Using Stylistic Features for Social Power Modeling

Rachel Cotterill

University of Sheffield, UK

UKcontact@rachelcotterill.com

Abstract. Social Network Analysis traditionally examines the graph of a communications network to identify key individuals based on the pattern of their interactions, but there is a limit to the level of detail which can be inferred from metadata alone. Message content is a richer source of data, and can provide an indication of the relationship between a pair of communicants. An individual's language use will vary depending on their relationship to the addressee, and this paper investigates a set of stylistic features which may be used to predict the nature of a relationship within an organizational hierarchy. Experiments are conducted on the Enron corpus for the sake of comparison with earlier results, and demonstrate successful classification of upspeak vs. downspeak using a small feature set.

Keywords. Social network analysis, social power modeling, stylistics, text mining.

El uso de características estilísticas para modelado del poder social

Resumen. El análisis de redes sociales examina tradicionalmente el grafo de una red de comunicaciones, con el fin de identificar personas clave basándose en el patrón de sus interacciones, pero existe un límite respecto al nivel de detalle que se puede inferir únicamente a partir de metadatos. El contenido de mensajes es una fuente más rica de datos y puede proporcionar la indicación de una relación entre un par de comunicantes. El uso de idioma en personas varía dependiendo de sus relaciones con los destinatarios, entonces este trabajo investiga un conjunto de las características estilísticas que pueden ser utilizados para predecir la naturaleza de una relación dentro de la jerarquía de una organización. Los experimentos se realizaron sobre el corpus Enron para comparar los resultados obtenidos con los anteriores, y mostraron la clasificación exitosa de mensajes dirigidos a personas en la posición más alta en la jerarquía (upspeak) vs mensajes dirigidos hacia abajo en la jerarquía (downspeak) utilizando un pequeño conjunto de características.

Palabras clave. Análisis de redes sociales, modelado del poder social, estilística, minería de texto.

1 Introduction

Communication does not take place in a vacuum. When an individual sets out to communicate, they must determine not only the information they need to convey but also the manner in which they wish to convey it. Such choices may not always be made at the conscious level, but it is a fundamental tenet of Gricean politeness theory that, in order to be polite, it is necessary to use an appropriate manner [6].

Brown and Levinson [2] expand on this idea with their own theory of politeness, in which polite strategies are employed to reduce the 'face threat' associated with communication. The level of this threat, and consequently the level of politeness required to mitigate it, depends on the interaction of three major contextual variables:

1. the degree of imposition of the message itself;
2. the symmetric relation: the social distance between the participants; and
3. the asymmetric relation: the power balance in the relationship, derived from the relative status of the participants.

These latter two factors, taken together, define the relationship between the interlocutors at a very basic level: how well do they know one another, and where does the power lie? Since politeness varies according to these aspects of a relationship it follows that, given an appropriate method of measuring politeness, it should be possible to infer something about a relationship from the levels of politeness employed by the individuals concerned.

This paper describes an attempt to address this goal, using features of linguistic style. Previous work is summarized in section 2. Section 3 describes our corpus, and section 4 outlines the set of features used. Section 5 gives details of the classification experiments, and section 6 goes into more detail on feature analysis, using a variety of statistical techniques to assess the value of individual features. The paper concludes with a few remarks on future directions.

2 Previous Work on Modeling Relationships

Social network analysis is a branch of data mining concerned with identifying 'important' individuals in a graph of communication events. These techniques make use of a variety of metadata features, from straightforward graph metrics such as the number of messages sent and received by an individual, or their centrality in the graph, through to temporal features, such as the response time between a message and any reply it engenders [13]. Other metadata-based studies have looked for 'roles' in the graph, by identifying groups who exhibit similar patterns of interaction, but without attempting to (automatically) assign any interpretation of these roles [4].

More recent work has begun to examine content features as indicative of social status. The term 'Social Power Modeling' was coined by Bramsen *et al*. [1] to describe the task of identifying relative social status based on language use. They trained classifiers on the Enron corpus, to differentiate 'UpSpeak' (messages addressed up the hierarchy) from 'DownSpeak' (messages going down the hierarchy). They experimented with a variety of n-gram models, using unigrams, word bigrams, and part-of-speech bigrams. They also tried a variety of different classifiers, obtaining best results using SVMs. Training data was partitioned by author to avoid picking up on features due to idiolect; without this partitioning, they found the results to be artificially inflated. Their best result was obtained by using n-grams binned into sets, with information gain measures used to filter out those sets which did not contribute much to the overall classification. This achieved an F- measure of 0.781 using a weighted test set. Without partitioning, their F-measure for 10-fold cross validation was 0.830.

Peterson *et al.* [12] use (in) formality features to show that, in accordance with Brown and Levinson's predictions, messages going up the hierarchy are likely to be more formal than those going downwards. For their measure of formality, they use a combination of informal wordlists, punctuation features, and case features to identify informal language. Using this analysis they also demonstrated that 'business' messages are more formal than 'personal' ones, pairs with lower social distance (measured by number of messages exchanged) are more informal, and messages including requests are more formal. These findings are all in line with politeness theory, and although Peterson *et al.* did not perform a classification task, their work supports the concept.

In a study by Duthler [3], students were asked to make a request of either low or high imposition, using either email or voicemail. The results showed that, while voicemail messages were almost equally polite for high- and low-imposition requests, email exhibits more variation. This effect may be due to having more time to think about (and edit) linguistic choices when composing an email, compared to leaving a voicemail. That email exhibits such variation makes it ideal for examining politeness in relationships.

3 The Enron Dataset

The Enron email corpus is a standard dataset for communications research, and is used in this paper for ease of comparison with earlier results such as Bramsen *et al.* [1].

We used the CMU version of the corpus [9], which has undergone some work to remove duplicate messages. The entire corpus contains around 200,000 message files from the mailboxes of 158 Enron employees. As ground truth for the organizational hierarchy, we used the organizational rank information made available by Peterson *et al.* [12], which is to our knowledge the only published hierarchy. Their categorization

ranks employees from 0 (general staff) to 4 (CEO), and gives a rank corresponding to 161 individuals (using 200 email addresses).

From the Enron corpus, we selected messages which met the following criteria:

– the message has a single recipient; and
– both sender and recipient are of known rank.

Peterson *et al.* constructed their dataset using the same criteria, resulting in a corpus of 3999 messages between two individuals of known rank. We used quoted message text to reconstruct missing messages, wherever a message was quoted in full with headers, and then removed duplicates based on their content. By this method we obtained a corpus size of 11,548 messages, an increase of 289%. Of these messages, 4812 (41.7%) are between individuals on the same level of the hierarchy, 3553 (30.8%) are addressed to someone of a higher rank, and 3183 (27.7%) are addressed to someone of lower rank.

We divided the corpus randomly into ten parts, in order to conduct ten-fold cross validation using the same split of the data on each run of the experiment (for example, testing out different combinations of features). We also created versions of the corpus partitioned by sender and by sender-recipient pairs, in order to test the difference made by having prior knowledge about an individual's personal style when building the models.

4 Initial Feature Selection

Features were selected using a combination of practical considerations and theoretical motivation. Bramsen *et al.* [1] demonstrated that this task can be addressed using n-gram models, but such models are highly language- specific, require a large amount of training data to generate, and the resulting models have a large number of poorly-explained features. By contrast, the present work uses a significantly smaller feature set, and the features selected are either language-independent, or depend only on the availability of general linguistic resources (such as word lists and part-of-speech taggers).

Many politeness strategies result in longer messages, with more complex sentences. Characters per word, words per sentence, and sentences per paragraph were calculated to capture the length and complexity of the message. At a character level, the percentage of letters, whitespace, numerals, and symbols were calculated. Within the category of letters, the proportions of uppercase glyphs and Latin-codeblock characters were calculated using the Unicode [5] class definitions. Case variation is predicted to be a telling feature as uppercase can be used for emphasis in informal text, or lazy writers may not use any capitals, whereas in formal English text the expectation would be normal 'sentence case' where only sentence beginnings and proper nouns are capitalized. Non-Latin script may indicate codeswitching, but this is not expected to feature strongly in the Enron dataset.

In addition to the proportion of symbols, more detailed punctuation features were also used, by calculating the percentage of each class of punctuation. Polite messages are more likely to be punctuated according to the grammatical norms of the language in question. Depending on the individual's style, informality may be expressed with more informal punctuation characters (e.g. strings of '!'s), or fewer punctuation characters if commas and periods are omitted.

Emoticons were identified by a series of regular expressions, looking for both western (e.g. :) :D =)) and eastern (e.g. o_0 >_>) styles. Emoticons are essentially paralinguistic features, expressed in text. They would not be used in formal documents, and so the use of emoticons indicates a certain level of informality. Future work may wish to consider the range of 'expressions' used, for example distinguishing smiley faces from frowns, but emoticons are rare in the Enron dataset so this is of limited value.

Codeswitching, slang, and typographical errors (typos) are all more prevalent in informal language. These categories were combined into a single class of 'non-dictionary' words, by using the MySpell English dictionaries to identify misspelt and out-of-vocabulary words. Future work could consider splitting these out into separate features. Both US and UK dictionaries were used, since Enron conducted international business with offices on both sides of the Atlantic.

Regular expressions were used to find alphanumeric words such as ‘l8r’ and ‘b4’. These are also indicative of informality, as they do not form part of standard language, and are only used in informal contexts. Words which are emphasised by affective lengthening (such as ‘reallllly’ or ‘coooool’) were also identified by use of regular expressions and the MySpell dictionaries.

Word lists were constructed for English, to enable identification of important classes of word. Five word lists were used: expletives, deictic expressions, politeness markers, hedges, and contractions. Expletives are indicative of informal language, as are contractions (although less strongly so). Deictic expressions require common knowledge for the interpretation of meaning, and as such, more deixis implies a closer relationship. Explicit politeness markers (such as ‘please’ and ‘thanks’) and less obvious politeness strategies such as hedging are used to mitigate face threat. This may be indicative of social distance, power imbalance, or both.

Heylighen and Dewaele [7] define formality in terms of the amount of contextual information included in a text. In brief: the more formal the document, the more detail is included explicitly, and the less background knowledge is taken for granted. They proceed to define a formality score based around parts of speech, and demonstrate that this varies between genres of text and speech; this F-score was adopted without modification as one feature for our model. Individual part-of-speech features were also examined for comparison, as the F-score has not been optimized for social relationship modeling.

The proportion of sentences which are questions, and the proportion of these which are tag questions (such as ‘isn’t it?’ ‘don’t you?’) were also calculated. In previous work Lakoff [10] has shown that women use more tag questions than men, which is viewed as an example of women using more polite language.

It should be noted that many of these features are not independent. To give one trivial example, the percentage of question marks is likely to be correlated with the percentage of questions in the data set but there are likely to be other, more subtle relationships between features.

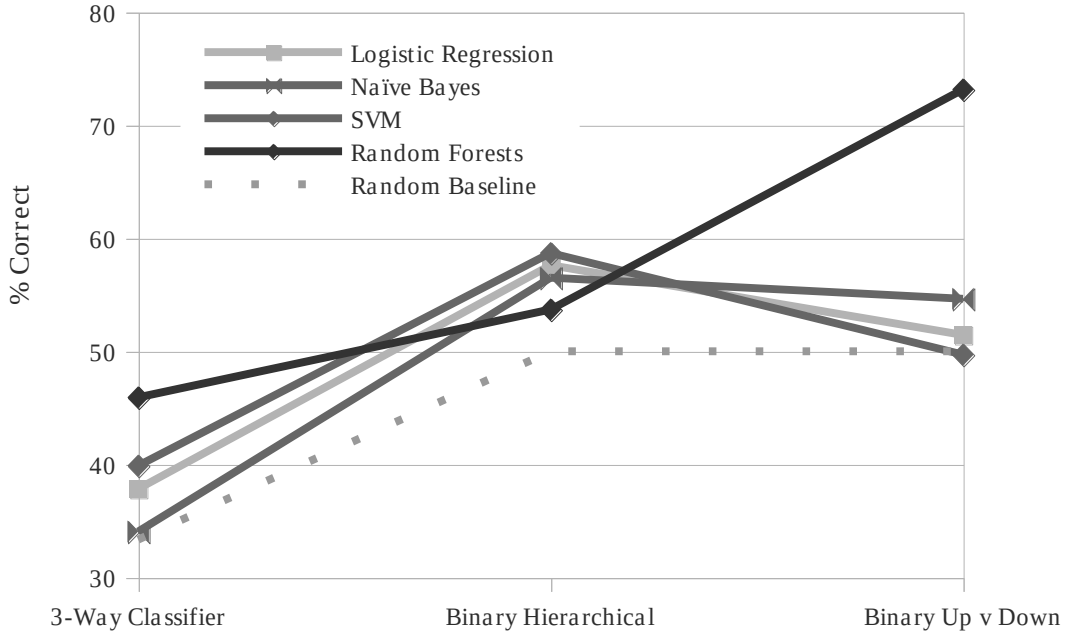

Fig. 1. Comparing classifiers

5 Classification

We have three classes in our data: 'up' for messages sent to someone higher in the hierarchy, 'down' for messages sent to someone lower in the hierarchy, and 'level' for messages sent to peers of the same rank.

We conducted two sets of experiments. The first addressed the three-class problem directly, by constructing a three-way classifier. The second approach was to break the problem down into two sequential stages: first, to determine whether a relationship is level or hierarchical, and second (in the case of a hierarchical relationship), to differentiate upwards from downwards communication. The up-down classifier has the advantage of being directly comparable to the results of Bramsen ***et al.***'s n-gram technique.

This results in three classification problems, one three-class and two binary. Each of these has its own natural baselines. For the three-way classifier, the random baseline is 33.3%, and the most common class baseline is 41.7% (the 'level' class). When considering hierarchical versus level communications, the random baseline is 50% and the most common class baseline is 58.3% (the 'hierarchical' class, which is the union of the 'up' and 'down' sets). For the up versus down classifier, the random baseline is still 50% but the most common class baseline is 52.6% (the 'up' class).

5.1 Comparing Classifiers

We compared four different statistical classifiers within the WEKA [16] framework: logistic regression, naïve Bayes, support vector machine (SVM), and J48 random forests. The results of this comparison can be seen in fig. 1, which plots the accuracy of each type of classifier, for each of our three classification tasks. The graph shows the case where data is partitioned by pairs, and normalized (see section 5.2). Similar results were found in the other cases, but space prohibits including all the examples in this paper.

For two of the three classifiers, random forests outperformed the other approaches. In the third case, that of the binary hierarchy classifier, the random forest has the worst performance, but none of the classifiers performs far above the baseline. For the remainder of this paper, we will continue to use random forests, as they also have the advantage of producing human-readable, meaningful models.

5.2 Effects of Normalizing and Partitioning Data

Different people have their own preferred styles of language use, and in many cases these differences in idiolect are likely to outweigh the variation due to social power relationships. To address this, we normalized each score relative to the mean and standard deviation of all communications originating with that particular sender, in the following manner:

$$x_i^A \longrightarrow \frac{x_i^A - \mu_i^A}{\sigma_i^A}, \qquad i \in \{1, \ldots, 45\}$$

where x_i^A represents a particular instance of person A using feature i, μ_i^A is the mean value of feature i across all communications originated by person A, and σ_i^A is the corresponding standard deviation across all of A's communications. This manner of normalization adjusts the feature scores for each sender to a mean of zero and standard deviation of one.

This transformation fails when σ_i^A is zero, in which case we set $x_i^A = 0$. The primary cause of this situation is when a particular individual simply does not use a feature; for example, there are numerous senders who never use expletives or emoticons.

As discussed in section 3, we constructed training and testing partitions of the dataset in three different ways:

1. completely at random;
2. ensuring no sender-recipient pair was represented in both training and testing data; and
3. ensuring no sender was represented in both training and testing data.

For each method of partitioning the data, we conducted ten-fold cross validation using both normalized and unnormalized data, in order to assess the value of normalization and the impact of partitioning.

As Figure 2 demonstrates, the results are significantly impacted by the choice of partitioning: in particular, the un-partitioned data gives unrealistically optimistic results. This is a natural consequence of messages from some senders to some recipients being available in both the train and test sets: essentially equating to these relationships being trained on themselves.

The difference between pairwise and sender partitioning is subtler, in both magnitude and implications. The question here is whether we wish to allow ourselves to train for a sender's particular style when communicating with known recipients, and use this to classify her other relationships with unknown parties. It is easy to imagine circumstances where this might be the best thing to do, for example if we knew about some of a sender's contacts and wished to make inferences about others. In other circumstances, however, we may be more interested in creating a "universal" classifier which can be used on completely unknown senders: a much harder problem on which we only narrowly beat the baseline.

Normalization also has a significant impact on accuracy, particularly in the case of upwards versus downwards communication, where the difference is almost twenty percentage points in the case of pairwise partitioning. Using sender-partitioned data, however, the effect of normalization is much lower (indeed, for the binary hierarchical classifier, it actually degrades performance in the sender-partitioned case).

Using normalized data, we obtain better results than Bramsen *et al.* for the non-partitioned case, but experience a steeper drop in results when the data is partitioned by sender. This may be interpreted as showing that stylistic features are more likely to vary at the individual level, compared to the n-grams used in the earlier paper.

As Bramsen *et al.* do not report results for a pairwise partitioning of the dataset, we are not able to compare our results with theirs in this case. However for the up-vs-down classifier, our result with pairwise partitioning is significantly above the baselines.

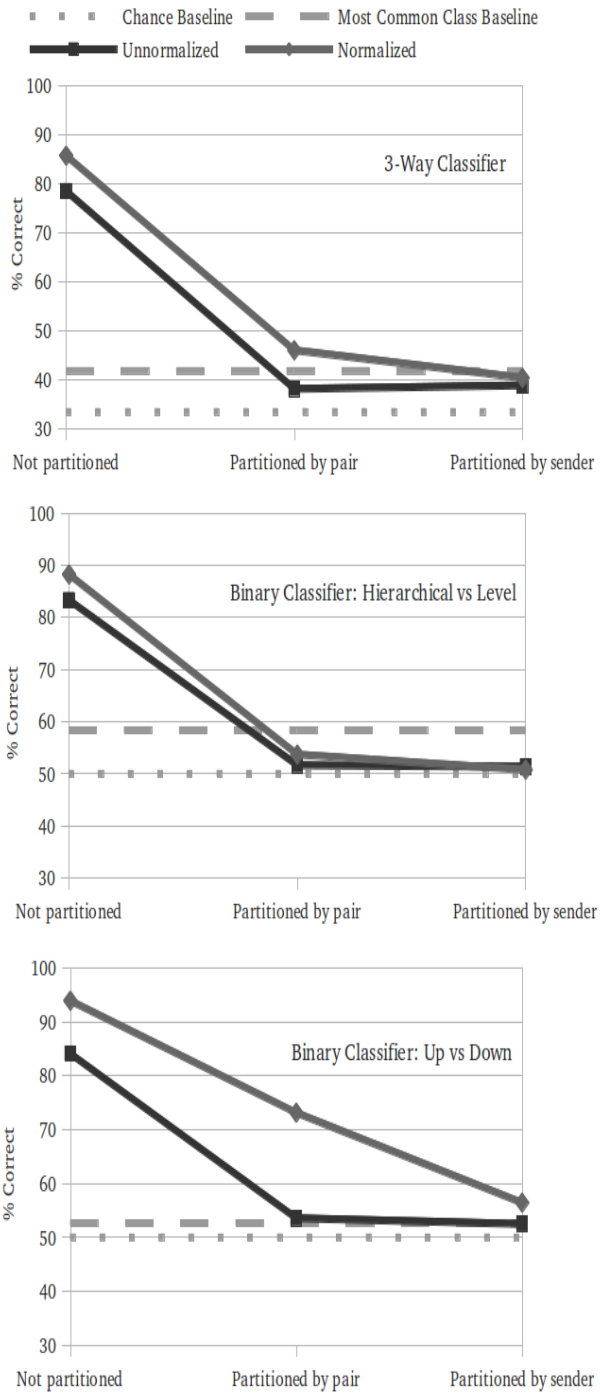

Fig. 2. The effects of partitioning and normalization on the classification accuracy

Table 1. Top ten features: information gain

3-way classifier	Binary: hierarchical-vs-level	Binary: up-vs-down
% Interjections	% Alphanumeric words	% Interjections
% Exclamation marks	% Interjections	% Exclamation marks
% Currency symbols	% Exclamation marks	% Semicolons
% Alphanumeric words	% Currency symbols	% Alphanumeric words
% Percent signs	% Percent signs	% Percent signs
% Maths symbols	% Semicolons	% Ampersands
% Ampersands	% Maths symbols	% Currency symbols
% Questions	% Ampersands	% Questions
% Semicolons	% Hash symbols	% Hash symbols
% Hash symbols	% Brackets	% Maths symbols

6 Feature Selection

Having experimented with classifiers using all available features, we then proceeded to examine the contribution of each feature in more depth. To this end, we applied a variety of statistical techniques: information gain, feature ablation, and principal component analysis.

In order to consider a representative set of results, without becoming overwhelmed by the number of variables, we shall for each classifier examine only the sender-partitioned, normalized data set. As the sender-partitioned set has the lowest accuracy, this scenario presents the greatest challenge.

6.1 Information Gain

Information gain was calculated for each of the three classification problems, using the WEKA [16] software package.

The top ten features for each data set are reported in Table 1. Some of these features feel more intuitive than others. Interjections and questions, being dialogue act categories, seem like logical features to appear in the top ten. Exclamation marks may be seen as indicative of emotionality, and alphanumeric words are non-standard language and therefore mark informality.

But on the other hand, surprisingly many punctuation-type features have made it on to all three lists: it is much harder to envision a link between use of maths symbols and social (hierarchical) role, and indeed this category was only included for completeness.

More generally, it is interesting to note that the three lists have nine of their top ten features in common, albeit in different orders.

6.2 Feature Ablation

Ablation is the process of removing each feature, in turn, from the classifier. By measuring which features, when removed, precipitate the greatest drop in accuracy, one obtains another measure of feature value. The results are displayed in Table 2.

Note that for the 9^{th} and 10^{th} ranking features for the binary hierarchical-vs-level classifier (marked with *), the results actually *improved* upon removal of these features. This is a dramatic contrast with the other two classifiers, for which the top-ten lists are taken from lists of 44 (three-way classifier) and 23 (up-down classifier) positive results. This adds to the evidence that the hierarchical-vs-level classifier is the least effective.

There is very little overlap between the features which appear to be of most importance to each classifier. This is in sharp contrast to what we observed for the information gain lists, and perhaps suggests that ablation is identifying more genuinely distinguishing features, although it is surprising that three-way classification does not have more in common with its two sub-problems.

Table 2. Top ten features: ablation

3-way classifier	Binary: hierarchical-vs-level	Binary: up-vs-down
% Verbs	% Pronouns	% Adjectives
% Maths symbols	% Modal	Number of Paragraphs
% Prepositions	% Questions	% Repeat Letter Words
% Non-dictionary words	Sentences Per Paragraph	% Prepositions
% Commas	% Commas	% Exclamations
% Letters	Words Per Sentence	% Questions
Number of Words	% Adjectives	% Emoticons
% Nouns	% Maths symbols	% Uppercase
Heylighen Dewaele	*% Percent signs*	% Determiners
% Symbols	*% Uppercase*	Words Per Sentence

6.3 Comparison of Approaches

There is surprisingly little overlap between the top features highlighted by these different methods, and further work on feature selection is required to identify the most useful combination of features for each classifier.

Principal component analysis projects a multi-dimensional feature space onto a smaller number of dimensions. Unlike information gain and ablation, PCA does not provide a simple list of features in order of importance. However, it does give the features which contribute to each component, and their relative contributions. When PCA was undertaken on these data, the first component contrasted nouns against pronouns and verbs, and a second component where length features contributed strongly.

Both PCA and feature ablation highlight the importance of part-of-speech tagging, as well as message length features. Meanwhile, the results for information gain suggest that punctuation features should not be overlooked, even when their relevance is not obvious. Further work should investigate possible refinements to the feature set, for instance by hillclimbing methods.

7 Conclusions

In this paper we have shown that it is possible to conduct relationship classification with a very limited feature set, using only stylistic features with a strong theoretical motivation, which are broadly applicable across languages, and which are easy and cheap to compute at scale. We have used these features to demonstrate the gain which can be achieved by normalizing all features by sender, and examined the impact of partitioning the training and testing data in different ways.

Previous work by Bramsen *et al.* [1] set a benchmark for relationship modeling in the Enron corpus. Our up-down classifier is the most appropriate analogue for comparison to their results, as they only address the problem of upspeak/downspeak classification. We have seen that the stylistic features perform slightly better for an un-partitioned dataset, but less well when the data is partitioned by sender. We also examined the case of partitioning by sender-recipient pairs, with promising results.

There are numerous possibilities for further work to improve relationship modeling. Moving on from message-level classification to relationship-level classification is the next essential stage. Each individual message can be taken as a piece of evidence about the relationship, and as our ultimate goal is to categorise the relationship, we can give less weight to messages which do not provide strong evidence; indeed, this is preferable to drawing unwarranted conclusions from low-confidence results. Further, we intend to aggregate the output of such a pairwise classifier, to construct a (partial) graph of the organizational hierarchy. This could then be combined with the

results of traditional SNA to develop a detailed graph.

This paper has deliberately focused on style, to the exclusion of subject matter features. However, we anticipate that topic will be an important feature for more advanced relationship modeling. Jabbari *et al.* [8] investigated the characteristics of business and personal emails at the message level, but this could easily be extended to categorise a relationship as 'mostly business' or 'mostly personal' in nature. Qualitative studies [11, 15] have found that greeting and closing words, address forms and sign-off names all vary with the relative status of message participants. Using zoning to identify sections of a message, there is scope to develop features of this kind.

References

1. **Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011).** Extracting Social Power Relationships from Natural Language. *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2011),* Portland, Oregon, 1, 773–782.
2. **Brown, P. & Levinson, S.C. (1987).** *Politeness: Some Universals in Language Usage.* Cambridge; New York: Cambridge University Press.
3. **Duthler, K.W. (2006).** The Politeness of Requests Made via Email and Voicemail: Support for the Hyperpersonal Model. *Journal of Computer Mediated Communication, 11(2), article 6*.
4. **Gallagher, I. (2010).** Bayesian Block Modeling for Weighted Networks. *Eighth Workshop on Mining and Learning with Graphs (MLG'10),* Washington, DC, 55–61.
5. ***Unicode 6.0.0.*** Retrieved from www.unicode.org/versions/Unicode6.0.0/
6. **Grice, H.P. (1975).** Logic and Conversation. *Syntax and Semantics, Volume 3: Speech Acts* (41–58). New York: Academic Press.
7. **Heylighen, F. & Dewaele, J.M. (2002).** Variation in the Contextuality of Language: an Empirical Measure. *Foundations of Science,* 7(3), 293–340.
8. **Jabbari, S., Allison, B., Guthrie, D., & Guthrie, L. (2006).** Towards the Orwellian Nightmare: Separation of business and personal emails. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006), Sydney, Australia*.
9. **Klimt, B. & Yang, Y. (2004).** Introducing the Enron Corpus. *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, California.
10. **Lakoff, R. (1973).** Language and a Woman's Place*. Language in Society,* 2(1), 45–80.
11. **Panteli, N. (2002).** Richness, Power Cues and Email Text. *Information and Management,* 40(2), 75–86.
12. **Peterson, K., Hohensee, M., & Xia, F. (2011).** Email Formality in the Workplace: A Case Study on the Enron Corpus. *Workshop* on Languages in Social Media (*LSM'11), Portland, Oregon,* 86–95.
13. **Rowe, R., Creamer, G., Hershkop, S., & Stolfo, S.J. (2007).** Automated Social Hierarchy Detection through email Network Analysis. *9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis,* San Jose, California, 109–117.
14. **Searle, J.R. (1969).** *Speech Acts: An essay in the philosophy of language.* London: Cambridge University Press.
15. **Waldvogel, J. (2007).** Greetings and Closings in Workplace Email. *Journal of Computer-Mediated Communication,* 12(2), article 6.
16. **WEKA 3**. (s.f.). Data Mining Software in Java. Retrieved from http://www.cs.waikato.ac.nz/~ml/weka.

Rachel Cotterill is studying for her Ph.D. at the University of Sheffield. Her particular interest is in characterizing communicants and their relationships, looking not only at what people say, but how they say it. Her research spans natural language processing, sociolinguistics, pragmatics, and stylometrics.

Article received on 08/12/2012, accepted on 17/01/2013.