ALL ABOUT MICROTEXT

A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing

Jeffrey Ellen
SPAWAR Systems Center Pacific, 53560 Hull St, San Diego, CA, U.S.A.

Keywords: Microtext, Natural language processing, Text classification, Semi-structured data, Information extraction, Sentiment analysis, Topic summarization.

Abstract: This paper defines a new term, 'Microtext', and takes a survey of the most recent and promising research that falls under this new definition. Microtext has three distinct attributes that differentiate it from the traditional free-text or unstructured text considered within the AI and NLP communities. Microtext is text that is generally very short in length, semi-structured, and characterized by amorphous or informal grammar and language. Examples of microtext include chatrooms (such as IM, XMPP, and IRC), SMS, voice transcriptions, and micro-blogging such as Twitter(tm). This paper expands on this definition, and provides some characterizations of typical microtext data. Microtext is becoming more prevalent. It is the thesis of this paper that the three distinct attributes of microtext yield different results and require different techniques than traditional AI and NLP techniques on long-form free text. By creating a working definition for microtext, providing a survey of the current state of research in the area, it is the goal of this paper to create an understanding of microtext within the AI and NLP communities.

1 INTRODUCTION

Information retrieval and extraction on free text (e.g. long form prose, newswire releases, emails, etc) is a relatively vibrant and burgeoning research area within the AI and NLP communities, but by comparison there is a lack of studies and experiments on shorter texts, especially where grammar is less formal and abbreviations are more common. One of the difficulties in organizing or tracking this type of research is there is not a common term differentiating these shorter, less formal texts. This paper suggests 'Microtext' as being an appropriate term for this type of text. As electronic communications become more prevalent, we expect Microtext sources to become more common, and more important in day-to-day operations within every industry.

Microtext sources include point to point instant messaging via any protocol (such as XMPP), Multi-User Chatrooms or MUCs (such as IRC), SMS (Short Message Service) common on mobile phones, transcriptions of voice conversations, and micro-blogging which has been popularized by Twitter and similar services.

In section 2 this paper will introduce a working definition for microtext, and characterize some common microtext examples. Section 3 will survey some NLP and AI papers that work on microtext sources, with varying degrees of acknowledgement or adjustment to the problem domain. This includes examples focused on classification, clustering, information extraction, sentiment analysis, etc. Section 4 briefly illustrates some counterexamples, and section 5 concludes the survey.

2 MICROTEXT DEFINITION

A definition of microtext is required for future research efforts. The definition is not strict, in the sense that it will not be defining an API or a protocol, but a solid definition will certainly help provide a 'stake in the ground' for future discussion of work. A definition will serve two purposed primarily. First, if adopted and utilized as a term or keyword, it will greatly aid scientists and engineers in locating similar research. Second, the definition will help assist future researchers by serve as the delineation between microtext and long form text.

Ellen J..
ALL ABOUT MICROTEXT - A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing .
DOI: 10.5220/0003179903290336
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 329-336
ISBN: 978-989-8425-40-9

Dalli, Xia, and Wilks (2004) presented a summary of the "unique characteristics of email" which consisted of essentially:

- Short messages between 2-800 words.
- Unconventional grammar & style (frequently).
- A cross between informal and traditional.
- Threading characteristics

This type of definition is what is necessary for microtext, however, 800 words is too long for many of the microtext sources. For a point of reference, there are approximately 700 words on this page. Additionally, since microtext is encountered in multiple media, not all of which include threading, this definition cannot be used as is.

2.1 Working Definition

Microtext is considered to have three main characteristics that separate it from the traditional documents used in text categorization:

- Individual author contributions are very brief, consisting of as little as a single word, and almost always less than a paragraph. Frequently the contribution is a single sentence or less.
- The grammar used by the authors is generally informal and unstructured, relative to the pertinent domain. The tone is conversational, and frequently unedited therefore errors and abbreviations are more common.
- The text is 'semi-structured' by traditional NLP definitions since it contains some meta-data in proportion to some free-text. At a minimum, all microtext has a minute-level timestamp and a source attribution (author).

This definition is subject to change with respect to precision. Through experimental validation, these definitions can be made more concrete.

In regards to the length, 'very brief' is not specific. It is suggested that future studies could help specifically quantify length either explicitly through experimentation, or implicitly through deriving where documents consisting of thousands of characters have different attributes from documents consisting of dozens of characters. Similarly, it is difficult to exactly ascertain a quantifiable metric for grammar. The two most similar widely known metrics, Flesch Reading Ease is based on the total number of words and syllables, which is muddled with abbreviations and acronyms. Flesch–Kincaid Grade Level is based on average sentence length and average syllables per word, which is also affected by acronyms and subject to extreme variety and outliers when considering 1 (or less) sentence documents (Flesch, 1948).

These three metrics were selected specifically because of their importance to the existing NLP algorithms. Brevity affects the performance of many NLP measures such as Term Frequency (TF). It is certainly the most unique aspect of microtext, and is reflected in the selection of the term itself. The informal language creates the most difficulty for NLP. The semi-structured nature of microtext is a definite advantage to be leveraged in processing, and is fairly unique. Generally, longer texts such as websites, newswire articles, etc are not specifically attributable to a single author or a single time. Microtext guarantees both. Even if an article has a single author or a timestamp, that generally covers hundreds or thousands of words, so the granularity or pedigree of individual thoughts or statements is not nearly as fine-grained or accurate as that of microtext.

Finally, the specific selection of the term 'Microtext' seems appropriate. The text is not only short, but often abbreviated. Most importantly, use seems to be clear. Other than a euphemism for very small physical printed text, the only other academic use of the term was decades ago (Bullen, 1972) for describing a finite state machine. The way seems clear for microtext to become adopted without conflict.

2.2 Microtext Characterization

Encoding thoughts into an electronic format continues to get easier. At first, capture and encoding was reserved for higher priority items, such as books, contracts, etc. As the internet expanded in parallel with computers becoming more prevalent and less expensive, the barrier was lowered to include essays, newswire articles, etc. The barrier continues to be lowered, in at least three dimensions: cost required to encode, accessibility to encoded work, and knowledge required to operate encoding technology. So text representations of thought and speech are becoming more prevalent daily, and as the cost goes down, so does the return on investment, and the messages and thoughts encoded tend to become more brief and less formal. The analog equivalent of 'micro-text' has always been a part of our modern, communal society, in the form of conversations, journal entries, etc. It's just that when these expressions occurred in spoken dialog, telephone conversations, and paper notebooks, they are not able to be as easily captured, archived, sorted, or discussed. There are many

examples of this digitization in the hands of the general public, including wikis, micro-blogs, SMS messages, and even voicemail transcriptions (through ad-supported free consumer technologies such as Google Voice or Jott).

Although they are both encoded as characters, micro-text varies in structure and content from long form text, as discussed in the introduction. It is not necessarily the case that micro-text is 'noisier' than regular text. From a semantic perspective, the 'signal' in microtext is very strong, the difficulty comes from lack of context, not too much 'noise'.

For example, O'Connor, et al (2010) collected a corpus of Twitter messages in 2009, and found that the average message length was eleven words, and that words rarely occur more than once in a message. Therefore, some of the standard NLP metrics such as Term Frequency (TF) and Document Frequency(DF) will need to be reworked.

Although mentioned a few different places throughout the document, it is useful to consider a list of available public and commercial technologies, services, and standards that would be considered microtext.

- *SMS* (aka Text Messages)
- *Instant Messaging* (point to point messages such as XMPP/Google Talk/Jabber, OSCAR/AIM/ICQ, Microsoft Messenger)
- *Multi-User Chatrooms* (aka MUCs, including IRC chatrooms, and communication within MMORPG and other online communities such as Second Life or World of Warcraft)
- *Voicemail Transcriptions* (Enterprise or government level, as well as consumer level technologies such as Google Voice or Jott)
- *Microblogs* (Twitter, Google Buzz, Identi.ca, FriendFeed, and other closed sources such as in-house or enterprise level microblogs such as the United States Department of Defense's 'Chirp' service, or private services such as Facebook)

There are some other sources which may potentially fit the definition of microtext, but may not. Generally this is because of the length of the author's individual statements. This includes email, wikis, 'regular' weblogs, website 'forums', UseNet, and RSS feeds.

For illustration, here is a non comprehensive list of some sample types of meta data (and specific values of those types in parenthesis):

- *Source Attribution* (Author, Screen Name, Originating Phone Number or Email Address)
- *Timestamp* (Almost always with minute-level accuracy)
- *Audience* (Public, Room or Chat channel for IRC or MUCs, one or more specific recipients of the Source Attribution type)
- *URL References* (both as a reply/threading mechanism and as a pointer to a longer reference)
- *Geo-location information* (Either specifically GPS coordinates, or through location tags)
- *Other meta-data tags* (Self selected topic tags i.e. #hashtags, Author's Mood, weather, etc. These include both author created and automatically generated)

Note that each of these types can be satisfied with its own rules including 'zero or more'. For example, in a Twitter reply (characterized by starting with '@username'), the Audience is both public and a specific recipient.

3 SURVEY OF CURRENT MICROTEXT RESEARCH

By far the primary difficult in conducting a survey of recent research in microtext is that since there is no common vocabulary or terminology, locating all of the research is non-trivial.

One metric is the references to the word 'Twitter' in peer reviewed publications. In the last three full years, (since 2007), the number of Association for Computing Machinery (ACM) journal articles with the word 'Twitter' are 11, 84, and 263. Through the first half of 2010, 284 have been published, so it would be reasonable to expect at least 568 articles to be published in 2010. In the same time period, the number of Association of Computational Lingustics (ACL) articles are 0, 0, 4, and 34. (68 projected). The number of INSTICC papers mentioning Twitter available in the SciTePress digital portal is 2, both in 2010. Since Twitter is one specific commercial product, extrapolation can be dangerous, but it is interesting to note the rapid and dramatic uptake within the academic community. Interest is obviously high, and growing fast.

Although 'Twitter' is mentioned in many papers, the intent of the research is widely varied. Obviously the hundreds of papers already published are beyond cataloguing in this survey paper, however, a sampling is presented in the following paragraphs. Note that in almost every paper, it is a single application that is being considered, rather than

attempting to define or derive a higher purpose or truth of the medium. This is a very industry or engineering centric approach, rather than a scientific approach.

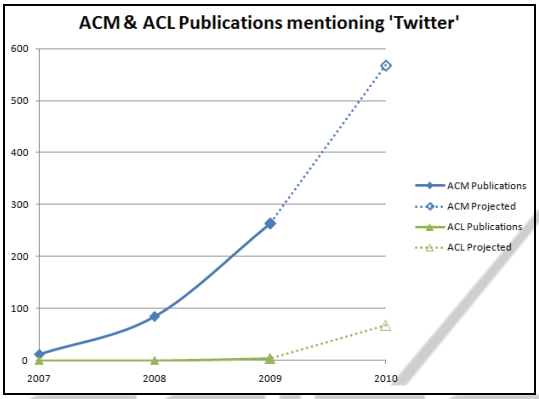

Figure 1: Twitter's growth in the academic community.

3.1 Topic Identification/Individual Summarization

There is some preliminary work (Adams, 2008) in topic detection within chat. Specifically looking at IRC chatrooms, these researchers illustrate some of the types of techniques that can be uniquely leveraged by microtext research, such as augmenting a typical TF-IDF based approach with temporal information.

Ranganath, Jurafsky, and McFarland (2009) were able to achieve 71.5% accuracy on a system designed to detect a speaker's intent to flirt using a spoken corpus of speed-dates. They also considered audible (prosodic) features, and their corpus was transcribed by humans and heavily annotated with extra information such as number of laughs, number of filled pauses, etc. The applicable part of this research where microtext is concerned is that their transcription/representation was very accurate as to the speech that actually occurred, including interruptions, pauses, laughter, backchannel utterances. (Examples include 'Uh-huh, Yeah, Wow, Excuse Me, Um, Uh). These types of attributes are not part of the formal written grammar that more traditional NLP approaches consider. Given that the results of this system were more accurate than human annotators, it is very notable and exciting to think that the informal grammar characteristic of microtext may be leveraged as an advantage over traditional free text.

Ritter, Cherry, and Dolan (2010) focus on modelling conversations using an unsupervised learning algorithm. In their collection of 1.3 million tweets, they note that Twitter postings tend to be "highly ungrammatical, and filled with spelling errors". They also note that 69% of the conversations in their data had a length of two. They find that a modification of the Latent Dirichlet Allocation overcomes the noisiness and brevity of their tweets that causes difficulty for named entity recognizers and noun-phrase chunkers. Although it was not the focus of their paper, it is precisely these types of discoveries that need to be re-used by the community.

3.2 Clustering/Mass Summarization

Examining the larger zeitgeist of the microtext sources, by considering them in aggregate, is an area with many commercial entities are pursuing, which are then commercial, proprietary, and closed, so there is no insight into their methods. One of the more academic approaches is TweetMotif (O'Connor, 2010). TweetMotif's website provides an elegant one sentence summary of the algorithm that "takes any word or phrase, finds tweets where people are talking about it, then groups them by statistically unlikely phrases that co-occur". This is a relatively standard NLP approach, and it would be interesting to compare results on Microtext vice longer text. Also, TweetMotif takes the important step of recognizing that many tweets are exact duplicates, ore essentially the same, and specifically "groups messages whose sets of trigrams have a pairwise Jaccard similarity exceeding 65%."

One drawback to clustering approaches in general for application in more serious matters, the current implementations do not seem comprehensive (nor do they claim to be). So their application would seem to be limited to more ephemeral uses, rather than rigorous or exhaustive. However, they do serve a useful purpose.

Another approach is taken by TWinner (Abrol, 2010) to attempt to cluster tweets by physical location, and then utilize this information to "improve the quality of web search and predicting whether the user is looking for news or not." Twitter is proposing an automated GPS tagging capability, and other microblogging services such as Google Buzz already support automatic or user specified location information, which will only improve the accuracy of algorithms such as TWinner. The TWinner paper also defines a 'Frequency-Population ratio), which is a ratio of the number of tweets per geographic location, normalizing with respect to population density.

The 'Phrase Reinforcement Algorithm', developed by Sharifi, et Al (2010), utilizes a different strategy. It provides a machine-generated summarization of 'trending topics' by examining a quantity of very similar updates, and normalizing them to produce a best possible summarization. This does not leverage the microtext aspect, but is leveraging the massive amount of human thought put into generating the content.

3.3 Classification

Phan (2008) proposes a "general framework for building classifiers that deal with short and sparse text & Web segments by making the most of hidden topics". The approach leverages a 'universal dataset' to augment the short and sparse text collected. This is a promising approach, and could be extended easily to include ontologies or language concepts and representation in the 'universal dataset.' So the bottleneck of this approach is essentially the same as the rest of the Natural Language community, the ability for the machine to understand human generated text.

Dela Rosa and Ellen (2009) have completed a series of experiments on classification of military chat posts. A number of different machine learning algorithms were evaluated, including SVMs, k-Nearest Neighbour, Rocchio, and Naive Bayes. Various feature selection methodologies were also considered, and Mutual Information (MI) and Information Gain (IG) were found to perform relatively poorly. K-NN and SVM were found to be the most suitable in a binary and four-way classification task.

3.4 Sentiment Analysis

Go and Bhayani (2010) perform sentiment analysis of Twitter messages. They are able to leverage emoticons as noisy labels, a technique first presented by Read (2005). Difficulties with less formal grammar constructs are also encountered. They attempt to perform clustering to assist with the analysis, and found that it unexpectedly hurt results.

Wilson, Wiebe, and Hoffmann (2005) examine contextual polarity (aka semantic orientation) of phrases in great detail. Their work attempts to deal with the paradox that in the English language, "Positive words are used in phrases expressing negative sentiments, or vice versa." One focus of the research is on feature selection, such as word features (e.g. what part of speech), the presence of nearby modifiers or negators, and other proximity features (e.g. whether the word is preceded by an adjective). The stated goal of this work is to provide insight into phrase-level sentiment analysis. Some microtext is not much more than a phrase in length, so this type of research is definitely applicable. A small question to be answered, however, is whether or not the informal grammar would interfere with the feature selection methods exploited in their work.

3.5 Question/Answer

Cong, et. Al (2008) attempt to leverage existing knowledge bases of questions and answers (i.e. website forums) to provide answers for new questions. While this is not specifically microtext related, it is interesting because of the implications. Social Search is a concept being explored by various companies and pundits (Google, Laporte, 2009); the idea is to focus search results to consider more highly authors that the searcher has a personal relationship with, under the guise that those recommendations or answers would be more appropriate, or authoritative. The majority of those 'social search' sources would be considered microtext, and therefore microtext extraction is crucial to these technologies succeeding.

3.6 Information Extraction

Marom and Zukerman (2009) study a corpus of paired question & response help desk emails with the intention of automating the process. The bulk of this research is focused on NLP tasks that are not applicable to microtext, such as meta-learning and semantic overlap. However, the study does investigate sentence level granularity for the purposes of generating hybrid or better tailored answers through combination. One thing specifically investigated is sentence cluster cohesion, a measure of the similarity of sentences to each other. This metric would be useful in microtext analysis because some microtext sources have an arbitrary character limit which forces the author to rapidly cycle between topics. Classifying the entire microtext 'document' will vary greatly depending on whether or not the individual sentences are cohesive.

Gruhl, et. al (2009) explore "statistical NLP techniques to improve named entity annotation in challenging Informal English domains". They achieve notably better results through application of SVMs. This paper illustrates the types of insight that can be gained through specific focus on microtext characteristics first and experimental validation

second. The majority of the research (both referenced in this survey, and otherwise reviewed and not referenced) centers on an experiment.

3.7 Semi-structured Data Exploitation

One of the most underutilized aspects of Microtext research is ignoring the semi-structured nature of the data. Kinsella, Passant, and Breslin (2010) examine the occurrences of hyperlinks in online message boards. They observe that not only is the use of hyperlinks increasing, but the hyperlinks themselves often reference "resources with associated structured data", and they discuss "the potential for using this data for enhanced analysis of online conversation".

Wang (2010) provides another example of utilizing the structure of the data in his research into identifying spammers on Twitter. He utilizes some of the relationship information available from twitter accounts to construct graphs and examine some typical directed graph features. Also, Wang makes the interesting choice of ignoring the NLP aspect of the tweets completely, and instead treating authors' contributions as strings of symbols, and compares them using Levenshtein distance, ignoring grammar and semantic content completely.

4 SPURIOUS MICROTEXT RESEARCH RESULTS

Not all papers that reference microtext sources are applicable to microtext characterization. For example, there are many instances where the microtext is utilized for some other purpose, such as using SMS to interface with other systems like FAQs (Kothari, 2009) or yellow pages (Kopparapu, 2007). Similarly, not all papers mentioning a microtext source are concerned with analyzing the content in any fashion. Mowbray (2010) publishes a paper on identifying spam in twitter, similar to the aforementioned paper by Wang, but unlike Wang focuses on automated use and abuse of the Twitter API and functionality, and other non-NLP, non-AI techniques.

There are also many interesting sociological applications and research to be performed on this type of data (which as stated earlier, used to be private, non-digital, or more expensive). There are dozens of papers on how to leverage these new sources of digital information, such as the influence of Twitter (Cha, 2010) (Lee, 2010), using Twitter to predict elections (Tumasjan, 2010), the stock market, or movie results, or the flu (Ritterman, 2009). While interesting and valuable in their own right, these papers do not provide insight into the mechanics of microtext, or leverage the characteristics that define microtext. While these works have an NLP aspect, it is really the publicness and ubiquity of the mechanisms that are being exploited, not the microtext.

Another example of this type of clever exploitation is Davidov, Tsur, and Rappoport (2010) who leverage emoticons in conjunction with user generated tags for sentiment analysis. Emoticons are by no means required by or limited to microtext sources, but they tend to appear more frequently. They examine the phenomenon that sometimes the user generated tags are overloaded and part of the grammatical/semantic content, such as "*I always enjoy the #Olympics*" and other times simply serve as metadata, for example "*I can't believe the USA just won the gold in hockey! #Olympics*". Twitter is leveraged as a large repository of sentiment, and "the obtained feature vectors are not heavily Twitter-specific". This is more of an exploration of the English language and the tagging and emoticon phenomenon than anything specifically about microtext, although the emoticon/sentiment analysis feature vectors could be leveraged as would any other ontology.

5 CONCLUSIONS

There are a growing number of papers being published on NLP and AI techniques as applied to brief, poorly formatted, semi-structured text. As presented in this survey, there are a number of interesting papers being published in the area. Much of the current work is more engineering than scientific in focus; they seek to provide anecdotal or experimental evidence about a single use case. So while not using a common terminology, these papers are providing the rough foundation for research on microtext.

There is some past NLP work on sentence and phrase level types of analysis that is partially relevant. Although the brevity condition is met, much of this work is relies on correct grammar and sentence structure, and to a lesser extent on a larger corpus. So, not all previous NLP work on concise expressions will translate to microtext.

It is the thesis of this paper that some discussion and meta-experimentation on the field itself would lead to greater insights, with a higher level of reuse. A first step in that direction is defining terminology,

'Microtext', so that researchers can have a common ground for future discussion.

Some of the scattered research surveyed in this paper has provided interesting insights as to the type of conclusions and methodologies that would be discovered and catalogued with a more focused effort. Some of these include: Leveraging an outside body of knowledge, leveraging non-traditional language features such as laughs and "uh/ums", and treating individual results as less important and focusing more on less granular trends. Overall, trend analysis and identification has the most research, and Information Extraction from microtext is particularly lacking.

In two different papers, SVMs were a successful strategy in dealing with informal grammars.

The next step is investigating and more rigorously quantifying the three attributes in the microtext definition. This would certainly provide reusable insights and help catalogue best performing techniques and unique quirks and advantages of microtext processing versus text processing. The goal of this paper is to create an understanding of microtext within the AI and NLP communities.

ACKNOWLEDGEMENTS

Thanks to the Office of Naval Research and the Space and Naval Warfare Systems Center Pacific for their financial support, and Dr. LorRaine Duffy for inspiration and motivation. This paper is the work of U.S. Government employees performed in the course of employment and no copyright subsists therein.

REFERENCES

Abrol, S. and Khan, L. 2010. TWinner: understanding news queries with geo-content using Twitter. In *Proceedings of the 6th Workshop on Geographic information Retrieval* (Zurich, Switzerland, February 18 - 19, 2010). GIR '10. ACM, New York, NY, 1-8

Adams, P., and Martell, C., 2008. Topic Detection and Extraction in Chat. In *International Conference on Semantic Computing,* IEEE.

Bullen, R.H. Jr., and Millen, J. K., 1972. Microtext: the design of a microprogrammed finite state search machine for full-text retrieval. In *Proceedings of the AFIPS Joint Computer Conferences*. ACM.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. Measuring user influence in twitter: the million follower fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, AAAI, Washington, D.C., 2010.

Chi, E. 2009 "Information Seeking Can Be Social," Computer, pp. 42-46, March, 2009. IEEE

Cong, G., et al. (2008). Finding question-answer pairs from online forums. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 467-474, New York, NY, USA. ACM.

Dalli, A., Xia, Y., and Wilks, Y., 2004. FASIL email summarisation system. In*Proceedings of the 20th international conference on Computational Linguistics* (COLING '04). ACL, Morristown, NJ, USA, Article 994.

Davidov, D., Tsur, O., Rappoport, A. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, In *Proceedings of the 23rd international conference on Computational Linguistics (COLING), 2010*.

Flesch, R. (1948); A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, pp. 221–233.

Go, A., Bhayani, R., and Huang, L. 2010. Exploiting the Unique Characteristics of Tweets for Sentiment Analysis. *CS224N Project Report*, Stanford.

Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., and Sheth, A. 2009. Context and Domain Knowledge Enhanced Entity Spotting in Informal Text. In *Proceedings of the 8th international Semantic Web Conference*. 260-276.

Kinsella, S., Passant, A., Breslin, J. 2010. Ten Years of Hyperlinks in Online Conversations. In *Proceedings of the Web Science Conference 2010.* WWW2010.

Lee, C., Kwak, H., Park, H., Moon, S., 2010. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM, New York, NY, USA, 1137-1138.

Laporte, Leo. 2009. [Internet Radio Broadcast] This Week in Google 13. October 24, 2009.

Kopparapu, S. K., Srivastava, A., and Pande, A. 2007. SMS based natural language interface to yellow pages directory. In *Proceedings of the 4th international Conference on Mobile Technology, Applications, and Systems and the 1st international Symposium on Computer H uman interaction in Mobile Technology* ACM. Mobility '07. ACM, New York, NY, 558-563.

Kothari, G., Negi, S., Faruquie, T. A., Chakaravarthy, V. T., and Subramaniam, L. V. 2009. SMS based interface for FAQ retrieval. In P*roceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th international Joint Conference on Natural Language Processing of the Afnlp*. Association for Computational Linguistics. Morristown, NJ, 852-860.

Marom, Y. and Zukerman, I. 2009. An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Computational Linguist.* 35, 4 (Dec. 2009), 597-635

Mowbray, M. 2010. The Twittering Machine. In *Proceedings of the 6th International Conference on*

Web Information Systems and Technologies (WEBIST 2010). INSTICC. 299-304.

O'Connor, B., Krieger, M., and Ahn, D. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media.* Washington, DC, May 2010

Phan, X.-H., et al. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pp. 91-100, New York, NY, USA. ACM.

Ranganath, R., Jurafsky, D., and McFarland, D. 2009. It's not you, it's me: detecting flirting and its misperception in speed-dates. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*. ACL.

Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop* ACL.

Ritter, A., Cherry, C. And Dolan, B. 2010 Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, Los Angeles, CA, 172-180.

Ritterman, J., Osborne, M., and Klein, E. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media - 13th Conference of the Spanish Association for Artificial Intelligence, 2009.* AEPIA (Asociación Española de Inteligencia Artificial).

Rosa, K. D. and Ellen, J. 2009. Text Classification Methodologies Applied to Micro-Text in Military Chat. In *Proceedings of the 2009 international Conference on Machine Learning and Applications* (December 13 - 15, 2009). ICMLA. IEEE Computer Society, Washington, DC, 710-71.

Sharifi, B., et al. (2010). Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 685-688, Los Angeles, CA. ACL.

Tumasjan, A., et al. 2010. Predicting Elections with Twitter: Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*, AAAI, Washington, D.C., 2010.

Wang, A. H. 2010. Don't follow me - Spam Detection in Twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010)*. INSTICC. 142-151.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, Morristown, NJ, 347-354.