Towards Analyzing Micro-blogs for Detection and Classification of Real-Time Intentions

Nilanjan Banerjee, Dipanjan Chakraborty, Anupam Joshi, Sumit Mittal, Angshu Rai
IBM Research - India, New Delhi, India
{*nilanjba, cdipanjan, anupam.joshi, sumittal, angshu.rai*}*@in.ibm.com*

B. Ravindran
Indian Institute of Technology, Madras, India
ravi@cse.iitm.ac.in

Abstract

Micro-blog forums, such as Twitter, constitute a powerful medium today that people use to express their thoughts and intentions on a daily, and in many cases, hourly, basis. Extracting 'Real-Time Intention' (RTI) of a user from such short text updates is a huge opportunity towards web personalization and social networking around dynamic user context. In this paper, we explore the novel problem of detecting and classifying RTIs from micro-blogs. We find that employing a heuristic based ensemble approach on a reduced dimension of the feature space, based on a wide spectrum of linguistic and statistical features of RTI expressions, achieves significant improvement in detecting RTIs compared to word-level features used in many social media classification tasks today. Our solution approach takes into account various salient characteristics of micro-blogs towards such classification – high dimensionality, sparseness of data, limited context, grammatical in-correctness, etc.

Introduction

Micro-blog forums like Twitter form a rich source of real-time status updates (Sakaki and Okazaki 2010; Banerjee et al. ; Java and Joshi 2007) where users make posts about their activities and future plans. Efficient classification of real-time intentions from this data is challenging, but at the same time, can significantly enhance the knowledge capturing potential of context-aware applications (Banerjee and Chakraborty 2009; Sakaki and Okazaki 2010) as well as community sensing systems (Miluzzo and Zheng 2008) that aim to exploit social data. Conceptually, we define a Real-time Intention (RTI) as *"text expression signifying an intent to perform an activity in near future"*. In particular, we focus on the two-class classification problem - is the update an RTI or a Non-Intention (NI)? Towards this, our goal is to design various feature extraction techniques for enabling efficient classification of RTIs from micro-blog feeds.

We begin by outlining several attributes that characterize micro-blog data and pose challenges for such classification tasks:

- *Limited Context Information:* Post size is restricted to a few characters (e.g. 140 in Twitter) – this gives a very short context window for traditional knowledge extraction algorithms (Han and Pei 2000) to be effective. Moreover, often the context is fragmented, making it difficult to mine the underlying intention, as in the tweet "Gorgeous evening. Out of work. Off to football. Life is sweet."

- *Richness of Exchange:* Postings are of several kinds: (1) daily activities (2) conversations, discussions (e.g. using hash-tags in Twitter) (3) URL mentions (4) Random thoughts (e.g. moods, feedbacks). For example, "Watching X-Men movies is fun" expresses an opinion, whereas "Can't wait to see the latest X-Men!!" suggests an activity intention. Efficiently segregating data pertaining to RTIs from other postings therefore becomes challenging.

- *High Dimensionality:* The use of language in a micro-blog is often informal and sometimes grammatically incorrect or ambiguous, contains mis-spellings, spoken acronyms and morphological variants (e.g. "me wanna play ice hokey nowwwwww!") . These factors along with inherent vastness of English vocabulary and proper nouns make the data highly dimensional in nature.

In this paper, we detect (and classify) RTIs in micro-blogs using various *feature extractors* that take into account micro-blog characteristics described above and provide a platform for efficient classification. Our feature extractors are built around the central intuition that relationships 'of several types' between word classes would yield better discriminatory power. This also reduces the dimensionality of the feature space. Ensemble classifiers are well suited for combining features that observe different aspects of the data. We propose a heuristic based ensemble technique to combine the proposed feature extractors. Our experiments show significant improvement in classifying RTIs and NIs over commonly used classification techniques that use word level features.

Related Work

Social networking websites, blogs and discussion forums are fundamentally different in nature from micro-blogs as they usually contain a large set of semantically interlinked statements, describing a view point. Prior works (Chen and Lin 2002) exploit this property and reveal relatively *static* or *long-term* trends in user interests, content generation patterns, etc. Our focus, on the other hand, is on capturing real-

time intentions in micro-blogs which are characterized by limited contextual information and spontaneous, incoherent expression of *short-term* to *real-time* thoughts and intentions.

Work on sentiment analysis (Godbole and Srinivasaiah 2007; Raaijmakers 2007; Pang and Lee 2008; Peng and Park 2011) looks at the problem of identifying positive or negative sentiments, typically from various kinds of document corporae. Our classification problem needs to handle wider range of expressions (emotions) than just positive and negative sentiment indicative terms, noting that intentions may not be necessarily correlated to any particular kind of sentiment. Moreover, our investigation is along the lines of uncovering suitable features that can be extracted within a short context window.

In micro-blogs, (Java and Joshi 2007) is one of the first works to explore various structural properties and long-term user interests and intentions expressed in them. However, investigation for short-term, real-time user intentions was not done here. In our earlier work (Banerjee et al.), we showed that a fractional (20%), but quantitatively significant bulk of micro-blog content contains keywords indicative of user intentions, and reported results on keywords and n-grams commonly observed in Twitter for expressing real-time context. This forms the inspiration for our current work.

Recently, (Sakaki and Okazaki 2010) exploited the concept of Twitter users as *Social Sensor* to detect real-time events (not real-time user intention) such as earth quakes, tornadoes, etc.

Tackling the RTI Classification Problem

In this section, we define various concepts related to *Real-Time Intention* (RTI), thereafter describing our overall approach of RTI detection and classification. **Content-indicative Word (CI word):-** "*keywords that carry the central subject in a RTI*." These are typically proper or common nouns (e.g. movie, football), but not necessarily restricted to these parts of speech (POS).

Usage-indicative Word (UI word):- "*Keywords that characterize the activity associated with a particular CI word*". These can be either *Temporal* keywords : T-Words (e.g. evening, morning) or *action* keywords: A-words (e.g. watch, go, see). We define T-words as words that describe the concept of *time* in a given statement. A-words are words that qualify the *action* associated with the CI word, normally verbs. Our definition is geared towards not imposing strong restrictions on the POS associated with these categories.

RTI *:- "A text expression containing* **one or more CI words** *providing* **class** *of intent; with* **one or more UI words** *that further qualify the intent, with no specific ordering"*. This provides a generic definition of intentions while safely avoiding ambiguous expressions like *"excited about fishing"*. Intuitively, this definition covers range of expressions to characterize RTIs in a single micro-blog, without grammatical constraints. Note that we define all the terms using loose semantics since we do not want the definitions to be strictly grammatical or linguistically binding in nature.

Labeled Training & Test Data Generation

We collected over 20 million publicly available tweets using twitter's APIs for our study. In this paper, we focus on five categories of intentions - *Movie, Sports, Music, Food, Dance*, which have been found to be the most popular topics for RTIs (Banerjee et al. ; Java and Joshi 2007). The choice of these categories form a good set of topics around which the users might perform activities and thereby leading to expressions of RTIs. To generate ground truth, we created a set of labeled micro-blogs containing a mix of RTIs and NIs and manually inspected and labeled each micro-blog as RTI or NI. This was done for different categories on a total of 13206 tweets. For each category, a tweet that means an intention to perform an activity strictly in the near future was marked as an RTI. All other tweets, including ones that indicate activities being performed at present (e.g. watching a movie) or past activities and other irrelevant ones were marked as NI. Table 1 gives a breakdown of the data manually labeled. Note that the exact meaning of 'future'could be application-specific (e.g. within the next hour versus within the next week). We assume that ground truths should reflect the definition adopted by the application. We believe our approach is generally applicable across these multiple interpretations.

Table 1: Labeled Dataset

RTI_{class}	No. Of Positives (RTI)	No. Of Negatives (NI)
Sports	1493	3714
Food	915	2503
Music	726	689
Movie	711	885
Dance	698	873
Total	4542	8664

Overall Approach

Figure 1 schematically depicts our overall approach of identifying distinctive features in the data pertaining to RTIs and using them to progressively reduce the high data dimensionality, ultimately leading to RTI classification. As shown in the figure, reduction of data dimensionality is done in a manner that preserves the relationships amongst words; our intuition being that these relationships would provide better feature sets for RTI classification. We describe below the various steps in our approach.

Dimensionality Reduction. First, we canonicalize the occurrences of words by reducing them to their base category representation, thereby reducing the dimensionality of the data in a simple and interpretable manner, while preserving the word positions, and hence relationships in the lower dimension. For example *"football"* is converted to *"sport"*, *"sushi"* to *"food"*, etc. These canonicalizations are used subsequently by several feature extraction techniques in the next step.

We follow an iterative learning process to create a vocabulary of CI and UI words that characterize presence of RTIs in micro-blogs. In particular, we first formulate a *Seed Set* - a bona fide (correctly spelled) set of well known and representative CI and UI words (covering both A-Words

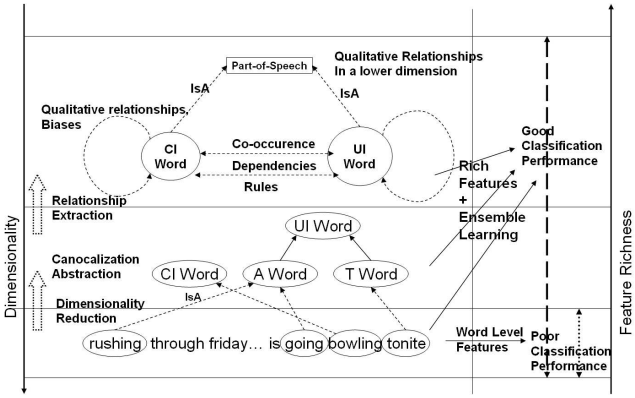

Figure 1: **Intention Classification in Micro-blogs**

and T-Words) based on knowledge of most frequently used words in the data-set, enriched further by manual inspection of several thousand tweets and consultation with the WordNet dictionary. Thereafter, we semi-automatically grow this set by looking for words and phrases appearing in similar context using the technique presented in (Godbole and Bhattacharya 2010) that reveals words and phrases that are either morphological variants or semantically equivalent to those in the Seed Set, including synonyms, hyponyms and hypernyms. Similarly, we create category specific vocabularies for the T-words and A-words. To keep the vocabulary up-to-date, we periodically repeat the process for newly arriving micro-blogs. Full details of this particular step are available in (Banerjee et al.). Due to lack of space, we skip the same in this paper.

Feature Extraction. The central idea here is to use linguistic relationships and statistical associations among words and phrases after representing data in the reduced dimension of CI and UI word categories, for RTIs as well as NIs. For this purpose, we design the following 'Feature Extractors' that capture distinguishing linguistic and statistical features from micro-blogs for classification, while tackling issues of limited context information, use of informal language and presence of noise.

- *Co-occurrence Feature Extractor:* This extractor analyses the micro-blogs based on the following intuition – if more relevant words co-occur in a micro-blog, likelihood of the micro-blog expressing an intent increases. For this, we find all the gappy bi-grams between relevant CI and UI words and compute the co-occurrences.

For each micro-blog, this feature extractor outputs: no. of CI-Words ($N_{ci-word}$), no. of A-Words (N_{a-word}), no. of T-Words (N_{t-word}), no. of CI–(A-Word) co-occurrences ($N_{ci-a-cooccur}$), no. of CI–(T-Word) co-occurrences ($N_{ci-t-cooccur}$), and no. of A-Word–T-Word co-occurrences ($N_{a-t-cooccur}$).

- *POS Feature Extractor:* This extractor exploits the fact that although micro-blogs often lack grammatical accuracy, at a sub-sentence level, a user is likely to arrange words in correct grammatical order. For each micro-blog, we first tag the POS; then this extractor provides a POS attribute to the relations between CI and nearby UI/additional relevant words and outputs: no. of verbs (N_{verb}), $position$, no. of nouns (N_{noun}),$position$, no. of past tense verbs ($N_{past-tense-verb}$), $position$, and no. of adverbs (N_{adverb}), $position$, where $position \in \{before, after\}$, indicating the position of occurrence relative to CI Word.

- *Rule-based Feature Extractor:* This extractor learns a set of conjunctive rules that capture words and phrases, containing CI/UI words, commonly used to express RTIs. Our motivation is to have a set of *intention favorable* rules (*RTI-Rules*) and *non-intention favorable* rules (*NI-Rules*), the rules themselves may not necessarily be grammatically correct. We use a conjunctive rule learner algorithm (Han and Kamber 2006) to learn word based rules for both classes and use presence of these rules as features.

For each incoming micro-blog, the output of this feature extractor is: no. of matched RTI-rules ($N_{RTI-rules}$) and no. of matched NI-rules ($N_{NI-rules}$).

- *Dependency Based Feature Extractor:* This feature extraction technique identifies generalized grammatical *relationship* patterns shared by words in RTI and NI sets, beyond what can be captured by simpler co-occurrence and POS extractors. The intuition is that words play *different roles* when used to express RTIs as when used for NIs.

We first parse the micro-blogs and for each dependency link (i.e. role) found, we generate tokens that contain relation, no. of gaps, i.e., number of words between the related words, and the words themselves. Through canonicalization (replace CI-words with the CI class name, A-words with "action" and T-words with "time"), we convert similar tokens to a common token representation; the token format being $word_i, relation_{i,j}, word_j, no.of gaps$.

Thereafter, we obtain sets of frequently occurring token patterns for both classes (RTI-Patterns, NI-Patterns), using the FP Growth algorithm from(Han and Pei 2000). Interestingly, we found RTI-Patterns $\cap$ NI-Patterns to be $null$, even though they were learnt independently. For each micro-blog, the output from this feature extractor becomes: - no. of matches found for pattern d (N_d), where $d \in$ *RTI-Patterns, NI-Patterns*.

- Δ*-TFIDF Feature Extractor:* This technique captures words whose usage is heavily biased towards either one of the sets. Δ-TFIDF driven SVM models have been shown to improve performance in document classification tasks (Martineau and Finin 2009). To use Δ-TFIDF, we compute at first, the TF-IDF values for different words separately for the RTI and NI sets for each category. Then the difference of two sets of TF-IDF values is assigned to each word as the Δ-TFIDF value. For non-discriminating words, Δ-TFIDF scores are nearer to 0. For each micro-blog, the output is: vector V=[Δ_1,Δ_2..Δ_n], where n= no. of distinct words in the micro-blog, and Δ_i=Δ-TFIDF value for word w_i.

Ensemble Classifier. Each feature extractor proposed in the previous section looks at a different representation of the data. An ensemble classifier is suited for our classification task since they can combine insights from extractors that inspect different views of each micro-blog. We present an ensemble approach using a heuristic that

considers a weighted sum of relevance measures from all feature extractors, and is computationally light weight. Each feature extractor adds a bias in the classification of a given micro-blog. For each extractor, we compute a relevance score R_{c_i}, which is the confidence given by the extractor to a given micro-blog to contain an RTI for class c_i:
- Co-Occurrence Feature: $R_{c_i} \propto (N_{c_i-word}) + (N_{a-word}) + (N_{t-word}) + N_{c_i-a-cooccur} + N_{c_i-t-cooccur} + N_{a-t-cooccur}$.
- POS Feature: $R_{c_i} \propto (N_{verb} + N_{noun}) - k * N_{past-tense-verb}$, where k is an internal constant parameter.
- Rule Based Feature: $R_{c_i} \propto k_1 * N_{RTI-rules} - k_2 * N_{NI-rules}$. where, k_1,k_2 are internal constant parameters.
- Dependency Based Feature:
$R_{c_i} \propto \sum_{d=1}^{Max_P} N_{patterns-d}$, where Max_P=|RTI-Patterns|.
- Δ-TFIDF Based Feature: [Δ_1,Δ_2..Δ_n] values from training data and the class values predicted by the SVM (trained using Δ-TFIDF features) are used as R_{c_i}.

Now, we compute combined relevance value S = $\sum_{j=1}^{5} w_j * R_{c_i}$, where w_j are the feature weights. A micro-blog is classified as an RTI of $classc_i$ if S > discrimination threshold τ; NI otherwise. To determine suitable values of the proportionality constants (and internal parameters of the features) for S, we compute the F1[1] score for different combinations of feature weights and internal parameters, for each feature extractor. The parameter combination that gives the highest F1 score is used. We observe performance of this approach by varying τ.

Experiments and Evaluation

We implemented the ensemble classifier in JAVA, using APIs provided by libsvm, openNLP and Stanford's Dependency Parser[2] to build the different modules. We used the maximum entropy based POS tagger from OpenNLP[3] toolbox as the POS tagger. We used a 10-fold cross validation process for performance evaluation. Experiments were performed on a 2.83GHz, 64 bit quad-core Intel processor system with 4GB RAM and 6MB L2 cache. We refer to the heuristic based ensemble classification system as "RTI Classifier". Along with this classifier (that makes use of all feature extractors), we also study the corresponding classifiers that utilize only a single feature extractor. This is to understand incremental contributions of different feature sets.

We use the area under the ROC curve to compare performance. This provides useful insights into expected performance of the classifier over the entire operating region.

Figure 2a shows that the ensemble approach, as expected, outperforms all individual feature extractor-based classifiers. Figure 2b presents details on the incremental performance improvement as ensemble approach employs more features to make decision. We observe that at lower values of τ, Rule-based features improve TPR (True Positive Rate) compared to combined discrimination capability of co-occurrence and POS analyzers, indicating rules are helpful for applications desiring high RTI detection precision (lower τ). Δ-TFIDF features contribute significant discriminating power to the heuristic approach.

[1] http://en.wikipedia.org/wiki/F1_score

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://incubator.apache.org/opennlp/

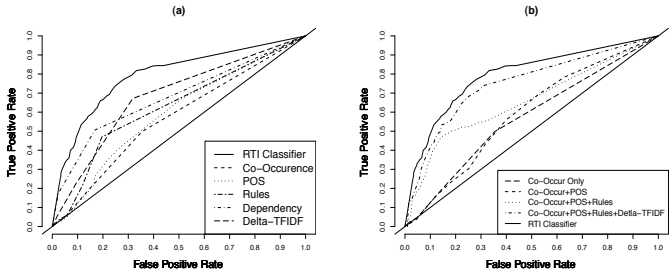

Figure 2: ROC curves for RTI Classifier using Ensemble Approach ($RTI_{class} = sports$) (a)Individual Feature Extractors Vs RTI Classifier, (b) Incremental Contributions of Feature Extractors to RTI Classifier Performance. Results for other categories were similar.

Conclusion

This paper carried out a thorough analysis of micro-blogs towards detection and classification of Real-Time Intentions (RTIs) of users. We presented feature extractors that obtain relational features in a reduced dimension of the complex micro-blog data along with the performance evaluation of ensemble heuristic based approach for combining these features. We believe our results on feature spaces and achievable accuracies offer significant insights to applications aiming to exploit free-text intentions from social media.

References

Banerjee, N., and Chakraborty, D. 2009. R-U-In? - Exploiting Rich Presence and Converged Communications for Next-generation Activity-Oriented Social Networking. In *Proceedings of MDM*.

Banerjee, N.; Chakraborty, D.; Dasgupta, K.; Mittal, S.; Joshi, A.; Nagar, S.; Rai, A.; and Madan, S. User interests in social media sites: an exploration with micro-blogs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, 1823–1826.

Chen, Z., and Lin, F. 2002. User Intention Modeling in Web Applications Using Data Mining. In *Proc. of WWW, vol. 5(2), pp. 181-191*.

Godbole, S., and Bhattacharya, I. 2010. Building Re-usable Dictionary Repositories for Real-world Text Mining. In *Proc. of CIKM*.

Godbole, N., and Srinivasaiah, M. 2007. Large Scale Sentiment Analysis for News and Blogs. In *Proc. of ICWSM*.

Han, J., and Kamber, M. 2006. Data Mining Concepts and Techniques. *Second edition, Chapter 6, pp. 318-327*.

Han, J., and Pei, J. 2000. Mining Frequent Patterns Without Candidate Generation. In *Proc. of SIGMOD*.

Java, A., and Joshi, A. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. of WebKDD/SNA-KDD*.

Martineau, J., and Finin, T. 2009. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proc. of ICWSM*.

Miluzzo, E., and Zheng, X. 2008. Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In *Proc. of SenSys*.

Pang, B., and Lee, L. 2008. Opinion Mining and Sentiment Analysis Foundations and Trends in Information Retrieval. *Vol 2, pp. 1-135*.

Peng, W., and Park, D. 2011. Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization. In *Proceedings of ICWSM*.

Raaijmakers, S. 2007. Sentiment Classification with Interpolated Information Diffusion Kernels. In *Proc. of ADKDD*.

Sakaki, T., and Okazaki, M. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. of WWW*.