Leveraging Cross-Lingual Transfer Learning in Spoken Named Entity Recognition Systems

Moncef Benaicha[1], David Thulke[2], M. A. Tuğtekin Turan[1]

[1]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany
[2]Machine Learning and Human Language Technology, RWTH Aachen University, Germany

Correspondence: moncef.benaicha@rwth-aachen.de

Abstract

Recent Named Entity Recognition (NER) advancements have significantly enhanced text classification capabilities. This paper focuses on spoken NER, aimed explicitly at spoken document retrieval, an area not widely studied due to the lack of comprehensive datasets for spoken contexts. Additionally, the potential for cross-lingual transfer learning in low-resource situations deserves further investigation. In our study, we applied transfer learning techniques across Dutch, English, and German using both pipeline and End-to-End (E2E) approaches. We employed Wav2Vec2 XLS-R models on custom pseudo-annotated datasets to evaluate the adaptability of cross-lingual systems. Our exploration of different architectural configurations assessed the robustness of these systems in spoken NER. Results showed that the E2E model was superior to the pipeline model, particularly with limited annotation resources. Furthermore, transfer learning from German to Dutch improved performance by 7% over the standalone Dutch E2E system and 4% over the Dutch pipeline model. Our findings highlight the effectiveness of cross-lingual transfer in spoken NER and emphasize the need for additional data collection to improve these systems.

1 Introduction

Named Entity Recognition (NER) identifies and classifies named entities within text, including persons, organizations, locations, and other predefined categories (Nadeau and Sekine, 2007). While substantial progress has been made in extracting entities from written text, adapting these techniques to spoken content has seen limited research. This is primarily due to the unique challenges associated with spoken text analysis (Tomashenko et al., 2019).

_____

This work was carried out while the first author was a research assistant at Fraunhofer IAIS.

Improved spoken NER has significant implications for various practical applications. Accurate entity recognition enhances user interaction in voice assistants by correctly identifying and responding to queries (Jehangir et al., 2023). In automatic transcription services, better NER improves the quality of transcriptions by correctly tagging entities, which is crucial for generating accurate and searchable text (Szymański et al., 2023). Spoken dialogue systems in customer service and virtual agents also benefit from enhanced NER by providing more context-aware and accurate responses.

Exploring spoken NER involves challenges due to the unpredictable nature of spoken language. Variabilities in pronunciation, speech disfluencies, and background noise present significant obstacles and, therefore, negatively impact system performance (Porjazovski et al., 2021). Additionally, the continuous flow of spoken language, with unclear word boundaries, adds complexity to the task (Chen et al., 2022). Despite these difficulties, spoken NER holds significant interest for its potential applications in areas like voice assistants, automatic transcription services, and spoken dialogue systems (Ghannay et al., 2018; Haghani et al., 2018; Serdyuk et al., 2018).

The introduction of Transformer-driven methodologies has advanced this field significantly. Notably, the End-to-End (E2E) modeling approach directly links speech patterns to transcriptions with embedded entity markers, which show promising results (Mdhaffar et al., 2022). These models effectively map temporal dependencies and manage the complexities of various spoken dialects. However, the research primarily focuses on high-resource languages like English, resulting in less effective model performance in data-scarce scenarios.

This paper addresses the issue of linguistic disparity by exploring cross-lingual transfer learning for spoken NER. We focus on using multilingual

language representation models to evaluate their effectiveness, especially in data-scarce environments where this term refers to the limited availability of high-quality, manually annotated datasets in comparison to more extensively studied languages like English. Our empirical studies cover three main languages: Dutch, English, German. We specifically examine transfers between languages with different resource levels, highlighting the strength of transfer learning in scenarios from zero to low resources. These languages were chosen for our study due to their varying resource levels and linguistic similarities, which provide a meaningful context for examining cross-lingual transfer learning.

Moreover, we provide a comparative analysis of different methodologies, including both pipeline and E2E approaches. The pipeline framework integrates the functionalities of Automatic Speech Recognition (ASR) systems with subsequent NER models. Initially, the ASR system transcribes spoken content into text, which is then tagged with entities by the NER system. Previous work has explored E2E strategies that simultaneously address ASR and NER tasks (Caubrière et al., 2020). This approach aims to refine ASR alongside Natural Language Understanding (NLU) and significantly reduce error propagation commonly caused by ASR limitations (Jannet et al., 2015).

A major challenge with E2E models is their need for extensive training datasets. The limited availability of audio-textual datasets with entity annotations emphasizes this issue, as creating large annotated speech datasets is both complex and costly. To address this, Pasad et al. (2022), used a labeling model to generate pseudo-annotations. Inspired by this approach, our paper includes custom pseudo-annotated datasets in Dutch, English, and German, created using the XLM-R$_L$-based NER model (Goyal et al., 2021). These pseudo-annotations were not manually corrected, and all evaluations were performed on these pseudo-annotated datasets. Furthermore, no gold-standard annotations were used.

We further examine the impact of various factors such as training data volume, language model choice, and target language, on spoken NER system performance. This paper enhances research in spoken NER by highlighting challenges and potential cross-lingual solutions. Our results advance spoken document retrieval and support the development of more sophisticated and accurate spoken NER

systems. Consistent with open research principles, all code, data, and results from this study will be publicly available[1].

To summarize our contributions:

- We comprehensively compare and analyze pipeline versus E2E strategies for spoken NER in Dutch, English, and German.

- We investigate transfer learning within both pipeline and E2E strategies for spoken NER.

- We move from a high-resource language, German, to a resource-scarce language, Dutch, resulting in a notable 10% improvement in spoken NER performance.

2 Related Work

Traditionally, NER from spoken content has utilized a pipeline approach, beginning with an Automatic Speech Recognition (ASR) phase followed by NER on the resulting transcriptions (Jannet et al., 2017). While such a system may seem intuitive, it has essential challenges. Specifically, it directly incorporates transcriptions annotated with entities within the ASR system (Cohn et al., 2019). By embedding such annotations, there's potential to refine the partial hypotheses, which often get overlooked or dismissed during the decoding process. A novel solution has been integrating specific entity expressions into the lexicon to enhance language model accuracy in recognizing these expressions (Hatmi et al., 2013).

In response to these challenges, interest in the E2E approach for spoken NER has increased. This method aims to simultaneously optimize ASR and NER processes, providing a potentially more efficient alternative to the traditional pipeline by leveraging deep neural networks' capabilities to manage long-range sentence dependencies.

Significant research into the E2E approach for spoken NER includes work with French datasets (Ghannay et al., 2018), adopting architectures similar to DeepSpeech (Amodei et al., 2016), guided by the Connectionist Temporal Classification (CTC) objective (Graves et al., 2006). Building on this, Yadav et al. (2020) developed a method tailored for English, introducing specific tokens in the ASR vocabulary to enhance NER tagging. Our research also incorporates this by incorporating unique symbols ('{', '[', '$', ']') in transcripts to

---

[1] https://github.com/moncefbenaicha/spoken-ner

assist in identifying entities such as organizations, persons, and locations, as depicted in Figure 1. Building on these foundational studies, more recent research (Shon et al., 2022; Pasad et al., 2022) has successfully employed this methodology in conjunction with the Wav2Vec2 model.

Consequently, the E2E strategy aims to align speech utterances with annotated transcriptions perfectly, facilitating direct entity extraction from spoken content. Empirical studies using French and English datasets have demonstrated the effectiveness of the E2E strategy, often surpassing traditional pipeline approaches, particularly against Long Short-term Memory (LSTM) based models, which no longer meet state-of-the-art standards (Vajjala and Balasubramaniam, 2022).

3 Methodology

3.1 Baseline Models

This paper introduces two distinct baseline systems designed for English, German, and Dutch. The first system adopts a conventional pipeline approach where ASR and NER models are trained separately. After training, these models are integrated during the inference phase to produce results. In contrast, the second system employs an E2E methodology involving more intricate processes. For the E2E system, the ASR model is fine-tuned using the robust pre-trained Wav2Vec2-XLS-R-300M model (Babu et al., 2022). During this fine-tuning, each language is treated separately, utilizing CTC-loss
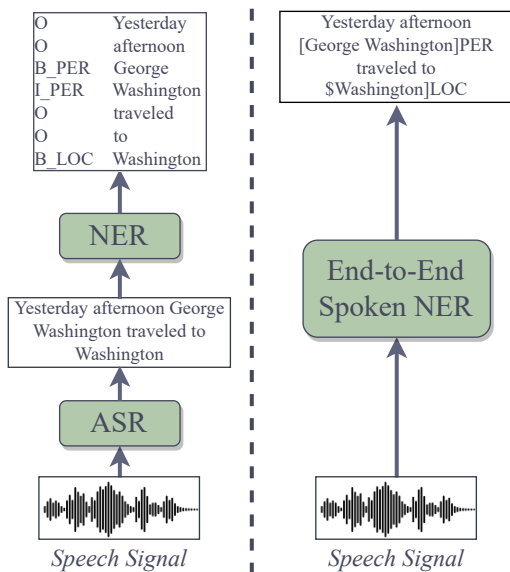
as the main objective function. Concurrently, the NER component of the pipeline is enhanced through modifications to the XLM-R_L language representation model. This enhancement includes the addition of a linear layer specifically designed to handle lower-cased tokens from the CoNLL 2002 and 2003 datasets (Tjong Kim Sang; Tjong Kim Sang and De Meulder, 2003).

To enhance the ASR model, we integrate a 4-gram language model trained on both the training and development datasets for each language. In the E2E system, we utilize the capabilities of the Wav2Vec2-XLS-R-300M model once more. For this iteration, we fine-tune using a specially augmented corpus. This corpus includes special tokens that explicitly indicate the start of an entity such as a person (PER), an organization (ORG), or a location (LOC) and mark the end of an entity mention, as shown in Figure 1.

3.2 Transfer Learning Models

In the pipeline approach, we leverage a dual-component architecture. Specifically, we replace the native NER model with a cross-lingual variant that is more universally applicable. In our transfer learning experiments, German serves as the source language, while English and Dutch are designated as target languages. The first step involves applying the pre-trained German NER model directly to the English and Dutch transcripts without any language-specific modifications. This initial approach is followed by a phase that combines transfer learning with model fine-tuning. During this phase, the German NER model forms the foundation. Fine-tuning is then performed using a subset (k) of the target language's training set, corresponding to either 10% or 20% of the original dataset. The model's performance is assessed using the test set of the target language to evaluate its effectiveness after fine-tuning.

In contrast, the E2E model adopts a more comprehensive transfer learning approach. Our experiments begin with a zero-shot transfer learning phase, where the capabilities of the German E2E spoken NER system are extended to English and Dutch. Following this initial phase, we apply an extensive fine-tuning using a portion of the target language's training dataset, typically 20% or 40% of the total. After completing this fine-tuning, the model is assessed against the test set of the target language. This evaluation yields critical insights into the effectiveness of the fine-tuning process.

Figure 1: The diagram on the left shows a two-stage pipeline, while the right shows the E2E approach.

	Training Set			Validation Set			Test Set		
	EN	**DE**	**NL**	**EN**	**DE**	**NL**	**EN**	**DE**	**NL**
#Sentences	527K	526K	42K	5K	5K	5K	5K	5K	5K
#Tokens	5.4M	5M	0.5M	46K	47K	46K	47K	45.6K	45K
#Tokens as LOC	211K	156K	588	1.2K	1.5K	241	1.3K	1.3K	314
#Tokens as ORG	177K	83K	95	1K	729	66	1.1K	621	88
#Tokens as PER	216K	144K	104	1.3K	1.5K	123	1.4K	1.2K	155
#Tokens as O	4.8M	4.6M	408K	43K	43K	45K	43K	42.4K	44.5K
Total Hours	840	839	54	8	8.5	6.5	8.5	8.5	7

Table 1: The statistics of pseudo-annotated data across English, German, and Dutch splits.

4 Corpus Overview

For our experiments, we utilized data from Common Voice[2], specifically selecting the validated corpus for each language. This corpus comprises various types of oral data, including read speech from diverse demographics. During preprocessing, we removed duplicate entries, retained essential punctuation, and converted non-Latin characters to Latin script, standardizing everything to lowercase. The processed data were then used with the XLM-R_L-based NER model, trained on the language-specific CoNLL dataset (Tjong Kim Sang; Tjong Kim Sang and De Meulder, 2003), to generate pseudo-annotations. Note that, in the English corpus, all punctuation except apostrophes was removed during the transcription process.

We encountered non-Latin scripts such as Cyrillic or Brahmic during preprocessing due to the inclusion of multilingual text data in the Common Voice corpus. These scripts were converted to Latin to maintain consistency across datasets. Detailed statistics of the training, development, and testing sets, including the number of sentences, tokens, and total hours, are presented in Table 1. This table illustrates the comprehensive scope and linguistic variability of the datasets used.

Additionally, Table 2 highlights the extent of entity overlaps within and between the languages studied. These were computed by comparing the exact match of entity spans across different languages. While overlaps within each language's training sets are complete, the inter-language overlaps vary significantly. For instance, the overlap between English and German is 36.5%, contrasting sharply with the minimal 0.1% overlap between

English and Dutch. These statistics underscore the challenges and considerations in developing multilingual NER systems that can effectively transfer learning across languages.

5 Evaluation Metrics

We employ various metrics to assess the performance of our spoken NER systems. The conventional Word Error Rate (WER) serves as the primary metric for evaluating the accuracy of ASR models. In addition, we utilize the Entity Error Rate (EER), which evaluates the specific accuracy of our spoken NER systems in the context of entity transcription. Unlike broader measures, EER focuses exclusively on the accuracy with which entities such as names of people, locations, and organizations are transcribed:

$$EER = \frac{N_{\text{Incorrectly Transcribed Entities}}}{N_{\text{Total Entities}}} \quad (1)$$

Here, $N_{\text{Incorrectly Transcribed Entities}}$ denotes the number of entities that the system has transcribed

		Train		
		EN	**DE**	**NL**
Train	**EN**	100.0	24.7	0.1
	DE	36.5	100.0	0.2
	NL	52.4	58.0	100.0
Test	**EN**	79.6	64.0	1.7
	DE	58.4	84.3	2.1
	NL	64.3	68.1	27.0

Table 2: The percentages of entity overlaps across different languages.

incorrectly, while $N_{\text{Total Entities}}$ indicates the total number of entities in the dataset. The EER measures the proportion of entities that were inaccurately transcribed, providing a direct indicator of the system's transcription accuracy.

Furthermore, to evaluate the effectiveness of the NER system, we calculate the micro-average F1-score, which is a harmonized measure of precision and recall, similar to the approach in (Pasad et al., 2022):

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad (2)$$

where true positives (TP) are the correctly identified entities, false positives (FP) are the incorrectly identified entities, and false negatives (FN) are the entities that were not identified. These values are computed based on the accuracy of entity transcription, their types, and the positions of these entities within the transcript. This score provides a balanced measure of the system's overall accuracy in recognizing and classifying entities.

6 Experiments and Results

In our comprehensive baseline experiments for the pipeline model, we utilized prominent open-source pre-trained models to enhance performance across various tasks. For ASR, we selected the Wav2Vec2-XLS-R-300M pre-trained model[3], which is noted for its efficiency. Simultaneously, for NER, we employed the robust XLM-R$_{\text{L}}$ pre-trained language model[4], well-known for its effectiveness in multilingual processing.

For fine-tuning the ASR model, we chose the AdamW optimizer, following recommendations by Loshchilov and Hutter (2019), with $betas = (0.9, 0.999)$ and a precision parameter of $eps = 10^{-8}$. The learning rate was carefully managed through a schedule that includes a warm-up phase covering one-third of the total training steps. During this phase, the learning rate gradually increases from a baseline to a peak of 10^{-4}. After this warm-up period, the learning rate linearly decreases through the remaining training sessions. It is important to highlight that we only froze the feature encoder of the Wav2Vec2 architecture and made no other structural changes to the model.

Regarding the NER model, we mostly maintained the same optimizer settings as those used for

the ASR model, with a slight adjustment to the maximum learning rate of the scheduler, setting it to 2×10^{-5}. Finally, for the E2E spoken NER model, we chose to maintain consistency by adopting the same set of hyper-parameters previously applied in the ASR component of pipeline framework.

6.1 Baseline Results

Table 3 presents the baseline performances across varying configurations and languages. It's evident from the results that the EER and the F1-scores consistently display a negative correlation across all three languages. This observed trend aligns with our expectations. Specifically, even if the ASR system exhibits any discrepancies in transcribing an entity accurately, it adversely affects the F1-score. This impact remains regardless of whether the entity boundaries and categories are delineated correctly.

Focusing on the German language, the E2E baseline results stand out. Out of a total of 5000 test utterances for German, 2928 have been transcribed with precision, translating to an accuracy rate of approximately 58.6% of the test set. Impressively, within this accurately transcribed set, a staggering 94% of utterances have been labeled correctly, a testament to the model's efficacy. When we pivot our attention to the subset of utterances that weren't transcribed with utmost accuracy, it's important to note that the model, despite the transcription issues, managed to label entities correctly in about 51.3% of these instances. Cumulatively, this implies that the E2E model adeptly labeled entities with accuracy in nearly 76% of the entire set of test utterances for German.

It's also worthwhile to note the relative performances of English and Dutch. For instance, when observing the pipeline approach with both ASR and NER components active, English has a WER of 16.7% and an F1-score of 40.7%, whereas Dutch, under the same conditions, recorded a WER of 9.3% and an F1-score of 40.0%. Such statistics offer nuanced insights into the distinct challenges and variances inherent to each language, emphasizing the importance of tailored strategies for each linguistic domain.

A comparable trend is also observed in the pipeline system, with the E2E model exhibiting a marginally superior performance. One of the E2E model's significant strengths lies in its ability to tag entities accurately, even in the presence of transcription errors. This proficiency translates to

---

[3]https://hf.co/facebook/wav2vec2-xls-r-300m
[4]https://hf.co/xlm-roberta-large

Lang	E2E	ASR	NER	WER	EER	F1
EN	no	no	yes	N/A	N/A	80.3
	no	yes	yes	16.7	48.0	40.7
	yes	no	no	16.5	46.0	41.8
DE	no	no	yes	N/A	N/A	87.2
	no	yes	yes	9.4	29.0	61.1
	yes	no	no	9.1	27.0	61.6
NL	no	no	yes	N/A	N/A	85.6
	no	yes	yes	9.3	49.0	40.0
	yes	no	no	9.2	47.0	37.4

Table 3: Performances of baseline models across different configurations and languages.

a performance uptick of around 2% when pitted against the traditional pipeline system. Exploring the performances across different languages, our analyses of English and Dutch mirror the patterns we uncovered for German. However, the Dutch distinguishes itself in one noteworthy aspect: an impressive accuracy rate of up to 99% in entity identification when the transcriptions are accurate. Across the board, the E2E spoken NER system demonstrates a propensity to fine-tune the WER, EER, and by extension, the F1-score. The sole deviation from this pattern is seen in Dutch. This anomaly can be attributed to the relatively limited pool of training data available for Dutch, especially when benchmarked against English and German.

6.2 Transfer Learning Results

Tables 4 and 5 present the results of transfer learning from German to English and German to Dutch, respectively. These results highlight patterns consistent with our baseline pipeline experiments.

In the German-to-English transition detailed in Table 4, the performance of E2E model in the zero-shot transfer learning scenario illustrates the adaptability of this approach. However, for the transition from German to Dutch, the situation is markedly different. Zero-shot transfer learning outcomes for German to Dutch align closely with baseline performances, with WER, EER, and F1-scores remaining relatively stable across the scenarios. Interestingly, the incorporation of 40% of the Dutch training data leads to a noticeable improvement in performance, particularly a roughly 2% increase in F1-scores as shown in Table 5.

Focusing again on the E2E model, it is apparent from the results that the German-to-Dutch transfer yields better performance metrics compared to the German-to-English transfer. An examination of Table 2 provides a potential explanation for this difference, indicating a more substantial overlap in entities between German and Dutch than between German and English.

A closer look at Table 5 reveals that fine-tuning the German E2E system with 40% of the Dutch training data significantly enhances the system's effectiveness in recognizing Dutch entities. This fine-tuning results in a performance increase of approximately 7% compared to the standalone Dutch E2E system and a 4% improvement over the Dutch pipeline system. The gains are particularly notable in the F1 scores within the PER and LOC entity categories, where there is an impressive 10% increase compared to the baseline Dutch E2E system. These findings underscore the efficacy of targeted training data in boosting system performance and highlight the benefits of cross-lingual transfer learning in multilingual NER systems.

7 Conclusions

In this paper, we explore spoken NER with a focus on cross-lingual transfer learning, employing both pipeline and E2E methodologies. Our findings indicate that the E2E approach to spoken NER generally outperforms the pipeline method in terms of both diverse evaluation metrics and overall parameter efficiency. Nevertheless, the pipeline approach retains its practical utility due to its

System	Transfer Learning		WER	EER	F1
	Source → Target	k			
Pipeline	N/A	N/A	16.7	48.0	40.7
	DE → EN	0%	16.7	48.0	38.5
		20%	16.7	48.0	39.6
		40%	16.7	48.0	40.0
E2E	N/A	N/A	16.5	46.0	41.8
	DE → EN	0%	52.5	66.0	20.8
		20%	21.8	53.0	35.8
		40%	22.9	54.0	35.8

Table 4: Performance of the pipeline and E2E models with German-to-English transfer learning, measured across various metrics.

System	Transfer Learning		WER	EER	F1
	Source → Target	k			
Pipeline	N/A	N/A	9.3	49.0	40.0
	DE → NL	0%	9.3	49.0	40.2
		20%	9.3	49.0	40.3
		40%	9.3	49.0	42.0
E2E	N/A	N/A	9.2	47.0	37.4
	DE → NL	0%	78.0	67.0	24.0
		20%	11.8	42.0	44.4
		40%	10.7	40.0	44.3

Table 5: Performance of the pipeline and E2E models with German-to-Dutch transfer learning, measured across various metrics.

flexibility in integrating various ASR and NER components.

Our investigations show that deploying a German NER model without fine-tuning in a Dutch or English context within the pipeline still allows the E2E spoken NER to achieve comparable or superior results to an NER model trained specifically for the pipeline's target language. This highlights the effectiveness of transfer learning in E2E spoken NER systems, which often surpass the performance of traditional pipeline systems. A key insight from our study is the robustness of the E2E model in tagging entities correctly, even when faced with transcription errors, slightly outperforming the pipeline approach.

Looking ahead, several promising directions for further research have emerged. One potential area involves refining the objective function of the ASR model to enhance focus on specific tokens within transcriptions that are of greater relevance to NER tasks. Another promising direction is the investigation of spoken NER within a multilingual framework that can accommodate a wide range of languages and dialects, potentially making significant advancements in the field. Additionally, creating and using human-annotated datasets, with consistent entity annotations across various languages, are crucial. We develop human-annotated datasets where such datasets would provide a solid foundation for evaluating spoken NER systems.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep Speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning (ICML)*, pages 173–182.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Language Resources and Evaluation Conference*, pages 4514–4520.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. AISHELL-NER: Named entity recognition from Chinese speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8352–8356. IEEE.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification-a new entity recognition task. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 197–204.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Naman Goyal, Jingfei Du, Myle Ott, Giri Ananthara-man, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 29–33. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International conference on machine learning (ICML)*, pages 369–376.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meignier. 2013. Incorporating named entity recognition into the speech transcription process. In *Annual Conference of the International Speech Communication Association (Interspeech)*.

Mohamed Ameur Ben Jannet, Olivier Galibert, Martine Adda-Decker, and Sophie Rosset. 2015. How to evaluate ASR output for named entity recognition? In *Annual Conference of the International Speech Communication Association (Interspeech)*.

Mohamed Ameur Ben Jannet, Olivier Galibert, Martine Adda-Decker, and Sophie Rosset. 2017. Investigating the effect of asr tuning on named entity recognition. In *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2486–2490.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Salima Mdhaffar, Jarod Duret, Titouan Parcollet, and Yannick Estève. 2022. End-to-end model for named entity recognition from speech without paired training data. In *Annual Conference of the International Speech Communication Association (Interspeech)*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Ankita Pasad, Felix Wu, Suwon Shon, Karen Livescu, and Kyu Han. 2022. On the use of external data for spoken named entity recognition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 724–737.

Dejan Porjazovski, Juho Leinonen, and Mikko Kurimo. 2021. Attention-based end-to-end named entity recognition from speech. In *International Conference on Text, Speech, and Dialogue*, pages 469–480. Springer.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we ner yet? artifacts of asr errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1746–1761.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Natural Language Learning at HLT-NAACL*, pages 142–147.

Natalia Tomashenko, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2019. Recent advances in end-to-end spoken language understanding. In *International Conference on Statistical Language and Speech Processing*, pages 44–55. Springer.

Sowmya Vajjala and Ramya Balasubramaniam. 2022. What do we really know about state of the art NER? In *Proceedings of the Language Resources and Evaluation Conference*, pages 5983–5993.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from English speech. *Annual Conference of the International Speech Communication Association (Interspeech)*, pages 4268–4272.