Bridging Vocabularies to Link Tweets and News

Tuan-Anh Hoang-Vu
NYU School of Engineering
tuananh@nyu.edu

Aline Bessa
Federal University of Minas Gerais
alineduartebessa@dcc.ufmg.br

Luciano Barbosa
IBM Research – Brazil
lucianoa@br.ibm.com

Juliana Freire
NYU School of Engineering
juliana.freire@nyu.edu

ABSTRACT

Social media has become a popular platform for publishing, sharing and consuming news. However, it is not a replacement for traditional sources of news — they are complementary. While news sites provide in-depth and comprehensive coverage of events and topics, social media postings include comments, opinions and rumors about facts publicized on the news. Social media can thus serve as a useful sensor for how popular a story (or topic) is, for how long, and people's sentiments about it. To use social media as a sensor, we first need to associate postings to news stories and topics. But doing so is challenging since postings are short and the vocabularies used in postings and in news can be very different. In this paper, we take a first step towards addressing this problem. We propose a framework that uses news as a proxy to build a topic model and associates Twitter postings (a.k.a. tweets) to the derived topics. Subsequently, to deal with vocabulary differences, we present a new strategy to adapt the topic model derived from news to tweets. We report the results of an experimental evaluation which indicates that our framework obtains high accuracy for a variety of topics, and that domain adaptation leads to significant gains in both precision and recall.

Keywords

Social media, domain adaptation, topic models

1. INTRODUCTION

Social media has become a popular platform for publishing, sharing and consuming news. A recent study showed that more than half of digital news consumers follow news from social media sites. However, they still prefer to go to traditional news outlets.[1] This choice is natural given that news sites provide in-depth coverage of events and topics, whereas social media postings are usually short and lack contextual information. On the other hand, social media postings contain comments, opinions or rumors about news stories (e.g., Figure 1 shows some comments on Twitter about the 2013 papal conclave). Social media is thus not a replacement for traditional news sources but complementary: it can serve as a useful social sensor for how popular a story (or topic) is and for how long [5].

Vatican announces the conclave date to select the next pope will be announced today at 1pm Vatican a gonna make u sweat for it!

Can't believe people can gamble on who the next Pope is going to be.

Picking the next Pope is March Madness for Catholics.

Bolt from the blue after pope resigns - Correspondent: http://t.co/PICMTNuo

Figure 1: Tweets related to news about the 2013 papal conclave.

However, for tweets to be used as sensors, they first need to be associated to a news story or topic. However, doing so is difficult for a number of reasons [7, 8, 10], notably, to due to the size of tweets and differences in the vocabularies used in tweets and in news. Since tweets are limited to a small number of characters, people often use abbreviated syntax. Furthermore, in contrast to editorially-managed news articles, tweets are written in colloquial language, e.g., they often includes slangs. Last, but not least, the postings are also noisy and may contain meaningless content.

Twitter provides hashtags to help label postings and categorize them.[2] Hashtags can thus be used to identify particular news stories or topics. However, not only is it difficult to obtain hashtags for the continuously growing Twitter corpus, but also they are not used consistently, and relevant postings may not contain an appropriate hashtag. There are many hashtags that do not relate to any event in particular (e.g., hashtags associated with organizations), or that are very general (e.g., #summer2013 can be associated with any summer 2013 event). Furthermore, some topics are associated multiple hashtags (e.g., #Christopher, #Dorner were used for the news story about a shooting in Los Angeles in 2013). Consequently, it is hard to identify which hashtags are the best ones to look for when trying to keep track of

[1] http://stateofthemedia.org/2012/mobile-devices-and-news-consumption-some-good-signs-for-journalism/what-facebook-and-twitter-mean-for-news

[2] https://support.twitter.com/articles/49309-what-are-hashtags-symbols

Seventeenth International Workshop on the Web and Databases (WebDB 2014), June 22, 2014 - Snowbird, UT, USA.

Table 1: Top terms in news and tweets in a topic

Topic ID	Top terms in News	Top terms in Tweets	Jaccard index
47	snow storm power friday boston connecticut plow massachusett sunday saturday	snow storm friday tomorrow #snow rain inch shovel hope play	0.818
50	immigr worker reform law illeg citizenship legal busi labor senat	immigr reform worker illeg american nativ act anti #immigr updat	0.708
52	polic los angel dorner depart offic suspect chief communiti cabin	los angel con todo por para las son como del	0.736
111	pope church benedict cathol priest gay vatican john cardin resign	pope gay resign benedict cathol xvi decis church year #pope	0.863
149	knick chandler rebound defens player anthoni minnesota guard team rugbi	miss knick roll lose game pick guard anthoni too chandler	0.857

a specific news event. Relying solely on hashtags to connect tweets to news events can lead to low recall and precision [11].

To address these challenges, we propose a new framework that aims to associate content in traditional news sources with tweets. The framework first uses Latent Dirichlet Allocation (LDA) [2] to build topic models over news articles. While there have been approaches that construct topic models directly over tweets (see e.g., [17]), we have opted to use news as the starting point for two important reasons: it matches our goal of using tweets as a sensor for news; and while topic modeling methods such as LDA work well for long, clean documents, they can be ineffective for short tweets [10]. However, this creates a challenge: the language (and vocabulary) used in news differs from the one used in tweets. Thus, relevant tweets may be incorrectly classified if they contain terms that are either infrequent or absent in news. Our approach deals with this problem by *adapting* the topic model from news to match tweets. It first selects tweets that match the news-derived topic model using a similarity measure, and then uses the terms in that appear in these tweets as a *bridge* to identify other relevant tweets.

To evaluate the effectiveness of our approach, we used 1,415 news articles and over 3 million tweets to derive a gold data set that consists of topics and associated tweets. We show that, for these data, our approach obtains high precision and recall, and the domain adaptation step leads to significant gains in recall.

2. ASSOCIATING TWEETS TO NEWS

Our main goal in this work is to associate Twitter postings (tweets) with news topics extracted from traditional news media. For this to be feasible, postings and news must have overlapping topics. Zhao et al. [17] have shown that, within a given time interval, a similar range of topics is covered by Twitter and news. Thus, the problem we address can be stated as follows: Given a set of postings Δ and a set of news documents D that appeared in a time window tw, and a set of topics Φ where each topic is characterized by a distribution over words in D, we aim to derive assignments $\delta \rightarrow^{tw} t$, where $\delta \in \Delta$ and $t \in \Phi$.

Solution Overview. We establish this association in an unsupervised way, by performing the following subtasks: (i) identify the set Φ of topics in D, (ii) adapt the news-derived topics to tweets, and (iii) use the adapted model to classify tweets and assign them to a topic. Figure 2 shows an overview of the end-to-end process. Initially, we run LDA over the news data (LDA-Train). LDA is an effective method for modeling topics in news [2]. It generates topic probability distributions over articles and word probability distributions over the topics. These word distributions tell us how important each word is for each topic, i.e., they define which set of words better represent the topic and provide a good summary for their content.

Note that, for a given topic, there exists some overlap in the vocabulary used. This is illustrated in Table 1, which shows the top-10 words in the same topic in news and tweets, and the Jaccard index between the two word sets. We exploit this observation and use distributions of words over topics learned from news data, to create an initial set of topics for tweets (Pre-Classification). The table also shows that there are considerable differences in these initial topics. Our approach addresses this through domain adaptation – it relies on word co-occurrence to both re-rank and expand the set of words in a topic (TwiRank). Finally, the adapted model is used to classify the remaining tweets (Classification).

2.1 LDA-Train: The Learning Model

The *LDA-Train* component takes a news corpus and a number T of topics as inputs, and generates a topic model that fits the data. The news corpus consists of a collection D of documents where each $d_i \in D$ is an arbitrary sequence of words. Words are the basic units of our system. The vocabulary V is the set of all words in D. The output of *LDA-Train* is a list Φ of topic vectors and each $\phi_t \in \Phi$ has size $|V|$, contains the probability distribution of words with respect to topic t. Note that the quality and granularity of the topic model depends on T.

2.2 Pre-Classification

We use the model Φ to bootstrap the creation of a Twitter-specific topic model. Since there is an overlap between the news and Twitter vocabulary, we use Φ to classify the tweets that contain the representative words from news.

Since each ϕ_t has size $|V|$, every word in the news vocabulary contributes to the formation of a topic. However, for any particular topic, only a small number of words make a significant contribution. Therefore, it is sufficient to consider only the most important words for each topic. First we reindex words based on the descending order of their distribution and convert the distribution into a discrete function of word indices. Then, we take the finite difference of this function and find the position where this difference falls below an empirical threshold. Finally, we discard all words having distribution lower than distribution of word at cut-off point. This process produces a new set of topic vectors, Φ'. Examples of topics in Φ' are given in Table 2.

The pre-classification is performed as follows. Given a Twitter corpus consisting of a set of tweets, $\Delta = \{\delta_1, \delta_2, ..., \delta_X\}$, we use a similarity-based approach to infer the initial assign-

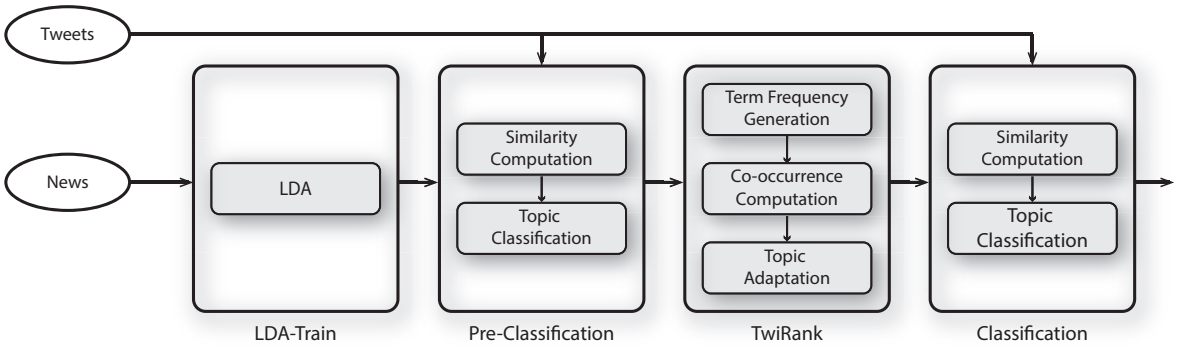

Figure 2: Processing framework for associating tweets to news

Table 2: Words and their weights for two sample topics: *"Christopher Dorner shootings and manhunt"* and *"Resignation of Pope Benedict XVI"*. Words are stemmed and weights are non-normalized.

Word	Weight	Word	Weight
polic	203.92	church	90.30
offic	72.57	pope	42.89
prison	66.89	benedict	40.98
depart	49.97	cathol	38.22
kill	48.59	februari	25.95
sentenc	47.35	member	24.38
dorner	33.99	servic	23.04
accus	31.86	priest	21.77
attack	30.39	2013	21.70
charg	29.53	rev	18.65

ment. Note that the vocabulary of Δ is not necessarily the same as vocabulary of D. To infer the initial assignment efficiently, we use a similarity-based approach. Given topic vector ϕ'_t in Φ' and a tweet δ, we compute the cosine similarity between the two:

$$z_t = \frac{\phi'_t \cdot TV(\delta)}{\|\phi'_t\| \|TV(\delta)\|} \tag{1}$$

where $TV(\delta)$ is the term frequency vector of tweet δ. A high similarity score between a tweet δ and a topic ϕ indicates that there is a high probability that δ belongs to ϕ. Once similarity scores are computed, the *Topic Classification* module assigns tweets with the highest similarity score to the corresponding topic. A tweet δ is assigned to topic ϕ^i if

$$z_i = \max_i^T(z_1, z_2, ..., z_T) \tag{2}$$

To avoid getting misclassified tweets, which may negatively impact the quality of the topic model, we discard tweets with similarity scores below a given threshold. As we discuss below, these tweets will be reconsidered and classified by the TwiRank algorithm, presented in section 2.3.

2.3 TwiRank

As discussed in Section 1, there are differences in the vocabularies used in news and tweets. For instance, in the topic *"Christopher Dorner shootings and manhunt"*, the word *"nigga"* has high frequency, but this word does not occur in the news articles about this topic. Therefore, a topic model derived from news is an imperfect classifier for tweets.

We propose Twirank, a new strategy that *adapts* Φ' to tweets by taking into account information about word co-occurrence in tweets. Given a word w^t that occurs in tweets, if w^t co-occurs often with words in a topic ϕ', TwiRank boosts the importance of w^t in ϕ'.

Several measures of co-occurrence have been introduced, such as Dice coefficient, Mutual Information, Lexicographers Mutual Information, log-likelihood test, Poisson significance measure, z-score and t-score. Bordag [3] did a comprehensive review of these measures and showed that when the frequency of most words is much smaller than the corpus size, which is the case for our corpus, Poisson significance measure outperforms the others. It is defined as:

$$CO_{Poisson}(A,B) \approx n_{AB}(\ln n_{AB} - \ln \lambda - 1) + \frac{1}{2}\ln 2\pi n_{AB} + \lambda \tag{3}$$

where n_{AB} is the co-occurrence frequency between two words A and B (i.e., the number of sentences having both A and B); and $\lambda = \frac{n_A \cdot n_B}{n}$ is the generalized likelihood ratio (n_A and n_B are word frequencies of A and B, and n is the corpus size).

Given a word A and a topic t, the co-occurrence score of word A w.r.t. t is defined as:

$$CO(A,t) = \sum_{X \in \phi'_t} CO_{Poisson}(A,X) \times w(X) \tag{4}$$

where ϕ'_t is the terms vector of topic t as described in Section 2.2, $w(X)$ is weight of word X in term vector ϕ'_t.

Using the equation above, we calculate scores for every word in the Twitter corpus with each topic t, rank them based on their scores and create a new term vector Φ^{TWEET}. It is important to note that Φ^{TWEET} is created based on Φ but is modified for tweets using the co-occurrence score.

2.4 Classification

In this step, we use the same approach described in Section 2.2, but apply the adapted model, Φ^{TWEET} to classify the tweets. This process derives revised similarity scores for tweets in topics. The Topic Classification module uses these scores to construct the final assignments of tweets to topics.

3. EXPERIMENTAL EVALUATION

3.1 Experimental Setup

Datasets. We collected news and tweets for a period of one week, from Feb 06 to to Feb 12, 2013. We obtained tweets

Table 3: Statistics for data sets used in evaluation

Corpora	# of docs	Avg. # of terms per doc	# of distinct terms
News	1415	304.74	28991
Twitter	3188497	6.73	117395

using the Twitter streaming API retrieving the maximum number of postings allowed by the API. We also developed a custom crawler which collected documents from 10 popular US news sites (NY Times, CNN, Fox News, NBC News, Washington Post, LA Times, Reuters, ABC News, USA Today and Huffington Post). Since these news articles contained many old pages, we implemented a script to extract the publishing date in the pages, and selected only articles whose dates fall within the time window of the Twitter collection. The crawler retrieved 5,000 news pages per day.

For both collections, we removed stopwords,[3] words with fewer than 3 characters, and words appearing in less than 4 documents. We also applied internal stemming using the Porter2 algorithm [12]. For tweets, we filtered out non-semantic terms such as URLs and user mentions (@user). We also observed that some hashtags were combinations of multiple words (e.g., *#christopherdorner* or *'#popebenedict'*). We split these composite hashtags using a dictionary. After pre-processing, we had a total of 1,415 news articles and over 3 million tweets. Statistics for the news and Twitter data sets are presented in Table 3.

Gold data. Given the size of the Twitter data set and the potentially large number of topics, evaluating the effectiveness of our approach is challenging. To scale the evaluation process, we need a robust and accurate method to (semi) automatically generate gold data. Such approaches have been used in previous works. For example, Jin et al. [9] relied on hashtags as hidden indicators for a topic. As we discuss below, we also use topic-related terms to select tweets, but we perform additional validation to rule out false positives and we expand the search using keywords.

To construct our gold data, starting from a list of topics derived by LDA for the news corpus, we manually selected a set of terms, hashtags and keywords, that are clearly associated with each topic. We scanned the Twitter corpus and select all tweets containing these terms. However, the presence of a term is not sufficient for determining the topic of a tweet. We observed both the inconsistent use of hashtags (i.e., a hashtag in a tweet that does not belong to the topic implied by the hashtag) as well as tweets that contain multiple, seemingly unrelated, hashtags. Thus, to improve the quality of the gold data, we selected from the set only tweets that also have a URL. Next, we retrieved the documents pointed to by the URLs, and for each document d associated with a tweet t, we computed the cosine distance between d and the terms of the candidate topic for t. We ignored all topics with fewer than 10 associated tweets, since these are unlikely to be significant.

For some topics, using only hashtags, we were unable to gather enough gold data candidates. For such topics, we repeated the process using keywords instead of hashtags. In the end of this process, we gathered a total of 2,147 tweets associated to 26 topics.

[3]http://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html

Table 4: Precision, Recall and F-measure for the different approaches

	T-S	J-LDA	LDA-C	LDA-T	LDA-T-I
Precision	0.340	0.88	0.77	**0.91**	**0.91**
Recall	0.20	0.27	0.68	**0.77**	**0.77**
F-measure	0.2478	0.4158	0.7204	**0.84**	**0.84**

Approaches Considered. To assess the effectiveness of our framework, we compared the following approaches:

- **Topic-Similarity (T-S)**: LDA is executed for news and tweets separately, and the resulting models are then paired up using Jensen-Shannon divergence. This is similar to the approach proposed by Zhao et al. [17].
- **Joint-LDA (J-LDA)**: LDA is executed for news and the resulting topic model is used to classify the tweets.
- **LDA-Cosine (LDA-C)**: After LDA is executed over news corpus, instead of using generated model to classify the tweets, we convert the generated model into vectors, use cosine similarity between these vectors and the tweets, and assign tweets to the topic with highest similarity score. This corresponds to the pre-classification step of our approach (see Section 2.2).
- **LDA-TwiRank**: Applies all the steps of our framework, including model adaptation (TwiRank) after the initial pre-classification step (see Figure 2).
- **LDA-TwiRank-I**: While LDA-TwiRank computes assignments for all tweets in the collection, LDA-TwiRank-I does so only for the tweets that were not considered in the pre-classification.

For all approaches, we ran LDA using the Stanford Topic Modeling Toolkit [13]. We selected $T = 150$ as number of topics and ran 1000 iterations using *Collapsed Variational Bayes Inference* (CVB0) [1]. For each method, we used the derived topic distribution and assigned tweets to the topic with the highest score.

Evaluation metrics. To compare the different approaches, we computed precision, recall and F-measure for each approach with respect to the gold data. Recall represents the number of correctly classified tweets in a topic ϕ as a fraction of all tweets in the gold data that belong to ϕ; and precision represents the number of tweets in topic ϕ as a fraction of the tweets that are classified by the approach as belonging to ϕ. The F-measure is the harmonic mean between precision and recall. The overall score for the different approaches was calculated by taking the arithmetic average of scores for all topics.

3.2 Effectiveness of TwiRank

Table 4 shows the average precision, recall and F-measure obtained by each approach for the 26 topics represented in the gold data. *LDA-TwiRank* and *LDA-TwiRank-I* obtained the highest precision (0.91), recall (0.77), and F-measure (0.84). *LDA-Cosine* obtained lower scores, but performed substantially better than the other two approaches. This underscores the effectiveness of domain adaptation: by adapting the topic model from news to tweets, there is a considerable gain both in precision and recall. In addition,

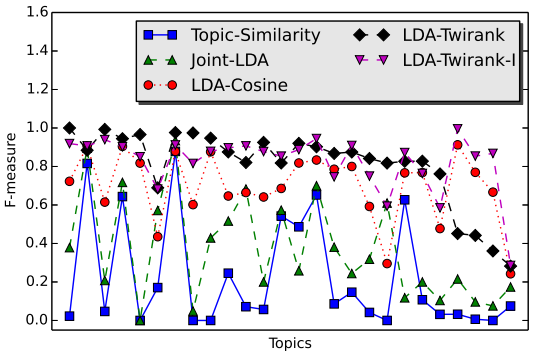

Figure 3: F-measure values obtained by the different approaches for each topic

these results are consistent with previous findings that LDA, when applied to tweets directly, derives low-quality topics.

We also examined performance of each approach on individual topics. The F-measure values are shown in the Figure 3. Our solutions outperformed baselines in all topics. One surprising fact that stands out in this plot is that even though on average, LDA-TwiRank and LDA-TwiRank-I obtained same F-measure values, for 3 out of 26 topics, LDA-TwiRank performs worse. These are actually examples of topics for which domain adaptation deviates from the topic. Because LDA-TwiRank applies the adapted model to all tweets in the collection, it ends up re-assigning tweets to the incorrect topic. As a result, its F-measure can be lower than that of LDA-Cosine. In contrast, LDA-TwiRank-I, which only applies the adapted model to tweets that were not considered in the pre-classification, always performs better than LDA-Cosine. Even though in some cases LDA-TwiRank outperforms LDA-TwiRank-I, these results indicate that the latter is more robust to topic drift. Upon further examination of the actual topics and associated tweets, we realized that a large proportion of the tweets assigned to these 3 topics were ambiguous: in the topic distribution, the probabilities for the top topics were very similar. This may explain why adaptation leads to topic drift. It also suggests that a better strategy is needed for topic assignment in situations where there is a substantial overlap between topics. For example, instead of just selecting a topic based on the highest score, ambiguous tweets may be presented to the user, who can decide the correct topic assignment.

3.3 Discussion

Limitations. As the experimental results show, our approach is effective at connecting tweets to news. But it also has limitations. Since we use topics derived by LDA as a starting point, if the word distribution does not properly represent the intended topic, the quality of the assignments is likely to be low. Also, as we discussed above, overlapping topics also lead to incorrect assignments. In future work, we plan to explore topic models that have been shown to perform better than LDA, including models that consider n-grams [16]. Another factor that impacts assignment quality is the size of the sets produced in the pre-classification step. While this is not a problem for popular topics, for topics that are not so popular, the pre-classification step may not generate a sufficient number of representative assignments. This may hamper the effectiveness of domain adaptation.

Table 5: Additional hashtags discovered by TwiRank

Defined Hashtags	Discovered Hashtags
#hurrican	#weather, #snow, #powdertrack, #snowstorm, #blizzard2013
#snowstorm	#snowpocalypse, #snow, #nemo, #snowstorm, #blizzard2013
#immgrant, #immigrantreform	#internship, #immigration, #job, #sotu
#popebenedict	#church, #pope, #churchflow, #catholic
#drone	#ReplaceSongTitleWithDrone, #drones
#chrisdorner	#dorner, #bigbear, #manhunt, #security, #losangeles
#breastcancer	#healthcare
#knick	#sportroadhouse, #lakers, #nba
#olympic	#SaveOlympicWrestling, #sportroadhouse, #ioc, #wrestling

Discovering Hashtags. While hashtags have the potential to help associate tweets to news as well as help in clustering tweets, it is difficult to obtain relevant hashtags due to the very dynamic nature of Twitter [10]. The approach we propose can be used to help identify hashtags that are associated with a given topic. From the assignments derived by our approach, for each topic t, we collected the set of all hashtags in tweets that belong to t. Table 5 shows some examples which include a popular hashtag and the set of related hashtags automatically derived by our approach. A cursory validation indicates that relevant hashtags are discovered. In future work, we plan to investigate this in more detail.

Alternative Methods and Features. We have experimented with a number of variations of the approach described in Section 2. For example, we used named entity recognition (NER) and represented documents and topics using entities, instead of relying solely on words. Our hypothesis was that entities might help better distinguish between topics that have a large overlap. We applied NER to both news and tweets using state-of-the-art algorithms (Stanford NER [4] for news and twitter_nlp [14] for tweets). However, for the topics in our gold data, the overlap between entities in news and tweets was very low.

To deal with the small size of tweets, we also tried to group tweets in a conversation: since conversations are likely to revolve around the same topic, they would provide more context (and words), what could potentially simplify the topic assignment task. Unfortunately, tweets provided by the Twitter API do not provide sufficient samples for conversations. Out of the 3 million tweets in our corpus, we were able to reconstruct only 40 thousand conversations that contained more than 1 tweet, and our experiments showed no significant improvements in the results.

4. RELATED WORK

Previous works have considered different problems related to connecting tweets and news. Gao et al. [5] proposed an approach based on topic modeling to generate summaries from news and tweets. Given an event, they collected rele-

vant news articles and tweets, and derived complementary summaries from both sources. Their goal is different from ours: while we aim to associate relevant news articles and tweets, they derive summaries from such associations. Guo et al. [6] used a graph-based model that connects news and tweets based on hashtags in tweets, named entities in news, and temporal constraints. As we discussed above, in our evaluation, we observed that hashtags are not used consistently. Thus, relying on hashtags can result in low recall. It would be interesting to explore the integration of our approach to derive hashtags for topics with Guo's model.

Zhao et al. [17] compared the content of tweets and news using unsupervised topic modeling. Their goal was to assess whether Twitter can be regarded as a "faster news feed that covers mostly the same information as traditional news media". They extracted topics from tweets and compared these against topics from the New York Times, taking into account differences in terms of proportions, types and categories. The goal of our work is different. We are not interested in comparing news and Twitter topics, instead we want to associate them so that news consumers (and providers) can be informed about the social media discussions regarding particular topics.

Given a news article and social media responses to that article, Štajner et al. [15] proposed a strategy to identify the most interesting responses to be displayed. Our approach is orthogonal – we aim to find responses to the articles (in an unsupervised fashion). We note that in an interface that presents the associations we discover to users, a strategy that identifies the most relevant tweets would be useful.

More closely related to our approach is the work by Kang et al. [10], which proposed to use transfer learning as a means to improve the effectiveness of topic modeling for tweets. They developed an extension of LDA that utilizes labeled documents in different domains, e.g., Yahoo! News and Wikipedia, as priors. However, their goal is to identify broad topics, such as science, business, health, sports, politics, world, technology and entertainment. While our approach can also handle these, we are primarily interested in specific events and fine-grained topics.

5. CONCLUSION

In order to use Twitter as a sensor for news, in this paper we proposed a new approach to associate tweets to news stories and topics. Our approach uses the content in news to derive a topic model, and to deal with the differences in vocabularies in the two sources, it adapts this model to tweets. We report the results of an experimental evaluation which indicates that our framework obtains high accuracy for a variety of topics, and that domain adaptation leads to significant gains in both precision and recall.

Acknowledgments. This work was supported in part by the National Science Foundation award CNS-1229185.

6. REFERENCES

[1] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI*, pages 27–34. AUAI Press, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[3] S. Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In *CICLing'08*, pages 52–63, 2008.

[4] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

[5] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *CIKM*, pages 1173–1182, New York, NY, USA, 2012. ACM.

[6] W. Guo, H. Li, H. Ji, and M. T. Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL*, pages 239–249, 2013.

[7] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *ICDM*, pages 153–164, 2012.

[8] B. Han and T. Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In *ACL HLT*, pages 368–378, 2011.

[9] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *ACM CIKM*, pages 775–784, 2011.

[10] J.-H. Kang, J. Ma, and Y. Liu. Transfer topic modeling with ease and scalability. In *ICDM*, pages 564–575, 2012.

[11] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how twitter is used to widely spread scientific messages. In *WebSci*, 2010.

[12] M. F. Porter, R. Boulton, and A. Macfarlane. The english (porter2) stemming algorithm. *Retrieved*, 18:2011, 2002.

[13] D. Ramage and E. Rosen. Stanford topic modeling toolbox, 2009.

[14] A. Ritter. twitter_nlp. https://github.com/aritter/twitter_nlp.

[15] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes. Automatic selection of social media responses to news. In *ACM KDD*, pages 50–58, 2013.

[16] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702, 2007.

[17] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR*, pages 338–349, 2011.