Why Aren't We NER Yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts

Piotr Szymański and **Łukasz Augustyniak** and **Adrian Szymczak**
Wroclaw University of Science and Technology, Poland
{piotr.szymanski,lukasz.augustyniak,adrian.szymczak}@pwr.edu.pl

Mikołaj Morzy
Poznan University of Technology Poland
mikolaj.morzy@put.poznan.pl

Krzysztof Surdyk

Piotr Żelasko
Meaning.Team Inc, USA
pzelasko@meaning.team

Abstract

Transcripts of spontaneous human speech present a significant obstacle for traditional NER models. The lack of grammatical structure of spoken utterances and word errors introduced by the ASR make downstream NLP tasks challenging. In this paper, we examine in detail the complex relationship between ASR and NER errors which limit the ability of NER models to recover entity mentions from spontaneous speech transcripts. Using publicly available benchmark datasets (SWNE, Earnings-21, OntoNotes), we present the full taxonomy of ASR-NER errors and measure their true impact on entity recognition. We find that NER models fail to recognize entity spans even if no word errors are introduced by the ASR. We also show why the F_1 score is inadequate to evaluate NER models on conversational transcripts[1].

1 Introduction

The performance of NLP models tends to deteriorate significantly when the models are applied to the raw outputs of the Automatic Speech Recognition (ASR) system. We coin the term *ASR-NLP gap* to describe this phenomenon. Despite unprecedented advances in modern language models, the transcript of a spontaneous human-human conversation remains an insurmountable challenge for most models. This is particularly true for Named Entity Recognition (NER) models, which struggle to retrieve even the most basic entity mentions from spontaneous speech.

Three primary factors contribute to the existence of the ASR-NLP gap. Firstly, the structure of spontaneous human conversations is diametrically different from the prescriptive written language used to train language models. These models can use the grammatical structure present in the training corpora, such as part-of-speech sequences, dependency trees, and dialog acts. On the other hand, spontaneous conversations lack sentence structure. They contain repetitions, back-channeling, phatic expressions, and other artifacts of turn-taking. The second challenge comes from the original ASR output containing neither punctuation nor sentence segmentation. These have to be restored by an auxiliary downstream model. Thus, NLP models trained on prescriptive written text or scripted conversations already have to process the out-of-domain input. The third problem stems from ASR systems injecting word errors into the transcript. Due to efficiency requirements, most ASR systems use unsophisticated language models such as n-gram models with limited vocabulary. Thus, many utterances in the input audio may be unrecognized and deleted from the output, while other utterances may cause substitutions or insertions of erroneous tokens into the output.

Consider the following sentence: "I am to see [Dr Smith]$^{\texttt{PERSON}}$ at [9 am]$^{\texttt{TIME}}$ on [Monday, May 14th]$^{\texttt{DATE}}$". The NER model[2] correctly recognizes three entity spans in the sentence. Compare this to the NER spans recognized in the sentence, which

[1] All code necessary to reproduce our results can be found in https://github.com/niedakh/asr-ner-eval-repository

[2] In this illustrative example we are using `spaCy` (Honnibal and Montani, 2017) trained on OntoNotes v5, Wordnet 3.0, and ClearNLP Constituent-to-Dependency Conversion (Choi et al., 2016).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1746–1761

is far more likely to be produced by the ASR: "I am to see doctor [Smith]$^{\texttt{PERSON}}$ at nine I am on [monday]$^{\texttt{DATE}}$ [uhm]$^{\texttt{ORG}}$ yeah [monday]$^{\texttt{DATE}}$ may for teen." Two entity spans have been cut short, an incorrect label has replaced one span's label, and the model recognized a filler *uhm* as the entity ORG! With a few more ASR errors and lowercase output, the model does not recognize a single entity in the output of the ASR: "I am to see doctor uhm doctor smith at nine I am on man day may for teen."

The main problem is that ASR errors are very "unnatural" from the point of view of the NER model because they tend to break the grammar of the sentence on which the NER model depends. One of the most consequential errors made by the ASR is the confusion about the part-of-speech tag. Consider possible ASR errors in the sentence "My [second]$^{\texttt{ORDINAL}}$ visit is [Wednesday]$^{\texttt{DATE}}$ at [half past one]$^{\texttt{TIME}}$." Changing the personal pronoun "My" to the noun "May" forces the NER model to recognize a DATE span, which is reasonable. But if the ASR changes the preposition "at" into a verb "add," the NER model loses the ability to recognize the utterance "half past one" as TIME because of the lack of the preceding preposition. Similarly, changing "half past one" to "[one thirty]$^{\texttt{TIME}}$" retrieves the TIME span, but an ASR error confusing the numeral "one" with the conjunction "when" produces "[Wednesday]$^{\texttt{DATE}}$ at when [thirty]$^{\texttt{DATE}}$." If, however, the same word is mistakenly recognized as the verb "want," the NER model produces "[Wednesday]$^{\texttt{DATE}}$ at want [thirty]$^{\texttt{CARDINAL}}$".

Unfortunately, the problems mentioned above cannot be easily solved. Word error rates (WER) of ASR systems remain high for spontaneous human conversations (Del Rio et al., 2021). Recently announced results claiming WERs at the level of 5% apply to conversations with digital assistants, where spoken utterances are imperative phrases with limited vocabulary. These results are not representative of spontaneous human open dialogues, which lack the rigid grammatical phrase structure and contain fillers, back-channeling, repetitions, hesitation markers, and other elements which are a part of spontaneous speech.

The interplay of two phenomena makes the processing of spontaneous speech transcripts with NLP models so challenging. On the one hand, every NLP model is inherently flawed and produces errors (such as not recognizing an instance of an entity). On the other hand, the ASR system injects errors in the form of insertions, deletions, and substitutions. This changes the structure and semantics of transcribed speech and introduces yet another source of errors: alignment. In order to measure the quality of the NER model on the transcript, one has to align tokens between gold transcripts and the ASR output to match entity spans. This process may produce artifacts that significantly skew the results of the evaluation.

The evaluation of the NER task is usually performed using precision, recall, and the F_1 score. Unfortunately, these measures are of limited use for processing spontaneous conversation transcripts because they confound two independent factors contributing to the errors mentioned above: the inability of the NER model to recognize a span as an entity and the word error introduced by the wrong transcription of a token.

Our paper is a reality check on the state of named entity recognition in spontaneous speech transcripts. Using popular benchmark datasets, we show how state-of-the-art language models fail to discover entity spans in transcripts of spontaneous speech. We identify several artifacts of ASR errors with respect to entity recognition. We measure the propensity of each type of artifact to influence the recognition of named entities. This approach brings us closer to understanding the true reasons for NER model failures on spontaneous speech transcripts. We argue that misalignment artifacts are essential characteristics of the performance of NLP models and should be considered when evaluating downstream NLP models on spontaneous speech transcripts.

2 Entity span alignment

We measure the loss of entity spans recognized in the ASR output compared to those recognized in the gold transcript. Thus, we must perform token alignment between the ASR output and the gold transcript, as they may differ in the number of tokens. Alignment is performed after diarisation (separating speakers' utterances into separate channels) for each channel independently. We use a greedy alignment procedure. We begin by running the NER model on the gold transcript and tagging each token in the transcript using the IOB scheme (B – beginning of an entity span, I – inside an entity span, O – outside of an entity span). Next, we collapse all adjacent I-tags so that each channel is represented by a sequence of B-tags

and O-tags. We repeat the same procedure for the ASR output and then align both transcripts. The alignment of gold transcripts, normalized gold transcripts, and the ASR output is performed by the `fstalign` (McNamara and Kokotov, 2021) and the `kaldialign` (Żelasko and Guo, 2021) libraries, with minor additional corrections. All transcripts are matched at the level of tokens.

In the remainder of the paper, we will use the following terminology (Pallett, 1985). For the ASR errors, we will distinguish the following types of errors:

- *insertion*: a token has been inserted into the ASR output which does not appear in the gold transcript,
- *substitution*: a token has been wrongly transcribed, the number of tokens in both transcripts is the same, but the values of tokens differ,
- *deletion*: the ASR has not recognized a token, the output sequence of the ASR is shorter than the original gold transcript.

In parallel, the NER model can introduce the following errors:

- *hallucination*: an entity tag has been produced in the ASR output which does not appear in the gold transcript,
- *replacement*: an entity tag has been added to the token, but the label of the entity class is different from the gold transcript,
- *omission*: the NER model does not produce an entity tag for a token tagged in the gold transcript.

Let us now describe in detail all possible combinations of the above ASR and NLP errors and their impact on the recognition of named entities. For the sake of clarity, we will only consider artifacts of the ASR-NLP gap within a single entity span. Detailed examples of every combination of ASR-NLP errors discovered in the *Earnings-21* dataset are presented in Appendix A.

Firstly, let us consider a scenario where the gold transcript and the ASR output are perfectly aligned, i.e., all tokens are correctly recognized. The gold transcript contains the utterance "second$^{\texttt{B-DATE}}$ quarter$^{\texttt{B-DATE}}$ twenty$^{\texttt{B-DATE}}$ twenty$^{\texttt{B-DATE}}$." The following entity span errors are possible (Table 1):

	second	quarter	twenty	twenty
A	B-DATE	I-DATE	I-DATE	I-DATE
B	B-DATE	I-DATE	I-DATE	I-DATE
C	O	O	O	O
D	B-CARD	I-CARD	I-CARD	I-CARD
E	B-DATE	I-DATE	B-CARD	I-CARD
F	B-DATE	I-DATE	O	O
G	B-DATE	I-DATE	O	B-CARD

Table 1: NER errors for fully aligned transcripts: (A) gold transcript tags (B) fully matched (C) fully omitted (D) fully replaced (E) partially replaced (F) partially omitted (G) partially replaced and omitted

- *full match*: each token in the ASR output receives the same entity tag as the gold transcript (row B),
- *full omission*: no entity tags are produced for tokens inside the gold transcript entity span (row C),
- *full replacement*: each token in the ASR output has a different entity tag from the gold transcript (row D),
- *partial match with replacement*: some tokens in the ASR output have different entity tags from the gold transcript (row E),
- *partial match with omission*: some tokens in the ASR output do not have entity tags (row F),
- *partial match with omission and replacement*: some tokens in the ASR output have a different entity class tag, and some tokens do not have entity tags.

Consider a situation where the ASR inserts a token into the gold transcript. Obviously, there is a mismatch in the number of tokens in the gold transcript and the transcription. Let us assume that the utterance "nextstart$^{\texttt{B-ORG}}$ group$^{\texttt{I-ORG}}$" has been mistakenly transcribed as "next door group." Table 2 summarizes possible combinations of ASR and NER errors.

- *full match*: tokens are tagged with the same entity class labels (row B),
- *full omission*: the introduction of a token by the ASR prevents the NER model from finding any entity tags (row C),

	nextstart next	door	group group
A	B-ORG		ORG
B	B-ORG	I-ORG	I-ORG
C	O	O	O
D	B-PROD	I-PROD	I-PROD
E	B-ORG	I-ORG	B-LOC
F	B-ORG	O	B-ORG
G	B-ORG	O	O

Table 2: NER errors for transcripts with ASR insertion: (A) gold transcript tags (B) fully matched (C) fully omitted (D) fully substituted (E) partially substituted (F, G) partially omitted

	american american	door <del>	bell bell	group group
A	B-ORG	I-ORG	I-ORG	I-ORG
B	B-ORG		I-ORG	I-ORG
C	O		O	O
D	B-GPE		B-ORG	I-ORG
E	B-ORG		I-ORG	O
F	B-GPE		O	B-ORG

Table 3: NER errors for transcripts with ASR deletion: (A) gold transcript tags (B) partially matched (C) fully omitted (D) partially replaced (E) partially matched with omission (F) partially matched with replacement and omission

- *full substitution*: tag introduced by the ASR forces the NER model to generate different entity labels (row D),
- *partial substitution*: some tokens in the ASR output are tagged with different entity class labels (row E),
- *partial omission*: some tokens in the ASR output do not have an entity tag, which may result in the multiplication of the entity span (row F) or shortening of the entity span (row G).

The ASR can delete a token from the gold transcript, resulting in a possible misalignment. In this scenario, full matching is impossible because the gold transcript will contain an unmatched token. Similarly, an entity span cannot be hallucinated or fully substituted. Let us assume that the gold transcript utterance "next$^{\texttt{B-ORG}}$ door$^{\texttt{I-ORG}}$ group$^{\texttt{I-ORG}}$" has been mistakenly transcribed as "next <del> group" (i.e., the ASR failed to recognize the "door" token). Table 3 presents possible combinations of ASR and NER errors.

- *partial match*: tokens not deleted by the ASR have correct entity tags,
- *full omission*: the deletion of a token by the ASR prevents the NER model from producing any entity tags,
- *partial replacement*: some tokens in the ASR output have the wrong entity tag,
- *partial omission*: the loss of token results in some of the tokens not being tagged with an entity tag,
- *partial replacement and omission*: some of the tokens receive correct entity tags, some receive wrong entity tags, and some do not receive any entity tags at all.

Finally, the NER model can hallucinate an entity span where the gold transcript has no entities.

As we can see, the number of possible mistakes is large, and it is not obvious which scenarios are common or rare. In other words, if we are to develop more robust models for named entity recognition in the transcripts of spontaneous speech, we need to understand which scenarios are the most impactful for the NER task. In the next sections, we present experiments that try to present a much more detailed and nuanced view of ASR and NER errors.

3 Datasets

We use three datasets in our experiments.

- *OntoNotes*: the LDC-released OntoNotes v5 (Weischedel et al., 2013) with texts from news, broadcast/telephone conversations, and web data annotated with 18 entity types.
- *SWNE*: data from Switchboard Dialog Acts Corpus annotated with entity tags following the OntoNotes v5 annotation scheme (Choi, 2020)
- *Earnings-21*: audio and transcriptions of 44 public phone calls which span almost 40 hours of recordings of human conversations, with 25 different entity classes annotated in transcripts (Del Rio et al., 2021).

We decided to omit the *CoNLL-2003/CoNLL++* (Tjong Kim Sang and De Meulder, 2003) dataset because it is annotated with only four classes of entities. Unfortunately, the three listed datasets are the only publicly available

datasets that contain audio segments and transcripts annotated with entity types. One may argue that these datasets are not representative of spontaneous conversations. For instance, *Earnings-21* transcripts sound heavily scripted, and the interlocutors present speeches rather than a free exchange of utterances. While this is true, at the same time, these three datasets present the closest that researchers can get to conversational audio transcripts with annotated entity spans.

There are datasets with audio recordings annotated with entity spans, but these datasets are not in the domain of spontaneous speech. In recent years we are observing significant progress in named entity recognition in transcripts of scripted speech. This progress is made possible mostly due to the publication of annotated datasets. Yadav et al. present a dataset consisting of TED talks, Mozilla Common Voice recordings, LibriSpeech audiobook recordings, and VoxForge recordings. As the authors observe, NER models achieve promising results on these transcripts (probably due to the fact that the input transcript is semantically similar to the typical training data for NER models). The same dataset is used by Zhang et al. to illustrate the error correction model. Recently, annotated transcripts of speech (albeit non-conversional) have been released for Scandinavian languages (Porjazovski et al., 2021), for French (Millour et al., 2022), and for Chinese (Chen et al., 2022). It is worth mentioning that NER task has been added to the recent Spoken Language Understanding Evaluation (SLUE) benchmark (Shon et al., 2022). Unfortunately, the annotation covers a small subset of the *VoxPopuli* dataset, which is not representative of spontaneous speech, the *VoxPopuli* is the set of recorded speeches in the European Parliament.

Entity classes annotated in the above datasets can be broadly divided into closed-domain and open-domain types. Closed-domain entity classes can be regarded as almost gazetteers, i.e., these are classes for which a vast majority of entities can be listed. Examples of closed-domain entity classes include geographical locations or first names (since the distribution of US first names follows a power law distribution (Hahn and Bentley, 2003), a relatively small number of first names represents the majority of first names encountered in the dataset). On the other hand, open-domain entity classes cannot be summarized using a gazetteer. This is the case with numbers, product names, money, or organizations.

entity	Earnings-21	SWNE	OntoNotes
CARDINAL	0.46	0.69	0.86
DATE	0.49	0.34	0.87
EVENT	0.12	0.37	0.74
FAC	0.07	0.32	0.77
GPE	0.63	0.87	0.97
LANGUAGE	0.00	0.94	0.75
LAW	0.02	0.36	0.67
LOC	0.56	0.45	0.76
MONEY	0.20	0.62	0.90
ORDINAL	0.79	0.00	0.86
ORG	0.49	0.62	0.92
PERCENT	0.66	0.00	0.86
PERSON	0.55	0.82	0.96
PRODUCT	0.10	0.58	0.79
QUANTITY	0.42	0.59	0.79
TIME	0.32	0.39	0.69
WORK_OF_ART	0.00	0.46	0.72
micro avg F1	0.37	0.51	0.83

Table 4: F-scores of the NER model on gold transcripts

Unfortunately, gazetteers are not a viable solution even for closed-domain entity classes because ASR errors may produce tokens outside the gazetteer. One possible solution would be to try to overcome ASR errors by retrofitting token representations using domain datasets. This technique has been successfully applied to static word embeddings to mitigate ASR errors by Augustyniak et al. (2020). It would be interesting to see the same technique applied to transformer-based embeddings.

4 Experiments

One might argue that the most important variable influencing the performance of downstream NLP tasks on a transcript is the choice of a particular ASR system. However, we do not find this to be the case. The ASR-NLP gap is equally pronounced for all major commercial ASR systems. In our experiments, we choose the ASR offered by Microsoft due to its lowest reported WER on the *Earnings-21* dataset (Del Rio et al., 2021).

4.1 Performance on gold transcripts

In our first experiment, we evaluate the state-of-the-art NER model on gold transcripts. We train a transformer using the Roberta-Large architecture (Liu et al., 2019) on the train split of the *OntoNotes* dataset [3]. The evaluation is performed on *Earnings-21*, *SWNE*, and the test split of the *OntoNotes* datasets. In order to make the comparison as fair

[3]We have also experimented with other models including BERT, DistilBERT, FLERT, and spaCy, we choose the best-performing model for the presentation of results

as possible, we normalize gold transcripts using a set of heuristics. Normalization changes all numbers into respective words. We unify the position of the currency indicator when spelling monetary values and the position of the percent sign. All gold transcripts are properly cased and punctuated. We report the results as measured by the micro F_1 score because the dataset is highly imbalanced, and we are interested in the overall performance of the NER model.

We must point out that the experimental setting is very favorable for the ASR. Not only is the transcript fully normalized, but the alignment procedure is fine-tuned to reduce the number of misalignments as much as possible. Furthermore, the NER model is applied to text fragments chunked according to punctuation in the gold transcripts and not to fixed-width sliding windows. In other words, the NER model is applied to the input text of much higher quality than should be expected from the commercial ASR.

Despite the fact that *OntoNotes* contains a significant amount of transcripts of unscripted human conversations, the accuracy of the model deteriorates dramatically on *SWNE* and *Earnings-21* datasets. For all entity classes, the recognition in *SWNE* and *Earnings-21* is much lower than for the *OntoNotes*. The NER model struggles particularly with open-domain entity classes. The complete failure to recognize `MONEY`, `PRODUCT` or `TIME` entities makes the NER model practically unusable in real-world scenarios. Leaving aside more exotic classes represented in the data by a few examples (`LANGUAGE`, `LAW`, `WORK_OF_ART`), we see that the NER model performs better (albeit not satisfactorily) for closed-domain classes, where it can to a certain degree memorize most of the instances of a class. For open-domain entity classes, the performance of the model is disappointingly bad. Please note that the NER model is applied to properly cased and punctuated transcripts of conversations and not to the ASR output, yet the F_1 scores are significantly lower than the scores obtained on the test split of the *OntoNotes* dataset.

4.2 Performance on ASR transcripts

In the second experiment, we run our NER model on the *Earnings-21* dataset, and we measure the number of occurrences of every error described in Section 2. Transcripts of *Earnings-21* recordings are produced by the Microsoft ASR. The results are presented in Table 5. The first column reports the number of occurrences of NER model errors when the ASR output is fully matched with the gold transcript (no ASR errors in the transcript). Subsequent columns report the number of occurrences of NER model errors when the ASR output is misaligned with the gold transcript due to token insertion, substitution, or deletion by the ASR. Please note that ASR insertion, substitution, and deletion errors often co-occur within a single entity span in the gold transcript, so a single entity span may contribute to multiple cells in the table. Our intention is to show the real impact of each type of ASR-NLP error.

The results presented in Table 5 clearly show the importance of the joint ASR-NLP model evaluation, as reflected by the breakdown of the two error sources[4]. First, the NER model makes mistakes on fully matched transcripts of spoken conversations, i.e., when the ASR manages to retrieve the gold transcript in the entity span without errors. These errors are responsible for approximately half of all recorded errors. Let us stress this result again: NER models are inherently incapable of processing the transcripts of spontaneous speech; even if the ASR introduces no errors, 37% of entity spans are partially or fully wrong (first column in Tab. 5)

We also see that the NER model is very sensitive to errors introduced by the ASR. It can correctly recognize only 18% of entities when the ASR substitutes a token inside the entity span, 6.8% of entities when the ASR inserts a token inside the entity span, and it fails to correctly recognize an entity when the ASR deletes a token inside the entity span. ASR errors are responsible for many hallucinated entities and the majority of omissions. In practice, the number of entity errors doubles compared to the number of errors made on fully matched transcript: ca. 6200 omitted entities in total vs. 3600 with perfect transcript and ca. 2000 hallucinated ones versus 1000 with the perfect transcript. Again, let us reiterate this finding: the NER model is helpless when ASR errors are introduced inside entity spans and cannot retrieve an entity when tokens are inserted, substituted, or deleted from entity spans. The results we obtained are vastly different from what one could infer from a WER of 15.8 and entity

[4]After deliberation, we have decided to report raw counts of NER-ASR errors instead of frequencies. The main reason is the fact that these results cannot be meaningfully summed up, and particular combinations of NER-ASR errors appear at different scales. This makes the analysis of results more challenging, but every simplification of the table leads to the loss of valuable insight.

	no ASR error	ASR insertion	ASR substitution	ASR deletion
correct tags	11408	64	1008	0
hallucinated	1039	784	958	200
omitted	3607	47	2649	709
replaced	1383	6	509	0
partially matched with replacement without omission	97	2	9	0
partially matched without replacement with omission	654	37	261	306
partially matched with replacement and omission	26	3	19	18

Table 5: Counts of different combinations of NER-ASR errors on the *Earnings-21* dataset

WER of 20.0 reported by (Del Rio et al., 2021)!

Finally, the case for partial matches, while smaller than hallucinated, replacement, and omissions, is of great importance. The true effect of entity hallucinations and omissions in a joint ASR-NLP system can only be measured on a downstream task. Usually, named entity recognition is a single step in a wider NLP task. This task may have a separate evaluation scheme with different metrics and business objectives. For example, in the task of intent retrieval and slot filling, hallucinating or omitting an entity span can lead to a situation where the intent is either not matched or matched in the wrong place. However, the effect of partial matches is more difficult to evaluate. With partial matching, the intent is caught, and the slot is filled, but most probably, the slot is filled with incorrect values. The scale of failures and the impact of upstream model improvements can only be measured by evaluating the entire NLP pipeline on a reference dataset with annotations of intents and slots. This observation strengthens our belief that measuring the increase in the scale of errors in a joint ASR-NLP system is more important than focusing on technical details of measures such as the F_1 score, WER, or entity WER.

5 Related Work

In our opinion, the NLP research community has an overly optimistic view of the WERs introduced by ASR systems. Recent experiments show that WERs in transcripts of spontaneous human speech is much higher than expected. For instance, Szymański et al. (2020) showed that a transcript of a standard GSM phone call conversation is subject to a 16%-20% error rate. Del Rio et al. (2021) confirm this result and report how WERs differ between different types of entity spans. Spans related to date, time, and ordinal numbers were observed to have a lower WER than entities related to proper names. Facility names, organizations, and personal names demonstrate a very high WER of 30%-50%. McNamara and Kokotov (2021) also released a library for using Finite State Transducers (FSTs) to account for different representations of the same entity (*2020* vs. *twenty twenty*) among ASRs.

These findings are in stark contrast to initial reports. For instance, Surdeanu et al. (2005) reported named entity recognition in Switchboard corpus to be within 5% from a system evaluated on clean textual data. Similarly, Béchet et al. (2002) claims to have achieved approximately 0.90 F_1 for recognizing phone numbers and 0.70 F_1 for recognizing money mentions in the transcripts from the AT&T *How may I help you?* system under 27.4% WER ratio. Favre et al. (2005) apply NER models to French corpora and achieve 0.74 F_1 for a relatively broad set of named entities.

Precision, recall, and F_1 scores are standard metrics for reporting NER model performance in NLP. However, these metrics can produce unreliable scores where entity spans are marked on spontaneous human conversation transcripts due to the presence of conversational artifacts (repetitions mentioned above, backchanneling, phatic expressions). An example of entity span tagging where the F_1 metric produces highly misleading scores is presented in Section 6.

To account for the presence of these artifacts, Message Understanding Conference (MUC) (Grishman and Sundheim (1996); Nadeau and Sekine (2007)) introduced metrics that allow for partial matching of an entity span. MUC defines six categories of partial matching based on the degree of span overlap, the type of the matched entity, and the strictness of expectations, as outlined by Batista (2020). Recently, this problem has been addressed by Caubrière et al. (2020) who argues for the use of slot error rates.

To the best of our knowledge, Hatmi et al. (2013) was the first to attempt to incorporate named entity recognition into the automatic speech transcription

process. The authors tagged the ASR dictionary with named entity tags (since ASR cannot produce any words not present in its dictionary). This initial approach has been superseded by methods aiming at training end-to-end joint models for ASR and NER, as proposed by Ghannay et al. (2018), Serdyuk et al. (2018), and Stiefel and Vu (2017). The authors train ASR systems to predict transcription tokens and their part-of-speech or named entity tags in these works.

6 Limitations

Obviously, the work presented in this paper is limited to transcripts of spontaneous conversations in English. Since we are investigating the problem of named entity recognition, we have to point out that there are practically no datasets of human conversations (both audio and transcripts) annotated with entity spans apart from *SWNE*, *OntoNotes* and *Earnings-21*, the three datasets used in our paper. These datasets are relatively small, and the distribution of the frequency of appearance of entity classes is extremely skewed, with several entity classes represented by a handful of examples.

Another significant limitation of the results reported in this paper is the choice of metric. Following the common practice in the NLP community, we have chosen the F_1 score as the primary metric of entity recognition. However, this metric is questionable in the context of NER recognition in ASR transcripts because it is highly dependent on two factors: the WER produced by the ASR and the definition of span alignment. Consider a gold transcript annotation "John$^{\texttt{B-PERSON}}$ F.$^{\texttt{I-PERSON}}$ Kennedy$^{\texttt{I-PERSON}}$" and the ASR output with "F." transcribed as "eh" annotated as follows: "John$^{\texttt{B-PERSON}}$ eh Kennedy$^{\texttt{B-PERSON}}$." Should this annotation be considered correct? The original person entity starting at "John" is only partially matched, and a new person entity starting at "Kennedy" is introduced in the ASR output. Consider another gold annotation of the following transcript: "second$^{\texttt{B-DATE}}$ quarter$^{\texttt{I-DATE}}$ twenty$^{\texttt{I-DATE}}$ twenty$^{\texttt{I-DATE}}$," which the NER model tags as follows: "second$^{\texttt{B-DATE}}$ quarter$^{\texttt{I-DATE}}$ twenty$^{\texttt{B-CARDINAL}}$ twenty$^{\texttt{I-CARDINAL}}$" (NER model trained on written language does not recognize "twenty twenty" as a valid date). Again, how should this scenario be scored by an accuracy metric? Unfortunately, the traditional definition of the F_1 score is too restrictive to produce a robust score that could paint a reliable picture of the model's performance. The design and implementation of a metric that could compute the alignment of entity spans in the presence of ASR errors would be a significant step in the direction of producing more robust NER models for spoken conversations.

We conduct experiments with the ASR on audio files from the *Earnings-21* dataset. These files are recorded at 11 kHz-44 kHz, while typical call center conversations are recorded at 8 kHz-16 kHz. Unfortunately, training datasets with recording characteristics resembling real-world usage scenarios are unavailable. We also do not address the problem of racial, gender, and age disparity (Koenecke et al., 2020) due to the lack of availability of sufficiently representative and inclusive datasets. It is, however, to be expected that the performance of the ASR deteriorates for the recordings of speakers other than male speakers of General American.

7 Conclusions

Our work provides a thorough, albeit pessimistic, reality check on the named entity recognition in conversational transcripts. Our first conclusion is straightforward: currently available NER models are not trained on representative data (due to the lack of annotated datasets), and their performance on transcripts of spontaneous conversations is much worse than their performance on written language. Importantly, this failure cannot be attributed solely to the presence of ASR word errors. As we show, NER models exhibit very high entity WERs even on gold transcripts, where no ASR errors are present. When the transcript contains ASR insertions, substitutions, or deletions, the entity recognition rates fall to the level where NER models become unusable in downstream tasks.

Secondly, we conclude that a completely new approach is required to meaningfully measure the quality of NER models on conversational transcripts. Traditional metrics, such as F_1 score or entity WER do not account for the intricate interplay of factors (NER errors, ASR errors, artifacts of spontaneous speech) and do not provide a useful insight into the model's performance. We need to design a more complex evaluation scheme that would take into account the token alignment errors, partial entity span matchings, ASR word errors, and NER errors.

8 Ethics statement

Following the ACM Code of Ethics and Professional Conduct we evaluate the ethical impact of the work presented in this paper. Our work aims at broadening the accessibility of communication technology. Spontaneous spoken language is the least limiting and exclusive mode of interacting with an information system. This mode does not require any digital competencies or expensive resources. The ability to correctly process spontaneous human conversations opens access to technology to stakeholders who might have been previously excluded. We strive to diminish discrimination resulting from biased training datasets, which may cause specific individuals to be disproportionally mistranscribed due to their accent, dialect, or speech impediments. As digital voice applications become increasingly integrated into society's infrastructure, we feel the need to improve the quality of statistical models processing spoken communications continuously.

The ability to better process and understand spoken human conversations carries the significant ethical risk associated with clandestine eavesdropping by adversarial agents. Correct recognition of spoken names of people, places, organizations, or events, can be malevolently used by authoritarian government agencies trying to suppress free speech. Recognition of names of products or services may be utilized by marketers for non-consensual profiling. Thus, it is in the best interest to foster public awareness and understanding of computing, the automatic processing of spontaneous speech, and its consequences.

References

Łukasz Augustyniak, Piotr Szymanski, Mikołaj Morzy, Piotr Zelasko, Adrian Szymczak, Jan Mizgajski, Yishay Carmiel, and Najim Dehak. 2020. Punctuation prediction in spontaneous conversations: Can we mitigate asr errors with retrofitted word embeddings?

David S. Batista. 2020. Ner evaluation. https://github.com/davidsbatista/NER-Evaluation.

Frédéric Béchet, Allen L Gorin, Jerry H Wright, and Dilek Hakkani-Tür. 2002. Named entity extraction from spontaneous speech in how may i help you? In *INTERSPEECH*.

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 12th Language Resources and Evaluation Conference*, pages 4514–4520.

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

Jinho D. Choi. 2020. Swne. https://github.com/emorynlp/swne.

Jinho D. Choi, Henry Chen, and Tomasz Jurczyk. 2016. Constituent to dependency conversion. https://github.com/clir/clearnlp-guidelines.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jette. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Benoît Favre, Frédéric Béchet, and Pascal Nocéra. 2005. Robust named entity extraction from large spoken archives. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 491–498.

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 *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Matthew W Hahn and R Alexander Bentley. 2003. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl_1):S120–S123.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meigner. 2013. Incorporating named entity recognition into the speech transcription process. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech'13)*, pages 3732–3736.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad

Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Quinn McNamara and Dan Kokotov. 2021. fstalign. *Software available from* `https://github.com/revdotcom/fstalign`.

Alice Millour, Yoann Dupont, Alexane Jouglar, and Karën Fort. 2022. FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 82–94, Avignon, France. ATALA.

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

David S Pallett. 1985. Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards*, 90(5):371.

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 *2018 IEEE 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Moritz Stiefel and Ngoc Thang Vu. 2017. Enriching asr lattices with pos tags for dependency parsing. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 37–47.

Mihai Surdeanu, Jordi Turmo, and Eli Comelles. 2005. Named entity recognition from spontaneous open-domain speech. In *INTERSPEECH*, pages 3433–3436.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

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 Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Fan Zhang, Mei Tu, Song Liu, and Jinyao Yan. 2022. Asr error correction with dual-channel self-supervised learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7282–7286. IEEE.

Piotr Żelasko and Liyong Guo. 2021. kaldialign. *Software available from* `https://github.com/pzelasko/kaldialign`.

A Examples of ASR-NLP errors from the *Earnings-21* dataset

In this section, we present several examples of alignments of the ASR output with the gold transcript with entity tags. In each table, the upper two rows present entity tags and word tokens present in the gold transcript, and the bottom two rows present word tokens generated by the ASR and entity tags produced by the NER model. A detailed description of each case is presented in the caption of each table. All examples are from the *Earnings-21* dataset.

O	O	B-PERSON	O	O	O
thank	you	anna	and	welcome	everyone
thank	you	anna	and	welcome	everyone
O	O	B-PERSON	O	O	O

Table 6: Full matching of word tokens and entity tags.

O	B-DATE	I-DATE	I-DATE	I-DATE	B-DATE	O
from	last	<ins>	years	comparable	quarter	results
from	last	year	’s	comparable	quarter	results
O	B-DATE	I-DATE	I-DATE	I-DATE	I-DATE	O

Table 7: Full matching of entity tags despite the insertion of a token by the ASR.

O	O	B-PERSON	I-PERSON	O	O	O
we	have	dominic	macklon	our	senior	vice
we	have	dominic	macklin	our	senior	vice
O	O	B-PERSON	I-PERSON	O	O	O

Table 8: Full matching of entity tags despite the ASR substitution of a token.

O	O	O	O	O	O
your	normal	mid	teens	revenue	growth
your	normal	mid	teens	revenue	growth
O	O	B-CARDINAL	I-CARDINAL	O	O

Table 9: Full matching of word tokens, the NER hallucinates the `CARDINAL` entity

O	O	O	B-ORDINAL
from	<ins>	perishables	first
from	paris	rivers	first
O	B-GPE	O	B-ORDINAL

Table 10: the ASR token insertion (due to wrong recognition of "perishables" as "paris rivers") makes the NER to hallucinate the `GPE` entity.

O	O	O	O	O	O	O	O
for	the	good	more	lean	work	to	help
for	the	good	<del>	morning	work	to	help
O	O	B-TIME	O	I-TIME	O	O	O

Table 11: The ASR deletes a token by recognizing "good more lean work" as "good morning work", causing the NER to hallucinate the TIME entity.

O	O	O	O	O
now	so	are	there	discernible
tina	so	are	there	discernible
B-PERSON	O	O	O	O

Table 12: The NER hallucinates the PERSON tag due to an ASR substitution

O	O	B-ORG	B-DATE	O
<ins>	see	nexstar's	annual	report
sing	next	cars	annual	report
O	O	O	O	O

Table 13: The DATE entity is missed due to the ASR insertion and replacement

B-PERSON	I-PERSON	O	O	O	O
shuang	liu	and	chief	financial	officer
strong	will	and	chief	financial	officer
O	O	O	O	O	O

Table 14: The ASR replaces tokens in the unrecognized person's name forcing the NER to omit the PERSON entity.

O	O	O	B-ORG	I-ORG	I-ORG
profile	to	the	s	m	e
profile	to	the	s	m	<del>
O	O	O	O	O	O

Table 15: The ASR deletes tokens related to the unrecognized name of the SME company, forcing the NER to omit the ORG entity.

O	B-DATE	I-DATE
in	twenty	nineteen
in	twenty	nineteen
O	B-CARDINAL	I-CARDINAL

Table 16: Full matching of tokens does not prevent the NER from replacing the DATE entity with the CARDINAL entity.

O	B-ORG	I-ORG	I-ORG	O	O
and	jj	<ins>	bistricer	chief	operating
and	jj	best	research	chief	operating
O	B-PERSON	I-PERSON	I-PERSON	O	O

Table 17: The ASR insertion results in the replacement of the ORG entity with the PERSON entity.

O	O	B-GPE	O	O	O
it's	not	mexico	for	example	right
he's	not	mexican	for	example	right
O	O	B-NORP	O	O	O

Table 18: The ASR substitution causes the full replacement of the GPE entity with the NORP entity.

B-DATE	I-DATE	I-DATE	I-DATE
twenty	twenty	second	quarter
twenty	twenty	second	quarter
B-CARDINAL	I-CARDINAL	B-DATE	B-DATE

Table 19: Example of a partial DATE entity match with the rest of the entity replaced by the CARDINAL entity despite the full matching of word tokens.

O	B-CARDINAL	I-CARDINAL	I-CARDINAL	O	O
and	one	twenty	eight	total	net
and	waterman	twenty	eight	dot	net
O	B-FAC	B-CARDINAL	I-CARDINAL	O	O

Table 20: Example of partial CARDINAL entity match with the replacement of the rest of the entity with FAC entity caused by the ASR substitutions.

O	B-ORG	I-ORG	O	O
while	ingersoll	rand	took	share
while	ingersoll	rand	took	share
O	B-ORG	O	O	O

Table 21: Example of the partial ORG entity match with parts of the entity span omitted despite the full matching of word tokens.

B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	O
the	<ins>	nextera	energy	inc	and
the	next	era	energy	inc	and
O	O	B-ORG	I-ORG	I-ORG	O

Table 22: Example of the partial ORG entity match with parts of the entity span omitted due to ASR insertion and substitution.

O	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG
present	that	to	the	florida	public	service	commission
present	that	to	the	florida	public	service	commission
O	O	O	O	B-GPE	B-ORG	I-ORG	I-ORG

Table 23: Example of the partial ORG entity match with parts of the entity span omitted or replaced despite the full matching of word tokens.

B-DATE	I-DATE	I-DATE	I-DATE	I-DATE	I-DATE	O
the	second	half	of	twenty	one	operating
the	second	half	i'm	twenty	one	operating
O	O	O	O	B-CARDINAL	B-DATE	O

Table 24: Example of the partial DATE entity match parts of the entity span omitted or replaced due to ASR substitutions.

O	B-DATE	I-DATE	I-DATE	I-DATE	B-MONEY	I-MONEY
to	june	30	twenty	twenty	$25.2	million
to	june	3020	twenty	<del>	$25.2	million
O	B-DATE	I-DATE	B-MONEY	O	I-MONEY	I-MONEY

Table 25: Example of the partial entity matches with replacements and omissions due to ASR substitutions and deletions.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations are described in Section 6.

☑ A2. Did you discuss any potential risks of your work?
Our work does not introduce new models or methods but provides a negative reality check on the state of the art in NER recognition from spoken transcripts. We address some of the potential risks of NER in conversational transcripts in Section 8 Ethics statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Main claims are presented in the Abstract.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We use three benchmark datasets with audio recordings and transcriptions.

☑ B1. Did you cite the creators of artifacts you used?
All benchmark datasets are properly cited.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We are using open benchmarks released on open licenses.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We use benchmarks exactly as they were intended to be used: to evaluate the efficiency of the NER model on the conversational transcript.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not collect any new data and we don't use our internal datasets. The only datasets used in the experiments were open benchmarks. We have assumed that it is the responsibility of the benchmarks' authors to remove personably identifiable information from the data properly.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We acknowledge the lack of diversity and inclusiveness of the benchmark dataset in Section 6 Limitations. We also point out to new benchmark datasets for languages other than English, but we do not use them in current evaluation.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We do not create any new data. We use benchmark datasets and follow their documented splits.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

The results of computational experiments are reported in Section 4.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Although we have experimented with several NER model architectures, our contribution is not in the development of SOTA models. Quite the contrary, we present negative results and we have decided to omit the details of benchmark model training to focus the paper on the presentation of a much more important aspect, namely, the deep dive into the relationship between ASR and NER errors.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
As above, the results of experiments only serve to illustrate a much more important and overlooked issue. We do not find the particular details of the trained NER model important. We provide the architecture and the training dataset. The training uses default values of hyper-parameters.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Our experiments involve the description of particularities of ASR-NER errors, we report on the number of occurrences of each error combination.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We use two packages for transcript alignment and we point to respective software repositories.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.