level of agreement among the annotators. Specifically, each example is assigned positive and negative scores, $p = n^+/N$ and $n = n^-/N$, where n^+ is the number of annotators that annotate the example as positive, and n^- for the negative. N is the total number of annotators. A system that outputs positive on the example results in p correct and n incorrect. The system gets p incorrect and n correct for predicting negative. Partial accuracy and F-measure can then be computed.

Formally, let $\underline{X} = \{x_i\}$ be a set of examples. Each example x_i is associated with positive and negative scores, p_i and n_i. Let $r_i = 1$ if the system outputs positive for example x_i and 0 for negative. The partial accuracy, recall, precision, and F-measure can be computed by:

$$pA = 100\times\sum_i(r_i p_i+(1-r_i)n_i) \,/\, \sum_i(p_i+n_i)$$

$$pR = 100\times\sum_i r_i p_i \,/\, \sum_i p_i$$

$$pP = 100\times \sum_i r_i p_i \,/\, \sum_i r_i$$

$$pF = 2\; pR\; pP/(pR+pP)$$

The maximum pA and pF may be less than 100 when there is disagreement between annotators. To achieve accuracy and F scores on a scale of 100, pA and pF are normalized using the maximum achievable scores with respect to the data.

$$npA = 100\times pA/max(pA)$$

$$npF = 100\times pF/max(pF)$$

[4] URLs and judgments are available by email.

5 System and Empirical Results

Our architecture is shown in Figure 1. We process a thread in three stages: (1) linguistic analysis of each message (post) to yield features, (2) Prediction of message-level properties using an SVM on the extracted features, and (3) Simple rules that predict language uses over the thread.

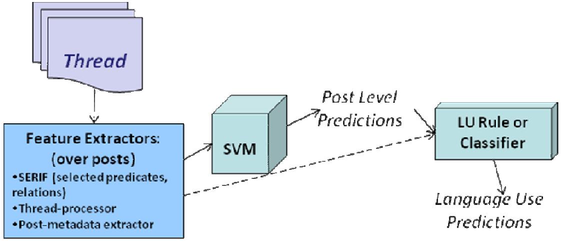

Figure 1: Message and LU Prediction

Phase 1: The SERIF Information Extraction Engine extracts features which are designed to capture different aspects of the posts. The features include simple features that can be extracted from the surface text of the posts and the structure of the posts within the threads. These may correlate directly or indirectly correlate to the language uses. In addition, more syntactic and semantic-driven features are also used. These can indicate the specific purpose of the sentences; specifically targeting directives, imperatives, or shows authority. The following is a partial list of features which are used both in isolation and in combination with each other.

Surface and structural features: average sentence length; number of names, pronouns, and distinct entities; number of sentences, URLs (links), paragraphs and out-of-vocabulary words; special styles (bold, italics, stereotypical punctuation e.g. !!!!), depth in thread, and presence of a quotation.

Syntactic and semantic features: predicate-argument structure including the main verb, subject, object, indirect object, adverbial modifier, modal modifier, and negation, imperative verbs, injection words, subjective words, and mentions of attack events.

Phase 2: Given training data from the message level (Section 3), an SVM predicts if the post contains evidence for an LU. The motivation for this level is (1) Posts provide a compact unit with reliably extractable, specific, explicit features. (2) There is more training data at the post level. (3) Pointing to posts offers a more clear justification for the predictions. (4) In our experiments, errors here do not seem to percolate to the thread level. In fact, accuracy at the message level is not directly predictive of accuracy at the thread level.

Phase 3: Given the infrequency of the Attempt-to-Persuade and Establish-Credibility LUs, we wrote a few rules to predict LUs over threads, given the predictions at the message level. For instance, if the number of messages with evidence for persuasion is greater than 2 from a given participant, then the system predicts AttemptToPersuade. Phase 3 is by design somewhat robust to errors in Phase 2. To predict that a poster is exhibiting the Attempt-to-Persuade LU, the system need not find every piece of evidence that the LU is present, but rather just needs to find sufficient evidence for identifying the LU.

Our message level classifiers were trained with an SVM that optimizes F-measure (Joachims, 2005). Because annotation disagreement is a major challenge, we experimented with various ways to account for (and make use of) noisy, dual annotated text. Initially, we resolved the disagreement automatically, i.e. removing examples with disagreement; treating an example as negative if any annotator marked the example negative; and treating an example as positive if any annotator marked the example as positive. An alternative (and more principled) approach is to incorporate positive and negative scores for each example into the optimization procedure. Because each example was annotated by the same number of annotators (2 in this case), we are able to treat each annotator's decision as an independent example without augmenting the SVM optimization process.

The results below use the training procedure that performed best on the leave-one-thread-out cross validation results (Table 2~~3~~ and Table 3~~4~~). Counts of threads appear in Section 4. We compare our system's performance (S) with two simple baselines. Baseline-A (A) always predicts absent for the LU/evidence. Baseline-P (P) predicts positive (present) for all messages/LUs. Table 4~~Table 3~~ shows results for predicting message level evidence of an LU (Phase 2). Table 5~~Table 4~~ shows performance on the task of predicting an LU for each poster.

The results show significantly worse performance in Arabic than English-- not surprising considering 5-10-fold difference in training examples. Additionally, Arabic messages are much shorter, and the phenomena is even more rare (as illustrated by the high npA, accuracy, of the A baseline).

	Persuade				Establish Credibility			
	npA		npF		npA		npF	
	En	Ar	En	Ar	En	Ar	En	Ar
A	72.5	83.2	0.0	0.0	77.6	95.0	0.0	0.0
P	40.4	29.7	61.1	50.7	33.9	14.4	54.5	30.9
S	86.5	81.3	79.2	61.9	86.7	95.5	73.9	54.0

Table 4~~3~~: Performance on Message Level Evidence

	Persuade				Establish Credibility			
	npA		npF		npA		npF	
	En	Ar	En	Ar	En	Ar	En	Ar
A	90.9	86.7	0.0	0.0	87.7	90.2	0.0	0.0
P	12.1	27.0	23.8	48.2	18.0	21.5	33.7	41.1
S	94.6	88.3	76.8	38.8	95.1	92.4	80.0	36.0

Table 5~~4~~: Cross Validation Performance on Poster LUs

Table 6~~Table 5~~ shows LU prediction results from an external evaluation on held out data. Unlike our dataset, each example in the external evaluation dataset was annotated by 3 annotators. The results are similar to our internal experiment.

	Persuade				Establish Credibility			
	npA		npF		npA		npF	
	En	Ar	En	Ar	En	Ar	En	Ar
A	96.2	98.4	0.0	0.0	93.6	94.0	93.6	0.0
P	13.1	4.2	27.6	11.7	11.1	10.1	11.1	22.2
S	96.5	94.6	75.1	59.1	97.7	92.5	97.7	24.7

Table 6~~5~~: External, Held-Out Results on Poster LUs

6 Related Research

Research in authorship profiling (Chung & Pennebaker, 2007; Argamon et al, in press; and Abbasi and Chen, 2005) has identified traits, such as status, sex, age, gender, and native language. Models and predictions in this field have primarily used simple word-based features, e.g. occurrence and frequency of function words.

Social science researchers have studied how social roles develop in online communities (Fisher, et al., 2006), and have attempted to categorize these roles in multiple ways (Golder and Donath 2004; Turner *et al.*, 2005). Welser *et al.* (2007) have investigated the feasibility of detecting such roles automatically using posting frequency (but not the content of the messages).

Sentiment analysis requires understanding the implicit nature of the text. Work on perspective and sentiment analysis frequently uses a corpus known to be rich in sentiment such as reviews or editorials (e.g. (Hardisty, 2010), (Somasundaran& Weibe, 2009). The MPQA corpus (Weibe, 2005) annotates polarity for sentences in newswire, but the focus of this corpus is at the sentence level. Both the MPQA corpus and the various corpora of editorials and reviews have tended towards more formal, edited, non-conversational text. Our work in contrast, specifically targets interactive discussions in an informal setting. Work outside of computational linguistics that has looked at persuasion has tended to examine language in a persuasive context (e.g. sales, advertising, or negotiations).

Like the current work, Strzalkowski, et al. (2010) investigates language uses over informal dialogue. Their work focuses on chat transcripts in an experimental setting designed to be rich in the phenomena of interest. Like our work, their predictions operate over the conversation, and not a single utterance. The specific language uses in their work (topic/task control, involvement, and disagreement) are different than those discussed here. Our work also differs in the data type of interest. We work with threaded online discussions in which the phenomena in question are rare. Our annotators and system must distinguish between the language use and text that is opinionated without an intention to persuade or establish credibility.

7 Conclusions and Future Work

In this work in progress, we presented a hybrid statistical & rule-based approach to detecting properties not explicitly stated, but evident from language use. Annotation at the message (turn) level provided training data useful for predicting rare phenomena at the discussion level while reducing the need for turn-level predictions to be accurate. Weighing subjective judgments overcame the need for high annotator consistency. For English, the system beats both baselines with respect to accuracy and F, despite the fact that because the phenomena are rare, always predicting the absence of a language use is a high baseline. For Arabic, more work is required, particularly since only 10-20% of the amount of training data exists so far.

This work has explored LUs, the implicit, social purpose behind the words of a message. Future work will explore incorporating LU predictions to predict the social roles played by the participants in a thread, for example using persuasion and credibility to establish which participants in a discussion are serving as informal leaders.

Acknowledgement

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the _____. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

References

Argamon, S., Koppel, M., Pennebaker, J.W., and Schler, J. (2009). “Automatically profiling the author of an anonymous text”. *Communications of the Association for Computing Machinery (CACM)*. Volume 52 Issue 2.

Abbasi A., and Chen H. (2005). “Applying authorship analysis to extremist-group web forum messages”. In *IEEE Intelligent Systems, 20(5)*, pp. 67–75.

Boyd, D, Golder, S, and Lotan, G. (2010). “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.” HICSS-43. IEEE: Kauai, HI.

Chung, C.K., and Pennebaker, J.W. (2007). “The psychological functions of function words”. In K. Fiedler (Ed.), Social communication, pp. 343-359. New York: Psychology Press.

Golder S., and Donath J. (2004) "Social Roles in Electronic Communities," presented at the *Association of Internet Researchers (AoIR)*. Brighton, England

Hovy E., Marcus M., Palmer M., Ramshaw L., and Weischedel R. (2006). “Ontonotes: The 90% solution”. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60. Association for Computational Linguistics, New York City, USA.

Joachims, T. (2005), “A Support Vector Method for Multivariate Performance Measures”, *Proceedings of the International Conference on Machine Learning (ICML).*

Kelly, J., Fisher, D., Smith, D., (2006) “Friends, foes, and fringe: norms and structure in political discussion networks”, *Proceedings of the 2006 international conference on Digital government research*.

NIST Speech Group. (2008). “The ACE 2008 evaluation plan: Assessment of Detection and Recognition of Entities and Relations Within and Across Documents”. http://www.nist.gov/speech/tests/ace/2008/doc/ace08 -evalplan.v1.2d.pdf

Ranganath, R., Jurafsky, D., and McFarland, D. (2009) “It’s Not You, it’s Me: Detecting Flirting and its Misperception in Speed-Dates” *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* pages 334–342.

Somasundaran, S & Wiebe, J (2009). Recognizing Stances in Online Debates. *ACL-IJCNLP 2009.*

Strzalkowski, T, Broadwell, G, Stromer-Galley, J, Shaikh, S, Taylor, S and Webb, N. (2010) “Modeling Socio-Cultural Phenomena in Discourse”. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1038–1046, Beijing, August 2010

Turner T. C., Smith M. A., Fisher D., and Welser H. T. (2005) “Picturing Usenet: Mapping computer-mediated collective action”. In *Journal of Computer-Mediated Communication*, *10(4)*.

Voorhees, E. & Tice, D. (2000)."Building a Question Answering Test Collection", *Proceedings of SIGIR*, pp. 200-207.

Welser H. T., Gleave E., Fisher D., and Smith M., (2007). "Visualizing the signatures of social roles in online discussion groups," In *The Journal of Social Structure*, vol. 8, no. 2.

Wiebe, J, Wilson, T and Cardie, C (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, volume 39, issue 2-3,* pp. 165-210.