Challenges of Analysing Speech Acts in Organisational Communication

Marcus Grattan
Computer and Information Science
Linköping University, Sweden
margr792@student.liu.se

Andrea Fried,
Management and Engineering
Linköping University, Sweden
andrea.fried@liu.se

Arne Jönsson
Computer and Information Science
Linköping University, Linköping, Sweden
arne.jonsson@liu.se

Abstract

The research presented in this paper addresses the challenges of analysing organisational stakeholder communication using speech act theory, supported by a novel analytical model. We present a large, automatically annotated dataset focused on present- and future-oriented speech acts in Swedish organisational discourse, specifically within the context of the ISO/IEC 27001 standard. To evaluate the dataset, we fine-tuned a multilingual XLM-Roberta model on the annotated data and tested it against a manually annotated gold standard. Although the results were inconclusive, the study contributes valuable information to the CLARIN infrastructure and advances research in organisational communication and information security.

1 Introduction

In a joint project between the Swedish SMS CLARIN K-centre[1] and management researchers, we investigate stakeholder communication on Swedish corporate websites related to preventive cyber innovation. A specialised corpus of corporate texts from Swedish company websites was created to analyse discourse surrounding information security with a focus on the ISO/IEC 27001 standards (Jönsson et al., 2024). This corpus contains data from 291 companies, a total of 5,960,328 tokens, with a vocabulary size of 204,312 and an average entry length of 24.3 tokens. The ISO/IEC 27001 standard requires organisations to establish, implement and audit the rules and authorities relating to the legal and ethical aspects necessary for the proper functioning and accountability of the entire data and algorithm lifecycle (Janssen et al., 2020, p.3), see also (Culot et al., 2021; Disterer, 2013; Mirtsch et al., 2021). ISO/IEC 27001 certification ensures information security and encourages good corporate behaviour by requiring compliance to be audited by external groups.

For analysis, we applied content analysis along with novel computational sentiment analysis to uncover differences in communication about preventive cyber innovations (Jönsson et al., 2024). Our findings revealed three distinct organisational perspectives on preventive cyber innovation:

- Adopters emphasise the relative advantage and complexity of ISO/IEC 27001 and actively communicate about their information security efforts, often through company websites.
- Stewards, in contrast, highlight trialability and, to a lesser extent, the relative advantage of preventive innovation. Their communication includes commissives and other future-oriented statements about their organisation's information security efforts.
- Advocates communicate less frequently about the efforts, achievements, and challenges related to the adoption of preventive innovations, adopting a more indirect approach similar to companies that directly influence their organisational reality of information management.

These categories help to understand the varying approaches of companies to information security communication and are added manually to the dataset.

[1] https://sprakbanken-clarin.lingfil.uu.se/centra/sms_en.html

To capture future-versus-present-oriented communication in relation to InfoSec standards, we conducted a speech acts analysis, which is presented in this paper.

2 Model building

Our working assumption was that organisations express their adherence to standards through intentional statements (Commissives) or descriptive statements (Assertives). This approach aligns with the concept of direction of fit (Searle, 1985), where Commissives represent a world-to-word fit, expressing future intentions and commitments that an organisation plans to fulfil. Conversely, Assertives represent a word-to-world fit, stating current or past conditions within the organisation. For the purposes of streamlined analysis and because of their overlap with Assertives in expressing current realities, Declaratives were grouped under the Assertive class. As a result, the defined speech act categories were Assertives and Commissives, with all other speech acts categorised as Other.

The construction of the Speech Act classification model began with selecting a large language model (LLM) for an initial round of annotation. Specifically, only quantised models were considered due to computing restraints. The selection was based on performance on a small, manually curated dataset, sampled from the ISO/IES 27001 corpus (Jönsson et al., 2024), containing approximately 100 instances per class. Model hyperparameters were then optimised using this same dataset.

The selected model, Llama3-70B-2.4BPW (quantised to an average of 2.4 bits per weight), was deployed to annotate a larger corpus. Annotations from this first round were manually validated, and correctly labelled instances were organised into class-specific datasets, each containing the original text and its corresponding Speech Act label. Ultimately, each class contained 550 validated entries. To ensure diversity and mitigate model bias, 50 instances per class were manually curated by reviewing and correcting the model's misclassifications. Due to data sparsity in the minority class, Commissive, ChatGPT-4 was employed for data augmentation. Forty carefully selected examples were expanded to better capture under-represented syntactic and semantic structures.

Next, a demonstration dataset was constructed for Clue and Reasoning Prompting (CARP) using the same LLM. Entries from the validated class collections were combined, and the LLM was prompted, following the procedure outlined in Sun et al. (2023), to generate supporting clues (e.g. keywords, phrases, syntactic features) for each entry. The model was then prompted to produce diagnostic reasoning based on the text, label, and generated clues. This process yielded a 1,650-entry demonstration dataset, evenly distributed across classes.

To support classification-sensitive demonstration retrieval in the second annotation round, an XLM-RoBERTa model (Conneau et al., 2020) was fine-tuned on the CARP demonstration dataset. Fine-tuning was essential for aligning the model not only with sentence embeddings but also with class annotations, enabling accurate retrieval of semantically and label-relevant examples for unseen sentences. The model was used to compute text embeddings, which were appended to each dataset entry.

A second round of annotation followed, using CARP. The Llama3 model was provided with task instructions, class definitions, and retrieved demonstrations via the fine-tuned retrieval model. Demonstrations were added up to the model's token limit, sorted in ascending order of cosine distance to the input sentence, as described by Sun et al. (2023).

This resulted in an automatically annotated dataset, comprising a random 56% subset of the full corpus. A final XLM-RoBERTa classifier was fine-tuned on this dataset. The model and dataset are available to the CLARIN community[2]. Training employed a linear learning rate scheduler with warm-up steps corresponding to 10% of batches per epoch and used the AdamW optimiser. To address class imbalance, the loss function incorporated class weights. Early stopping was triggered if validation loss increased for three consecutive evaluations. Weight decay was set to 0.01, and the learning rate was set to a conservative value of 4 x 10^{-5} to prevent premature convergence. A batch size of 16 was chosen to encourage generalisation. An 80/20 train/validation split was used, with evaluation occurring three times per epoch. Model selection was based on the weighted F1 score. The final model was trained for 2.33 epochs. Entries labelled as "No match found" were excluded from the training data.

[2] https://huggingface.co/MarcusGrattan/classifier-model/tree/main

To evaluate the classifier, a test set of 5,000 entries was manually annotated. An average of 1,000 sentences were labelled per day, split across two sessions. The final test set comprised 665 Commissive, 3,228 Assertive, and 1,107 Other instances. Following Abercrombie et al. (2023), intra-annotator agreement was measured on 10% of the test set one week after completion, yielding an agreement score of 0.8, indicating strong consistency.

For more details regarding the model-building process, see Grattan (2024).

3 Classifier model results

The classifier model achieved an accuracy of 0.73 and a weighted F1-score of 0.75 on the test dataset, outperforming the most frequent class baseline, which had an accuracy of 0.65 and a weighted F1-score of 0.51. The model's weighted precision and recall were 0.80 and 0.73, respectively. For the Commissive class, the model achieved precision, recall, and F1-scores of 0.34, 0.68, and 0.45, respectively. In the Assertive class, the model performed better with precision of 0.88, recall of 0.74, and an F1 score of 0.80. The Other class achieved a precision of 0.84, recall of 0.74, and F1 score of 0.79.

The automatically annotated dataset contained 138,052 entries, excluding instances with no annotation. Among these, 32,458 entries (24%) were classified as Commissives, 80,377 entries (58%) as Assertives, and 25,215 entries (18%) as Other. Notably, the model failed to annotate 301 entries.

In the context of Speech Act classification, large language models (LLMs) can significantly reduce the reliance on manually annotated data while producing high-quality, large-scale datasets. State-of-the-art results from similar domains, such as the Switchboard corpus and the MRDA corpus, report accuracy scores of 0.85 (Colombo et al., 2020) and 0.92 (Chapuis et al., 2021), respectively. The classifier model presented in this paper, with an accuracy score of 0.73, achieved these results using approximately 80% fewer manually labelled data compared to similar studies (Vosoughi & Roy, 2016; Zhang et al., 2011). These results suggest that leveraging techniques like Retrieval Augmented Prompting (RAG)(Lewis et al., 2021) and CARP(Sun et al., 2023) enables LLMs to not only generate large datasets but also achieve substantial quality.

4 Applying the model and discussion

The model was applied to the corpus of corporate websites of all Swedish companies that communicate with their stakeholders (e.g. clients, customers, government, non-profit organisations and other) about their ISO/IEC 27001 implementation efforts (Jönsson et al., 2024). The results are shown in Table 1 and visually in Figure 1, revealing little difference between the analysis categories Adopters, Stewards, and Advocates.

	Assertive	Commissive	Other
Adopters	55%	27%	18%
Stewards	55%	23%	22%
Advocates	61%	24%	15%

Table 1: Speech act distribution

Interestingly, contrary to expectations, Advocates exhibit a similar frequency of Assertive speech acts as the other categories. Manual analysis of these Assertive acts reveals that they primarily describe the actions of third parties. This poses a challenge for the model, as it struggles to determine whether a statement reflects the organization's own reality or that of others, which impacts classification results.

Moreover, the performance of the classification model revealed that it often confused Commissives with Assertives, and to a lesser degree, vice versa. The precision and recall metrics for the Commissive class suggest that the model had difficulty accurately identifying Commissive speech acts. Specifically, the precision was low (0.34), indicating that many instances labeled as Commissives were incorrect. On the other hand, the recall for Commissives was higher (0.68), suggesting that the model was relatively effective at identifying true Commissive statements, though not always accurately.

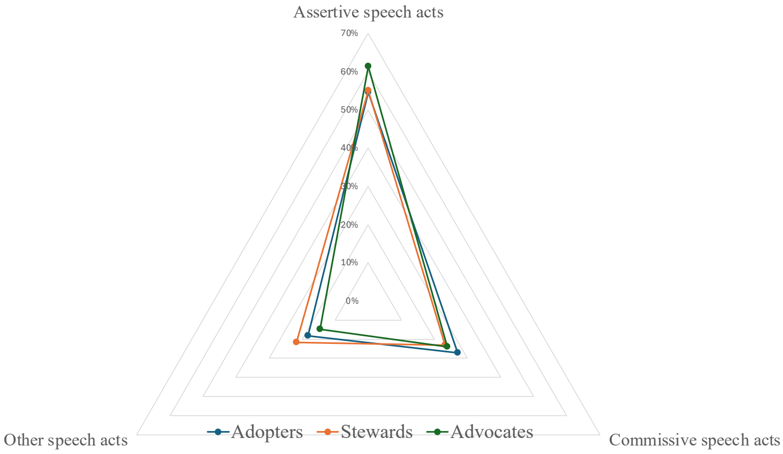

Figure 1: Diagram of the distribution

During manual annotation, it was noted that few Commissive instances were explicit commitments; instead, many were implicit guarantees or commitments. For instance, *The contracted supplier must be certified according to ISO 27001, which is a globally renowned information security standard.* is formulated as assertive, but not for the speaker itself, but for others, in this case suppliers.

This made it particularly challenging to differentiate between Assertives and Commissives. It is likely that the LLM used for automatic annotation faced similar difficulties in distinguishing between these two categories. Additionally, the indirect nature of Commissives may have contributed to ambiguities in the demonstration dataset, influencing the model's performance.

5 Conclusions

The research presented in this paper demonstrates how large language models (LLMs) can be effectively leveraged to reduce reliance on manually annotated datasets in the domain of Speech Act classification. Utilising techniques such as Clue and Reasoning Prompting (CARP) (Sun et al., 2023) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2021). The research successfully generated a substantial dataset of 138,052 entries. This dataset facilitated the training and evaluation of and a classifier, which achieved an accuracy of 73% and a weighted F1-score of 0.75.

This work builds upon existing frameworks and proposes novel approaches for dataset generation in text classification tasks, particularly within the context of organisational communication related to ISO/IEC 27001 standards. Beyond its technical contributions, the study also adds value to the CLARIN infrastructure and to the broader field of organisational communication research. Furthermore, it highlights the potential of quantised LLMs for Speech Act classification, offering valuable insights into their performance, especially in low-resource language settings.

However, the results of the model were not helpful for the management researchers to show differences in how the ISO/IEC 27001 standard was adopted by Swedish companies.

An expanded LLM selection process, incorporating a wider array of models more data and focused tests specifically aimed at indirect and/or context dependent commisives, could improve the selection process and extend to better Commissive classification results. Additionally, while LLM selection was restricted to quantised models, using un-quantised models while keeping similar model parameter size, would likely improve performance.

References

Abercrombie, G., Rieser, V., & Hovy, D. (2023). Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

Chapuis, E., Colombo, P., Manica, M., Labeau, M., & Clavel, C. (2021). Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Colombo, P., Chapuis, E., Manica, M., Vignon, E., Varni, G., & Clavel, C. (2020). Guiding attention in sequence-to-sequence models for dialogue act prediction. *arXiv preprint arXiv:2002.08801*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Culot, G., Nassimbeni, G., Podrecca, M., & Sartor, M. (2021). The iso/iec 27001 information security management standard: Literature review and theory-based research agenda. *The TQM Journal*, *33*(7), 76–105.

Disterer, G. (2013). Iso/iec 27000, 27001 and 27002 for information security management. *Journal of Information Security*, *4*(2), 92–100.

Grattan, M. (2024). *Now or later: Classifying future- and present-oriented speech acts in organizational communication* [Bachelor's Thesis]. Linköping University.

Janssen, M., Brousa, P., Estevez, E., Barbosad, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, *37*. https://doi.org/https://doi.org/10.1016/j.giq.2020.101493

Jönsson, A., Bandyopadhyay, S., Dragisic, S. P., & Fried, A. (2024). Analyses of information security standards on data crawled from company web sites using SweClarin resources. *Selected papers from the 2023 CLARIN Annual Conference*. https://doi.org/https://doi.org/10.3384/ecp210

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Mirtsch, M., Blind, K., Koch, C., & Dudek, G. (2021). Information security management in ict and non-ict sector companies: A preventive innovation perspective. *Computer & Security*, *109*. https://doi.org/109.https://doi.org/10.1016/j.cose.2021.102383

Searle, J. R. (1985, November). *Expression and meaning*. Cambridge University Press.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Vosoughi, S., & Roy, D. (2016). Tweet acts: A speech act classifier for twitter [Number: 1]. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 711–714. https://doi.org/10.1609/icwsm.v10i1.14821

Zhang, R., Gao, D., & Li, W. (2011). What are tweeters doing: Recognizing speech acts in twitter. *Analyzing Microtext, Papers from the 2011 AAAI Workshop*.