TED-EL: A Corpus for Speech Entity Linking

Silin Li[1], **Ruoyu Song**[1], **Tianwei Lan**[1], **Zeming Liu**[2], **Yuhang Guo**[1]*
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
{lisilin, songruoyu, lantianwei, guoyuhang}@bit.edu.cn zmliu@buaa.edu.cn

Abstract

Speech entity linking amis to recognize mentions from speech and link them to entities in knowledge bases. Previous work on entity linking mainly focuses on visual context and text context. In contrast, speech entity linking focuses on audio context. In this paper, we first propose the speech entity linking task. To facilitate the study of this task, we propose the first speech entity linking dataset, TED-EL. Our corpus is a high-quality, human-annotated, audio, text, and mention-entity pair parallel dataset derived from Technology, Entertainment, Design (TED) talks and includes a wide range of entity types (24 types). Based on TED-EL, we designed two types of models: ranking-based and generative speech entity linking models. We conducted experiments on the TED-EL dataset for both types of models. The results show that our ranking-based models outperform the generative models, achieving an F1 score of 60.68%.

Keywords: Speech Entity Linking, TED-EL, Entity Linking

1. Introduction

Entity linking (Cucerzan, 2007a; Dredze et al., 2010; Le and Titov, 2018; Liu et al., 2023) involves accurately resolving the identity of a named entity within a given text and mapping it to the corresponding entity within a knowledge base while avoiding any potential ambiguity (Shen et al., 2014). This task is crucial in natural language processing, information retrieval, knowledge engineering, and data mining, as it facilitates the linking of knowledge bases and plays an important role in the field of various downstream applications, such as knowledge base population (Ji and Grishman, 2011; Yu et al., 2017), content analysis (Michelson and Macskassy, 2010), information extraction (Li et al., 2022) and question answering (Asai et al., 2020; Ye et al., 2022). Current entity linking tasks mainly rely on textual information. However, entities usually exist in textual, audio, and visual modalities context in real-world data such as social media and video websites. Therefore, we propose a speech entity linking task in this paper. Figure 1 shows an example of speech entity linking. This motivation arises from two aspects:

On the one hand, speech entity linking is important for many practical applications, for example, voice assistants. By providing voice assistants with more information in the knowledge base, speech entity linking has the potential to enable voice assistants to answer user queries better.

On the other hand, as shown in Table 1, despite the recent success in entity linking, the inclusion of audio modality has been completely overlooked. Given the widespread dissemination of short videos worldwide, it is necessary and urgent to investigate speech entity linking.

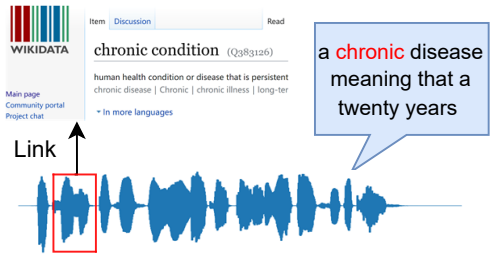

Figure 1: Speech entity linking: recognize mentions from speech and link them to entities in the knowledge base.

In this study, we have undertaken the following endeavors to advance speech entity linking: First, we construct a large-scale human-annotated speech entity linking with textual and acoustic contents, named TED-EL[1]. Specifically, we annotate all occurrences of 24 entity types in 2,351 documents originating from the transcripts of TED LIUM 3 (Hernandez et al., 2018), a corpus that has been widely employed in automatic speech recognition (ASR) (Karita et al., 2019; Ravanelli et al., 2021).

Second, we developed a cascade model based on the pipeline approach (mention detection followed by entity disambiguation). However, the excessive number of cascaded components in the pipeline can lead to significant error propagation. To mitigate this issue, we introduced joint training in mention detection and entity disambiguation to reduce error propagation. Additionally, we explored the possibility of performing mention detection di-

*Corresponding author

[1]We conducted an NLPCC shared task(Song et al., 2022) using this dataset, and the data has been released at https://github.com/BITHLP/TED-EL.

LREC-COLING 2024, pages 15721–15731

Dataset	Modality	Speech	source	Mention	Document	Types	KB
MSNBC (Cucerzan, 2007b)	text ⇒ text	✗	News	656	20	Multiple	Wikipedia
AIDA-CONLL (Hoffart et al., 2011)	text ⇒ text	✗	News	27,816	1,393	Multiple	Wikipedia
AQUAINT (Milne and Witten, 2008)	text ⇒ text	✗	News	727	50	Multiple	Wikipedia
TAC-KBP (Ji et al., 2010)	text ⇒ text	✗	News,Web	2,250	2,231	Multiple	Wikipedia
ACE2004 (Ratinov et al., 2011)	text ⇒ text	✗	News	257	35	Multiple	Wikipedia
Snap (Moon et al., 2018)	visual,text ⇒ text	✗	Social Media	-	12,000 captions	Multiple	Freebase
Twitter (Adjali et al., 2020)	visual,text ⇒ text	✗	Social Media	1,678	4M tweets	PER,ORG	Twitter Users
Movie (Gan et al., 2021)	visual,text ⇒ text	✗	Movie Reviews	181,240	1,000 reviews	PER	Wikipedia
Weibo (Zhang et al., 2021)	visual,text ⇒ text	✗	Social Media	-	25,000 posts	PER	Baidu
WikiDiverse (Wang et al., 2022)	visual,text ⇒ text	✗	News	7,969	8,000 captions	Multiple	Wikipedia
TED-EL (ours)	**speech**,text ⇒ text	✓	TED Talks	65,873	2,351	Multiple	Wikidata

Table 1: Overview of EL and MEL datasets. The symbol "-" indicates that the article of the dataset doesn't provide information on the size of its mentions.

rectly on the speech signal to further mitigate error propagation.

Third, based on the GENRE (De Cao et al., 2020) framework, we also developed a cascade model using an encoder-decoder architecture. Furthermore, we explored the feasibility of an end-to-end speech entity linking model by replacing the encoder with a speech feature extraction model.

In summary, this work makes the following contributions:

- To our best knowledge, we first propose the task of speech entity linking, which aims to recognize mentions from audio and link them to the corresponding knowledge base.
- To facilitate the study of this task, we create the first speech entity linking dataset, TED-EL, which is a parallel corpus for speech, the transcribed text, and mention-entity pairs.
- To address this task and support future research, we introduce two distinct model types in our study: ranking-based and generative models. The ranking-based model with joint ASR and mention detection yields SOTA results on TED-EL.

2. Related work

Text Entity Linking To facilitate entity linking research, several datasets have been developed in previous works. The MSNBC (Cucerzan, 2007b) dataset is one of the earliest datasets that can be used for entity linking tasks, with labels derived from news articles in ten different areas. AQUAINT (Milne and Witten, 2008) contains documents collected from the Xinhua News Agency, the New York Times, and the Associated Press, where the first mentioned entity is manually annotated to Wikipedia. TAC-KBP 2010 (Ji et al., 2010) is created for the Text Analysis Conference (TAC) based on news and web sources. The AIDA-CoNLL (Hoffart et al., 2011) dataset is a widely used public dataset for entity disambiguation and entity linking tasks. In this dataset, entities are identified using YAGO2 entity names, Wikipedia URLs, or Freebase MID. To achieve entity linking task, popular earlier methods address the mention detection and entity disambiguation stages of entity linking separately (Daiber et al., 2013; Kannan Ravi et al., 2021) while modern techniques leverage their mutual dependency (Kolitsas et al., 2018; Zhang et al., 2016). A new line of work (De Cao et al., 2020) departs from linking mentions using a vector space and instead uses large language models fine-tuned with a generative objective. However, as stated in De Cao et al. (2020), numerous methods have achieved comparable and impressive results in the past three years. One possible explanation is that it may simply be near the ceiling of what can be achieved for these datasets, and it is difficult to conduct further research based on them.

Multimodal Entity Linking In recent years, with the increasing importance of multimodal data, there has been a shift towards extending the research of entity linking from monomodality to multimodality. Moon et al. (2018) address the MEL task with a zero-shot framework that extracted textual, visual, and lexical information for EL in social media posts. However, their proposed dataset is unavailable due to GDPR rules. Adjali et al. (2020) propose a framework for automatically building the MEL dataset from Twitter. Although this dataset has limited entity types and mentions ambiguity, it is still a useful resource for MEL research. Zhang et al. (2021) study a Chinese MEL dataset collected from the social media platform Weibo, which mainly focuses on person entities. Gan et al. (2021) release a MEL dataset collected from movie reviews and propose to disambiguate both visual and textual mentions. This dataset mainly focuses on characters and persons in the movie domain. Zhou et al. (2021) propose three MEL datasets built from Weibo, Wikipedia, and Richpedia information, and use CNDBpedia, Wikidata, and Richpedia as the corresponding knowledge bases. There have also been some recent remarkable works on the task of multi-modal entity linking. For example, Adjali et al. (2020) propose a model that integrates visual,

textual, and statistical information to perform MEL. Zhang et al. (2021) introduce a two-stage mechanism that initially determines the relations between images and texts to mitigate the negative impacts of noisy images, followed by the disambiguation process.

However, different from them, we aim to explore an unexplored territory in this work, which is speech entity linking with both speech and textual contents. To our best knowledge, TED-EL has the highest number of mentions in multi-type entity-linking datasets.

3. Dataset Construction

In this section, we present the task of multimodal entity linking with speech and the dataset construction procedure.

3.1. Speech Entity Linking Task

Given a $talk = A$ and a knowledge base (ε), where A is an audio segment. The task is to detect a set of entity mentions $M = \{m_0, m_1 \dots m_n\}$ in the audio A and link each mention m_i to the corresponding entity entry e_i in the ε.

3.2. Dataset Collection

Data Selection To ensure the data diversity and size of the dataset, careful consideration was given to the selection of the speech-text pairs. TED was founded in 1984 as a conference on technology, entertainment, and design, covering almost all topics and over 2,600 video demonstrations, and is considered the best source of our corpus. As for the knowledge base, the widely-used Wikidata was utilized, with all of the annotated content in Wikidata being made available to facilitate flexible and convenient research.

Data Source To obtain the speech-text pairs, we utilized the TED presentation dataset, which comprises 2,351 talks available on the official TED website, and for the transcribed text, we used the TED LIUM Release 3. These talks cover a wide range of topics commonly encountered in the real world. For the knowledge base, we employed Wikidata, which boasts a large entity set comprising all entities in the main namespace.

3.3. Annotation

3.3.1. Annotation Guidelines

The objective of entity linking is to link mentions to their corresponding entities in Wikidata. Thus, annotators were tasked with identifying mentions from the text and assigning each detected mention to the appropriate entity in the form of a Wikidata ID. If there is no corresponding entity in Wikidata for a detected mention, it is labeled as "NULL" (5.38% in all entities). The TED-EL dataset contains no nested named entities. Therefore, the first guideline for annotating entity boundaries is the longest match principle which means that annotations are conducted at the level of complete entities. When annotating entity boundaries, a complete word must be annotated without splitting it, and overlapping annotations are not allowed. For example, in the phrase "Harry Potter and the Deathly Hallows" according to the longest match principle, it should be annotated as a single named entity without separately annotating "Harry Potter". When annotating entity boundaries, even if there are multiple repeated entities in a sentence, their positions in the sentence may contain different information. To avoid missing cases where the same word represents different entities, all entities in the sentence need to be annotated. Each annotated entity needs to include its position information in the sentence, i.e., the $offset$, which allows for distinguishing the annotated entities based on their positions.

3.3.2. Quality Control

In the annotation process, manual annotation was employed to control the quality of annotations. Multiple annotators were involved, and they were guided and supervised in a phased manner throughout the entire annotation process. In the initial stage, annotators underwent rigorous training based on annotation guidelines and were given test data to annotate. They were qualified to annotate the actual dataset only when their annotation results met the guidelines. During formal annotation, a multi-annotator approach was adopted, where two annotators independently annotated the same data, and their results were compared. If there were discrepancies between the two annotations, a third experienced annotator with good annotation quality would intervene to re-annotate and obtain the final annotation result. To ensure quality control during annotation, two researchers who specialized in entity linking were involved in different stages to assess the quality of the annotations. It was required that the F1 score of the data exceed 0.95. The annotation process was conducted in batches, with each batch of data being checked by researchers through random sampling. If the quality of a batch did not meet the standard, the data in that batch would need to be re-annotated, and any newly identified issues were incorporated into the subsequent annotation guidelines.

Data Quality After the completion of annotation, the dataset underwent a cleaning process

primarily targeting errors in the formatting of entity ID labels. A random sampling check was conducted on the entire dataset to ensure an F1 score of above 0.95. Only after completing all the aforementioned tasks, the quality control process for the dataset could be formally concluded.

3.4. Dataset Analysis

A single annotated data instance in the dataset consists of a dictionary with six elements: $\{text\_id, file\_name, start, end, text, mention\_data\}$. The $text\_id$ indicates the position of the data within a document, the $file\_name$ indicates the source speech file, $start$ and end indicate the corresponding audio segment's start and end positions, $text$ represents the correct transcription of the audio, and $mention\_data$ contains information about the entities mentioned in the audio segment. The $mention\_data$ is a list comprising $\{kb\_id, mention, offset, type\}$ elements, where kb_id refers to the corresponding QID of the entity in Wikidata, and $type$ denotes the entity type.

Entity Mention	Num	65,873
	Avg Length	1.55
NULL Entity	Num	3,546
	Percent	5.38%
Sentence	Num	47,058
	Avg Length	21.95
Audio	Total Duration	94.7h
	Avg Utterance Duration	7.24s

Table 2: Statistics of TED-EL

Size and Distribution We divide TED-EL into train data and test data with a ratio of 9:1. Train data has 1,695 speakers and 1,936 documents. Test data has 144 speakers and 151 documents. It is worth noting that the two datasets have non-overlapping speakers. We conducted statistics on the TED-EL dataset from three aspects: entity mentions, sentences, and audio. The statistical information is presented in Table 2. There are 65,873 entity mentions in total, with an average length of 1.55 words. For entities that do not exist in the knowledge base, they are identified as NULL entities. Out of the 65,873 entity mentions in the TED-EL dataset, there are only 3,546 NULL entities, accounting for 5.38% of all entities. There are a total of 47,058 sentences containing mentions, with an average sentence length of 21.95 words. After text and audio alignment, the total available audio length is 94.70 hours, with an average sentence audio length of 7.24 seconds.

4. Our Approach

As shown in Figure 2, we have established three ranking-based speech entity linking models (i.e. RBSEL and its joint variations RBSEL-J1 and RBSEL-J2) and two generative speech entity linking models (GSEL and its joint variation GSEL-J). RBSEL is a cascaded model consisting of ASR, mention detection, and entity disambiguation components. RBSEL-J1 is a variation based on RBSEL, which jointly trains mention detection and entity disambiguation. RBSEL-J2 is also a variation based on RBSEL, which jointly trains ASR and mention detection. GSEL is a cascaded model based on a generative approach, consisting of ASR, encoder, and decoder. GSEL-J is an end-to-end generative model based on GSEL, where the ASR model is combined with the encoder, and a length adapter is added before the decoder. The reason for building joint models (i.e. RBSEL-J1, RBSEL-J2, and GSEL-J) is that we found the cascaded approach suffers from cascaded errors in our experiments.

4.1. Mention Detection

We employed the BERT (Devlin et al., 2018) pretrained model by adding $[CLS]$ and $[SEP]$ tokens at the beginning and end of the input sentence, which was then fed into the BERT model. Subsequently, we utilized logistic regression to classify each token into BIO labels.

4.2. Joint ASR and Mention Detection

We refer to joint ASR and mention detection as speech mention detection and it can be considered as entity perception without modifying the ASR system structure. It only requires adding special token symbols to the ASR vocabulary to identify entity mentions. The special token symbols (i.e.,"[]") are added on both sides of the entity mention to mark it. The entity-aware ASR learns the alignment between speech and transcribed text. During decoding, the entity mentions in the transcribed text are bound to their corresponding special token pairs. Therefore, the entity-aware ASR can detect entity mentions during decoding. The speech mention detection module outputs a set of entity mentions, $M = \{m_1, m_2, ..., m_n\}$, where each entity mention m_i is input to the entity disambiguation module.

4.3. Entity Disambiguation

Entity disambiguation consists of two components: candidate entity generation and candidate entity ranking.

Candidate Entity Generation Given a knowledge base ε, a text input, and a set of entity men-

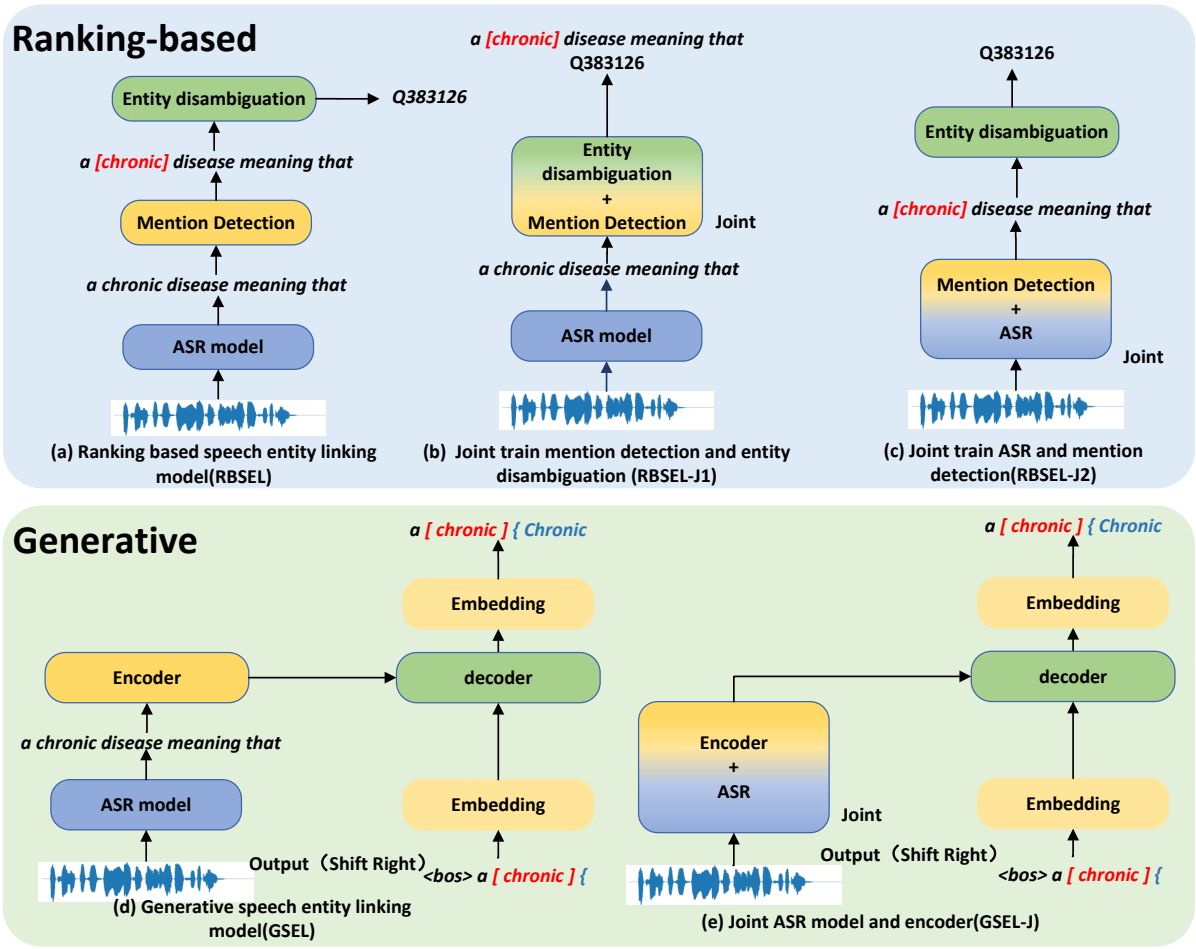

Figure 2: The architecture of five speech entity linking models based on ranking and generative models.

tions M, the objective of candidate entity generation is to generate, for each entity mention m_i, the top K candidate entities in ε that m_i could potentially be linked to. This involves outputting a list of mention-candidate entity pairs $(m_i, Cand_i), i \in [1, n]$, where $Cand_i = \{c_1, ..., c_k\}, c_j \in \varepsilon$. The candidate entity generation model employs a dual-tower encoder, consisting of two independent BERT models. The representations of the context and entity mention are constructed based on the surrounding context of the mention and the mention itself. To enable the model to understand the current mention being disambiguated, we adopt the approach of inserting mention markers to highlight the mention. Each input context is constructed as follows:

$$P = BERT([CLS]ctx_l[M]men[\backslash M]ctx_r[SEP]), \quad (1)$$

where men, ctx_l, ctx_r correspond to the mention itself, the preceding context, and the succeeding context, respectively. $[M]$ and $[\backslash M]$ are special tokens placed before and after the mention, respectively. The input to the entity encoder consists of:

$$E = BERT([CLS]des[SEP]), \quad (2)$$

where des is the entity description provided by the knowledge base. The context p and candidate entity e_i are encoded into vectors P and E. The score of the candidate entity e_i is calculated based on the dot product between the encoded context and entity vectors :

$$score_g(p, e_i) = P^{\mathrm{T}} \cdot E \quad (3)$$

Using this score, we select the top K candidate entities.

Candidate Entity Ranking For candidate entity rerank, we employ a cross-encoder that allows for deep cross-attention between the context and entity descriptions, enabling a better understanding of the relationship between them. The context and entity are encoded into a vector representation H:

$$En([CLS]ctx_l[M]men[\backslash M]ctx_r[SEP]des[SEP]) \quad (4)$$

To score the candidate entities, we apply a multi-layer perceptron with a sigmoid activation function to the last layer's $[CLS]$ representation H_1 to map it to the range $(0, 1)$:

$$o = \sigma\left(H_1^T W\right) \in R^d \quad (5)$$

$$score_r\left(p, e_i\right) = sigmoid\left(o^T w\right), \tag{6}$$

where $W \in R^{h \times d}$ is a learnable matrix, and both vectors o and w have a dimensionality of d. Based on the computed scores, we perform candidate entity rerank according to the $score_r$ and select the entity with the highest score.

4.4. Joint Mention Detection and Entity Disambiguation

This module implements a dual-encoder structure based on BERT, which is capable of simultaneous mention recognition and entity disambiguation. Specifically, this module represents entity in the entity library with e and token in the transcribed text t, respectively:

$$\mathbf{x}_e = \mathrm{BERT}_{[CLS]}\left([\mathrm{CLS}]t\left(e_i\right)[\mathrm{ENT}]d\left(e_i\right)[\mathrm{SEP}]\right) \in R^h \tag{7}$$

$$\left[\mathbf{t}_1 \cdots \mathbf{t}_n\right]^\top = \mathrm{BERT}\left([\mathrm{CLS}]t_1 \cdots tn[\mathrm{SEP}]\right) \in R^{n \times h} \tag{8}$$

In the above formula, $t(e_i)$ represents the name of the entity, and $d(e_i)$ represents the description text of the entity. To detect mentions in a sentence, we consider all substrings whose length does not exceed L (predefined). We then calculate scores for each token at the start, end, or middle of a mention:

$$\begin{aligned} s_{\text{start}}\left(i\right) = & w_{\text{start}}^T t_i, \quad s_{\text{end}}\left(j\right) = w_{\text{end}}^T t_j, \\ & s_{\text{mention}}\left(m\right) = w_{\text{mention}}^T t_m, \end{aligned} \tag{9}$$

where w is a learnable vector that subsequently derives the score of the mentioned substring:

$$p([i,j]) = \sigma\left(s_{\text{start}}\left(i\right) + s_{\text{end}}\left(j\right) + \sum_{m=i}^{j} s_{\text{mention}}\left(m\right)\right) \tag{10}$$

For entity disambiguation, we obtain the representation of a mention by taking the average of token representations in the mention:

$$y_{i,j} = \frac{1}{(j-i+1)} \sum_{k=i}^{j} t_k \in R^h \tag{11}$$

Further, we compute similarity scores $s(e,[i,j])$ between mentions and entities, and the probability $p(e|[i,j])$ of entity e corresponding to text $[i,j]$:

$$s(e,[i,j]) = x_e^T y_{i,j} \tag{12}$$

$$p(e \mid [i,j]) = \frac{\exp(s(e,[i,j]))}{\sum_{e' \in \mathcal{E}} \exp\left(s\left(e',[i,j]\right)\right)} \tag{13}$$

During training, joint training is achieved by summing the loss of mention recognition and entity disambiguation. The loss for mention recognition and entity disambiguation are as follows:

$$\begin{aligned} L_{MD} = -\frac{1}{N} \sum_{1ij\ \min(i+L-1,n)} & \left(y_{[i,j]}\ \mathrm{logp}([i,j]) + \left(1 - y_{[i,j]}\right) \log(1 - p([i,j]))\right) \end{aligned} \tag{14}$$

$$L_{ED} = -\log p\left(e_g \mid [i,j]\right) \tag{15}$$

The loss for mention recognition uses binary cross-entropy loss. If $[i,j]$ is the correct mention span, then $y_{[i,j]}$ being 1, otherwise it is 0.

4.5. Generative Speech Entity Linking

GSEL The part of the text entity linking model was proposed by De Cao et al. (2020). It transforms the entity linking task into a translation task, generating labeled text to achieve mention recognition and entity disambiguation. To constrain the decoding space, a method of dynamically computing the decoding constraint trie is employed. At each generation step, the decoder is either generating a mention span, generating a link to a mention, or continuing from the input source. When outside a mention/entity step, the decoder has only two options: (i) to continue by copying the next token from the input source, or (ii) to generate the "start of mention" token (i.e., "["), which makes the decoder enter the mention generating phase. While generating a mention, the decoder has either to continue with the next token in the input source or to generate the "end of mention" token (i.e., "]"), which makes the decoder enter the entity generating phase. Finally, when generating an entity, the decoder employs the entities trie such that it can only output a valid entity identifier.

GSEL-J We formulate GSEL-J as a speech-to-text task that requires a speech encoder and a text decoder. We employ the Hubert-large (Hsu et al., 2021) as our encoder. We take the decoder component of BART-large as the text decoder. As same as GSEL, we use a Markup annotation where span boundaries are flagged with special tokens "[]" and accompanied by their corresponding entity identifiers. Simply combining them can lead to optimization problems because they are pre-trained on different modalities that differ significantly in length. To address this issue, we introduce a length adapter made of n number of 1-D convolutional layers, each parameterized with kernel size p, stride s, and padding p. We follow the partial training strategy used by Gállego et al. (2021) and train the length adaptor together with part of the encoder and decoder (including encoder self-attention, encoder-decoder cross-attention, and layer normalization) while freezing the rest of the parameters. The trained parameters account for 20% of the entire model. This training strategy

is efficient while retaining performance in speech translation.

5. Experiments and Results

5.1. Evaluation Metrics

We have chosen commonly used metrics in text entity linking: precision, recall, and F1 score (Li et al., 2020; Kannan Ravi et al., 2021; De Cao et al., 2021). However, due to the possibility of disparities between the text recognized by speech recognition and the correct text, we evaluate by not using mention boundaries as the correct answers for mention identification. Instead, we consider the mention itself as the correct answer. For the ASR model in the cascaded architecture, we evaluate its performance using Word Error Rate (WER).

5.2. Implementation Details

RBSEL, RBSEL-J1, RBSEL-J2, and GSEL utilize the Wav2vec2-Conformer[2] model for their ASR component, with pre-trained model weights initialization. The ASR model is trained on the TED-EL dataset. The audio data is sampled at a rate of 16,000, with a batch size of 2 during training. The learning rate is set to 1e-5, and the AdamW (Loshchilov and Hutter, 2017) optimizer is used with a warmup step count of 500 and a gradient accumulation step count of 2. The training cycle consists of 10 epochs for speech mention detection and 7 epochs for ASR.

For GSEL-J, the speech encoder incorporates both Hubert[3] and Wav2vec2.0[4] (Baevski et al., 2020), while the text decoder adopts the BART[5] (Lewis et al., 2020) model. To address coreference resolution challenges, the approach from GENRE is referenced, and the BART model is pre-trained on the BLINK dataset. The three CNN layers of the length adaptor have kernel size, stride, and padding set to 3, 2, and 1, respectively. The models are trained for 8k updates with early stopping after 20 updates. The training employs a label-smoothed cross-entropy loss function with a smoothing parameter of 0.2. During training, an efficient fine-tuning strategy is employed, where certain modules have fixed parameters. The fixed modules include the feature extractor of the encoder, the feed-forward neural network of the encoder, the self-attention module of the decoder, and the feed-forward neural network of the decoder. The length adaptor, encoder self- attention, encoder-decoder cross-attention, and layer normalization are trained modules. The GSEL-J experiments are conducted using the fairseq framework, while the remaining experiments utilize the PyTorch framework. All models are evaluated on the best-performing checkpoint on the validation set. All experiments are conducted on a TITAN GPU.

[2] https://huggingface.co/docs/transformers/main/en/model_doc/wav2vec2-conformer

[3] https://dl.fbaipublicfiles.com/hubert/hubert_large_ll60k_finetune_ls960.pt

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[5] https://dl.fbaipublicfiles.com/fairseq/models/bart.large.tar.gz

Method	P	R	F1	Extra Data
RBSEL	53.38	50.98	52.15	✗
RBSEL-J1	**61.39**	53.98	57.45	✗
RBSEL-J2	61.08	**60.29**	**60.68**	✗
GSEL	56.61	37.08	44.81	✓
GSEL-J	52.44	52.79	52.61	✓

Table 3: Precision (%) , Recall (%) and F1 score (%) of our proposed models on TED-EL.

5.3. Main Result

Table 3 shows the results of our proposed models on TED-EL. According to the experimental results, we can find that:

(1) RBSEL-J1 achieves a 5.3% higher F1 score compared to RBSEL, indicating that joint training of mention recognition and entity disambiguation is effective in reducing cascaded errors.

(2) RBSEL-J2 achieves an 8.53% higher F1 score compared to RBSEL, demonstrating that speech mention recognition can improve the overall performance of entity linking. In the next chapter, we analyze the reasons behind this performance improvement in more detail.

(3) GSEL-J outperforms GSEL by 7.8%, indicating that combining ASR with the encoder can reduce cascaded errors and improve the overall performance of entity linking.

(4) Compared to RBSEL, GSEL shows inferior overall performance. This is attributed to the encoder-decoder framework's need to handle both mention recognition and entity disambiguation, which increases the model's complexity. Additionally, the encoder-decoder architecture fails to fully leverage the acoustic information of entities, posing significant challenges for entity disambiguation.

6. Analysis

6.1. Speech Mention Detection

To further demonstrate the effectiveness of speech mention detection, we compared the performance of RBSEL-J2 and RBSEL in mention detection. Additionally, we compared the mention detection per-

Error Type	Golden Text	RBSEL-J2 Predict Text
Boundary Error	the head of { cascadia } corridor that stretches	the head of { cascadia corridor } that stretches
ASR Error	an order from { wal mart }	an order from { walmart }
Identifiers Mismatch	president of { the united states } { brooks }	president of { the { united states brooks }
Mention Undetected	{ himalayan kingdom } nestled	himalayan kingdom nestled

Table 4: Four error types of RBSEL-J2's prediction on mention detection compared with the golden mentions. Green color means right mention, and red color means wrong mention.

Method	WER	P	R	F1
RBEL	-	83.51	82.16	82.83
RBSEL	14.36	57.89	55.29	56.56
RBSEL-J2	16.03	59.03	58.27	58.64

Table 5: Comparison among the RBEL, RBSEL, and RBSEL-J2 on mention recognition, where RBEL is the text entity linking model of RBSEL with input as the correct transcribed text.

Method	P	R	F1
RBEL	69.61	68.48	69.04
RBEL-J1	70.03	66.72	68.33
GEL	61.26	63.51	62.37

Table 6: Comparison among the RBEL, RBEL-J1, and GEL on text entity linking, where RBEL, RBEL-J1, and GEL are the text entity linking models of RBSEL, RBSEL-J1, and GSEL, respectively.

formance of RBSEL-J2 on the correct transcripts. As shown in Table 5, there is a significant difference of 26.27% in terms of F1 score between RBEL and RBSEL. This phenomenon indicates that context is a key factor influencing mention detection, and errors in the context have a severe impact on mention detection performance. RBSEL-J2 achieves the best performance in speech mention detection with an F1 score of 58.64%, which is 2.08% higher than RBSEL. However, RBSEL-J2 has a WER of 16.03%, slightly higher than that of the ASR system, mainly due to identifier recognition errors.

To better understand the source of errors in RBSEL-J2, we sampled a portion of error cases and compared the predicted results with the gold standard, as shown in Table 4. We categorized these errors into four types: boundary recognition errors, ASR errors, unmatched identifiers, and undetected mentions. Boundary recognition errors result in boundary errors in the subsequent entity disambiguation phase. ASR errors affect the semantic context, for example, a proper noun like "wal mart". The pre-trained model originally knew the current mention of "wal mart", but identifying it as "walmart" tests the fault-tolerant ability of the entity disambiguation model. The problem of identifier mismatch and undetected mentions is more serious, as the cascaded method directly leads to the entity disambiguation module not disambiguating the currently mentioned ambiguity.

6.2. Entity Linking on Golden Text

We evaluated RBSEL, RBSEL-J1, and GSEL on golden text, and the results are shown in Table 6. The ranking-based entity linking models consistently outperformed the generation-based models. This further demonstrates the difficulty of achieving excellent mention recognition and entity disambiguation capabilities simultaneously in the encoder-decoder framework. To further enhance the effectiveness of generation-based entity linking models, it is necessary to explore how to leverage the semantic information of entities better.

During the experiment, we observed that when conducting experiments on transcribed text, the performance of the joint model (including mention detection and entity disambiguation) was relatively inferior to the cascade model. However, when conducting speech experiments, the performance of the joint model was relatively better. We explain this phenomenon as follows:

The RBSEL model consists of three modules: ASR, MD, and ED. Due to the identification error of ASR, the results of MD are degraded, which in turn affects the performance of ED. However, in the RBSEL-J1 model, the MD and ED components are jointly modeled, which helps to mitigate the impact of ASR errors on the overall performance to a certain extent.

7. Conclusion

We introduce a novel task of speech entity linking for the first in this paper, which is to link entities in speech to their corresponding entries in a knowledge base. To accomplish this task, we construct a large-scale manually annotated speech entity linking dataset, named TED-EL, which is derived from TED talks. Based on this dataset, we establish ranking-based and generative speech entity linking models. Furthermore, we propose RBSEL-J1, RBSEL-J2, and GSEL-J to reduce cascaded errors, and experimental results demonstrated that the three joint models can improve the performance of speech entity linking.

Limitations

We did not fully leverage the information contained in speech signals, such as speaker style and emotion, which could benefit the task. There is still a gap in performance between speech-based entity linking and the results achieved on correct text. Additionally, the audio inputs are limited to a maximum of 20 seconds, as processing long-duration audio signals poses challenges due to computational resource limitations, a known issue in the speech domain. Addressing these three limitations is part of our future research plan.

Ethical Considerations

We ensure that the collection of TED-EL is conducted in a manner consistent with the terms of use of any sources, as well as the intellectual property and privacy rights of the original texts. Crowd workers are treated fairly, including aspects such as fair compensation, informed consent, and voluntary participation with awareness of any potential risks. For further details on the specific characteristics and collection process of TED-EL, please refer to Section 3.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments. This work is supported by the National Key R&D Program of China (No. 2020AAA0106600) and National Natural Science Foundation of China (No. U21B2009).

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Building a multimodal entity linking dataset from tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4285–4292, Marseille, France. European Language Resources Association.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Silviu Cucerzan. 2007a. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Silviu Cucerzan. 2007b. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th international conference on semantic systems*, pages 121–124.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Highly parallel autoregressive entity linking with discriminative correction. *arXiv preprint arXiv:2109.03792*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive Entity Retrieval. *arXiv e-prints*, page arXiv:2010.00904.

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*.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 277–285, Beijing, China. Coling 2010 Organizing Committee.

Gerard I Gállego, Ioannis Tsiamas, Carlos Escolano, José AR Fonollosa, and Marta R Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: Upc at iwslt 2021. *arXiv preprint arXiv:2105.04512*.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 993–1001, New York, NY, USA. Association for Computing Machinery.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 782–792.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1148–1158.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third text analysis conference (TAC 2010)*, volume 3, pages 3–3.

Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. 2021. CHOLAN: A modular approach for neural entity linking on Wikipedia and Wikidata. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 504–514, Online. Association for Computational Linguistics.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient One-Pass End-to-End Entity Linking for Questions. *arXiv e-prints*, page arXiv:2010.02413.

Xiangsheng Li, Jiaxin Mao, Weizhi Ma, Zhijing Wu, Yiqun Liu, Min Zhang, Shaoping Ma, Zhaowei Wang, and Xiuqiang He. 2022. A cooperative neural information retrieval pipeline with knowledge enhanced automatic query reformulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 553–561, New York, NY, USA. Association for Computing Machinery.

Yi Liu, Yuan Tian, Jianxun Lian, Xinlong Wang, Yanan Cao, Fang Fang, Wen Zhang, Haizhen Huang, Weiwei Deng, and Qi Zhang. 2023. Towards better entity linking with multi-view enhanced distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9743, Toronto, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Matthew Michelson and Sofus A Macskassy. 2010. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Ruoyu Song, Sijia Zhang, Xiaoyu Tian, and Yuhang Guo. 2022. Overview of the nlpcc2022 shared task on speech entity linking. In *Natural Language Processing and Chinese Computing*, pages 294–299, Cham. Springer Nature Switzerland.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6032–6043, Dublin, Ireland. Association for Computational Linguistics.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. *arXiv preprint arXiv:1704.06194*.

Li Zhang, Zhixu Li, and Qiang Yang. 2021. Attention-based multimodal entity linking with high-quality images. In *Database Systems for Advanced Applications*, pages 533–548. Springer International Publishing.

Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2016. A joint model for question answering over multiple knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. 2021. Weibo-MEL, Wikidata-MEL and Richpedia-MEL: Multimodal Entity Linking Benchmark Datasets. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 315–320, Singapore. Springer Singapore.