# Classifying short messages using collaborative knowledge bases: Reading Wikipedia to understand Twitter

Yegin Genc, Winter Mason, Jeffrey V. Nickerson

Stevens Institute of Technology
{ygenc, wmason, jnickerson}@stevens.edu

## 1 Introduction

To detect concepts from tweets, we leverage the content of Wikipedia. This is a form of semantic transformation: ideas that emerge in short texts are mapped onto more extensive texts that contain additional structure. This additional structure is used to amplify the signal in the short text. This idea is rooted in our previous research [1, 2], as well as in the work of other authors pursuing similar goals [3-5].

Our method has two main stages. First, we recognize candidate concepts—parts-of-tweets—that may be valid entities in the tweet. These concepts are then classified into four categories: Locations, People, Organizations, and Miscellaneous. Candidate concepts are identified by mapping tweets to Wikipedia pages, and the networks of these concepts in Wikipedia are used for filtering and classification. We believe this technique can be applied more generally to the understanding of many forms of short messages, not just tweets, utilizing many forms of collaborative knowledge bases, not just Wikipedia.

## 2 Concept Recognition

Automatically determining whether a word in a tweet represents a concept is not trivial, because the words may be stop words or personal or idiosyncratic concept. Wikipedia titles, on the other hand, can be viewed as representing concepts. Moreover, Wikipedia pages are situated in a network, so that the semantics of a page title can be utilized to classify the concept. Thus, as a first step, we look for parts-of-tweets that match a Wikipedia title. Specifically, concept words are extracted and submitted as search criteria against the page titles of Wikipedia articles using the Wikipedia API. To this end, we segmented each tweet in two ways: First, using Natural Language Processing toolkits , we extracted sentences and then noun phrases from each sentence. Second, we removed punctuation and extracted n-grams (n up to 4) from the entire tweet using a sliding window. To meet Wikipedia's title conventions required for matching search results, we normalized the parts-of-tweets (noun phrases and n-grams) by capitalizing the first letter and changing the rest to lower case. For the parts-of-tweets that didn't match a Wikipedia title after normalization, we also searched for a match after capitalizing each word in the text. When a part-of-tweet


Making Sense of Microposts Workshop @ WWW'13, May 13th 2013, Rio de Janeiro, Brazil

landed on a Wikipedia title, we ignored all the other parts-of-tweets that are its subsets. For example, when 'Sarah Palin' occurs in a tweet, and maps to Wikipedia page containing 'Sarah Palin', 'Sarah' and 'Palin' are not processed.

## 3 Filtering And Classification

For classification and filtering, we utilized the concept network in Wikipedia, which consists of categories and category containers. Wikipedia pages are tagged with categories they belong to and these categories are linked to one another in a graph structure. Container-categories are special categories that contain only other categories and are not referenced by any page. They arguably serve as meta-level tags for the pages that belong to its sub-graph of categories. Moreover, their titles capture the mutual themes that run through the children categories. For example, Container Category: 21st Century people by their nationality holds categories that are used to tag pages, or other categories about people. Therefore, we labeled the container-categories with the entity labels from the contest (Locations, People, Organizations, Miscellaneous) using simple keyword searches. The keywords we selected for each label are shown in Table 1. Using this keyword search process, we labeled 1,560 of the 4,227 containers. Based on our tests, we later included 9 manually selected categories from Wikipedia to our list to improve our results. We provide more detail in section 3.

For the parts-of-tweets that match a Wikipedia page title, we traverse up the page's category graph and count how many of the categories within 3 levels of the original page fall immediately under a labeled container-category. We label the Wikipedia page, and hence the part-of-tweet, with the container label that holds the maximum number of the categories from the page's category graph. If the categories from the traversal of the page's category graph don't fall under any of the labeled containers, we ignore the concept.

## 4 Using The Training Set

One benefit to our method is that both the concept extraction and the classification are completely unsupervised. However, we found it was possible to improve our classification results for this contest by leveraging the training set to refine our category selection, as well as to decrease the run time. homogeneous as possible.

**Table 1. Keywords used to label container categories**

| Locations | People | Orgs. | Misc. |
|---|---|---|---|
| cities<br>provinces<br>states<br>countries<br>continents<br>facilities<br>buildings<br>counties | people<br>men<br>women<br>doctors<br>musicians<br>government officials<br>actors<br>actresses<br>champions<br>officials<br>athletes<br>alumni<br>rappers<br>soccer- players<br>sportspeople<br>members<br>comedian | organizations<br>companies<br>colleges<br>businesses<br>enterprises | films<br>television series<br>awards<br>events |

### 4.1 Category Selection

During our test runs, we realized that our method works well with entities that are explicit mentions of people or locations, e.g., Sarah Palin. However, for mentions of more generic entities—e.g., Louis, Clint, or Sue—despite successfully finding a matching Wikipedia page, they are dismissed during the classification process. We observe that for such ambiguous parts-of-tweets the matching Wikipedia pages tended to be lists of its many possible meanings; such pages are called disambiguation pages. Disambiguation pages are also categorized in a graph-like structure, however their classification scheme is distinct from the other category pages and serves only to organize disambiguation pages. Therefore, we labeled 5 of the top 26 disambiguation-categories and added them to our containers list. Finally, since the MISC category includes 'Programming Languages', we included 'Computer Languages' category to our list. These manually added containers are shown in Table 2.

**Table 2. Additional Categories**

| Category | Label |
|---|---|
| Disambiguation pages with given-name-holder lists | PER |
| Disambiguation pages with surname-holder lists | PER |
| Human name disambiguation pages | PER |
| Place name disambiguation pages | LOC |
| Educational institution disambiguation pages | ORG |
| Computer Languages | MISC |

## 5 Discussion And Concluding Thoughts

The approach to classification described here takes advantage of information that has been created and curated by many thousands of people. The contest task illustrated the complexity of classifying short messages. For example, a noun such as "Canada" might be classified as a place, or as an organization. It is far from obvious that people will agree on such a classification. Tests might be run to determine the consistency of human judgment on this and related short message classification tasks; we might learn from the diversity of human judgment when such tasks are ambiguous, and, with further research, how such ambiguity might modeled in machine classification tasks. More generally, the task of classifying entities is one that is not only context dependent, but also may admit to differing degrees of certainty. If our goal is to classify as humans do, we ideally should understand the distribution of human responses. Thus, we suggest two paths for future research: one that continues to study how classification can be improved by using collaborative data stores, and another that examines human performance on such tasks, so that we may further understand and augment the still-mysterious process of sense making.

## References


[1] Genc, Y., Mason, W., and Nickerson, J.V. 2012. Semantic Transforms Using Collaborative Knowledge Bases, *Workshop on Information in Networks*

[2] Genc, Y., Sakamoto, Y., and Nickerson, J.V. Discovering context: Classifying tweets through a semantic transform based on Wikipedia, In D. Schmorrow and C. Fidopiastis (Eds). *Foundations of Augmented Cognition: Directing the Future of Adaptive Systems*, Lecture Notes in Computer Science, 6780 LNAI, Springer, Berlin, 2011, 484-492.

[3] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: a first look," *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pp. 73–80, 2010.

[4] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research*, vol. 34, no. 1, pp. 443–498, 2009.

[5] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 2, pp. 1419–1424, 2006.