Constructing A Corpus Of Figurative Language For a Tweet Classification and Retrieval Task

Guofu Li
Computer Science and Informatics
University College Dublin
Belfield, Dublin, Ireland
li.guofu.l@gmail.com

Aniruddha Ghosh
Computer Science and Informatics
University College Dublin
Belfield, Dublin, Ireland
arghyaonline@gmail.com

Tony Veale
Computer Science and Informatics
University College Dublin
Belfield, Dublin, Ireland
tony.veale@ucd.ie

ABSTRACT

Twitter is an intriguing source of topical content for tasks involving the detection of phenomena such as sarcasm and metaphor. The hashtags that users employ to self-annotate their own micro-texts can often facilitate the targeted retrieval of texts with the desired characteristics. Though tweets tagged with #sarcasm are highly likely to be sarcastic, the lack of a topic model for sarcastic tweets makes it difficult to detect when such tags are used in the expected way, or indeed, to retrieve tweets that are not explicitly tagged in this way. In this study, we explore how a tweet-retrieval and classification system can benefit from a topic model when constructing a task-specific Twitter corpus, such as for irony, sarcasm or metaphor detection.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Query Information, Tweet Retrieval, Hashtag Analysis.

General Terms

Algorithms, Performance, Tweet, Reliability, Experimentation, Languages.

Keywords

Information retrieval, Latent Semantic Analysis, Query Expansion, Figurative Language, Twitter.

1. INTRODUCTION

Twitter[1] is a rich source of texts for NLP research, not least because of its topicality, each of access, and the presence of self-annotation in the form of hashtags. As an immediate and very social use of language, Twitter micro-texts, or tweets, are rich in figurative language phenomena, making them a valuable resource for the study of metaphor, irony, and sarcasm.

However Twitter content also has weaknesses of its own to which Natural Language systems must adapt. For instance, tweets are short in length. Thus contractions, abbreviations and other special forms such as "u r" are ubiquitous. The tendency of users to self-tag their sarcastic tweets with hashtags like #sarcasm – to avoid misunderstandings that can escalate in a highly connected social setting, allows such tweets to be directly retrieved with Twitter API queries. Yet there has been relatively little advance in searching at a semantic or topic level in Twitter. Thus, to construct a corpus of tweets rich in sarcasm, a common approach is to search for explicit mentions of the hashtag #*sarcasm* [1]. Meanwhile, truly relevant tweets may not carry any of the hashtags expected by the corpus developer in advance. It is thus useful to explore how topic modeling can support a more robust collection process.

The rest of the paper is organized as follows: Section 2, we describe our approach to improve the tweet retrieval system based on LSA; Section 3 further studies on the value of hashtags in tweet retrieval task; Section 4 gives a brief review on the related ideas and works to our study; The final conclusions and possible future work will be presented in Section 5.

2. RETRIEVING TASK-SPECIFIC TEXTS

Twitter hashtags have their origins in earlier blogging systems, in which users tagged their own articles with category labels. Such tags carry rich semantic information when non-textual multimedia objects such as photos and videos are shared. Such tags provide a toehold for NLP techniques to be applied to non-language objects.

Twitter exploits hashtags to identify and concisely label "trending topics" amongst its users [2], while users themselves use hashtags to emphasis key parts or themes in their texts [3]. Since there is no central control over the usage of hashtags on Twitter, users may choose arbitrary tags to convey an attitude or feeling, or choose to reuse a trending tag to echo the sentiments of an emerging group. As invented and used by the tweet creators, hashtags represent concentrated bursts of information in an already concise text, making them extremely useful for tweet search and retrieval [4].

2.1 Hashtag-Based Searching

Each hashtag is thus a potential vehicle for the expression of key topics. By searching for a specific (set of) hashtag(s), one can retrieve a large set of tweets to suit a specific task. For example, Riloff *et al.* [1] use the hashtag #*sarcasm* to construct a corpus of relevant utterances for the training and testing of their sarcasm detection system. More recently, Reyes, Rosso and Veale [5] employ a fixed set of hashtags, together with syntagmatic patterns, to build a tweet collection that is rich in figurative language (including sarcasm, irony and metaphor).

Nonetheless, different users are free to choose different tags for conveying the same topic, while many tweets contain no hashtags at all. In this study, we aim to retrieve tweets that are considered sarcastic, and use, as a baseline system, a basic model that simply retrieves tweets that explicitly contain the tag #*sarcasm*. Such a baseline system cannot know whether such tweets really are sarcastic (as opposed to, say, talking about the sarcasm of others),

[1] https://twitter.com/

and cannot retrieve any sarcastic tweets that are not so marked. We compare this baseline to a more sophisticated approach that also employs topic modeling to appreciate sarcasm in tweets.

2.2 Retrieving Sarcastic Tweet with LSA

Twitter provides a convenient API that allows users to find tweets with keyword-based search, but a major limitation is the lack of a semantic retrieval capability. A keyword-based search system will only return tweets that contain raw-matches of a target keyword. Following the distributional hypothesis by Harris [6][7], hashtags that commonly appear in the same context are assumed to carry similar topics. A good topic modeling should thus be able to retrieve documents that fall into similar topic categories suggested by the query hashtag, even without a direct match of that hashtag.

A commonly used semantic model in Information Retrieval is Latent Semantic Analysis (LSA) [8]. LSA is a vector-space model based on matrix factorization. A common approach is to factorize a term-document matrix (with TF-IDF weights [9]) via SVD (Singular Value Decomposition) into the product of three matrices. In this study, we prepare the term-document matrix by regarding each tweet as a document and each token in these tweets as a term. We implemented the search process after applying SVD to the term-document matrix using *Jama* [10].

Conventional topic modeling techniques in IR are typically designed to capture the literal use of topics [9]. For instance, one may find documents that are literally about the topic *sport* using the query "football". We know of no system that uses topic modeling to retrieve the figurative use of topics (e.g. politics as a blood sport), or to retrieve uses of language that are figurative – classes of texts that are united by their pragmatic rather than their purely semantic qualities. In this study we aim to see whether LSA topic modeling can be productively applied to the retrieval of a specific kind of figurative text: sarcastic tweets.

Some tweet words are hash-tagged inline, while most hashtags are meta-level *add-ons* to a text, of a kind that is not available in other kinds of Web texts. We are still interested in the value of hashtags like #*sarcasm* when exploring topic models and LSA, and thus employ three different configurations of the term-document matrix, so as to determine the relative performance of the system when including/excluding hashtags or normal words from tweets.

2.3 Model Configurations

We employ three different configurations of the LSA term-document matrix; each configuration highlights or downplays a different source of textual knowledge. The configurations are W+H, W+#*sarcasm* and H.

2.3.1 Configuration 1: W+H.

This configuration simply uses the full contents of each tweet. In theory, this configuration should provide the most information to LSA topic modeling, and should presumably yield the best result, as the term document matrix contains both normal words and hashtags (which we denote as **W+H**).

2.3.2 Configuration 2: W+#sarcasm.

In this configuration, we eliminate support from explicit hashtags, in order to determine their actual contribution. We thus remove all the hashtags from the term-document matrix, except for the query key term "#*sarcasm*". This configuration is thus denoted as **W+ #sarcasm**.

2.3.3 Configuration 3: H.

The converse of Configuration 2 is the case in which we eliminate all normal text words and retain only the hashtags. This configuration is thus denoted simply as **H**. If the hashtags within each tweet can fully capture its meaning and topics, then an LSA topic model using this configuration should also be able to produce a result as good as Configuration **W+H**.

2.4 Testing Tweet Set

A tweet-set is prepared to evaluate different configurations of the LSA topic-model when used for the retrieval of sarcastic tweets – a representative type of figurative language. We have collected 1.5 million figurative language tweets. Among the corpus, 26K tweets are likely to be sarcastic by retrieving tweets that contain one or more of the following tags which is manually selected by observing previous sample of sarcastic tweet data:

{#*sarcasm,* #*sarcastic*, #*yeahright,* #*not*}

Then, we randomly sampled 2,500 tweets from this 26K tweet repository, to form a tweet-set S (for *sarcastic*). Roughly 45% of tweets in S explicitly contain the hashtag "#sarcasm": we denote this subset as S_1. The remainder of S is denoted as S_2. We cannot realistically expect all users to explicitly mark their sarcastic tweets in this way. We thus randomly remove #*sarcasm* from 50% of the tweets in S_1, yielding a subset S_1^h. The remainder of S_1 is denoted S_1^r.

We construct another set of 2,500 tweets, randomly sampled from normal tweets that are deemed as non-sarcastic (denoted as *NS*). So the overall tweet-set for testing is a mixture of both relevant and irrelevant tweets (i.e., $S \cup NS$), where relevance is defined by the presence of sarcastic intent. To introduce the noise that is expected when one speaks about #*sarcasm* without having a sarcastic intent, we randomly add an extra "#sarcasm" to 10% of tweets in *NS* (i.e., positive noise). Figure 1 below provides an overview of the composition of this tweet-set for evaluation.

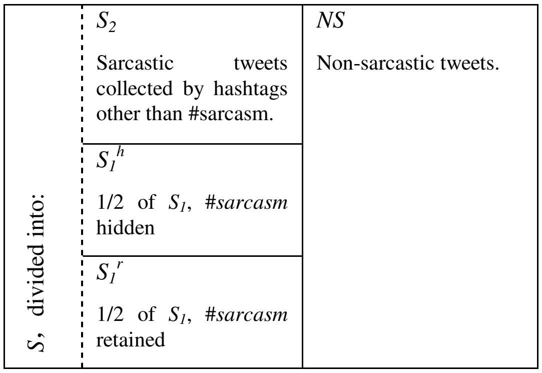
S, divided into:
S_2
Sarcastic tweets collected by hashtags other than #sarcasm.
S_1^h
1/2 of S_1, #*sarcasm* hidden
S_1^r
1/2 of S_1, #*sarcasm* retained
NS
Non-sarcastic tweets.

Figure 1. Overview of the composition of the testing tweet-set.

A-priori knowledge of the composition of the corpus yields a gold standard for evaluating a tweet retrieval system. During evaluation, tweets that are expected to be sarcastic (from set S) are deemed to be relevant tweets for this specific tweet-retrieval task.

2.5 Empirical Evaluation

We regard the entire tweet-set S as the relevance set when measuring the performance of different retrieval approaches. The baseline system is a naïve approach that returns every tweet that contains a explicit mention hashtag *#sarcasm*, which is how the current Twitter search API works. The precision of this baseline can be estimated as 0.658, while its recall is just 0.192.

By applying the four different configurations to the evaluation task, we obtain the following P/R/F performance on each set-up:

	Precision	**Recall**	**F-Score**
Baseline	0.658	0.192	.297
W + H	0.539	**0.498**	.517
W + #sarcasm	0.526	0.466	0.494
H	**.695**	0.466	**.578**

Table 1. P/R/F scores of models based on LSA topic modeling.

The baseline system clearly has difficulty in finding tweets that lack an explicit *#sarcasm* tag, so recall falls significantly below that of the other three systems. After removing the support of all other hashtags in configuration 2 (**W+*#sarcasm***), this configuration yields reduced performance overall, suggesting that hashtags other than the usual suspects are important carriers of information regarding the figurative status of a tweet.

Indeed, configuration 3 (**H**) reveals that the best performance is achieved by ignoring all *normal* words in a tweet, and by focusing only on the hashtags. This stripped-down configuration yields a precision that is higher than that of the baseline system, which offers a very credible baseline in terms of precision alone. Hashtags do indeed seem to crystalize the meaning of all other words in a tweet, and do so in a way that is less sensitive to noise.

3. QUERYING WITH HASHTAGS

The result given in Section 2.5 should perhaps not be so surprising. Hashtags are used by tweet creators not just to annotate the *normal* words of a tweet, but to comment on other hashtags themselves. An end-of-tweet hashtag like *#sarcasm* may refer to the meaning expressed in the body of a tweet, or it may refer to the figurative status of the hashtag that directly precede it (as in *"#wonderful #yeahright*").

3.1 Hashtag Suggested Tweet Clusters

As originally conceived by Twitter, tweets on topics that follow a developing trend will tend to carry the same hashtag. A hashtag often represents a concise distillation of a social movement, in which users signal allegiance by citing a designated tag (such as *#cancelcolbert*, a hashtag that emerged to capture the controversy surrounding a provocative TV comic/commentator). When users employ tags in a more sophisticated manner, e.g. to comment on uses of a hashtag by others or to mention but not endorse a tag, they are likely to use constellations of hashtags that have their own internal relationships and implied linguistic structure. We can connect tweets with common hashtags to form a graph, and thus explore how tweets are clustered by their inter-connections. Following the Harris hypothesis [6][7], we expect tweets of a similar type should carry similar hashtags too. We regard the connectivity within and between groups of tweets as a tool to study how hashtags will suggest clusters of tweets that employ similar uses of figurative language, such as sarcasm or irony.

Formally, we define a tweet linkage graph $G=<V, E>$ where each v in V represents a tweet in the dataset. An edge $e=<u, v>$ is added to E whenever two tweets u and v has at least one shared hashtag. We again use the tweet-set described in Section 2.4 for testing. To avoid the biasing effect of the seed hashtags that were used to retrieve tweets, we carefully exclude all edges that are suggested by these tags.

Specially, we are interested in three connectivity scores: the internal density of sarcastic tweets V_S (denoted as C_S), the internal density of non-sarcastic tweets V_{NS} (denoted as C_{NS}), and the inter-connectivity between V_S and V_{NS} (denoted as $C_{S,NS}$). Observing that a fully connected graph will have $|E| = 0.5|V|(|V| - 1)$, the graph density score that represent internal connectivity can be calculated as follows:

$$C = \frac{2|E|}{|V|(|V|-1)}$$

Since the fully connected bi-partite graph between V_S and V_{NS} has $|V_S||V_{NS}|$ edges, the connectivity between V_S and V_{NS} is given by:

$$C_{S,NS} = \frac{|E|}{|V_S||V_{NS}|}$$

We again use the dataset described in Section 2.4 for this analysis. To avoid the biasing effect of the seed hashtags used to build it, we carefully hide these tags when constructing the graph G. The following tables compare these three density scores:

	Density Score
C_S (Sarcastic Internal)	**0.15723**
C_{NS} (Non-Sarcastic Internal)	0.01416
$C_{S,NS}$ (Inter-connectivity)	0.02979

Table 2. Connectivity within and between groups of tweets.

Due to the sparseness of hashtags, it is difficult to find a high-density connectivity for each of them. However, it is still obvious that sarcastic tweets are much more densely connected than the other two kinds, and these tweets thus form a coherent cluster. We are encouraged then to believe that commonly shared hashtags for certain categories of tweets can facilitate the retrieval of even more tweets of the same type. In other words, the clustering of hashtags is a useful query expansion technique for tweet retrieval.

3.2 Expanding Query Hashtags

Query expansion is a widely-used technique to improve the recall of a given query [11], by adding related terms to a query that provide more hooks with which to retrieve additional texts. Ideally, expansion should increase recall by adding terms that do not simultaneously reduce precision in any significant way.

Previous topic modeling with Configuration **H** (hashtags only) offers good estimate of the relatedness of each hashtag to the original query word (i.e., *#sarcasm*), which in turn provides a simple way to expand the query for the task. In our study, we randomly collected 80 tweets for each of the following queries:

1. The seed hashtags used previously, as listed in Section 2.4;

2. The top-5 ranked hashtags suggested by **Configuration H** as described in Section 2.5; (i.e., *{#Oops, #fuckyall, #ISIS, #HandleThat, #imgoingtohell* });

3. 5 random hashtags that were chosen from tweets in *NS* as described in Section 2.4.

We manually evaluated the 240 tweets that are collected using queries expanded by these three sets of hashtags. The following table summarizes the precision of each query type. The seed hashtags that were manually chosen for the task (1) unsurprisingly yield the best precision. However, auto-generated hashtags yield almost identical performance, which supports the idea of using related hashtags for query expansion for this type of retrieval task.

Query Basis	*Precision*
Seed Hashtags	**0.837**
Top-5 ranked Hashtags	**0.712**
5 random hashtags	0.112

Table 3. Precision for tweets with different query hashtags.

4. RELATED WORK AND IDEAS

Naveed *et al.* [12] summarize the challenges that arise with text retrieval in a micro-blogging environment, and suggest that search algorithms must adapt to meet the nature of this new type of text source. Posch *et al.* [13] explore the connection between hashtags and semantic categories, and while not particularly interested in how hashtags can support search tasks, suggest that hashtags can be categorized by their pragmatic and lexical properties. Massoudi *et al.* [4] proposes a micro-blog retrieval model based on dynamic query expansion that identifies – via a language model derived from tweets – useful terms (though not specifically hashtags) that can be appended to a search query.

Work similar to that described here is reported by Efron [14], who specifically studied hashtags in a micro-blogging environment. Efron explored three perspectives on hashtags: tag following, result display, and query expansion. IDF (inverse document frequency) was employed as a primary indicator of the information content of a hashtag, while KL-divergence (a Dirichlet prior [15] for smoothing) was used to measure the relevance of a tag to be expanded to the query. Rather than test the value of hashtag-based query expansion on an ecological task, Efron's evaluation is based on 29 manually chosen query topics that are tailored by the author.

5. CONCLUSIONS AND FUTURE WORK

Hashtag-based tweet retrieval has already proved its value in providing corpora for various kinds of NL research. The background need of this study is to construct a tweet collection that is rich in figurative language use, such as sarcasm, irony and metaphor, and we have focused here on one of these figurative phenomena, sarcasm. We employ LSA as a topic-modeling tool, and employ three different configurations of the tool to determine the relative value of different information-bearing parts of a tweet. A performance comparison suggests the true value of hashtags as concise vehicles of tweet meaning and user intent, which motivates us in turn to develop and evaluate a query expansion technique for hashtags. The top 5 hashtags suggested by Configuration H yielded a retrieval precision (.712) that is encouragingly close to that yielded by hand-picked query tags. Building on this pilot study, research can be continued along several dimensions, including an evaluation and comparison of other topic modeling and query expansion techniques. It remains to be seen if a combination of topic modeling and hashtag-based query expansion will yield benefits that improve on each in isolation.

6. REFERENCES

[1] Riloff, E., Qadir, A., Surve, P., Silva, L.D., Gilbert, N., Huang, R. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. *Proceedings of the 2013 Conference on EMNLP* pages 704-714. 2013, Seattle..

[2] Kwak H., Lee C., Park H., and Moon, S. 2010. "What is Twitter, a Social Network or a News Media?" *WWW 2010*, Raleigh.

[3] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. *In Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 173–178, New York, 2010.

[4] Massoudi K., Tsagkias M., Rijke M., and Weerkamp W. Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. *Advances in Information Retrieval - 33rd European Conference on IR Research*. 2011, Dublin.

[5] Reyes, A. and Rosso, P. and Veale, T. 2013. A multidimensional approach for detecting irony in Twitter. *Journal of Language Resources and Evaluation*. March 2013, Volume 47, Issue 1, pp 239-268. Springer Netherlands.

[6] Harris, Z. S. 1954. Distributional structure. *Word. Journal of the linguistic circle of New York*. 10, 2–3, 146–162.

[7] Harris, Z. S. 1957. Co-occurrence and transformation in linguistic structure. *Language* 33, 3, 283–340.

[8] Dumais, S. T. (2005). Latent Semantic Analysis. *Annual Review of Information Science and Technology* 38: pages 188-230. doi:10.1002/aris.1440380105.

[9] Rajaraman, A., Ullman, J. D. 2011. Data Mining. *Mining of Massive Datasets*. pages 1-17.

[10] http://math.nist.gov/javanumerics/jama/

[11] Efthimiadis, E. N. Query Expansion. In: Martha E. Williams (ed.), *Annual Review of Information Systems and Technology (ARIST)*, v31, pp 121–187, 1996.

[12] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. 2011. Searching Microblogs: Coping with Sparsity and Document Quality. *CIKM'11*, 2011, Glasgow.

[13] Posch, L., Wagner, C., Singer, P., and Strohmaier, M. 2013. Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter. *WWW 2013 Companion*, May 13–17, 2013, Rio de Janeiro, Brazil.

[14] Efron, M. 2010. "Hashtag Retrieval in a Microblogging Environment". *SIGIR 2010*. Geneva.

[15] Hazewinkel, Michiel, ed. 2001. Dirichlet distribution, *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-01