Chapter 1

Introduction

This thesis began as an investigation of ways of extracting information from informal communication. By "informal communication", we mean unedited textual communication which has at most minimal formal structure. Some examples include e-mail, mailing lists, bulletin boards, blogs and instant messages. We think this kind of communication is interesting for two reasons: (1) the volume of it which is available, and (2) the quickness with which it can reflect current events or new information. With the advent of the Internet and digital communications networks, text has become increasingly popular as a means of communication. As such, textual communication has become less formal and the range of information available in such communication has become extremely broad. Another trend has been the increased ease with which individuals can produce textual communication. Whereas access to the Internet was once quite limited and required specialized knowledge, now individuals can publish content via a web browser or even a cellular phone. The result of this trend is that informal communication can provide specialized, up-to-date information beyond that of more formal sources of communication (such as newspapers, guides and "portal" web sites). However, informal communication has obvious drawbacks: it is poorly organized and more difficult to understand than structured or well-edited text. So, locating and extracting information in which we might be interested can be quite difficult. Yet, the volume and timeliness of such information suggests to us that the extra effort may be worthwhile.

One possible approach to this task of extracting information from informal communication is to construct a system which at least partially solves the problem. This would require substantial domain-specific knowledge and engineering effort. We feel that before such a system will be valuable, advancements need to be made on the core information extraction problems. So, instead of building a system, we focus our efforts on tackling what we feel are key sub-problems of the information extraction task. For each sub-problem, we identify a deficiency in existing techniques' ability to handle informal communication and provide an alternate approach or attempt to advance the state-of-the art. We feel that this will allow us to make lasting contributions to this problem.

We consider four sub-problems, devoting a chapter to each. Our depth of analysis and contribution varies, from a relatively small extension of an existing co-reference resolution algorithm (Chapter 3), to development of and large-scale evaluation of a new framework for learning user preferences (Chapter 5). We use restaurant discussion boards as a running example. Restaurant discussion boards exhibit many of the benefits and drawbacks that we find in general informal communication. They provide a wealth of up-to-date information on local restaurants. But, they can be so unwieldy and free-form so as to make finding information difficult.

We begin with one of the most fundamental problems in information extraction: named entity extraction. People, places, locations, organizations, etc. almost always play a role in information that we want to extract. One type of information we find on restaurant discussion boards is opinions about specific restaurants, e.g. "French Laundry was outstanding." Before we can extract the opinion, it is helpful, if not necessary, to identify "French Laundry" as the name of a (restaurant) entity. Identification of such names is the first step toward being able to extract information about them. An entity (particularly after it has been mentioned), may be referred to in many different ways—with the full name, using an abbreviation, or via a pronoun or descriptive phrase. To be able to extract all of these information about an entity, we must be able to resolve all these various mentions—we must perform "co-reference resolution", which is the association of multiple mentions of the same entity. If we can solve these two tasks, named entity extraction and co-reference resolution, we will be able to identify and resolve all explicit entity mentions. However, sometimes information is provided indirectly, without explicit mention of an entity. For example, in reviewing a restaurant, someone might say, "The swordfish was excellent," which is a comment on the food served at a particular restaurant. Association of this comment with the restaurant requires that we be able to track context. We must be able to follow the "topic" of conversation. A final sub-problem that we address involves adding value to the morsels of information that we extract from text. Whereas formal communication tends to focus on factual information, informal communication often is filled with expressions of opinions and preferences. For example, restaurant boards are typically filled with user reviews of restaurants. Individually, such reviews and opinions are of limited value. But, collectively, they can be used to characterize differences between restaurants (or other named entities) and may also be used to predict unobserved opinions—whether an individual will like a restaurant she hasn't experienced yet.

1.1 Chapter Summaries

Most work on named entity extraction has focused on formal (e.g. newspaper) text. As such, systems tend to rely heavily on titles ("Mr."), keywords ("Inc."), capitalization and punctuation. However, capitalization and punctuation are noisy in informal communication. And, titles and keywords are used relatively rarely in informal communication, if they are used at all. Some named entity types (e.g. restaurant names) do not have keywords or titles. One aspect of names that is not fully utilized is that they are often involved in the "topic" of discussion. As such, words in names are often distributed like topic-oriented or informative words. If we can characterize the distribution of topic-oriented words, then we can use this as an additional feature for extracting named entities. Our contribution is exactly that: a new score which estimates a word's "informativeness" or "topic-orientedness". The score captures two aspects which we believe to be typical of the distribution of topic-oriented words: modality and rareness. Experiments indicate that our score can be used to improve NEE performance. Details can be found in Chapter 2.

Two categories of co-reference resolution algorithms have emerged. The first treats each noun phrase mention as referring to a single other mention; learning involves training a classifier to identify the antecedent for each noun phrase mention. The other framework treats co-reference resolution as a clustering problem; mentions are grouped together which have high average “similarity”. We view neither approach as being the answer. The classification approach treats each mention as referring to exactly one other mention (if any). Pronouns and other non-proper nouns do typically refer to other mentions in this way, but reference for proper nouns is less constrained. The classification approach also has the disadvantage of being greedy, making locally optimal reference decisions. The clustering approach requires that each mention in a cluster have (relatively) high average “similarity” with the other mentions in that cluster. This reflects how proper nouns tend to co-refer (string similarity), but is less appropriate for other nouns, which heavily depend on context and locality. Our contribution is a new co-reference resolution algorithm which is a hybrid of these two approaches. We extend a probabilistic clustering-style algorithm to utilize the clustering approach for proper nouns and a classification approach for other nouns. Details can be found in Chapter 3.

Topic or context changes in formal or structured text are often indicated by formatting or markup. Not so with informal communication, where word meaning may be the only clue that context has changed. Thus, we treat the problem of tracking context in informal communication as a sentence clustering problem. One theory for how text is generated within a topic is that it corresponds to a low-dimensional subspace of the probability simplex. Neither of the two popular clustering frameworks, mean/centroid and spectral/normalized cut, can discover low-dimensional subspaces in noisy data. Our contribution is the development of a new clustering framework which simultaneously identifies cluster assignments and the subspace of variation corresponding to each cluster. Details can be found in Chapter 4.

We address two related preference learning problems. Individuals typically express their opinions as partial orderings or ratings. Yet, we think that limited attention has been paid to algorithms which learn ordered categories. First, we consider the problem where a single user rates items and feature information is available on each of the items which might be rated. This is known as ordinal regression, which is an generalization of binary classification to multiple, ordered classes. We introduce a loss function which extends large margin classification theory: our loss function bounds the ordinal classification error. Our contribution is a set of experiments which show that it greatly outperforms other loss functions used for ordinal regression. Details can be found in Section 5.2. The second problem we address is when we have multiple users, but no information about the items. This is known as collaborative filtering. Most approaches to this problem have utilized a rank constraint to force the model to uncover similarities between users and items. However, a rank constraint yields poor solutions due to local minima which are introduced. We instead utilize a soft trace norm penalty for regularization which encourages a low-rank solution without the creation of local minima. We contribute a new way to optimize the trace norm which allows us to scale the framework to large collaborative filtering problems. Experiments on two large collaborative filtering data sets validate our approach. Finally, we show how to extend our preference learning framework in various ways. Details can be found in Section 5.3.