Blog Link Classification

Justin Martineau
University of Maryland, Baltimore County
jm1@umbc.edu

Matthew Hurst
Microsoft Live Labs
mhurst@microsoft.com

Abstract

Blog links raise three key questions: Why did the author make the link, what exactly is he pointing at, and what does he feel about it? In response to these questions we introduce a link model with three fundamental descriptive dimensions where each dimension is designed to answer one question. We believe the answers to these questions can be utilized to improve search engine results for blogs. While proving this is outside the scope of this paper, we do prove that knowing the rhetorical role of a link helps determine what the author was pointing at and how he feels about it.

Introduction and Related Work

Based upon the insights gain from Rhetorical Structure Theory we assert that links have a rhetorical role in posts forming relationships between their two ends. With slight modifications, the three basic assumptions of RST(Thompson & Mann 1987) apply to blogs:

1. Links form an organized hierarchy of clauses in a conversation.
2. Links can be described by the purpose of the writer, his assumptions about his audience, and the organization of his message.
3. Links are asymmetric placing unequal importance upon the source and the sink of the link. In RST it is possible for a relation to exist between three or more spans of text, links however only go from the source to the target making them a special case of relations within RST.

Taken with the traditional strengths of SVMs in the general field of text classification(Joachims 1998) ranging from sentiment detection in movie(Pang, Lee, & Vaithyanathan 2002) and product(Dave, Lawrence, & Pennock 2003) reviews, to splog detection(Kolari, Finin, & Joshi 2006) SVMs are the tool of choice for our work. While Joachims recomends removing stop words in (Joachims 1998) I have chosen not to since conjunctions and other such stop words are correlated to rhetorical structures as implied by the converse of Mann and Thompson's assertion that RST relations are "useful in predicting other facts about the text, such as the kinds of conjunctions that will appear in certain places."(Mann & Thompson 1986)

Proposed Link Model Dimensions

- Reference - What the link points to in the sink url and how it is being used by the source url.
  - Item: Objects and places.
  - Event: Actions in the past, present, or future.
  - Entity: People and organizations.
  - Idea: Facts, theories, hypotheticals, arguments, and abstract concepts.
- Rhetorical Intention - Why the author created the link, and how it is being used.
  - Topic: To continue talking about another post as the main topic of this post.
  - Evidence: To provide support for a claim.
  - Define: To define, or specify which term or thing.
  - Cite Source: To credit the original source.
  - Provide Background: Provide info tangential to the topic.
- Sentiment - How the author feels about the link.
  - Positive Sentiment: Both explicit and implicit.
  - Negative Sentiment: Both explicit and implicit.
  - No Sentiment: No discernible sentiment.

Results

We present 12 experiments testing our dimensions as multi-value classification tasks, and 12 experiments as binary classification tasks since "Kappa is an average and that, as such, it may hide the fact that one category accounts for most of the misclassification. Moreover, it should be appreciated that choosing among alternative nominal classification schemes (e.g., white/black/other vs. non-Hispanic white/Hispanic/black/ Asian/other), the Kappa for the more detailed classification scheme will be lower."(Maclure & Williett 1987). All experiments use SVMs, with a uni-gram bag of words[1] occurring in a fixed window around the link.

Reference Dimension

When determining what a link points to, knowing the rhetorical intent of the link is much more important than knowing the surrounding sentiment and words. In-fact, except for event detection (81.1518 % Accuracy, .4147 Kappa, 51.3

[1]Word boundaries are defined by whitespace, punctuation marks, bracketing marks, and slashes are included in the feature set. The bag size was limited to the top thousand words.

Xpt	Dim	Additional Features	% Accuracy	Kappa
#1	Ref	None - A Baseline	46.3351	0.247
#2	Ref	Sent	49.2147	0.2962
#3	**Ref**	**Rhet**	**68.5864**	**0.5668**
#4	Ref	Sent + Rhet	66.7539	0.5395
#5	Rhet	None - A Baseline	50.7853	0.1998
#6	**Rhet**	**Ref**	**73.2984**	**0.5856**
#7	**Rhet**	**Sent**	**61.7801**	**0.3856**
#8	**Rhet**	**Sent + Ref**	**76.4398**	**0.6332**
#9	Sent	None - A Baseline	59.6859	0.3417
#10	Sent	Ref	63.089	0.4012
#11	**Sent**	**Rhet**	**71.2042**	**0.5429**
#12	Sent	Ref + Rhet	68.0628	0.4966

Table 1: Baseline SVM (uni-gram features only) tests for each multi-valued dimension, plus tests for how the dimensions influence each-other.

Label	Accuracy	Kappa	Recall	Accuracy	Kappa	Recall
	Reference Dimension Baseline			Reference Dim with additional Rhetorical Intent class labels		
Item Ref	73.0366	0.2611	0.408	**82.199**	**0.5302**	**0.643**
Event Ref	78.534	0.2714	0.338	79.8429	0.3828	0.500
Entity Ref	81.9372	0.0997	0.140	**90.8377**	**0.6365**	**0.684**
Idea Ref	70.1571	0.3648	0.592	**80.1047**	**0.5819**	**0.755**
	Rhetorical Intent Dimension Baseline			Rhetorical Intent Dim with additional ref dim class labels		
Topic	60.733	0.2105	0.628	**79.8429**	**0.5938**	**0.816**
Context	96.0733	0.6135	0.684	**97.1204**	**0.6707**	**0.632**
Evidence	88.2199	0.3376	0.333	**90.8377**	**0.5174**	**0.511**
Cite	91.8848	0.0298	0.04	91.8848	0.0799	0.08
Define	83.7696	0.0364	0.085	**90.0524**	**0.5553**	**0.638**

Table 2: The presence of each labels was tested with SVMs in a binary classification task.

Recall), knowing the sentiment around the link does not significantly outperform the baseline SVM as shown in experiment 2. Experiment 3 in table 1 along with table 2 show an across the board improvement in reference detection along all measures for all reference types when using rhetorical labels relative to the baseline. Entity references were the biggest winner, note their vastly improved recall and kappa statistics in table 2.

While sentiment features are effective at identifying events (81.1518 % Accuracy, .4147 Kappa, 51.3 % Recall), combining sentiment features with rhetorical features generally does not produce an improvement over rhetorical features. Sentiment features seem to be largely independent of the reference dimension.

Rhetorical Intent Dimension

A weak correlation exists between the words around a link and its rhetorical usage, with certain types of rhetorical uses showing stronger correlations than others. Context links did particularly well with the baseline SVM bag of words.

The rhetorical role of a link is strongly affected by its Reference Dimension, except in the case of citations. Including rhetorical information in experiment 6improved overall accuracy by over 20 percentage points and nearly tripled the baseline's (experiment 5) Kappa. The sentiment around the link also affects the rhetorical role, to a lesser extend.

Sentiment Dimension

Our base line SVM recall results of 41.3% positive recall, 59% negative recall, and 67.6% neutral recall were better (when taken as a whole) than Urseanu's 70.5% positive recall, 51% negative recall, and 31% neutral recall in (Urseanu 2007) for their advanced system with valence shifters. Adding in rhetorical knowledge we achieved 66.7% positive recall, 69.7% negative recall, and 74.1% neutral recall. Our overall accuracy was much better than theirs. They had a 51.5% accuracy with their valence shifters versus our 59.6859% accuracy with our baseline SVM, and 71.2042% accuracy against our SVM with rhetorical knowledge.

Conclusion

Our results show that knowing the Rhetorical Intent behind the link is very helpful in determining what kind of thing the author is pointing at and how he feels about it. Knowing what the author is pointing at is more important than knowing how he feels about it when trying to determine his rhetorical intent. Knowing what the author feels about the link does not help determine what he is pointing at and vice versa. Our results show that existing techniques for ternary sentiment detection can be improved with rhetorical knowledge.

References

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW*, 519–528.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In Nédellec, C., and Rouveirol, C., eds., *Proc of ECML-98, 10th European Conf on Machine Learning*, number 1398, 137–142. Chemnitz, DE: Springer Verlag, Heidelberg, DE.

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *Proc of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*. AAAI Press.

Maclure, M., and Willett, W. C. 1987. Misinterpretation and misuse of the kappa statistic. *Journal of Epidemiolgy* 126(2):161–169.

Mann, W. C., and Thompson, S. A. 1986. Assertions from discourse structure. In *HLT '86: Proc of the workshop on Strategic computing natural language*, 257–270. Morristown, NJ, USA: Association for Computational Linguistics.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc of EMNLP 2002*.

Thompson, S. A., and Mann, W. C. 1987. Rhetorical structure theory: A framework for the analysis of texts. *IPRA Papers in Pragmatics* 1(1):79–105.

Urseanu, A. A. S. B. M. 2007. All Blogs Are Not Made Equal:. In *Proc of the ICWSM 2007*.