# Detecting Informative Blog Comments using Tree Structured Conditional Random Fields

Wei Jin, Shafiq Joty, Giuseppe Carenini and Raymond Ng
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada V6T1Z4
{kimijin,rjoty,carenini,rng}@cs.ubc.ca

## 1 Introduction

The Internet provides a variety of ways for people to easily share, socialize, and interact with each other. One of the most popular platforms is the online blog. This causes a vast amount of new text data in the form of blog comments and opinions about news, events and products being generated everyday. However, not all comments are informative. Informative or high quality comments have great impact on the readers' opinions about the original post content, such as the quality of the product discussed in the post, or the interpretation of a political event. Therefore, developing an efficient and effective mechanism to detect the most informative comments is highly desirable. For this purpose, sites like Slashdot, where users volunteer to rate comments based on their informativeness, can be a great resource to build such automated system using supervised machine learning techniques.

Our research concerns building an automatic comment classification system leveraging this freely available valuable resources. Specifically, we discuss how informative comments in blogs can be detected using Conditional Random Fields (CRFs) [6]. Blog conversations typically have a tree-like structure in which an initial post is followed by comments, and each comment can be followed by other comments. In this work, we propose to use Tree-structured Conditional Random Fields (TCRFs) to capture the dependencies in a tree-like conversational structure. This is in contrast with previous work [1] in which results produced by linear-chain CRF models had to be aggregated heuristically. As an additional contribution, we present a new blog corpus consisting of conversations of different genres from 5 different blog websites.

## 2 Related Work

CRFs have been applied to many NLP tasks including Part of Speech (POS) tagging, chunking, syntactic parsing [8], word segmentation and meeting utterances classification [5]. The most relevant to our work is [1], where the authors use a linear-chain CRF (LCCRF) to detect the informative blog comments through the exploration of conversational and topical features. They apply the LCCRF model to each path (i.e., from the root to the leave) of the conversation's tree structure. However, this approach has two main limitations: (1) it ignores many hierarchical dependencies between the output tags/labels and (2) the common internal nodes fall in multiple paths, which cause them to be classified multiple time, and possible inconsistent classifications have to be combined heuristically. Our use of Tree-structured CRFs (TCRFs) is inspired by [2] and [3]. [2] applies a TCRF model to better handle the dependencies across the hierarchically laid-out information on the web in the task of *semantic annotation* and show that TCRF outperforms Support Vector Machines (SVMs) and LCCRFs. According to the work done by [3], in the task of *semantic role labelling*, TCRFs outperform LCCRFs due to the inherent tree structure which captures the semantic and syntactic dependencies better.

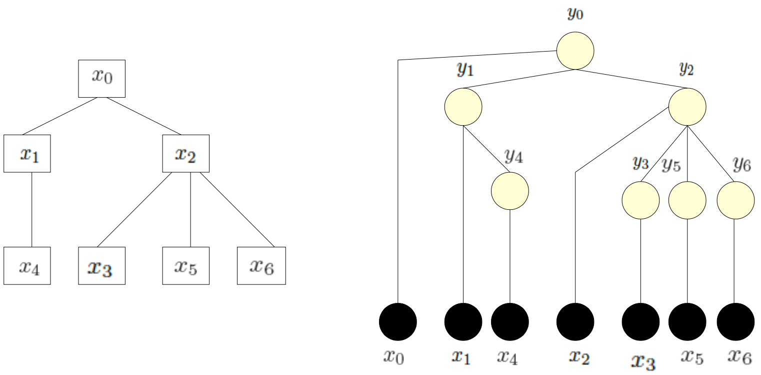
## 3 Datasets

We compiled a new corpus comprised of the Slashdot (http://slashdot.org) dataset collected by [1] and conversations from four other blog websites namely, DailyKos (www.dailykos.com) , AndroidCentral (www.androidcentral.com), BusinessInsider (www.businessinsider.com), Macrumors (www.macrumors.com) and TSN (www.tsn.ca). Table 1 shows their key properties and basic statistics. These blog sites cover a variety of genres such as technology, business. politics and sports; they also have different conversation structure (sequential or tree). For example, Slashdot conversations have a tree-like structure; users can directly reply to a given comment, and their reply will be placed underneath that comment in a nested structure. On the contrary, the comments in a conversation from AndroidCentral form a single linear thread. Notice that TCRF will cover both conversation with a tree-like structure and the ones with a single linear thread. The TCRF will simply degenerate to an LCCRF in the linear case.

The definition of informativeness in our work is assumed to be the moderation score received by a comment. For example, comments on Slashdot are moderated by users of the site, meaning that each comment has a score from $-1$ to $+5$, indicating the total score of moderations assigned, where each moderator is able to modify the score of a given comment by $+1$ or $-1$. The comments with scores greater than 0 are considered as GOOD and others are considered as BAD.

| Name | Structure | Rating Method | Genre | Articles | Comments | Avg. comments per Article |
|---|---|---|---|---|---|---|
| Slashdot | Tree | -1 to 5 | Technology | 3300 | 425000 | 128.8 |
| DailyKos | Tree | positive - negative | Politics | 500 | 31200 | 62.4 |
| AndroidCentral | Sequential | 0 - 3 Stars | Technology | 500 | 4150 | 8.3 |
| BusinessInsider | Tree | positive - negative | Business News | 500 | 12400 | 24.8 |
| Macrumors | Sequential | positive - negative | Technology | 500 | 7500 | 15.0 |
| TSN | Tree | positive - negative | Sports | 500 | 27000 | 54.0 |

*Table 1:* Dataset description



*(a)* Sample Conversation

*(b)* Corresponding TCRF

*Figure 1:* Sample Conversation and corresponding TCRF

## 4 Proposed Solution

Conditional Random Fields (CRFs) are undirected graphical models that allow large numbers of arbitrary local and global features in predicting labels for a given observation. For many NLP tasks, the underlying graphical model is usually implemented as a linear chain [1, 6], where observations come in a linear sequence and dependencies exist between two consecutive labels. However, since the structure of blog comments is often a tree, we propose to apply TCRFs to take advantage of the dependencies across hierarchically laid-out information.

The structure of the tree CRF will mirror the structure of the conversation as shown in Figure 1. On the left side, we have a sample conversation initiated by the post $x_0$, which received two comments $x_1$ and $x_2$, and so on. On the right side, you can see the corresponding TCRF, in which all the posts are represented as observed variables $x_i$ and their labels as the hidden variables $y_i$. Notice that the TCRF models both the parent-child dependencies (e.g., $y_0$-$y_1$ and $y_0$-$y_2$) and the sibling dependencies (e.g., $y_5$ and $y_6$).

We will learn the parameters of our TCRF model from our annotated corpora described in Section 3. As features for each comment, we are considering the ones used in [1], which include bag-of-words features, similarity features between the current comment and adjacent comments, and conversational features (e.g., number of comments replying to the current comment). When the TCRF model is applied for inference, the observations $x_i$ in the TCRF correspond to all the comments of one conversation, while the inferred values for the $y_i$ correspond to the classification of these comments.

We are currently experimenting with the GRMM (Graphical Models in Mallet) toolkit [4] to design our TCRF model. GRMM is a general machine learning software package for implementing probabilistic graphical models providing various fitting/learning algorithms like Limited Memory BFGS (LBFGS) and inference algorithms like Belief Propagation, Tree-based re-parameterization (TRP).

## References

[1] Nicholas FitzGerald, Giuseppe Carenini, Gabriel Murray and Shafiq Joty. (2011) Exploiting Conversational Features to Detect High-Quality Blog Comments. *In Proceedings of the Canadian Conference on Artificial Intelligence (CAI) 2011. St. Johns, Newfoundland.*

[2] Jie Tang, Mingcai Hong, Juanzi Li and Bangyong Liang. (2006) Tree-structured conditional random fields for semantic annotation. *In Proceedings of 5th International Conference of Semantic Web (ISWC2006)*

[3] Trevor Cohn and Philip Blunsom. (2005) Semantic Role Labelling with Tree Conditional Random Fields. *In Proceedings of the 9th Conference on Natural Language Learning (CoNLL), pp. 169172, Ann Arbor, USA, 2005*

[4] Charles Sutton. 2006. GRMM: A Graphical ModelsToolkit. http://mallet.cs.umass.edu.

[5] Galley, Michel (2006) A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2006. Sydney, Australia*

[6] J. Lafferty, A. McCallum, and F. Pereira. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning, 2001.*

[7] Koller, D. and Friedman, N. Probabilistic Graphical Mod-els: Principles and Techniques. MIT Press, 2009.

[8] Finkel, Jenny Rose and Kleeman, Alex and Manning, Christopher D. (2008) Efficient, Feature-based, Conditional Random Field Parsing. *Proceedings of ACL-08: HLT, Columbus, Ohio, USA*