Anywhere but Here
Enron's Emails in the Midst of a Crisis

Corey Taylor[1,2], Richard Leibbrandt[1] and David Powers[1]
[1]*School of Computer Science, Engineering and Mathematics, Flinders University, PO Box 2100, Adelaide 5001, South Australia*
[2]*Institute of Pharmaceutical Chemistry, Philips-University Marburg, Marbacher Weg 6, 35032 Marburg, Germany*

Keywords: Sentiment Analysis, Conditional Random Field, Topic Model, Information Retrieval, Text Processing, Enron.

Abstract: The emotional states of employees under stress are likely to manifest themselves in ways other in inter-personal interactions, namely, emails. The Enron Email Corpus was mined by both supervised and unsupervised methods to determine the degree to which this was true for Enron employees whilst the corporation was under investigation. Changes in language patterns were then compared against the timelines of the investigation. The method as described validates both the use of a subset of a very large corpus and the use of tagging methods to understand the patterns in various phrase types as used by Enron employees.

1 INTRODUCTION

There is now considerable evidence that perceived stress at work is widespread and associated with ill health at work. Work stress has been reported as highest in the middle-aged, educated to degree level and those separated or divorced (Smith, 2001). Additionally, employees affected by work-related stress are often underwhelmed by their employers' attempts to address the issue. It has been reported that employees in smaller organisations (i.e., less than 100) rated their work environments significantly higher than larger organisations on job satisfaction and also seems relatively clear that the level of trust between management and workers, joint resolution on ways to manage stress and organisation commitment to deal with issues causing stress are crucial for the management of workplace negativity (Buys et al., 2010) This is a finding replicated elsewhere with similar conclusions regarding the effect of workplace stress on a worker's health, well-being and productivity, even with highly professional environments such as a university (Shikieri et al., 2012).

Coupled with this are the effects of a singular stressful event upon companies in changing worker attitudes toward work and how that can manifest itself in non-work-related behaviours. Particularly in the short-term after an event, behaviours such as 'communal bereavement' become apparent (Hurley-Hanson, 2011). 'Incivility' and mistreatment generally has also been reported to impact on organisational mood and subsequent behaviour as persons within an organisation express their displeasure with each other and, left unchecked, can result in significant numbers of persons within an organisation displaying negative affect and subsequent impact on performance and the health of the organisation generally. This can result in withdrawal from participating in organisational activities or more direct behaviour (e.g. undermining a fellow employee, petty behaviours designed to humiliate them). Generally, colleagues will not overtly confront their co-workers but will find expression of their displeasure in other media, namely, email (Pearson et al., 2001; Walinga et al., 2013). An exploration of an email corpus to discern whether (and by how much) workers' attitudes manifest themselves in their emails is therefore warranted.

The only significant public email corpus available to explore the emotional content of emails is the Enron Email Dataset and, thus far, analysis of the emotional content of text in emails is non-existent, the research mainly focussed on testing researchers' machine learning models and algorithms. Researchers have mainly concentrated their efforts on analysis of network clustering and

Taylor C., Leibbrandt R. and Powers D..
Anywhere but Here - Enron's Emails in the Midst of a Crisis.
DOI: 10.5220/0005174502470253
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (ICAART-2015), pages 247-253
ISBN: 978-989-758-074-1

their evolution into social communities. One study found via directed graph/Singular Value Decomposition (SVD) methods that clustering of this nature tends to reflect seniority and pay structure (Chapanond et al., 2005; Carley et al., 2005). Analysis of network changes during the course of the investigation has also been conducted and from a study was reported that organisational communication, even the sample that exists in email, changes not only in volume but in who is communicating with whom during periods of organisational change and crisis and this reflects changes in communication norms and the way in which groups present themselves (Diesner et al., 2005). Another study, again using SVD methods, found that word use within the Enron dataset is correlated to function within the organisation and changes in word usage can reflect the increased influence of 'key players' in major company events (Keila et al., 2005).

Missing from the literature is analysis of emotional and attitude content changes in the email of employees. As suggested, the real moods of employees may manifest themselves via indirect communications such as email. The purpose of this study was therefore to extract this information from emails using both supervised and unsupervised methods, with some basic assumptions, and to link these patterns to investigation timelines. It was hoped that changes in the use of language reflecting employee's moods would reflect the very public and stressful nature of the investigation into and subsequent financial collapse of the Enron Corporation. Computational methods used to extract this meaning from the corpus were a Conditional Random Field used for tagging phrases of interest and Probabilistic Topic Modelling to extract semantic data from emails.

1.1 Conditional Random Field

A Conditional Random Field (CRF) is an undirected discriminatively-trained graphical model, a special case of which is a linear-chain CRF which is trained to maximise the conditional probability of a label sequence given an input sequence (Schneider, 2006). This CRF has an exponential form allowing it to integrate many over-lapping non-independent features and avoiding the *label-bias problem* (Le-Hong et al., 2006).

Let $x = x_1 \ldots x_T$ be an input sequence and $y = y_1 \ldots y_T$ be a corresponding state (or label) sequence. A CRF with parameters $\Lambda = \{\lambda, \ldots\}$ defines a conditional probability for y given x to be:

$$P_\Lambda(y \mid x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

where Z_x is a normalisation constant that makes the probabilities of all label sequences sum to one, $f_k(y_{t-1}, y_t, x, t)$ is a feature function, and λ_k is a learned weight associated with f_k. A feature function indicates the occurrence of an event consisting of a state transition $y_{t-1} \rightarrow y_t$ and a query to the input sequence x centered at the current time step t.

Inference in CRFs is done by finding the most probable label sequence, y*, for an input sequence, x, given the model in Equation 1:

$$y^* = \arg\max_y P_\Lambda(y \mid x) \quad (2)$$

During training, the weights λ_k are set to maximise the conditional log-likelihood of a set of labelled sequences in a training set $D = \{(x^{(i)}, y^{(i)}) : i = 1, \ldots, M\}$:

$$LL(D) = \sum_{i=1}^{M} \log P_\Lambda(y^{(i)}, x^{(i)}) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2} \quad (3)$$

$$= \sum_{i=1}^{M} \left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}^{(i)}, x^{(i)}, t) - \log Z_{x(i)} \right) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2}$$

The term $\sum_k \frac{\lambda_k^2}{2\sigma_k^2}$ is a Gaussian prior used to penalise the log-likelihood in order to avoid over-fitting and σ_k^2 is a variance. Maximisation of Equation 3 corresponds to matching the expected count of each feature according to the model to its adjusted empirical count:

$$\sum_{i=1}^{M} \sum_{t=1}^{T} f_k(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)}, t) - \frac{\lambda_k^2}{\sigma_k^2} \quad (4)$$

$$= \sum_{i=1}^{M} \sum_{y'} P_\Lambda(y' \mid x^{(i)}) \sum_{t=1}^{T} f_k(y'_{t-1}, y'_t, x^{(i)}, t)$$

The term $\frac{\lambda_k^2}{\sigma_k^2}$ is used to discount the empirical feature counts. Finding the parameter set Λ that maximises the log-likelihood in Equation 3 is done using an iterative procedure called limited-memory quasi-Newton (L-BFGS). Since the log-likelihood function in a linear-chain CRF is convex, the learning procedure is guaranteed to converge to a global maximum.

1.2 Topic Model

Given the definition that a document is a mixture of

topics, that P(z) is a distribution of topics over a document and P(w|z) represents the probability distribution of words w given topic z, a model specifying the following distribution over words can be formed. As $P(z_i = j)$ is the probability that the *j*th of T topics was sampled for the *i*th word token and $P(w_i \mid z_i = j)$ is the probability of w_i for topic *j*:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \quad (5)$$

To obtain topic and word information from the topic model, we must first extract the topic-word distributions, ϕ, and topic distributions θ. As this is computationally expensive, we must sample from the posterior distribution over z using a Gibbs Sampling algorithm.

For this purpose, the collection of documents by a set of word indices w_i and document indices d_i, for each word token *I* is represented. The Gibbs sampling algorithm considers each word token within the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on topic assignments to all other word tokens. From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token. This conditional distribution is $P(z_i = j \mid Z_{-i}, w_i, d_i, \cdot)$, where $Z_i = j$ represents the topic assignment of token i to topic j, Z_{-i} refers to the topic assignments of all other word tokens, and "·" refers to all other known or observed information such as all other word and document indices w_{-i} and d_{-i}, and hyperparameters α and β (Steyvers et al, 2006).

$$P(z_i = j \mid z_{-i}, w_i, d_i, \bullet) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \quad (6)$$

Here C^{WT} and C^{DT} are matrices of counts of words with dimensions WxT and DxT respectively; $C^{WT}_{w_i j}$ contains the number w_j of times word w is assigned to topic j, not including the current instance i and $C^{DT}_{d_i j}$ contains the number of times topic j is assigned to some word token in document d, not including the current instance i.

Each Gibbs sample consists the set of topic assignments to all *N* word tokens in the corpus from a single pass through all documents. During the 'burn-in' period, the Gibbs samples have to be discarded because they are poor estimates of the posterior. After the burn-in period, the successive Gibbs samples begin to approximate the target distribution (i.e., the posterior distribution over topic assignments).

From this process, estimation of ϕ and θ is relatively straightforward. They correspond to the predictive distributions of sampling a new token of word *i* from topic *j*, and sampling a new token (as of yet unobserved) in document *d* from topic *j*. They are also the posterior means of these quantities conditioned on a particular sample *z*.

$$\phi_i^j = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \quad (7)$$

$$\theta_i^d = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dj} + T\alpha} \quad (8)$$

2 METHODS

The corpus used for analysis was the Enron Email Dataset, originally purchased by Leslie Kaelbling and made publicly available for download by the Federal Energy Regulatory Commission on Carnegie-Mellon University's servers.[1] The dataset contains ~500K emails from senior management executives at the Enron Corporation collected during the time of investigation, covering approximately 2000-2004.

Emails in this ~2.2Gb dataset are in MBOX format and arranged by email folders (e.g. inbox, sent items, personal folders). As many of these folders contain emails that are not relevant to this study (e.g. spam) or contain significant amounts of duplicated text (e.g. nested replies in long discussions), the decision was made to only include emails in users' 'sent items' for analysis. The rationale for this decision was that items such as spam, mailing list digests, etc. were far less likely to be in a person's sent items. Pre-processing scripts stripped out email headers, quoted and forwarded messages, HTML and formulaic block text (e.g. signatures). The resultant dataset contained ~90K emails split into months for tagging.

The MAchine Learning for LanguagE Toolkit (MALLET) was used for computation of the Enron dataset. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modelling,

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

information extraction, and other machine learning applications to text.[2] A modified version of the SimpleTagger class was used for tagging of emails with other methods written to instantiate, train and query topic models.

A linear-chain Conditional Random Field (CRF) was instantiated and trained in binary (on/off) fashion on the following types of phrases:

1) Slacker language – 24 phrases
2) Aggressive language – 26 phrases
3) Panic language – 14 phrases

In each case, the most 'descriptive' word was chosen as the word to tag as one of the above, the rest of the words in a given sentence tagged with MALLET's default 'O'. Example sentences for training the CRF were obtained from one month of email data and manually tagged commensurate with the above speech categories. For example, "let's do lunch" was tagged as 'slacker' language, "I need it ASAP" tagged as 'aggressive' language and "we have to get onto this immediately" as 'panic' language. Sentence-ending full stops were removed from each text file to avoid tagging of each phrase type only in the instance where it is used as a completed sentence. The frequency of each tagged phrase type in a given month of emails was then calculated.

Topic models were trained on 100 topics using each month's emails as its corpus. Alpha (i.e. sum over topics) and beta (e.g. single dimension of Dirichlet prior) hyperparameters were set to 0.01. Information subsequently extracted for each month was the top-10 highest probability topics given documents (i.e. emails) for the month and the top-10 words given topics in ranked order.

Counts of each phrase type per month were used in Pearson correlations with the monthly closing Enron stock price from that time. This data and phrase count data as elucidated from the CRF was added to a static plot of frequencies. The plot was generated using the JChart2D toolkit.[3]

3 RESULTS

Figure 1 is representative of the frequency of slacker/aggression/panic phrases by month. Additionally, the closing stock price at the end of the month is plotted.

Output from the topic model was in the form of topics with probabilities and words under those topics in ranked order. Pearson correlation of phrase type with stock price is not suggestive of a relationship between 'slacker' and 'aggressive' phrases but a strong negative correlation with 'panic' phrases was evident.

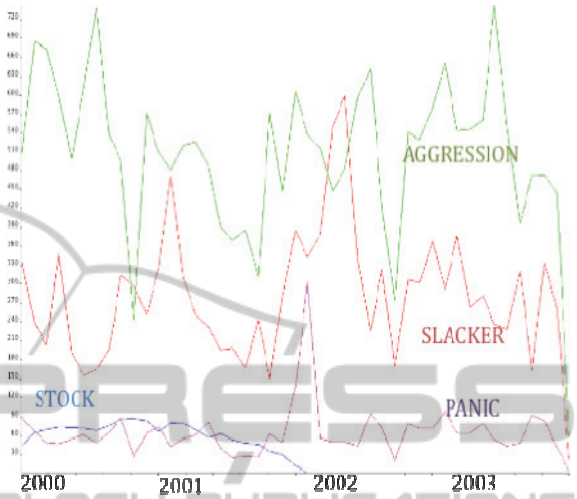

Figure 1: Frequencies of language types and stock price over time.

Table 1: Pearson correlations of counts of language types with Enron stock price.

Phrase type	R^2
Slacker language	-0.032
Aggressive language	-0.0004
Panic language	-0.797

4 DISCUSSION

The frequencies of various language types may possibly reflect events during the course of the investigation into Enron. Highlighted examples as in Figure 2 may be indicative of the following events as detailed in court documents and re-published in various books about Enron.

1: Spikes in 'aggressive' language during 2000 are approximately prior to Enron's all-time high stock price of $90.56 recorded on August 23rd 2000. This corresponds to the activities of influential head of trading Timothy Belden. Some of his activities were found to have played a role in precipitating the California energy crisis, a shortage of electricity supply caused by illegal market manipulations, eventually resulting in his conviction for conspiracy to commit wire fraud. This was also reflected in output from the topic model, for eample in June 2000, the word 'California' or other places in the

[2] http://mallet.cs.umass.edu
[3] http://jchart2d.sourceforge.net/

California area were used in 4 of the top-10 topics for that month under topics about regulation and energy.

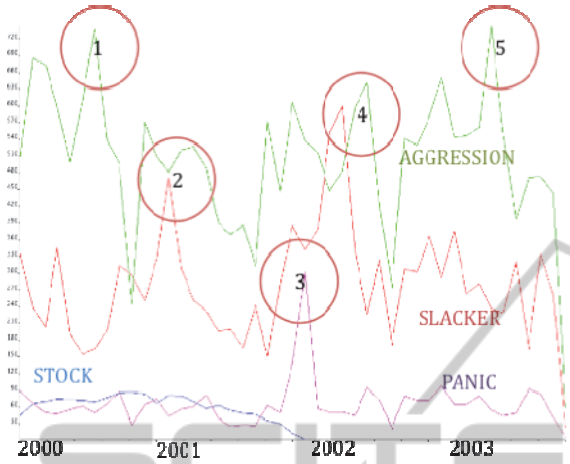

Figure 2: Frequencies of language types and stock price over time with highlighted events.

Table 2: Use of the word 'California in topics – June 2000.

power gas **california** market price energy electricity prices electric company
san francisco chronicle angeles diego los wine city sgovenar@govadv.com times
california commission ferc energy utilities puc public file order federal
call meeting conference week office number time sellers discuss **california**
davis bill governor dow assembly **california** committee energy jones state

2: A spike in 'slacker' language approximately corresponds to executives such as Tom White cashing out their stocks, retiring and making large purchases.

3: Spikes in 'panic' language correspond to a precipitous drop in Enron's share price following on from the beginning of an informal probe by the Securities and Exchange commission commenced on October 17 2001 and followed by Enron declaring bankruptcy and massive layoffs on December 2nd 2001.

4: Indicative spikes in 'slacker' and, subsequently, 'aggressive' language as resignations and senate hearings involving executives such as Ken Lay commence in early 2002. Other key events around this time include the suicide of Cliff Baxter, former vice chairman, and the release of The Powers Report on Feb 2nd 2002, a 218-page summary of internal investigations into Enron's collapse by the University Of Texas School Of Law Dean William Powers. This report publicly named Enron executives and held them responsible for the collapse of the company.

5: A spike in 'aggressive' language in early 2003 corresponds to announcement on March 19th 2003 detailing that Enron will keep its North American and international pipeline and power assets in an attempt to emerge from bankruptcy. The topic model for the month of February 2003 suggested increases in the use of the word 'Enron', as it was in the top-10 words for 4 of the top-10 topics for that month in relation to topics such as trading and communications. In most other months, 'Enron' appears either not at all or only in one of the topics for that month.

Table 3: Use of the word 'Enron' in topics – February 2003.

call questions group issues comments credit **enron** 145 mark discuss contact
ferc rto market transmission **enron** process meeting model epsa filing
agreement draft master isda guaranty **enron** form letter send doc
enron program entity kaiser trading product power tax forward products

A precipitous drop in 'aggressive language corresponds to charges being laid against Chief Financial Officer Andrew Fastow and his wife on May 1st 2003 for activities which resulted in concealment of Enron's massive financial losses. For these activities, he subsequently served 6 years in prison.

As the above is merely a descriptive analysis, the magnitude and direction of any association is unknown and conclusions drawn are entirely subjective. However, it can be said with some certainty that, to some degree, the emotional states of employees in Enron were reflected in language patterns they chose to use in sending emails. This is broadly supportive of the contention as stated earlier that the language used in emails by employees working within a dysfunctional and uncivil work environment, whilst perhaps responding in direct ways too, will reflect both the emotional state of the employee and the work environment in the use of what is ostensibly a work tool.

The strong negative correlation between panic phrases and stock price is also indicative of the validity of the method. This provides a cautious validation of the use of CRFs to tag phrases based on semantic considerations, as opposed to syntactic or for Named-Entity Recognition purposes, even with a lightly-trained model comprised of only 24 'slacker phrases, 26 'aggressive' phrases and 14

'panic' phrases. This is as opposed to training sets comprised of hundreds or sometimes thousands of training phrases, as is more usually observed in the use of CRFs. That the use of relatively broad categorisations of phrases was able to approximately reflect the timelines of the investigation into Enron means the method could be extended in many ways.

There are many limitations with the approach as detailed within this study. Selection of phrases corresponding to the three categories studied was entirely subjective and therefore there was a risk of bias in model training. Additionally, the nature of the corpus meant that although there were extensive attempts to clean the dataset, many artifacts of email in its raw form remain (e.g. spam, multiple quoting biasing counts). The precise nature of the association between phrase use and actual events in Enron's history can only be guessed at, more information regarding the detailed course of events would be required to validate the accuracy and sensitivity of the association detailed here. The *a priori* nature of CRF model training in this instance virtually guarantees bias.

There are also general limitations in probabilistic topics models which may affect inferred results; topic models are prone to overfitting, as in, the mode by which an individual document's topic mixture is established is not robust enough to handle the addition of new documents to the trained corpus. Related, the number of free model parameters increases linearly with the number of training documents, making re-training a computationally expensive exercise.

Possible extensions to the software are many and varied. Results from what was a relatively lightly trained CRF model seemed reasonable but it was trained only on binary data with a first-order model. The use of higher-order model will likely increase the precision of tagging of phrases as more information about context is modelled. This may also allow a finer grained model training of more specific phrases as 'slacker', 'aggressive', etc. are relatively broad terms for the language being modelled.

Instead, sub-types of slacker/aggressive/panic phrases could be tagged. The results of a topic model could also be used to inform the tagging of phrases rather than the *a priori* method as detailed in this study. Ideally, a formal evaluation of tagging predictive accuracy could be conducted on non-Enron emails or on Enron emails with a k-folds cross-validation methodology with attendant measures of fit (e.g. positive predictive value).

5 CONCLUSIONS

The method as detailed provides a broad method for the descriptive analysis of email data by tagging of phrases that are semantically interesting. That this exercise even broadly reflects the timeline of investigation validates the use of both a sub-set of the full Enron corpus as well as the method used to tag information of interest. This is suggestive that the performance of even a lightly-trained model may be acceptable on a far smaller test set than would be the case were it exhaustively trained on the full Enron Email Dataset.

REFERENCES

Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM , 55* (4), 77-84.

Buys, N. M. (2010). Employees' Perceptions of the Management of Workplace Stress. *International Journal of Disability Management, 5* (2), 25-31.

Chapanond, A. K. (2005). Graph Theoretic and Spectral Analysis of Enron Email Data. *Computational & Mathematical Organization Theory, 11*, 265-281.

Chekina, L. G. (2013). Exploiting label dependencies for improved sample complexity. *Machine Learning, 91*, 1-42.

Dahl, C. (2004). Pipe Dreams: Greed, Ego, and the Death of Enron/Anatomy of Greed: *The Energy Journal, 25* (4), 115-134.

Diesner, J. C. (2008). Conditional random fields for entity extraction and ontological text coding. *Computer and Mathematical Organisation Theory, 14*, 248-262.

Diesner, J. F. (2005). Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory, 11*, 201-228.

Dreijer, J. H. (2013). Left ventricular segmentation from MRI datasets with edge modelling conditional random fields . *BMC Medical Imaging, 13*, 1-24.

El Shikieri, A. M. (2012). Factors Associated with Occupational Stress and Their Effects on Organizational Performance in a Sudanese University . *Creative Education , 3* (1), 134-144.

Hayashida, M. K. (2013). Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso . *BMC Systems Biology, 7* (Suppl 2), 1-11.

Hurley-Hanson, A. G. (2011). The Effect of the Attacks of 9/11 on Organizational Policies, Employee Attitudes and Workers' Psychological States. *American Journal of Economics and Business Administration, 3* (2), 377-389.

Hutton, A. L. (2006). Crowdsourcing Evaluations of Classifier Interpretability. *AAAI Technical Report SS-12-06 Wisdom of the Crowd*, 21-26.

Jahanian, R. T. (2012). Stress Management in the

Workplace. *International Journal of Academic Research in Economics and Management Sciences , 1* (6), 1-9.

Keila, P. S. (2005). Structure in the Enron Email Dataset. *Computational & Mathematical Organization Theory , 11*, 183-199.

Klimt, B. Y. (2005). Introducing the Enron Corpus. 2.

Le-Hong, P. P.-H.-T. (2006). On the Effect of the Label Bias Problem in Part-Of-Speech Tagging. *Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, 103-108.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. *http://mallet.cs.umass.edu* .

Ozcaglar, C. (2008). CLASSIFICATION OF EMAIL MESSAGES INTO TOPICS USING LATENT DIRICHLET ALLOCATION. *MSc. thesis*, 1-37.

Pearson, C. A. (2001). When Workers Flout Convention: A Study of Workplace Incivility. *Human Relations , 54* (11), 1387-1419.

Priebe, C. C. (2005). Scan Statistics on Enron Graphs. *Computational & Mathematical Organization Theory , 11*, 229-247.

Robinson, S. (2009). The Nature of Responsibility in a Professional Setting. *Journal of Business Ethics, 88*, 11-19.

Schieman, S. W. (2006). The Nature of Work and the Stress of Higher Status. *Journal of Health and Social Behaviour , 47* (3), 242-257.

Schneider, K.-M. (2006). Information extraction from calls for papers with conditional random fields and layout features . *Artificial Intelligence Review, 25*, 67-77.

Smith, A. (2001). Perceptions of Stress At Work. *Human Resource Management Journal, 11* (4), 74-86.

Steyvers, M. G. (2006). Probabilistic Topics Models. In L. Erlbaum, *Latent Semantic Analysis: A Road to Meaning* (pp. 1-15).

Styler, W. (2011). The Enronsent Corpus. *University of Colorado - Boulder* , 1-7.

Sun, C. G. (2007). Rich features based Conditional Random Fields for biological named entities recognition . *Computers in Biology and Medicine, 37*, 1327-1333.

Sun, X. H.-Y. (2011). Chinese New Word Identification: A Latent Discriminative Model with Global Features . *Journal of Computer Science and Technology, 26* (1), 14-24.

Walinga, J. R. (2013). Transforming stress in complex work environments Exploring the capabilities of middle managers in the public sector . *International Journal of Workplace Health Management, 6* (1), 66-88.

Wallach, H. (2001). Efficient Training of Conditional Random Fields. *Masters's thesis*, 1-8.

Wu, J. L. (2010). Chunk Parsing and Entity Relation Extracting to Chinese Text by Using Conditional Random Fields Model. *Journal of Intelligent Learning Systems & Applications, 2*, 139-146.