Does Size Matter? When Small is Good Enough

A.L. Gentile*, A.E. Cano, A.-S. Dadzie, V. Lanfranchi, and N. Ireson

Department of Computer Science,
University of Sheffield,
Sheffield, United Kingdom
{a.l.gentile,a.cano,a.dadzie,v.lanfranchi,n.ireson}@dcs.shef.ac.uk

Abstract. This paper reports the observation of the influence of the size of documents on the accuracy of a defined text processing task. Our hypothesis is that based on a specific task (in this case, topic classification), results obtained using longer texts may be approximated by short texts, of micropost size, i.e., maximum length 140 characters. Using an email dataset as the main corpus, we generate several fixed-size corpora, consisting of truncated emails, from micropost size (140 characters), and successive multiples thereof, to the full size of each email. Our methodology consists of two steps: (1) corpus-driven topic extraction and (2) document topic classification. We build the topic representation model using the main corpus, through k-means clustering, with each k-derived topic represented as a weighted number of terms. We then perform document classification according to the k topics: first over the main corpus, then over each truncated corpus, and observe the variance in classification accuracy with document size. The results obtained show that the accuracy of topic classification for micropost-size texts is a suitable approximation of classification performed on longer texts.

Keywords: Short Messages; Email Processing; Text Processing; Document classification.

1 Introduction

The advent of social media and the widespread adoption of ubiquitous mobile devices has changed the way people communicate: fast, short messages and real time exchange are becoming the norm. This phenomenon was first manifested with the introduction of SMS (Short Messaging Service) capabilities on mobile phones. Despite the technical restrictions the size limit of 160 characters imposed, SMS was quickly adopted by users, thanks to ease of use and very short delivery time. The widespread adoption has had significant impact on the language used and the way people communicate; as pointed out by Grinter and Eldridge [8], users tend to adapt media to make themselves understood. In the case of SMS this meant modifying language to condense as much information as possible into 160 characters.

* to whom correspondence should be addressed

At the same time Instant messaging (IM) services such as MSN[1], Yahoo[2] and Jabber[3] rose in popularity, offering another platform with low barrier to entry and use, for real time, text-based *chatting* and communication. Newer social media services and applications such as Twitter[4] adopted this interaction paradigm (restricting even more, messages to 140 character chunks), evolved to support real-time communication within social networks. FourSquare[5], Facebook[6] and MySpace[7] posts, while using relatively longer feeds, also follow the general trend of using small chunks of text, i.e., microposts, to carry out (asynchronous) conversations.

While a large amount of this information exchange is social, micropost services are also used to exchange information in more formal (working) environments, especially as collaboration crosses wide geographical borders, bandwidth increases and the cost of electronic services decreases [10, 11]. Twitter, for instance, is currently one of the most widely used methods for exchanging up to date information about ongoing events, and topical discussion in professional and social circles [23, 25]. However, while the usage of Twitter and similar services in the workplace is increasing, it is sometimes perceived negatively, as they may be seen to reduce productivity [22], and/or pose threats to security and privacy.

The impact of text-based SMS and IM has however been such that where restrictions to use are in place, alternatives are sought that obtain the same benefits. Individuals in such environments often adopt the same communication patterns in alternative media, e.g., both desktop-based and mobile email usage often follow the same pattern. Further, empirical evidence suggests that even where IM and social media services are available, individuals may employ email as a short message service for communication via, e.g., mailing lists. This is often done in order to reach a wider audience that includes both the initiator's personal networks and other individuals with shared interests and who may be potential sources of expertise. Because mailing lists are in essence based on communities of practice (CoPs) with shared, specialised interests [20], both detailed and quick, short requests posted to mailing lists tend to receive quick replies from colleagues and more distantly related members of a network or CoP. Such email exchanges converge to a rapidly evolving conversation composed of short chunks of text.

The aim of this paper is twofold. We first consider a corpus of emails exchanged via an internal mailing list (over a period of six months), and perform statistical analysis to determine if email is indeed used as a short messaging service. Secondly, we analyse the content of emails as microposts, to evaluate to what degree the knowledge content of truncated or abbreviated messages can be compared to the complete message. Further, we wish to determine if the knowl-

[1] `http://explore.live.com/windows-live-messenger`
[2] `http://messenger.yahoo.com`
[3] `http://www.jabber.org`
[4] `http://twitter.com`
[5] `http://foursquare.com`
[6] `http://www.facebook.com`
[7] `http://www.myspace.com`

edge content of short emails may be used to obtain useful information about e.g., topics of interest or expertise within an organisation, as a basis for carrying out tasks such as expert finding or content-based social network analysis (SNA).

We continue the paper with a review of the state of the art in section 2. We then describe, in section 3, the corpus we employ, followed by our experimental methodology (section 4) and the results of the text classification experiments used to extract and compare the knowledge content of different size emails (sections 5.1 and 5.2). We conclude the paper in section 6, and discuss briefly the next stages of our research.

2 Related Work

Expertise identification and knowledge elicitation, key components of effectiveness and competitiveness in formal organisations, are often achieved via informal networks or CoPs [5, 20]. Email is a common tool for quick exchange of information between individuals and within groups, both on a social basis, but especially also in formal organisations, both for co-located and dispersed communication [6]. Email content, and addressee and recipient, often provide clues about the existence of CoPs and the interests and expertise of participants [2]. Quantitative data from email traffic (e.g. frequency of exchange) is useful in inferring social networks, and mining email content complements this by supporting the exploration and retrieval of organisational knowledge and expertise.

Exchange Frequency In the panorama of work on extracting social networks from email, the frequency of email exchange has been widely used as the main indicator of relevance of a connection. In some cases the effort is on determining frequency thresholds [24, 7, 1, 3], while in others time-dependent threshold conditions are defined to detect dynamic networks [4, 15]. Diesner et al. [6] construct a social network via weighted edges over a classical dataset, the *Enron* corpus[8], a large set of email messages made public during the legal investigation of the Enron corporation. They reported the emergence of communication subgroups with unusually high email exchange in the period prior to the company becoming insolvent in 2001, when email was a key tool for obtaining information especially across formal, inter-organisational boundaries. Diesner et al. [6] also observed that variations in patterns of email usage were influenced by knowledge about and reputation of, in addition to, formal roles within the organisation.

Content-Based Analysis Email content analysis has been used for different purposes: determining expertise [20], analysing the relations between content and people involved in email exchange [2, 12, 17, 26], or simply extracting useful information about names, addresses, phone numbers [16]. Schwartz et al. [20] derived expertise and common interests within communities from email exchange. While acknowledging the value of the results obtained, Schwartz et al. [20] note the risk to privacy in mining emails.

[8] `http://www.cs.cmu.edu/~enron`

Campbell et al. [2] exploit addressee and recipient information, in addition to information obtained from clusters of emails created through supervised and unsupervised keyword extraction, to create networks of expertise. McCallum et al. [17] recognise the contribution of Machine Learning (ML) and Natural Language Processing (NLP) to SNA, in order to retrieve the rich knowledge content of the information exchanged in such networks, and better interpret the attributes of nodes and the types of relationships between them. By running their experiments on the Enron email dataset and that of an employee in a research institution, [17] highlight a phenomenon that is becoming increasingly common – the blurring of the lines between inter-communication on purely professional and social levels. This underlines the importance of the analysis of the content of email documents in the derivation and verification of roles (a significant attribute of nodes) and relationships within communication networks, when used for expertise determination or topic extraction, for instance.

Keila et al. [12] investigate the use of domain-specific terms and the relationships between these and roles or activity in organisations, using the Enron email dataset. They conclude that e-mail structure and content is influenced by users' overall activity, e.g., when engaged in unusual activities. They, as do [6], who reported the emergence of communication sub-groups, observed alterations in patterns in email usage in the lead up to the failure of Enron, with similarity influenced by organisational roles. Zhou et al. [26] perform textual analysis of the Enron dataset to discover useful patterns for clustering in a social network. They found that individuals communicate more frequently with others who share similar value patterns than with those exhibiting different ones. They however could not draw definite conclusions about whether or not individuals who communicate more frequently with each other share similar value patterns.

Laclavík et al. [16] observe that enterprise users largely exploit emails to communicate, collaborate and carry out business tasks. They design a pattern-based approach to information extraction (IE) from and analysis of enterprise email communication, and exploit the data obtained to create social networks. The test sets (one in English containing 28 emails, and a second in Spanish with 50) consist of mainly formal emails exchanged between different enterprises. Their experimental design follows the classic IE approach: they automatically extract information such as names, telephone numbers and addresses from the email corpus, and compare results against a gold standard, the same email corpus, manually annotated. The results obtained indicate that emails are a valid means for obtaining information across formal, inter-organisational boundaries.

The work we present in this paper, on the other hand, makes use of a test set containing more informal email exchange in an internal mailing list for an academic research group, for a pilot, exploratory set of experiments. Rather than carrying out a classic IE evaluation task, we wish to determine if relatively short and informal texts can be used to aid the understanding of the content of the conversations carried out via email, and depict the variety of topics discussed using this communication medium.

Corpora The Enron corpus is a preferred test set in this field. The original corpus contains 619,446 messages belonging to 158 users, but [14], among others, suggest that cleaning is needed, for a number of reasons, including the fact that some of the folders are computer-generated (such as "discussion threads" and "notes inbox"), others contain duplicate email messages (such as the folder "all document"), and yet others constitute delivery failures and repeated attempts to deliver the same message.

Depending on the task being performed, accurate cleaning is required to avoid misleading results; [6, 12, 17] all perform cleaning and merging of data to increase the accuracy and reliability of the results of analysis. While the Enron corpus is valued as a widely available test set that aids replication of experiments in the field, we do not use it at this stage in our research. The main reason for this is that our experiments currently examine the usage of email as a tool for sharing information within a fixed community, as an alternative to social publishing services, and explore phenomena observed in such environments. The internal mailing list we use as a starting test set meets this requirement. A statistical analysis of our corpus is provided in section 3.

3 Email Corpus

The corpus used for analysis and knowledge content extraction is an internal mailing list of the OAK Group[9] in the Computer Science Department of the University of Sheffield. The mailing list is used for quick exchange of information within the group on both professional and social topics.

We use all emails sent to the mailing list in the six month period from July 2010 to January 2011, totalling 659 emails. For each we extracted the email body: the average length of which is 351 characters (just shorter than 2.5 microposts), with a standard deviation of 577 characters. We refer to this corpus as *mainCorpus*. Detailed statistics on document length are shown in Fig. 1. The percentage of messages of micropost size (up to 140 characters) constitutes more than 35% of the whole corpus. Considering emails up to two micropost sizes increases the percentage to ∼65%. Very few emails (around 4%) are really long (above 1000 characters).

These statistics indicate that the corpus largely consists of *micro-emails* – which we define as short email messages exchanged in rapid succession about a topic. We carried out a number of experiments on this corpus, to understand the knowledge content of the (micro-)emails. Future work will consider how this corpus varies from other email corpora of the same type (mailing lists) and what generic assumptions could be made about the existence and use of micro-emails.

4 Dynamic Topic Classification Of Short Texts

One of our main goals is to evaluate to what degree the knowledge content of a shorter message can be compared to that of a full message. Our hypothesis

[9] `http://oak.dcs.shef.ac.uk`

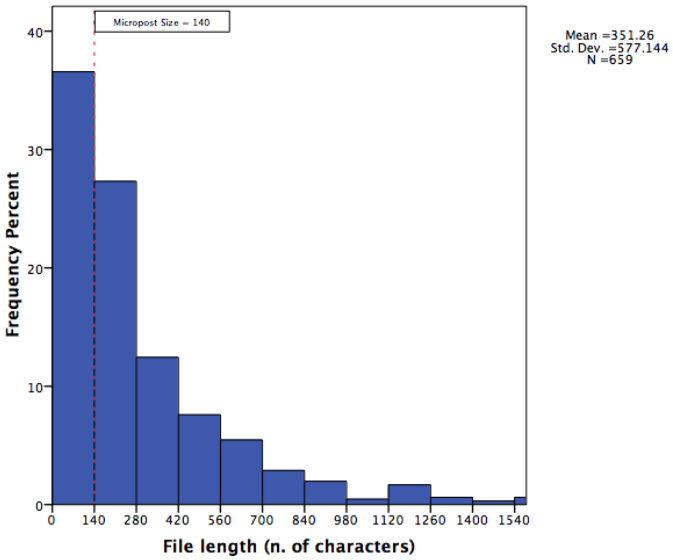

Fig. 1. Email length distribution

is that based on a specific task, results obtained using short texts of micropost size approximate results obtainable with longer texts. The task chosen for the evaluation is text classification on non-predefined topics. The test bed is generated by preprocessing the email corpus (see section 3) to obtain several fixed-size corpora, as detailed in section 5.1. The overall method consists of two steps:

corpus-driven topic extraction: a number of topics are automatically extracted from a document collection; each topic is represented as a weighted vector of terms;

document topic classification: each document is labelled with the topic it is most similar to, and classified into the corresponding cluster.

4.1 Topic Extraction: Proximity-based Clustering

Given a document corpus D, we represent it using a vector space model; each document in the corpus $d = \{t_1, ..., t_v\}$ is represented as a vector of weighted terms (using tf-idf weights). Using an inverted index of the documents we generate clusters of terms. Each cluster in $C = \{C_1, ..., C_k\}$ is represented as a weighted vector of terms $C_k = \{t_1, ..., t_n\}$, selecting n terms t for each cluster with highest tf-idf weight. Each cluster ideally represents a topic in the document collection. To obtain the clusters we apply a K-Means algorithm [9], using as feature space the generated inverted index of document terms (i.e., for each term we define which document contains it). Starting with k random means (centroids), each vector is assigned to the nearest centroid. By minimising the euclidean distance between each point and the centroids the process is repeated until convergence is reached.

4.2 Email Topic Classification

We then use cosine similarity to determine similarity between documents and clusters; we will explore further, in the next stage of our research, alternative similarity functions and their impact on the results obtained. For each document we calculate the similarity $sim(d, C_i)$ with each cluster. The labelling process $labelDoc : D \rightarrow C$ consists of mapping each document d to the topic C_i, which maximises the similarity $sim(d, C_i)$. The complete procedure is shown in Fig. 2.

labelDoc procedure

Input: Collection of documents $\{d_1, ..., d_{|D|}\}$, set of clusters $C = \{C_1, ..., C_k\}$, term representation for each cluster $C_k = \{t_1, ..., t_n\}$

Step 0: Obtain a document d_i's feature vector, of tf-idf weighted terms.

Step 1: Apply cosine similarity between a document d_i's feature vector and the k clusters and generates a vector of similarities $S_i = \{S_{i0}, ..S_{ik}\}$, $S_{ij} = sim(d_i, C_j)$.

Step 2: Label d_i with the highest weighted cluster in S_i.

Output: All classified documents.

Fig. 2. *labelDoc*: Topic classification procedure

5 Experiments

5.1 Dataset Preparation

In this experiment we artificially generate comparable corpora starting from the mailing list described in Section 3. The notions of *comparable corpora* and the strongly related alternative, *parallel corpora*, are very common in multi-language IE. Parallel text corpora contain the same documents with different content representation. An example is parallel language resources [19, 27], where a corpus consists of a set of documents, each of which is an exact translation of the original document in a different language. Comparable corpora [21, 13], however, do not contain document-level or sentence-level alignment across corpora, but talk about the same important facts. An example of comparable corpora for different languages is the multi-lingual Wikipedia [18], where the same articles are available in different languages, but which are not necessarily literal translations of each other, e.g., an article may be richer in one language than in another.

We produce comparable corpora using the following process: starting form *mainCorpus* we generate different corpora, each containing documents of fixed maximum length, by chunking the email body in multiples of 140 characters. We generated 8 comparable corpora, as shown in Table 1.

Table 1. Automatically generated comparable corpora.

Corpus Name	Maximum text length of each document
*corpus*140	email body truncated at length 140 if longer than 140 characters, full text otherwise
*corpus*280	email body truncated at length 280 if longer than 280 characters, full text otherwise
*corpus*420	email body truncated at length 420 if longer than 420 characters, full text otherwise
*corpus*560	email body truncated at length 560 if longer than 560 characters, full text otherwise
*corpus*700	email body truncated at length 700 if longer than 700 characters, full text otherwise
*corpus*840	email body truncated at length 840 if longer than 840 characters, full text otherwise
*corpus*980	email body truncated at length 980 if longer than 980 characters, full text otherwise
mainCorpus	full email body

5.2 Experimental Approach

As described in section 4, the set of topics for categorising the initial documents is not predefined, but corpus-driven. We use the *mainCorpus* for topic extraction. Since the k in the k-means clustering approach must be approximated, we repeated the clustering process several times. We applied the procedure *labelDoc* over the *mainCorpus*, varying each time the input clusters, from 3 to 15. The cardinality of clusters providing the widest distribution of classified documents on *mainCorpus* was 10; we therefore selected this as the optimal number of clusters for the final experiment on document classification. The main keywords in each cluster are shown in Fig. 3.

Using the 10 clusters obtained from the main corpus, we apply the *labelDoc* procedure to the different comparable corpora, including *mainCorpus*. Results obtained for the classification of *mainCorpus* are considered as the gold standard, and used for comparing results of all the other corpora.

5.3 Results and Discussion

We evaluate the performance of the topic classification using standard Precision (P), Recall (R) and F-Measure (F). Given the number of classes for classification (10) we calculate P, R, and F by micro-averaging results on the classification confusion matrix. Results for all text size corpora are shown in Table 2.

As expected, it is recall rather than precision with a bigger decrease as text length is reduced. If we relax the limitation of 140 characters and consider the next size corpus (280) the drop in performance is much lower.

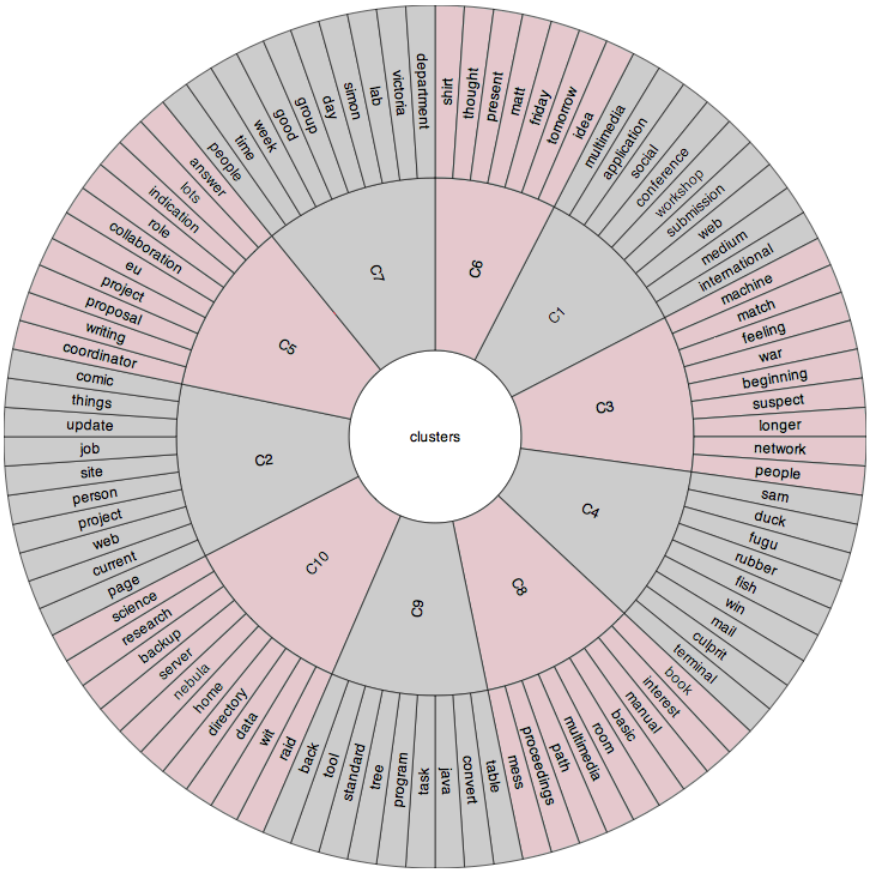

Fig. 3. Visualisation of topic clusters.

Table 2. Precision, Recall and F-measure values for topic classification for each corpus

	Precision	Recall	F-Measure
*corpus*140	0.86	0.66	0.74
*corpus*280	0.93	0.88	0.90
*corpus*420	0.95	0.94	0.95
*corpus*560	0.98	0.97	0.97
*corpus*700	0.99	0.98	0.99
*corpus*840	0.99	0.99	0.99
*corpus*980	0.99	0.99	0.99

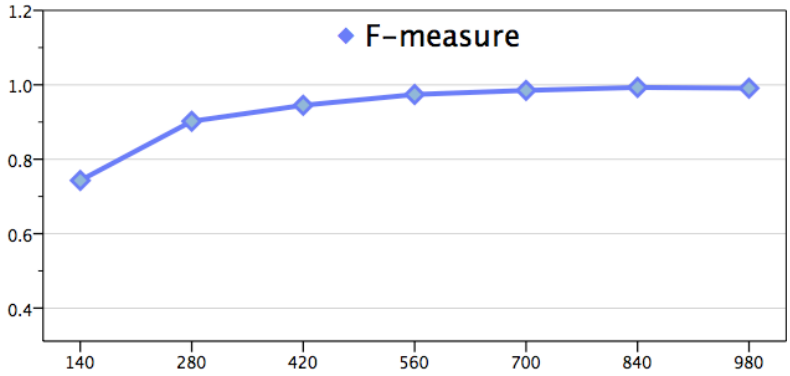

Fig. 4. F-Measure trend on different text size corpora

Considering the results obtained for *mainCorpus* as the upper boundary for classification, the trend of F-Measure over the different size corpora in Fig. 4 shows the impact of using shorter texts for topic classification. As expected, the trend increases monotonically with the size of texts, which means that reducing the text size directly affects the classification performance. What is interesting is that the performance is not significantly affected by reduction in text size.

At one micropost size there is a drop in F-Measure, to 74%. However with an increase to only two micropost sizes this improves significantly, to 90%. The larger drop at one micropost size may be explained by the method of truncation we use; among others, where a greeting exists this takes up a fair portion of the first micropost block. We are currently exploring the use of a sliding window to determine how best to chunk the e-mail content and identify the most salient region(s) of each, as a way of improving recall.

6 Conclusions

We have presented in this paper exploratory work on the usage of email as a substitute for online social publishing services. We explore how this kind of data may be exploited for the knowledge discovery process and how document size influences the accuracy of a defined text processing task. Our results show:

1. that a fair portion of the emails exchanged, for the corpus generated from a mailing list, are very short, with more than 35% falling within the single micropost size, and ∼65% up to two microposts;
2. for the text classification task described, that the accuracy of classification for micropost size texts is an acceptable approximation of classification performed on longer texts, with a decrease of only ∼ 5% for up to the second micropost block within a long e-mail.

These results are indicative of the convenience in communication using microposts in different environments and for different purposes. Because the research at this stage is still exploratory we refrain from generalising to other datasets.

However, our test corpus, which contains emails talking about both formal work and social activities, is not atypical in the workplace (see, for instance, [17]). We therefore believe that this work does provide a starting point from which to carry out more extensive analysis, using other standard email corpora such as the Enron corpus, in addition to other enterprise mailing lists similar to the corpus we analyse in this paper. This will allow us to explore what generic assumptions could be made on the creation and use of micro-emails.

A second hypothesis we wish to examine is whether enriching the micro-emails with semantic information (e.g., concepts extracted from domain and standard ontologies) would improve the results obtained using unannotated text. We also plan to investigate the influence of other similarity measures.

One area we wish to explore more fully is the application to expert finding tasks, exploiting dynamic topic extraction as a means to determine authors' and recipients' areas of expertise. For this purpose a formal evaluation of topic validity will be required, including the human (expert) annotator in the loop.

Acknowledgements A.L Gentile and V. Lanfranchi are funded by the Siloet project. A.E. Cano is funded by CONACyT, grant 175203. A.-S. Dadzie is funded by SmartProducts (EC FP7-231204). A.L. Gentile, A.-S. Dadzie, V. Lanfranchi and N. Ireson are also funded by WeKnowIt (EC FP7-215453).

References

1. L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187 – 203, 2005.
2. C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: 12th international conference on Information and knowledge management*, pages 528–531, 2003.
3. M. D. Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts. Inferring relevant social networks from interpersonal communication. In M. Rappa et al., editors, *Proc., 19th International Conference on World Wide Web*, pages 301–310, 2010.
4. C. Cortes, D. Pregibon, and C. Volinsky. Computational methods for dynamic graphs. *Journal Of Computational And Graphical Statistics*, 12:950–970, 2003.
5. A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *CEAS 2004: Proc., 1st Conference on Email and Anti-Spam*, 2004.
6. J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the Enron email corpus "It's Always About the People. Enron is no Different". *Compututational & Mathematical Organanization Theory*, 11(3):201–228, 2005.
7. J. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.
8. R. E. Grinter and M. A. Eldridge. y do tngrs luv 2 txt msg? In *Proc., 7th European Conference on Computer Supported Cooperative Work*, pages 219–238, 2001.
9. J. Hartigan. *Clustering Algorithms*. John Wiley and Sons, New York, 1975.
10. J. D. Herbsleb, D. L. Atkins, D. G. Boyer, M. Handel, and T. A. Finholt. Introducing instant messaging and chat in the workplace. In *Proc., SIGCHI conference on Human factors in computing systems*, pages 171–178, 2002.

11. E. Isaacs, A. Walendowski, S. Whittaker, D. J. Schiano, and C. Kamm. The character, functions, and styles of instant messaging in the workplace. In *Proc., ACM conference on Computer supported cooperative work*, pages 11–20, 2002.
12. P. S. Keila and D. B. Skillicorn. Structure in the Enron email dataset. *Computational & Mathematical Organization Theory*, 11:183–199, 2005.
13. A. Klementiev and D. Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *Proc., main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 82–88, 2006.
14. B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In J.-F. Boulicaut et al., editors, *ECML 2004: Proc., 15th European Conference on Machine Learning*, pages 217–226, 2004.
15. G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
16. M. Laclavik, S. Dlugolinsky, M. Seleng, M. Kvassay, E. Gatial, Z. Balogh, and L. Hluchy. Email analysis and information extraction for enterprise benefit. *Computing and Informatics, Special Issue on Business Collaboration Support for micro, small, and medium-sized Enterprises*, 30(1):57–87, 2011.
17. A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.
18. A. E. Richman, P. Schone, and F. G. G. Meade. Mining wiki resources for multilingual named entity recognition. *Computational Linguistics*, pages 1–9, 2008.
19. E. Riloff, C. Schafer, and D. Yarowsky. Inducing information extraction systems for new languages via cross-language projection. In *COLING '02: Proc., 19th international conference on Computational linguistics*, pages 1–7, 2002.
20. M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.
21. R. Sproat, T. Tao, and C. Zhai. Named entity transliteration with comparable corpora. In *Proc., 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 73–80, 2006.
22. TNS US Group. Social media exploding: More than 40% use online social networks. `http://www.tns-us.com/news/social_media_exploding_more_than.php`, 2009.
23. T. Turner, P. Qvarfordt, J. T. Biehl, G. Golovchinsky, and M. Back. Exploring the workplace communication ecology. In *CHI '10: Proc., 28th international conference on Human factors in computing systems*, pages 841–850, 2010.
24. J. Tyler, D. Wilkinson, and B. Huberman. E-Mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005.
25. J. Zhang, Y. Qu, J. Cody, and Y. Wu. A case study of micro-blogging in the enterprise: use, value, and related issues. In *CHI '10: Proc., 28th international conference on Human factors in computing systems*, pages 123–132, 2010.
26. Y. Zhou, K. R. Fleischmann, and W. A. Wallace. Automatic text analysis of values in the Enron email dataset: Clustering a social network using the value patterns of actors. In *HICSS 2010: Proc., 43rd Annual Hawaii International Conference on System Sciences*, pages 1–10, 2010.
27. I. Zitouni and R. Florian. Cross-language information propagation for arabic mention detection. *ACM Transactions on Asian Language Information Processing*, 8:17:1–17:21, 2009.