Speech de-identification data augmentation leveraging large language model

Priyanshu Dhingra
Rajiv Gandhi Institute of Petroleum Technology
priyanshudhingra1@gmail.com

Satyam Agrawal
National Institute of Technology Karnataka
lmcsatyam@gmail.com

Chandra Sekar Veerappan
Singapore Institute of Technology
chandra.veerappan@singaporetech.edu.sg

Thi Nga Ho
Nanyang Technological University
ngaht@ntu.edu.sg

Eng Siong Chng
Nanyang Technological University
aseschng@ntu.edu.sg

Rong Tong
Singapore Institute of Technology
tong.rong@singaporetech.edu.sg

***Abstract*—This work addresses the challenge of limited real-world speech data in speech de-identification, the process of removing Personally Identifiable Information (PII). We formulate speech de-identification as a named entity recognition (NER) task specifically for spoken English. To overcome data scarcity and enhance NER performance, we propose a data augmentation approach. This approach leverages a large language model to generate synthetic speech style text data enriched with diverse PII entities. The generated data undergoes an iterative process using a customized NER model for semi-automatic PII annotation. Our analysis demonstrates the effectiveness of this data augmentation strategy in significantly improving NER performance on spoken language text. Furthermore, to gain deeper insights into the specific errors made during NER, we employ performance analysis using alternative evaluation metrics.**

***Index Terms*—speech de-identification, data augmentation, named entity recognition, large language model**

I. Introduction

The exponential growth of digital applications has transformed the way people live. Those digital applications have facilitated a substantial increase in data generation and dissemination across digital networks. This has inevitably led to the collection of a vast amount of personal identifiable information, raising concerns about potential misuse, which can potentially lead to privacy violations, identity theft, and other harms.

De-identification refers to the process of removing personally identifiable information from data to ensure the remaining information cannot be linked back to a specific individual. PII can be embedded in various media formats, including text, images, speech, video, or even a combination of these. In this paper, we focus on the challenges of de-identifying PII carried within speech content.

The field of speech processing has seen significant efforts in de-identifying speaker biometrics. Initiatives like the ASVSpoof Challenge [1] and the Voice Privacy Challenge [2] have been ongoing for several years, focusing on detecting spoofed speech and masking speaker identity. However, less focus has been placed on de-identifying non-biometric information within speech content. Automatic de-identification of speech content, therefore, remains an under-explored area with significant potential.

Audio recordings may convey a variety of personally identifiable information. Customer service calls, for instance, often involve the collection of a user's registered ID number, phone number during the verification process. This information must be redacted before the recording is used for customer service evaluation or training purposes. However, the current practice of manual PII removal is labor-intensive and time-consuming. An automatic de-identification system would be highly desirable to streamline this crucial task.

Fig. 1: Example of a speech de-identification system

Figure 1 depicts a typical speech de-identification system. This system aims to protect sensitive information by identifying the presence of PII entities, locating their positions within the speech, and subsequently removing or redacting them.

Automatic speech de-identification typically involves processing the speech signal, transcribing the content, detecting PII entities, and finally obfuscating or removing them from the original audio data. However, constructing such a system necessitates speech data with extensive PII annotations. Publicly available datasets often prioritize common identifiers like names and public information, neglecting less frequent PIIs such as passport numbers or IDs. Conversely, generating large speech datasets with labeled PII is hampered by privacy concerns, as individuals are understandably reluctant to share recordings containing their personal information [3].

To address the scarcity of real-world PII annotated data problem, this paper proposes a data augmentation approach. Our two-step approach first leverages a large language model (LLM) to generate synthetic speech data that closely resembles natural spoken Singaporean English. Second, we cast the

979-8-3315-4085-2/24/$31.00

PII detection problem within the framework of named entity recognition. By employing customized NER models in an iterative process, we achieve semi-automatic annotation of PII entities within the synthetic data.

II. Related work

A. Text based system

The concept of text de-identification has been explored previously. Notably, [4] presents the first-ever approach utilizing Artificial Neural Networks (ANNs) for this purpose. The authors applied this approach to de-identify personal information within patient notes, demonstrating its effectiveness on text-based data. Their method outperformed the state-of-the-art Conditional Random Field (CRF) systems at the time, achieving superior results in terms of F1 score, precision, and recall metrics.

Liu et al. [5] proposed a hybrid system for de-identifying clinical notes. This ensemble system comprised several individual subsystems: a Bi-directional Long Short-Term Memory (BLSTM) subsystem, a Conditional Random Field (CRF) subsystem, and a rule-based subsystem. The system combined the outputs from these subsystems to generate the final de-identified results.

B. Pipeline system

Early speech de-identification methods often employed a pipeline approach. This involved two main steps:

- Automatic Speech Recognition (ASR): The audio data was first processed using ASR to convert it into text.
- Named Entity Recognition: The generated text was then analyzed using NER techniques to identify and remove sensitive information.

A significant focus of this research area has been on building NER models robust to the inherent noise present in speech recognition outputs. For instance, [6] utilized an SVM-based model for NER on Japanese speech recognition outputs. Similarly, [7] explored the use of Conditional Random Fields (CRFs), demonstrating the robustness of their proposed tree-structured named entities to noisy speech inputs.

To address the challenges posed by out-of-vocabulary (OOV) words, Parada [8] proposed augmenting Named Entity Recognition systems within speech recognition. This approach involves incorporating terms indicative of OOV words into the speech recognition system. Consequently, any OOV word encountered would be treated as part of an OOV region, allowing the NER system to rely on the surrounding context for entity detection.

Ido Cohn et al. [9] investigated the task of removing sensitive information from audio data in various domains, including medical settings. They introduced a novel evaluation metric based on partial value recall and precision (denoted by ρ) and demonstrated that their approach achieved results comparable to text de-identification methods. Szymanski et al. [10] conducted a critical analysis of ASR-NER systems, acknowledging the limitations of both ASR and NER components. They identified shortcomings in traditional evaluation metrics like F1, precision, recall, and accuracy for the task of Named Entity Recognition in speech data. Additionally, they highlighted the scarcity of representative annotated datasets that adequately capture the complexities of spontaneous speech.

C. End-to-end system

Recent advancements in speech de-identification have seen the exploration of end-to-end approaches. In contrast to the traditional pipeline approach, these methods aim to learn both speech recognition and named entity recognition tasks simultaneously within a single neural network architecture.

One such example is presented in [11], where a deep neural network is first trained for the ASR task. Subsequently, the final layer of the network is reinitialized and fine-tuned for the NER task. This allows the model to leverage the learned features from speech recognition during the NER stage.

The progress in speech NER has been fueled by the development of publicly available annotated datasets. Notably, Yadav et al. [12] introduced the first publicly available English speech dataset with NER annotations. This dataset employs special symbols to mark entity boundaries, facilitating the joint optimization of ASR and NER during model training. Similarly, Chen et al. [13] presented a Chinese speech NER dataset. Their experiments demonstrate that combining entity-aware ASR with pre-trained NER taggers can significantly improve speech NER performance.

D. Data augmentation

Real-world language applications face a variety of challenges, including typos, informal language, slang, and diverse sentence structures. These complexities can hinder the performance of models trained on clean, annotated data. Data augmentation tackles this issue by injecting these variations into the training set, fostering robustness and adaptability to different writing styles and potential errors encountered in real-world scenarios.

A chain-of-prompt method is applied to a large language model for document-level relation extraction [14]. A data augmentation method with a sanity-checker is proposed to improve the performance of transformer-based NER models [15]. Graph Propagated Data Augmentation (GPDA) framework [16] is proposed for Named Entity Recognition, which leverage graph propagation to build relationships between labeled data and unlabeled natural texts.

III. Proposed Solution

A. Challenges

Speech processing presents unique challenges compared to written text. Speech is inherently informal, non-standard formats, expressed with slang, contractions, and incomplete sentences. Additionally, disfluencies like ”um” and ”uh” further disrupt the flow of information. These characteristics can significantly hinder the accurate isolation and identification of Personal Identifiable Information within spoken language. Consider these two spoken sentences as examples:

1) *The day before yesterday, Ram received another email from r e m y at outlook dot sg.*
2) *My phone number is (uh) eight five eight two nine three one one.*

In the first sentence, the email address: ***r e m y at outlook dot sg*** is spelled out phonetically, which is a format rarely encountered in written text. Similarly, the second sentence presents challenges for phone number recognition due to the inclusion of the filler word ***(uh)*** and the phone number's representation.

Singapore English (Singlish), while sharing similarities with standard English, possesses distinct characteristics [17]. Singlish grammar can be simpler than standard English, with features like the omission of subject pronouns or the use of "can" to express possibility. Singlish incorporates words and phrases from Malay, Mandarin, Tamil, and other languages spoken in Singapore, creating a unique blend not found in Standard English. Generating synthetic Singapore English data allows for training speech processing models specifically tailored to this unique linguistic environment.

On the other hand, training data for Automatic Speech Recognition and named entity recognition often focus on common identifiers like names, addresses, and phone numbers. This focus can lead to under-representation of less frequent PII types, such as passport numbers. Furthermore, creating large, high-quality speech datasets with labeled PII is difficult due to privacy concerns. Individuals are understandably hesitant to share recordings containing their personal information.

B. Data augmentation

Figure 2 illustrates the proposed data augmentation flow chart. Our solution utilizes a seed dataset specifically crafted to capture the characteristics of Singapore English and the desired PII categories. This seed dataset is then fed into a large language model for the generation of a substantial amount of synthetic data. A custom NER model pre-annotates this synthetic data, identifying potential PII entities. The pre-annotated data undergoes manual review and correction to ensure accuracy, ultimately transforming it into high-quality training data for PII detection. After annotation, this data is used to augment the training data for NER models.

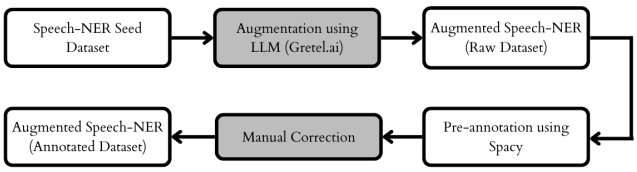


Fig. 2: Proposed data augmentation process

1) Seed data creation: To gain a deeper understanding of the characteristics of Singaporean English, we analyzed spontaneous speech examples extracted from Part 6 of the National Corpus of Singapore (NSC) corpus [18]. The NSC corpus is an open-source speech dataset containing English speech recordings collected in Singapore. This corpus was specifically developed to promote research and development in Singaporean English. Each audio recording is accompanied by human-labeled text transcriptions. We leverage these transcriptions to analyze the characteristics of Singaporean English. Part 6 of the NSC corpus features simulated conversations across six everyday scenarios, including hotel booking, holiday agency interaction, restaurant reservation, banking services, insurance services, and telecommunication services.

Given the focus on common domains in Singapore English conversations, the NSC corpus primarily contains entities such as PERSON, GPE (Geopolitical Entity), and CARDINAL. To address this limitation and enrich our data with a wider range of PII categories, we defined additional PII categories relevant to the Singaporean context. Table I lists the customized entity categories represent PII information under Singapore context.

TABLE I: Customized PII categories

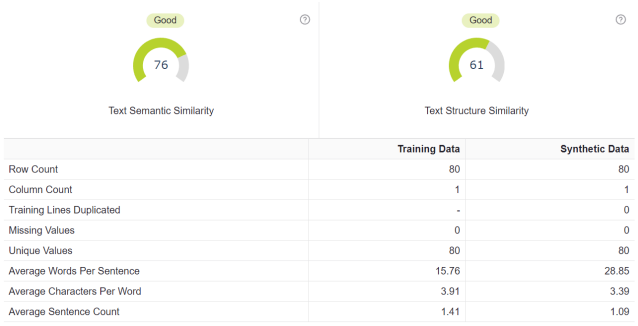
Entity name	**Description**
NRIC	Start and end with alphabet, 7 digits
PASSPORT_NUM	Start with K, 7 digits, 1 alphabet
PHONE	Telephone number, 8 digits, starts with 6, 8 or 9
EMAIL	Email address
CREDIT_CARD	Credit number, 16 digits, first digits is 3,4,5 or 6
CAR_PLATE	Car plate number, S with 2-3 alphabets 4 digits
BANK_ACCOUNT	Singapore bank account number

To create a seed dataset for fine-tuning the large language model, we revised randomly selected sentences from the NSC corpus. We specifically focused on incorporating new PII entities (listed in Table I) without altering the original sentence structures. This seed dataset will be used as input data for large language model fine tune. We hope the fine-tuned model to learn the characteristics of the data and subsequently generate new synthetic data containing these PII entities.

2) Synthetic data generation: This work utilizes the Gretel.ai open-source toolkit [19] for synthetic data generation. Gretel provides an integrated environment to perform fine tune on the pre-trained large language models. The fine-tuning process adjusts the model's parameters to enable the generation of synthetic text that closely resembles the statistical properties of the original data while introducing essential variations. Gretel offers a diverse selection of language model structure, for example LSTM, ACTGAN, and GPT. We opted for the Gretel-GPT structure, the pre-trained large language model we used is Mistral-7B-Instruct-v0.2[1].

Figure 3 shows a sample report generated by the Gretel.ai synthetic text quality evaluation tool. Gretel assesses the quality of synthetic text using a combination of two metrics: Text Semantic Similarity and Text Structure Similarity. The Text Semantic Similarity Score (0-100) indicates how closely the meaning of the synthetic text aligns with the original data. The Text Structure Similarity Score measures how well the sentence length, average words per sentence, and character distribution within the synthetic data mirror the original dataset.

[1]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2



	Training Data	Synthetic Data
Row Count	80	80
Column Count	1	1
Training Lines Duplicated	-	0
Missing Values	0	0
Unique Values	80	80
Average Words Per Sentence	15.76	28.85
Average Characters Per Word	3.91	3.39
Average Sentence Count	1.41	1.09

Fig. 3: Synthetic data generation report

3) Synthetic annotation and correction: We further refine the synthetic data generated by Gretel using a semi-automatic approach. First, a custom NER model is trained on the base spaCy[2] model and our seed data, which is rich in PII annotations. This customized model then annotates named entities within the synthetic data.

Following the automatic annotation, human experts review the output to identify and correct any errors. Through this process, we obtain a high-quality dataset containing PII information with accurate annotations.

The augmented dataset may serve as a new seed dataset for an iterative data augmentation process. This iterative approach involves using the augmented data to further refine the NER model, which in turn improves its ability to identify PII entities. This cycle can be repeated to progressively enhance model performance.

4) Evaluation metric: we argue that traditional named entity recognition evaluation metrics, such as F1 score, may not be sufficient for analyzing the specific errors encountered during entity detection on spontaneous speech data. Here's some reason on why F1 score might not be ideal for PII detection tasks:

1) Equal weighting of precision and recall: F1 score treats precision and recall equally. In PII detection tasks, the relative importance of each metric can vary depending on the specific context and risk tolerance.
2) Insensitivity to class imbalance: F1 score doesn't account for the frequencies of different entity types. Accurately detecting some entities might be more critical than others. For instance, in PII detection, a credit card number maybe more critical than email.
3) Limited granularity: F1 score is an aggregate measure that lacks details about specific error types.

We propose a set of alternative evaluation metrics specifically designed to provide a more granular analysis of named entity recognition model performance in identifying PII entities from speech data. These metrics offer a deeper understanding of error types compared to the conventional F1 score approach.

- Perfect (%): The percentage of entities correctly identified without any errors.
- Partial (%): The percentage of entities where some tokens are correctly tagged but others are not.
- Substitution (%): The percentage of entities where the model misidentified the entity type.
- Hallucination (%): The percentage of entities incorrectly identified which are not present in the ground truth data.
- Omission (%): The percentage of entities the model fails to recognize altogether.

[2]https://spacy.io/api/entityrecognizer/

IV. Experiments and Analysis

A. Experiment setup

1) Seed Data: Building upon the sample data extracted from NSC samples, we enriched it with diverse PII entities based on patterns defined in Table I. This revised dataset, containing 992 sentences, was divided using an 80/20 split. Eighty percent (796) of the data serves as the seed data for the data augmentation process, while the remaining twenty persent (196) is reserved for testing. An example sentence is illustrated in Figure 4, where the highlighted PIIs were synthesized by human expert using these defined patterns.

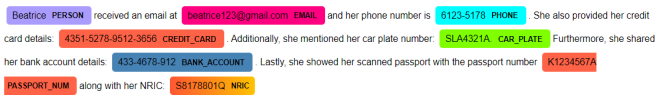


Fig. 4: Sample seed data

2) Synthetic Data: To capture the nuances of spoken Singapore English, we leverage a large language model (LLM) trained on the seed data. This training process allows the LLM to learn the characteristics and style of spoken Singapore English. Subsequently, the fine-tuned LLM can generate new sentences that mimic this natural language style.

Specifically, we employed the mistralai/Mistral-7B-Instruct-v0.2 model as the base architecture. During fine-tuning, a batch size of 4 was used for 3 epochs with a weight decay of 0.01 and a warmup step of 100. The initial learning rate was set to 0.0002 and a linear learning rate scheduler was adopted. AdamW was employed for optimizer updates.

Following fine-tuning, the Gretel tool was employed to generate 2,000 synthetic sentences. These synthetic sentences effectively captured the characteristics of the seed dataset, resulting in new sentences that mirrored the original data's style. These synthetic sentences were then subjected to a two-step annotation process. First, they underwent automatic annotation, followed by human expert review to refine the PII annotations.

B. Experiment results and analysis

The refined synthetic data enriched with PII annotations is then employed as the training set for a named entity recognition model, denoted as Augmented model. For comparison, the NER model trained on seed data only base model is denoted as Base Model. To evaluate the effectiveness of the data augmentation approach, both models are assessed on the hold-out test set.

Table II reports the performance of Base model and Augment model using the conventional F1 measure. As expected, incorporating augmented data leads to an overall improvement in NER performance as reflected by the F1 score.

TABLE II: PII detection results with F1 measure

Entity/F1	Base Model %	Augmented Model %
NRIC	0.72	0.69
EMAIL	0.66	0.88
CREDIT_CARD	0.74	0.70
PHONE	0.67	0.80
PASSPORT_NUM	0.61	0.85
CAR_PLATE	0.42	0.80
BANK_ACCOUNT	0.9	0.79
Overall	0.67	**0.80**

To gain deeper insights into the specific errors made during detection, we employ the alternative evaluation metrics proposed in the previous section. Table III presents the NER results for the model trained solely on the seed data, while Table IV reports the results for the model trained on the augmented data.

TABLE III: PII detection results on base model

Entity	Perfect %	Partial %	Substitution %	Hallucination %	Omission %
NRIC	77.27	0.00	0.00	24.00	22.73
EMAIL	57.14	14.29	5.71	3.85	22.86
CREDIT_CARD	90.91	0.00	9.09	6.25	0.00
PHONE	60.71	10.71	17.86	13.04	10.71
PASSPORT_NUM	57.89	21.05	10.53	11.76	10.53
CAR_PLATE	36.84	21.05	0.00	14.29	42.11
BANK_ACCOUNT	86.67	0.00	13.33	0.00	0.00
Overall	66.77	**9.59**	8.07	**73.19**	**15.56**

TABLE IV: PII detection results on data augmented model

Entity	Perfect %	Partial %	Substitution %	Hallucination %	Omission %
NRIC	54.55	0.00	31.82	7.69	13.64
EMAIL	80.00	0.00	2.86	3.45	17.14
CREDIT_CARD	63.64	0.00	27.27	0.00	9.09
PHONE	85.71	0.00	7.14	3.13	7.14
PASSPORT_NUM	89.47	0.00	0.00	0.00	10.53
CAR_PLATE	73.68	0.00	10.53	12.50	15.79
BANK_ACCOUNT	73.33	6.67	20.00	0.00	0.00
Overall	**74.34**	0.95	**14.23**	26.77	10.48

1) Overall analysis: Comparing the NER results of two models, we have the following observations:

- Improved entity recognition accuracy: The augmented model achieves a significantly higher average perfect match rate (74.34%) compared to the base model (66.77%). This indicates that the augmented model is more adept at accurately identifying entities, precisely matching the annotations in the gold standard data.
- Reduced hallucination: The augmented model exhibits a substantially lower average hallucination rate (0.71%) compared to the base model (2.14%). This reduction in hallucinations suggests that the augmented model is less prone to identifying non-existent entities, leading to more reliable entity recognition.
- Reduced omission rate: The augmented model exhibits a lower average Omission rate (10.48%) compared to the base model (15.56%). This decrease indicates the augmented model Augmented Model is more successful in identifying entities, leading to a more comprehensive recognition performance.
- Improved entity completeness: The average partial match metric is significantly lower in the augmented model (0.95%) than the base model (9.59%). This reduction suggests the augmented model is more likely to correctly identify entire entities, rather than just a portion of them.

2) Trade-Offs and Entity-Specific Performance: While the augmented model exhibits overall improvements, a closer look reveals some trade-offs and entity-specific performance variations that provide valuable insights.

- NRIC and CREDIT_CARD entities: The base model achieves a higher perfect match rate for NRIC (77.27%) and CREDIT_CARD (90.91%) entities compared to the augmented model. However, this seemingly higher accuracy is likely inflated by a higher hallucination rate in the Base Model. This means the model might be incorrectly tagging non-existent entities, leading to misleading Perfect match scores.
- EMAIL and PHONE entities: The augmented model demonstrates significantly improved perfect match rates for EMAIL (80.00%) and PHONE (85.71%) entities compared to the base model. These improvements suggest that the data augmentation techniques effectively help the model generalization.
- Balanced performance across entities: Despite some increase in the substitution rate (misidentified entities), the augmented model demonstrates a more balanced performance across different entity types. This is achieved by reducing both false positives (hallucinations) and false negatives (omissions).

In summary, the augmented model outperforms the base model across key metrics, primarily due to its higher overall perfect match rate, lower hallucination rate, and lower omission rate. While the base model achieves higher perfect match rates for specific entities, these scores are likely inflated by false positives due to hallucinations. The augmented model's improvements reflect a more accurate and reliable approach to NER in oral transcripts, highlighting the effectiveness of the data augmentation strategies employed.

V. Conclusions and Future Works

In conclusion, this work addressed the challenge of limited real-world speech data for speech de-identification, particularly for Singaporean English. We proposed a two-step approach that leverages synthetic data generation and semi-automatic PII annotation. A large language model was employed to create synthetic spoken text, and a customized

NER model was used in an iterative process to annotate PII entities within this data. The experimental results confirmed the effectiveness of this data augmentation strategy, demonstrating its positive impact on NER performance. Our analysis using alternative evaluation metrics provided deeper insights into the specific errors made during NER, enabling a more comprehensive understanding of model performance. This work paves the way for the development of more robust speech de-identification systems for Singaporean English.

In the future works, we would like to explore alternative data augmentation techniques suitable for speech de-identification task. We will also explore the following directions: speech generation from synthetic text Data, speech de-identification system development and evaluation.

ACKNOWLEDGMENT

This work is supported by Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant R-R12-A405-0009 and Ignition grant R-IE2-A405-00006.

REFERENCES

[1] Héctor Delgado et al. "ASVspoof 5 Evaluation Plan". In: 2024.

[2] Natalia Tomashenko et al. "The VoicePrivacy 2024 Challenge Evaluation Plan". In: *arXiv preprint arXiv:2404.02677* (2024).

[3] Martin Flechl et al. "End-to-end speech recognition modeling from de-identified data". In: *INTERSPEECH*. 2022.

[4] Franck Dernoncourt et al. "De-identification of patient notes with recurrent neural networks". In: *Journal of the American Medical Informatics Association : JAMIA* 24.3 (2017), pp. 596–606. DOI: 10.1093/jamia/ocw156. URL: https://doi.org/10.1093/jamia/ocw156.

[5] Zengjian Liu et al. "De-identification of clinical notes via recurrent neural network and conditional random field". In: *Journal of biomedical informatics* 75 (2017), S34–S42.

[6] Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. "Incorporating speech recognition confidence into discriminative named entity recognition of speech data". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 2006, pp. 617–624.

[7] Christian Raymond. "Robust tree-structured named entities recognition from speech". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 8475–8479.

[8] Carolina Parada, Mark Dredze, and Frederick Jelinek. "OOV sensitive named-entity recognition in speech". In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[9] Ido Cohn et al. "Audio De-identification - a New Entity Recognition Task". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 197–204. DOI: 10.18653/v1/N19-2025.

[10] Piotr Szymański et al. "Why Aren't We NER Yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1746–1761. DOI: 10.18653/v1/2023.acl-long.98.

[11] Sahar Ghannay et al. "End-to-end named entity and semantic concept extraction from speech". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 692–699.

[12] Hemant Yadav et al. "End-to-end named entity recognition from english speech". In: *arXiv preprint arXiv:2005.11184* (2020).

[13] Boli Chen et al. "Aishell-ner: Named entity recognition from chinese speech". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 8352–8356.

[14] Qi Sun et al. "Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction". In: *Proceedings of the ACM on Web Conference 2024*. 2024, pp. 4407–4416.

[15] Aluru VNM Hemateja et al. "Novel data augmentation for named entity recognition". In: *International Journal of Speech Technology* 26.4 (2023), pp. 869–878.

[16] Jiong Cai et al. "Improving Low-resource Named Entity Recognition with Graph Propagated Data Augmentation". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2023, pp. 110–118.

[17] *Wikipedia - Singapore English*. https://simple.wikipedia.org/wiki/Singapore. 2024 (accessed May 20, 2024).

[18] Jia Xin Koh et al. "Building the singapore english national speech corpus". In: *Malay* 20.25.0 (2019), pp. 19–3.

[19] *Gretel Synthetics*. https://github.com/gretelai/gretel-synthetics. 2024 (accessed May 20, 2024).