S2E: TOWARDS AN END-TO-END ENTITY RESOLUTION SOLUTION FROM ACOUSTIC SIGNAL

Kangrui Ruan[⋆] Columbia University

Xin He[⋆]*, Jiyang Wang, Xiaozhou Zhou, Helian Feng, Ali Kebarighotbi* Amazon Alexa

ABSTRACT

Traditional cascading Entity Resolution (ER) pipeline suffers from propagated errors from upstream tasks. We address this issue by formulating a new end-to-end (E2E) ER problem, Signal-to-Entity (S2E), resolving query entity mentions to actionable entities in textual catalogs directly from audio queries instead of audio transcriptions in raw or parsed format. Additionally, we extend the E2E Spoken Language Understanding framework by introducing a novel dimension to ER research. We adapt three public datasets for the S2E task, and propose a novel solution, which aligns the multimodal signals via an effective retrieval co-attention mechanism and refined multimodal objectives. Despite 42% smaller in terms of the total model size, the proposed design outperforms the cascading baseline by 2.6%, 47.0%, and 73.3% across the three datasets respectively with different acoustic conditions.

***Index Terms*—** Entity Resolution, End-to-End, Multimodal Alignment

1. INTRODUCTION

Entity Resolution maps entity mentions in customer utterances to actionable entities stored in catalogs [1] and serves as a crucial component for voice assistants (VAs) such as Amazon Alexa and Apple Siri [2]. Within these VAs, the long prevalent standard is a cascading design which typically consists of three major components: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and Entity Resolution (ER). Specifically, ASR first transcribes customer audio signal into text, NLU next comprehends customer's intention and processes the generated transcriptions. Ultimately, ER resolves the entity mention identified by NLU.

Based on this design, ER has traditionally focused on textual information retrieval [3] and similarity measures [4]. However, this design makes ER particularly vulnerable to upstream errors. For instance, ASR transcription errors and NLU Named Entity Recognition (NER) spanning and slot typing errors can all be detrimental to ER. Although upstream error robust ER has been proposed [5], the effectness is still limited. Whereas end-to-end (E2E) ER solution from audio signal allows us to directly avoid upstream error propagation and use richer customer signals. Recent successes in E2E Spoken Language Understanding (SLU) research have showcased the possibility of integrating ASR and NLU [6, 7], which inspires us to apply the E2E approach to ER.

S2E allows to directly optimize ER performance instead of locally optimizing a series of relevant tasks. Compared to the SLU problem, we 1) intake multimodal signals: audios and textual catalogs, 2) aim to retrieve the relevant entity mentions. The multi-modality aspect makes our problem more challenging than the classic text-based ER problem. Other cross-modality studies such as CLIP [8] aligning image and text and CLAP [9, 10] aligning text and audio pair, which assume a one-to-one mapping relationship, our problem handles multiple-to-one relationships with multiple audios mapping to the same entity. Moreover, S2E offers practical advantages of reducing coordination effort, saving model maintenance cost and being footprint economical. It reduces the necessity for additional components, such as ASR transcription decoder, separate NER head and becomes advantageous in environments with limited resources (e.g., edge devices).

Our main contributions are summarized as follows: 1) we formulate a new end-to-end ER research problem: S2E and adapt three public datasets for this problem, covering a various catalog sizes, entity types and acoustic recordings conditions; 2) we propose a novel S2E solution by enhancing modality alignment between audio and entity text signals via an effective retrieval co-attention mechanism and refined training objectives composed of a modality alignment loss and an audio discriminative loss; 3) we conduct comprehensive experiments to demonstrate the efficacy of the proposed design and attribute the contribution of each individual component. Remarkably, the proposed method surpasses the cascaded solution, with ER recall@1 improvements of 2.6%, 47.0%, and 73.3% across three datasets respectively, although being 42% smaller in terms of model parameters.

2. PROBLEM FORMULATION

Here we formally formulate the signal-to-entity (S2E) problem. Given a catalog with a set of entities $\boldsymbol{\mathcal{E}} = \{e_1, e_2, \ldots, e_E\}$, where E denotes the catalog size and e_j denotes an textual entity. Let X denote the query audios and M denote the entity mentions in X. S2E's goal is to create a function $\mathcal{F} : M \rightarrow \boldsymbol{\mathcal{E}}$, which maps the entity mention in a given query audio to the correct textual catalog entity label. We only consider the in-catalog entity mentions and assign a separate label for queries with no entity mention during evaluation, similar as [3].

3. METHODOLOGY

Our proposed solution shown in Figure 1 consists of three modules: multimodal embeddings extraction, retrieval co-attention and multimodal losses. For embedding extraction, we leverage

[⋆] Equal contribution. Kangrui performed the work while interning at Amazon Alexa.

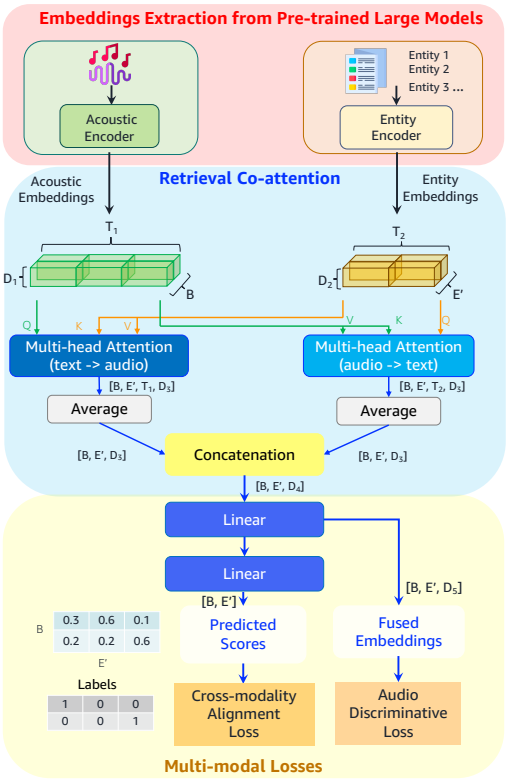

Fig. 1: The overall architecture of the proposed method.

pretrained large models, respectively Whisper encoder [11] for audio and BERT [12] for catalogs. We utilize the output of the whisper encoder and take the average of hidden states of the last four layers from BERT. This approach is adopted because the hidden states from these layers have demonstrated superior efficacy in generating embeddings, as opposed to relying on the [CLS] token [12]. Additionally, [13] highlights that higher layers in BERT capture more complex semantics, crucial in representing entities.

3.1. Retrieval Co-attention

The primary challenge in S2E lies in aligning the audio and text signals. To address this issue, we introduce a novel retrieval co-attention mechanism, detailed in Fig. 1, which allows for more effective alignment and fusion of multimodal information. Specifically, we have two distinct multi-head attention where in one the acoustic embedding serves as the query, utilizing the entity embeddings as both the key and value, and in another, the entity embeddings become the query, and the acoustic embedding become both the key and the value, inspired by [14]. Algorithm 1 summarizes how a forward pass of an uni-directional retrieval attention is done. With the attention outputs from both directions, we take the average across the temporal dimension and concatenate them together.

We derive the fused embeddings through a linear projection layer and the corresponding score matrix through another projection layers. To obtain each fused embedding, we extract across E' dimension by the matched entity index, where E' is the size of a catalog subset. The subset catalog contains the ground truth entity e_j and a random sample of negative entities. We adopt $E' \leq E$ to make training more memory efficient.

Algorithm 1: PyTorch-style pseudocode for the core implementation of a forward pass of one of the retrieval co-attention mechanism

```
# Embedding 1 shape:  [s1, t1, d1]
# Embedding 2 shape:  [s2, t2, d2]
# Number of heads:  nh
# Head dimension:  d
q = q_linear(embeddings1) # query
k = k_linear(embeddings2) # key
v = v_linear(embeddings2) # value
q = q.view(s1,t1,nh,d).permute(0,2,1,3)
k = k.view(s2,t2,nh,d).permute(0,2,1,3)
v = v.view(s2,t2,nh,d).permute(0,2,1,3)
# Calculate attention weights
 [s1,s2,nh,t1,t2]
qk = torch.einsum('bntd,enld->bentl', (q, k))
attn_w = F.softmax(qk / (d**0.5), dim=-1)
# Calculate attention output
attn_out =
 torch.einsum('bentl,enld->bentd',(attn_w, v))
attn_out = permute(attn_out, dims=(0, 1, 3, 2,
 4)).reshape(s1, s2, t1, -1)
```

3.2. Multimodal Losses

As detailedly shown in Figure 2, the proposed training objective inlcludes a cross-modality alignment loss $\mathcal{L}_{\text{cross}}$ and an audio discriminative loss $\mathcal{L}_A$. The cross-modality alignment loss trains the model to identify embeddings that represents aligned cross-modal signals (i.e., audio and ground-truth entity). Additionally, the audio discriminative loss focuses on distinguishing the fused embeddings with entity positive or negaitve pairs. As our problem setting follows a many-to-one relationship, intuitively, the audio discriminative loss enhances the clustering of audios corresponding to the same ground truth entities, thereby improving resolution accuracy. Specifically, it pulls two fused embeddings with the same entity index closer, and those with different entity indices further away.

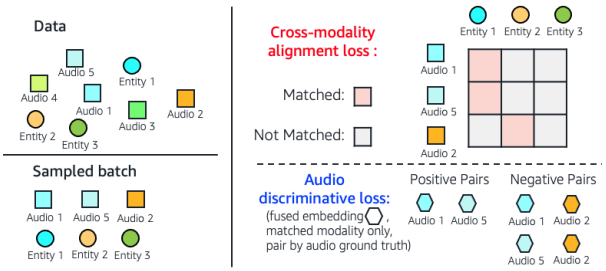

Fig. 2: Difference between the cross-modality alignment loss $\mathcal{L}_{\text{cross}}$ and the audio discriminative loss $\mathcal{L}_A$.

We explore two variants of the audio discriminative loss: triplet loss [15] and n-pair loss [16]. The triplet loss is defined below:

$$\max(0, d(f, f^+) - d(f, f^-) + m) \quad (1)$$

where f is the anchor, f^+ is the positive example corresponding to f, and f^- is the negative example corresponding to f. d

denotes the distance function, e.g., L2 distance. The n-pair loss tries to generalize the triplet loss (Eq. (1)) by jointly comparing more than just one negative examples [16]. Specifically, the n-pair loss is defined as

$$\frac{1}{N}\sum_{i}\log(1+\sum_{j\neq i}\exp(f_i^\top f_j^+ - f_i^\top f_i^+)) \quad (2)$$

where f_i^+ is the positive example, and all $j \neq i$ are the negative examples. Afterwards, the proposed method is jointly optimized by the total loss below:

$$\mathcal{L} = \mathcal{L}_{\text{cross}} + \mathcal{L}_A \quad (3)$$

where $\mathcal{L}$ is the sum of the cross-modality alignment loss $\mathcal{L}_{\text{cross}}$ and audio discriminative loss $\mathcal{L}_A$.

In summary, our proposed co-attention method primarily addresses the Signal-to-Entity task, extending beyond SLU, and it efficiently retrieves the corresponding entity by attending to the significant parts of an utterance with multimodal fused embeddings. Furthermore, our novel loss function, which includes cross-modality alignment loss and audio discriminative loss, enhances alignment by incorporating both intramodal and intermodal modeling, extending beyond traditional cross-modality methods. This approach is innovative in its comprehensive integration of different modalities, offering a more effective solution for entity retrieval in speech processing.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We adapt three public datasets for S2E: Fluent Speech Commands (FSC) [6], Snips SmartLights (SSL) [17] and Snips SmartSpeaker (SSS) [17]. They contain human audios and corresponding transcriptions. For each data set, we extract the entity mentions from the transcriptions and compile into them a catalog with unique entities.

1. FSC contains $\sim 30,000$ audios from 97 distinct speakers [6]. We identify a catalog of 20 unique entities, e.g., kitchen lights or bedroom heat. We adopt the predefined data partitions. **2. SSL** contains more diverse smart home commands with $\sim$ 2,000 audios [17]. We create a catalog with 81 unique entities and partition the dataset where the test set is excluded from both training and evaluation phases, given no predefined partitions. **3. SSS** contains diverse queries to play music. Each audio references one artist, and each artist has only two associated audios: near field and far field. Far-field audio recordings typically exhibit degraded quality due to reduced signal power and potential reflections from walls [18]. We extract 1,278 unique artist names for the catalog. For data partitions, training set includes both near and far-field audios, test set only includes the samples of the far-field audios excluded from both training and evaluation phases.

4.2. Experimental setup

Semi End-to-end This pipeline imitates a cascading architecture with intermediate audio transcriptions but relax the need for additional NER. We first generate transcriptions with Whisper and leverage BERT to respectively embed the transcription and entities. For Semi E2E (Linear + CE), a linear projection layer is utilized so that the utterance embeddings are projected to the entity space for comparison and cross-entropy is utilized as the cross-modality alignment loss.

End-to-end This is the proposed S2E solution described in Section 3. Each audio recording is re-sampled to a rate of 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram representation is utilized. We apply AdamW optimizer [19] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$, and the weight decay $\lambda = 0.01$.

4.3. Results

We demonstrate the proposed *E2E pipeline (Co-attention + CE + N-pair loss)* outperform the semi-E2E approach by 2.6%, 47.0%, and 73.3% across three datasets respectively and being 42% smaller in the number of model parameters. See details in Table 2.

We examine the cases where the semi E2E pipeline fails but the proposed *E2E pipeline (Co-attention + CE + N-pair loss)* succeeds. For example, given an utterance with ground truth transcription: "turn on the lights in the bedroom", the ASR model wrongly predicts the transcription "turn on the kits and lights in the back", thereby identifying an incorrect entity: "kitchen lights". However, the proposed is capable of accurately identifying the correct entity "bedroom lights", as it can directly match audio embeddings with entity embeddings, bypassing intermediate transcriptions.

Moreover, to assess the contribution of each component, we conduct structured ablation study, detailed in Table 1, and present results in Table 2. To find an effective form for the cross-modality alignment loss, we compare *E2E (Linear + CE)* and *E2E (Linear + Triplet loss)* and choose cross-entropy as the default option due to its marginally superior performance. To validate the alignment module design, we compare three methods: *E2E (Linear + CE), E2E (Attention + CE) and E2E (Co-attention + CE)*. The results show *E2E (Co-attention + CE)* performs the best, suggesting the co-attention mechanism effectively utilizes the multimodal signal compared to linear and uni-directional attention (i.e. use acoustic embedding as query). To validate the audio discriminative loss, we compare *E2E (Co-attention + CE), E2E (Co-attention + CE + Triplet loss), and E2E (Co-attention + CE + N-pair loss)*. The results indicate the one with the N-pair loss achieves the best performance.

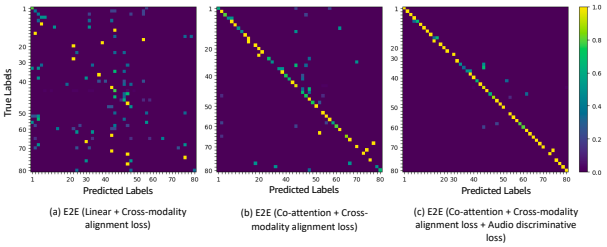

Fig. 3: The normalized confusion matrix for the retrieval entities. Method with more diagonal elements has a better performance.

Further, we present qualitative analysis visualizations Figure 3 and Figure 4. As depicted in Figure 3, the method with more diagonal elements performs better. The E2E pipeline enhanced with retrieval co-attention significantly outperforms the

Table 1: Model names and corresponding components of the proposed method and baselines.

Model	Components		
	Alignment Methods	Cross-modality Alignment Loss	Audio Discriminative Loss
Semi E2E (Linear + CE)	Linear	Cross-entropy	✗
E2E (Linear + CE)	Linear	Cross-entropy	✗
E2E (Linear + Triplet loss)	Linear	Triplet loss	✗
E2E (Attention + CE)	Attention	Cross-entropy	✗
E2E (Co-attention + CE)	Retrieval co-attention	Cross-entropy	✗
E2E (Co-attention + CE + Triplet loss)	Retrieval co-attention	Cross-entropy	Triplet loss
E2E (Co-attention + CE + N-pair loss)	Retrieval co-attention	Cross-entropy	N-pair loss

Table 2: Experimental results of all methods on the adapted three datasets, FSC, Snips Smart Lights and Snips Smart Speakers.

Model	Size	Performance		
		Fluent Speech Command (audios:∼30,000; catalog size: 20)	Snips Smart Lights (audios:∼2,000; catalog size: 81)	Snips Smart Speakers (audios:∼2500; catalog size: 1278)
Semi E2E (Linear + CE)	353M	0.971	0.611	0.363
E2E (Linear + CE)	201M	0.983	0.345	0.054
E2E (Linear + Triplet loss)	201M	0.983	0.343	0.008
E2E (Attention + CE)	203M	0.994	0.777	0.512
E2E (Co-attention + CE)	205M	0.993	0.855	0.551
E2E (Co-attention + CE + Triplet loss)	205M	0.994	0.885	0.531
E2E (Co-attention + CE + N-pair loss)	205M	**0.996**	**0.898**	**0.629**

linear projection layer, demonstrating its significance in improving accuracy. Additionally, Figure 3 (c) shows incorporating audio discriminative loss (N-pair loss) yields the best confusion matrix result.

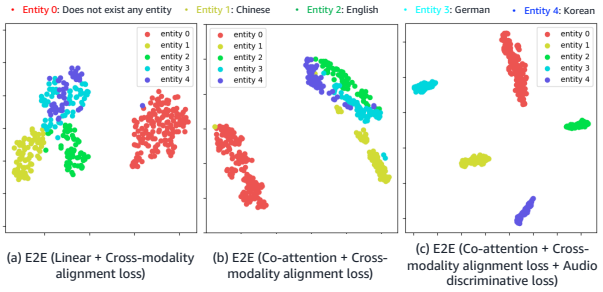

Fig. 4: The t-SNE visualizations of extracted embeddings. The embeddings in (a) are extracted from the projection layer, and embeddings in (b) and (c) are the fused embeddings.

In Figure 4, each point represents a fused embedding corresponding to a unique customer audio query. Queries sharing the same entity ground truth are represented by the same color. In Figure 4 (a) and (b), without audio discriminative loss, fused embeddings corresponding to entities 1 to 4 exhibit a strong tendency to cluster together. This observation is intuitive as the scenarios of using different languages are rather similar. After adding the audio discriminative loss (N-pair loss) in Figure 4 (c), the proposed method is able to distinguish more effectively among the four clusters.

To summarize, the proposed E2E pipeline (Co-attention + CE + N-pair loss) achieves the best performance because it bypasses the error-prone intermediate transcription step, and effectively utilizes the multimodal signal with the help of the retrieval co-attention and the proposed losses. However, for the semi end-to-end pipeline, even an optimized ASR module does not guarantee the performance of the downstream ER task, as the cascaded architecture prevents ER recovering from the upstream ASR errors.

5. RELATED WORK

Spoken Language Understanding (SLU) SLU infers the semantics of spoken utterances and typically involves intent classification and slot tagging tasks. End-to-end SLU solutions have been shown and evolved in various work, where [6] leverage the transcribed speech data to pre-train the lower levels of the model; [20] overcome the lack of available SLU datasets via transfer learning. [21] utilize a shared latent space, learning to guide the text embedding closer to the paired acoustic embedding. Compared to SLU, S2E directly optimizes ER performance and reduces the necessity for the NER task.

Entity Resolution (ER) The importance of ER has grown in multiple tasks, such as question answering [1, 22], knowledge base population [23], causal inference [24, 25], transportation [26, 27], and finance [28, 29]. Generally, there are two types: discriminative and generative. Discriminative ER methods predict relevancy score between each mention and entity labels. ReFinED [3] leverages fine-grained entity descriptions and excels in zero-shot settings. On the other hand, generative ER methods, like GENRE [30], retrieve entities by autoregressively generating entity names. Compared to the classical text-based ER problem, S2E directly considers multimodal inputs, i.e., audio and textual catalogs, which makes the problem more challenging. Additionally, unlike traditional ER tasks, which often leverage a rich repository of good quality data from sources like Wikipedia [30, 3], our methodology encounters unique hurdles. The challenge lies in the scarcity of datasets that pair audio with corresponding entity information, a limitation that significantly complicates the data collection process in our experiments.

6. CONCLUSION

We formulate a new end-to-end ER problem: S2E, and adapt three public datasets for this problem. We propose a novel solution, which helps to augment modality alignment between acoustic signals and entity texts. We demonstrate the efficacy of the proposed end-to-end design, and attribute the contribution of each component.

7. REFERENCES

[1] Vassilis Christophides, Vasilis Efthymiou, et al., "An overview of end-to-end entity resolution for big data," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–42, 2020.

[2] Ross McGowan, Jinru Su, et al., "Smaller: Scaling neural entity resolution for edge devices," in *Interspeech 2021*, 2021.

[3] Tom Ayoola, Shubhi Tyagi, et al., "Refined: An efficient zero-shot-capable approach to end-to-end entity linking," *arXiv preprint arXiv:2207.04108*, 2022.

[4] Jian Xie, Cynthia He, et al., "Simtde: Simple transformer distillation for sentence embeddings," in *SIGIR*, 2023.

[5] Xiaozhou Zhou, Ruying Bao, and William M. Campbell, "Phonetic Embedding for ASR Robustness in Entity Resolution," in *Proc. Interspeech 2022*, 2022, pp. 3268–3272.

[6] Loren Lugosch, Mirco Ravanelli, et al., "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[7] Loren Lugosch, Brett H. Meyer, et al., "Using speech synthesis to train end-to-end spoken language understanding models," in *ICASSP 2020*, 2020, pp. 8499–8503.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[9] Yusong Wu, Ke Chen, Tianyu Zhang, et al., "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[10] Benjamin Elizalde, Soham Deshmukh, et al., "Clap learning audio concepts from natural language supervision," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[11] Alec Radford, Jong Wook Kim, et al., "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

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

[13] Ian Tenney, Dipanjan Das, and Ellie Pavlick, "Bert rediscovers the classical nlp pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

[14] Yao-Hung Hubert Tsai, Shaojie Bai, et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[15] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of machine learning research*, vol. 10, no. 2, 2009.

[16] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[17] Alaa Saade, Joseph Dureau, et al., "Spoken language understanding on the edge," in *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition*. IEEE, 2019, pp. 57–61.

[18] Reinhold Haeb-Umbach, Jahn Heymann, et al., "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.

[19] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] Antoine Caubrière, Natalia Tomashenko, et al., "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," *arXiv preprint arXiv:1906.07601*, 2019.

[21] Bhuvan Agrawal, Markus Müller, et al., "Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding," in *ICASSP*. IEEE, 2022, pp. 7157–7161.

[22] Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang, "Retrieval, re-ranking, and multi-task learning for knowledge-base question answering," 2021.

[23] Raphael Hoffmann, Congle Zhang, et al., "Knowledge-based weak supervision for information extraction of overlapping relations," .

[24] Joan Heck Wortman and Jerome P Reiter, "Simultaneous record linkage and causal inference with propensity score subclassification," *Statistics in Medicine*, vol. 37, no. 24, pp. 3533–3546, 2018.

[25] Kangrui Ruan, Junzhe Zhang, et al., "Causal imitation learning via inverse reinforcement learning," in *The Eleventh International Conference on Learning Representations*, 2023.

[26] Farman Ali, Daehan Kwak, et al., "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowledge-Based Systems*, vol. 174, pp. 27–42, 2019.

[27] Xiaobo Ma, *Traffic Performance Evaluation Using Statistical and Machine Learning Methods*, Ph.D. thesis, The University of Arizona, 2022.

[28] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras, "Finer: Financial numeric entity recognition for xbrl tagging," *arXiv preprint arXiv:2203.06482*, 2022.

[29] Shun Liu, Kexin Wu, Chufeng Jiang, Bin Huang, and Danqing Ma, "Financial time-series forecasting: Towards synergizing performance and interpretability within a hybrid machine learning approach," *arXiv preprint arXiv:2401.00534*, 2023.

[30] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni, "Autoregressive entity retrieval," *arXiv preprint arXiv:2010.00904*, 2020.