Are We NER Yet? Measuring the Impact of ASR Errors on Named Entity Recognition in Spontaneous Conversation Transcripts

Anonymous ACL submission

Abstract

Transcriptions of spontaneous human conversations present a significant obstacle for traditional NER models trained on prescriptive written language. The lack of grammatical structure of spoken utterances, combined with word errors introduced by the ASR, makes downstream NLP tasks challenging. In this paper, we examine the impact of ASR errors on the ability of NER models to recover entity mentions from transcripts of spontaneous human conversations in English. We experimentally compare several commercial ASR systems paired with state-of-the-art NER models. We use both publicly available benchmark datasets (Switchboard Named Entity Corpus, SWNE), as well as the proprietary, real-life dataset of gold (human-transcribed) phone conversation transcripts. To measure the performance of NER models on ASR transcripts, we introduce a new method of token alignment between transcripts. Our findings unequivocally show that NER models trained on the written language struggle when processing transcripts of spontaneous human conversations. The presence of ASR errors only exacerbates the problem.

1 Introduction

The term *ASR-NLP gap* refers to the significant deterioration of the performance of NLP models when applied to the raw outputs of the Automatic Speech Recognition (ASR) system. 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.

Two primary factors contribute to the existence of the ASR-NLP gap. The structure of spontaneous human conversations is diametrically different from the prescriptive written language used to train large language models. These models can use the grammatical structure present in the training corpora, such as part-of-speech sequences, dependency trees, dialog acts. On the other hand, spontaneous conversations miss sentence structure, contain repetitions, back-channeling, phatic expressions, and other artifacts of turn-taking. Original ASR output contains neither punctuation nor sentence segmentation. These have to be restored by a dedicated model. Thus, NLP models trained on written text or scripted conversations already have to process the out-of-domain input. To further exacerbate the problem, ASR systems introduce inherent errors to the transcript. Errors can come as insertions, deletions, or substitutions, making them more confusing for downstream NLP models.

To better understand how complex these problems are, let us review some examples of how spontaneous speech combined with ASR errors can confuse the NER model. In all following examples, we will be using the NER model included in the `spaCy` library (Honnibal and Montani, 2017). The model was trained on OntoNotes v5, Wordnet 3.0, and ClearNLP Constituent-to-Dependency Conversion (Choi et al., 2016). We assume that an external model has correctly restored the casing of the ASR output. Otherwise, the task of the NER model becomes even more challenging.

Consider the following sentence: "*I am to see* $\boxed{\textit{Dr Smith}}^{\texttt{PERSON}}$ *at* $\boxed{\textit{9 am}}^{\texttt{TIME}}$ *on* $\boxed{\textit{Monday, May 14th}}^{\texttt{DATE}}$". The NER model correctly recognizes three entity spans in the sentence. Compare this to the NER spans recognized in the sentence which is far more likely to be produced by the ASR: "*I am to see doctor* $\boxed{\textit{Smith}}^{\texttt{PERSON}}$ *at nine I am on* $\boxed{\textit{monday}}^{\texttt{DATE}}$ $\boxed{\textit{uhm}}^{\texttt{ORG}}$ *yeah* $\boxed{\textit{monday}}^{\texttt{DATE}}$ *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`! Let us allow for a few more ASR errors, and 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 forteen*".

The main problem is the fact 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 of the part-of-speech tag. Let us consider possible ASR errors in the sentence "*My* $\boxed{\textit{second}}^{\texttt{ORDINAL}}$ *visit is* $\boxed{\textit{Wednesday}}^{\texttt{DATE}}$ *at* $\boxed{\textit{half past one}}^{\texttt{TIME}}$". Changing the personal pronoun "*My*" to a 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 looses 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 "$\boxed{\textit{one thirty}}^{\textit{TIME}}$" retrieves the `TIME` span, but an ASR error confusing the numeral *one* with the conjunction *when* produces "$\boxed{\textit{Wednesday}}^{\texttt{DATE}}$ *at when* $\boxed{\textit{thirty}}^{\texttt{DATE}}$". If, however, the same word is mistakenly recognized as the verb *want*, the NER model produces "$\boxed{\textit{Wednesday}}^{\texttt{DATE}}$ *at want* $\boxed{\textit{thirty}}^{\texttt{CARDINAL}}$" (not to mention that an unlikely transcription of *one* as *wand* produces "$\boxed{\textit{Wednesday}}^{\texttt{DATE}}$ *at* $\boxed{\textit{wand}}^{\texttt{GPE}}$ *thirty*").

Unfortunately, the problems mentioned above cannot be easily solved. Word error rates (WER) of ASR systems remain high for spontaneous human conversations. 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.

One possible solution might be to train or fine-tune NER models on transcripts of spontaneous conversations. The main obstacle is the lack of sufficient training datasets. Obtaining gold transcripts (i.e., transcripts manually tagged by human annotators) is prohibitively expensive. Additionally, annotated entity spans are not likely to generalize across application domains. NER models need to generalize patterns that appear in the vicinity of entity spans. In other words, a NER model needs to focus on the systematic regularities around entity spans. However, these spans contain many personal properties of individual speakers, their mannerisms, sociolinguistic artifacts, and regional dialect characteristics. It is highly unlikely that one can compile a training dataset representative of the majority of speakers in a given domain.

This paper investigates the true size of the ASR-NLP gap, which concerns the downstream task of recognizing named entities. Using a combination of benchmark and internal datasets, we show how state-of-the-art language models fail to discover entity spans for primary classes of named entities in transcripts of spontaneous human conversations. Our second contribution is the introduction of a new method of joint evaluation of ASR and NER models. We observe that traditional NLP metrics are not suited for measuring the performance of models on ASR transcripts. Inspired by DARPA's Message Understanding Conferences, we developed a new metric that is much more robust in measuring the performance of the NER model under transcript alignment.

2 Related Work

Word Error Rate (WER) remains the primary metric used to evaluate ASR systems. Over the years, many alternatives and amendments have been proposed. Nanjo and Kawahara (2005) introduced the idea of weighting word errors by the importance of words in the corpus. The authors develop several error weighting schemes, resulting in new metric definitions of Weighted Word Error Rate (WWER), Keyword Error Rate (KER), and Weighted Keyword Error Rate (WKER). To calculate WWER, classical TF-IDF weights are applied to words prior to counting insertions, deletions, and substitutions. In the KER scheme, only words considered to be keywords contribute to the error rate. These two schemes are combined to produce WKER, where only keywords are considered, but the weights of keywords vary. A practical example of keyword-based error rate estimation is presented in Cohn et al. (2019). Using a NER annotation scheme, the authors annotated a subset of Fisher and Switchboard datasets with Personal Health Identifier (PHI) annotation spans. The resulting metric evaluated transcription quality only within PHI spans, effectively turning all tokens within PHI spans into keywords. A very similar proposal comes from Del Rio et al. (2021) where WER is calculated only within entity spans, but these spans are not limited to a single entity type. However, another measure reported in the literature is the Slot Error Rate (SER) (Makhoul et al., 1999) defined as the ratio of the number of all slot errors (substitutions, deletions, and insertions) divided by the total number of slots.

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-score for recognizing phone numbers and 0.70 F-score 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-measure for a relatively broad set of named entities.

Precision, recall, and F-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). 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). The MUC scheme influences our method of measuring the performance of NER models on ASR transcripts.

To the best of our knowledge, Hatmi et al. (2013) were 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 both transcription tokens and their part-of-speech or named entity tags in these works.

3 Experiment

3.1 Datasets

Ruder (2021) remarks that the state-of-the-art models for Named Entity Recognition are most often evaluated on two datasets:

- the CoNLL-2003/CoNLL++ shared task (Sang and De Meulder, 2003) with annotations of persons, locations, organizations, and *misc* entity types in news stories, and
- the LDC-released OntoNotes v5 (Weischedel et al., 2013) with 18 entity types annotated in news, broadcast/telephone conversations, and Web contents.

Apart from benchmark datasets, we have used a proprietary dataset of 66 real-world call center conversations. These are multi-domain English calls recorded in standard telephony quality amounting to over 2 hours of spontaneous dialogues. The dataset has been manually transcribed and annotated with named entities, including date and time spans, mentions of persons, organizations (including brand names and facility names), locations (addresses, geopolitical entities), money, and percentages. All entity types have been mapped to CoNLL-03 and OntoNotes v5 annotation schemes. Table 1 presents the number of entity instances per entity type in the merged training set.

3.2 Entity span alignment

We measure the loss of entity spans recognized in the ASR transcript as compared to the entity spans

entity type	CoNLL-03	OntoNotes v5
outside of entity	63846	62250
ORGANIZATION	388	388
LOCATION	250	250
PERSON	240	240
MONEY		705
PERCENT		214
TIME		677

Table 1: Counts for every entity type annotation in the training set

recognized in the gold transcript. Thus, we have to perform entity span alignment between ASR and gold transcripts as they differ in the number of tokens. Alignment is performed after diarisation for each channel separately. We begin by running a NER model on the gold transcript and tagging each word 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 entity spans to only the beginning word. As a result, each channel is represented by a sequence of `B` and `O` tags. We repeat the same procedure for the ASR transcript and then we align both transcripts. The alignment is computed using the `kaldialign` library (Żelasko and Guo, 2021).

Consider the following sentence appearing in the gold transcript: "*I have called Cleveland Clinic Hospital three days ago*". There are five possible cases for entity span alignment with the output of the ASR.

3.2.1 Full alignment

Tokens in both sequences are aligned, entity spans have been correctly recognized in the ASR transcript, even if some minor ASR errors have been inserted into the transcript. This scenario is depicted in Table 2.

3.2.2 Inserted or removed tag

Due to a WER in the ASR transcript, a tag was either inserted or removed. Table 3 presents an extreme case of such a scenario.

3.2.3 Missing tag

An ASR error may have caused the entity span to shrink. In such a case, the gold transcript has a `B`-tag, and the ASR transcript has an `O`-tag. As a result, an entity span has been lost in the ASR transcript. An example of such a scenario is presented in Table 4.

gold token	NER	ASR token	NER
I	O	I	O
have	O	ϵ	
called	O	called	O
Cleveland	B-ORG	Cleveland	B-ORG
Clinic	I-ORG	Clinic	I-ORG
Hospital	I-ORG	Hospital	I-ORG
three	B-DATE	tree	B-DATE
days	I-DATE	days	I-DATE
ago	I-DATE	ago	I-DATE

Table 2: Full alignment, entity spans are recognized correctly despite the fact that ASR has changed "*three*" to "*tree*" and did not recognize the word "*have*".

gold token	NER	ASR token	NER
I	O	I	O
have	O	called	O
ϵ		Cleveland	B-GPE
called	O	hmm	O
Cleveland	B-ORG	Clinic	B-ORG
Clinic	I-ORG	Hospital	I-ORG
Hospital	I-ORG	ϵ	
three	B-DATE	three	B-DATE
days	I-DATE	days	I-DATE
ago	I-DATE	ago	I-DATE

Table 3: Inserted tag: ASR includes a backchannel "*hmm*" which confuses the NER model and divides original `ORG` entity span into `GPE` and `ORG` spans.

3.2.4 Spurious tag

An ASR error may have introduced a word that the NER model recognizes as an instance of an entity, when in fact, there is no entity span in that part of the transcript. In other words, the gold transcript has an `O`-tag, and the ASR transcript has a `B`-tag, which means that an entity span has been hypothesized in the ASR transcript. This is illustrated in Table 5.

3.2.5 Incorrect tag

However, another possibility is that an ASR error forces the NER model to recognize another type of entity in a given span. This situation occurs when both transcripts have a `B`-tag, but entity labels are different. Table 6 illustrates this scenario.

Using the MUC scheme, we can characterize the last three scenarios as missing, spurious, and incorrect, respectively. Depending on the domain

gold token	NER	ASR token	NER
I	O	I	O
have	O	ϵ	
called	O	called	O
Cleveland	B-ORG	clean	O
Clinic	I-ORG	land	O
Hospital	I-ORG	cleaning	O
ϵ		hospital	O
three	B-DATE	three	B-DATE
days	I-DATE	days	I-DATE
ago	I-DATE	ago	I-DATE

Table 4: Missing tag: ASR incorrectly transcribes "*Cleveland Clinic*" as "*clean land cleaning*", as the result the entire ORG entity span is removed.

Gold token	NER	ASR token	NER
I	O	I	O
have	O	have	O
called	O	called	O
Cleveland	B-ORG	Steve	B-PERSON
Clinic	I-ORG	Lannic	I-PERSON
Hospital	I-ORG	Hospital	I-PERSON
three	B-DATE	three	B-DATE
days	I-DATE	days	I-DATE
ago	I-DATE	ago	I-DATE

Table 6: Incorrect tag: ASR has changed an entity ORG into PERSON by erroneously transcribing "*Cleveland Clinic*" to "*Steve Lannic*".

Gold token	NER	ASR token	NER
I	O	I	O
have	O	Eve	B-PERSON
called	O	called	O
Cleveland	B-ORG	Cleveland	B-ORG
Clinic	I-ORG	Clinic	I-ORG
Hospital	I-ORG	Hospital	I-ORG
three	B-DATE	three	B-DATE
days	I-DATE	days	I-DATE
ago	I-DATE	ago	I-DATE

Table 5: Spurious tag: ASR has hypothesized an entity PERSON by changing "*have*" to "*Eve*".

of applications, some types of misalignment may be more expensive and consequential than others. When presenting experimental results, we will refrain from normalizing the errors and present raw counts of observed errors for each entity type.

4 Results

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 have evaluated five state-of-the-art ASR systems, choosing a telephony model whenever possible. Unfortunately, commercial ASR licenses prohibit the public evaluation of these systems on non-public datasets, and we cannot disclose the names of evaluated products. This section reports results obtained for the ASR system with the lowest WER on the training set. Standard ASR output is lower-cased without punctuation, and the ASR performs the output segmentation into tokens. In a real-world scenario, one would first apply a punctuation model to restore commas, periods, question marks, and exclamation marks. Then, one would apply a true-casing model to restore text casing. We focus on the ASR-NLP gap in this work, so we do not use auxiliary models but apply NER models directly to the raw ASR output.

4.1 Performance on SWNE

Recently the NLP team at Emory University released the subset of the well-known Switchboard Dialog Acts data annotated with entity spans. This subset is called SWNE. As the data set is annotated with the OntoNotes v5 entity labeling scheme, we evaluate two NER models trained on OntoNotes v5 (spaCy[1] and Flair[2]), and compare their performance with the Ontonotes v5 performance baseline.

The results presented in Table 7 show a general decline in macro-averaged F-scores by 36-44 percentage points against the baseline OntoNotes v5 on Switchboard transcripts which retain punctuation. Running the model on standard ASR output of lowercase text without punctuation costs an additional 10 to 15 percentage points, lowering the F-scores from an impressive 0.8-0.9 range to a poorly performing 0.3-0.5 range. The average loss would be even higher were it not for the language label, which denotes any named language. Number-related entities (cardinals, money, quantities) suffered a performance drop of 20-30 percentage points. Location-related entities were subject to 20-40 percentage point performance degradation,

[1] en_core_web_lg
[2] flair-ontonotes-large

and proper names (people, products, and organizations) suffered a 25-45 percentage point drop on readable transcripts. We should stress that these results are obtained for transcripts with restored punctuation and casing. The drop of F-scores for lower-cased transcripts reached 60-70 percentage points, rendering the results of the NER model completely useless in practical applications.

We are also observing a significant degradation of the date-related entity recognition. This degradation is consistent for both correctly-cased and lower-cased transcripts. Date and time-related entity spans are notoriously hard to recognize due to multiple ways to represent dates in spontaneous speech. Dates can be defined as relative ("*in three days*") or absolute ("*on Monday, May second*"). There are often hesitation markers and repetitions in the speech around dates. Many speakers confuse prepositions producing grammatically dubious utterances.

Switchboard is among the most popular resources used to train ASR models. It is safe to assume that major commercial ASRs used in our experiments have been trained on the entire data set, including the subset annotated with entity spans as the SWNE. Evaluating NER models on Switchboard would lead to an overly optimistic estimation of performance. This assumption is partially validated because we are observing much lower WERs on Switchboard compared to our internal benchmark data set. For this reason, we evaluate the size of the ASR-NLP gap using our internal benchmark data set by comparing entity recognition on gold transcripts and ASR output.

4.2 Performance on real-world conversations: gold transcripts

In the first experiment, we evaluate five state-of-the-art NER models (Wolf et al., 2020; Devlin et al., 2018) on gold transcripts. The models are evaluated using the F-score as calculated by the `seqeval` library by Nakayama (2018). As we can see in Table 8, NLP models trained on the correctly cased written text fail spectacularly in the NER task. The difference between the performance of cased vs. uncased models is striking. Both for CoNLL-03 and OntoNotes v5, the models trained on the cased data severely underperform. We also note that all models perform significantly worse than the F-score range of $(0.8 - 0.9)$, often reported as the expected performance level of NER models.

All models tend to perform better for `LOC` and `PER` entity types, but struggle to recover `ORG` entities. We hypothesize that `LOC` and `PER` entity types are easier to recognize because they are based on proper nouns. The same argument does not apply to `ORG` entities because the training set contains several rare organizations which pose a challenge to language models. The recognition of `MONEY`, `PERCENT`, and `TIME` entities is relatively poor due to the diversity of number transcriptions. Some numbers may be transcribed using digits ("*I called at 4 p.m.*"). In contrast, other numbers may be spelled out ("*My order number is one zero twelve five*"), and important entity indicators may be absent from spontaneous speech ("*Let's meet, how about four?*").

4.3 Performance on real-world conversations: ASR transcripts

To perform named entity recognition in ASR transcripts, we choose the ASR with the lowest WER on the training data, and we feed the output of the ASR to the Flair large model (Schweter and Akbik, 2020) trained on OntoNotes v5. The results are presented in Table 9. We see a dramatic drop in performance. Only 50% of `LOC` entities and 38% of `PERSON` entities are correctly matched. For `ORG` entities, the model could correctly match only 15% of spans from the gold transcript. Recognition of `MONEY`, `TIME`, and `PERCENT` is slightly better, but remains unsatisfactory. We can see that the ASR errors, which are more pronounced inside entity spans, significantly degrade the performance of the NER model. An important insight can be gained from analyzing the number of hypothesized entities. As we can see, non-existent entity spans are hypothesized mostly for `PERSON` and `TIME` entities. We attribute this behavior of NER models to the fact that they are poorly equipped to handle confused word sequences – an atypical bigram can be easily confused with the haphazard nature of person mentions. Consider an ASR error when the utterance "*how may I help you?*" is erroneously transcribed as "*how maya help you?*", from the point of view of the NER model, the term "*maya*" is a good candidate for a `PERSON` entity span. Interestingly, for each entity type, more entity spans are hypothesized than lost. It may suggest that NER models trained on the prescriptive written language are too eager to recognize entity spans.

	S-punct	S-no-punct	S-onto	F-punct	F-no-punct	F-onto
CARD	0.57	0.55	0.84	0.63	0.64	0.86
DATE	0.36	0.34	0.87	0.33	0.30	0.88
EVENT	0.21	0.05	0.41	0.38	0.22	0.71
FAC	0.17	0.05	0.36	0.40	0.24	0.79
GPE	0.83	0.82	0.91	0.86	0.80	0.97
LANG	0.90	0.79	0.63	0.97	0.87	0.74
LAW	0.36	0.00	0.38	0.27	0.22	0.62
LOC	0.37	0.35	0.64	0.48	0.36	0.78
MONEY	0.63	0.63	0.90	0.62	0.61	0.91
NORP	0.83	0.83	0.90	0.88	0.81	0.96
ORG	0.45	0.34	0.82	0.54	0.19	0.91
PERSON	0.66	0.64	0.91	0.72	0.64	0.96
PROD	0.36	0.20	0.38	0.35	0.15	0.81
QUAN	0.47	0.45	0.67	0.53	0.53	0.81
TIME	0.44	0.38	0.71	0.40	0.37	0.67
WOA	0.10	0.03	0.36	0.41	0.15	0.71
F1[macro]	0.41	0.34	0.85	0.46	0.37	0.82

Table 7: F-scores of spaCy (S) and Flair (F) models on Switchboard NER annotated gold transcripts with punctuation (punct), without punctuation (no-punct), and on the non-conversational OntoNotes v5 baseline (onto).

model	LOC	ORG	PER	MONEY	PERCENT	TIME	F-score
DistillBERT, cased, CoNLL-03	0.22	0.00	0.14				0.12
DistillBERT, uncased, CoNLL-03	0.67	0.27	0.74				0.56
BERT, cased, CoNLL-03	0.15	0.02	0.09				0.09
BERT, cased, CoNLL-03	0.25	0.02	0.26				0.17
BERT, uncased, CoNLL-03	0.68	0.32	0.83				0.61
Flair, CoNLL-03	0.71	0.37	0.59				0.56
Flair, OntoNotes v5	0.70	0.30	0.81	0.60	0.54	0.32	0.55
spaCy, OntoNotes v5	0.40	0.05	0.07	0.55	0.58	0.38	0.34

Table 8: F-scores of NER models on gold transcripts of spontaneous conversations.

5 Conclusions

In this paper, we investigate the implications of the ASR-NLP gap using as an example the problem of recognizing named entities in ASR transcripts. We find the performance of NER models to significantly deteriorate not only on ASR transcripts but also on gold transcripts. The characteristics of NER errors are consistent with the WER distribution across entity spans, as reported by Del Rio et al. (2021). In our opinion, this fact strengthens the claim that the research community should give the ASR-NLP gap more attention.

Our experiments show that cased language models trained on the prescriptive written language are not suited to transcripts of spontaneous human conversations. We attribute this to the unique characteristics of spontaneous speech and the artifacts of the psychology of conversation. Additionally, the presence of back-channeling, phatic expressions, repetitions, interjections, and the lack of sentence structure confounds NLP models and impacts prediction quality. Even the most performant language model cannot recover almost 50% of all entity spans annotated in the gold transcript when applied to the ASR transcript. To answer the question posed in the title of the paper: are we NER yet? No, we are not. Despite significant progress in NLP, spontaneous speech still poses a considerable challenge to ASR systems and downstream NLP models.

	O	LOC	ORG	PERSON	MONEY	PERCENTAGE	TIME
total	30340	68	140	144	128	40	175
matched	27485	34	25	89	81	25	81
deleted	2587	18	100	46	44	14	77
substituted	0†	4	3	1	0	0	2
lost	268‡	12	12	8	3	1	15
inserted	92	3	8	3	9	1	20
hypothesized	0	16	19	69	19	1	144

Table 9: NER model performance on ASR transcripts. Counts relate to words with either `O` or `B` tags. † indicates that `O` cannot be substituted - only lost, substitutions happen between two entity labels. ‡ indicates that the number of lost `O` tags is the sum of all labels hypothesized by the NER model on the ASR output - these are reported in label breakdown in the hypothesized row. Note that the difference in counts from Table 1 comes from the fact that here we only count the `B-` parts of each sequence.

References

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*.

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

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. *arXiv preprint arXiv:1903.07037*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

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.

John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252. Herndon, VA.

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

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

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. *Software available from https://github. com/chakki-works/seqeval*.

Hiroaki Nanjo and Tatsuya Kawahara. 2005. A new asr evaluation measure and minimum bayes-risk decoding for open-domain speech understanding. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1053. IEEE.

Sebastian Ruder. 2021. Nlp-progress. https://github.com/sebastianruder/NLP-progress.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

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.

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. *arXiv preprint arXiv:2010.03432*.

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.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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