

Semantic Role Labeling from Chinese Speech via End-to-End Learning

Huiyao Chen*, Xinxin Li*, Meishan Zhang†, Min Zhang

Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), China

chenhy1018@gmail.com, lixinx0714@gmail.com

mason.zms@gmail.com, zhangmin2021@hit.edu.cn

Abstract

Semantic Role Labeling (SRL), crucial for understanding semantic relationships in sentences, has traditionally focused on text-based input. However, the increasing use of voice assistants and the need for hands-free interaction have highlighted the importance of SRL from speech. SRL from speech can be accomplished via a two-step pipeline directly: transcribing speech to text via Automatic Speech Recognition (ASR) and then applying text-based SRL, which could lead to error propagation and loss of useful acoustic features. Addressing these challenges, we present the first end-to-end approach for SRL from speech, integrating ASR and SRL in a joint-learning framework, focusing on the Chinese language. By employing a Straight-Through Gumbel-Softmax module for connecting ASR and SRL models, it enables gradient back-propagation and joint optimization, enhancing robustness and effectiveness. Experiments on the Chinese Proposition Bank 1.0 (CPB1.0) and a newly annotated dataset AS-SRL based on AISHELL-1 demonstrate the superiority of the end-to-end model over traditional pipelines, with significantly improved performance.

1 Introduction

Semantic role labeling (SRL), as one of the core tasks in natural language processing (NLP, Gildea and Jurafsky, 2000), concentrates on analyzing semantic relationships between predicates and their corresponding arguments in sentences, such as who did what to whom, as well as when, where and how it occurred (Màrquez et al., 2008). Semantic information has been exploited in a variety of complex downstream NLP tasks including information extraction (Surdeanu et al., 2003; Niklaus et al., 2018), question answering (He et al., 2015;

* Authors contributed equally.

† Corresponding author: Meishan Zhang.

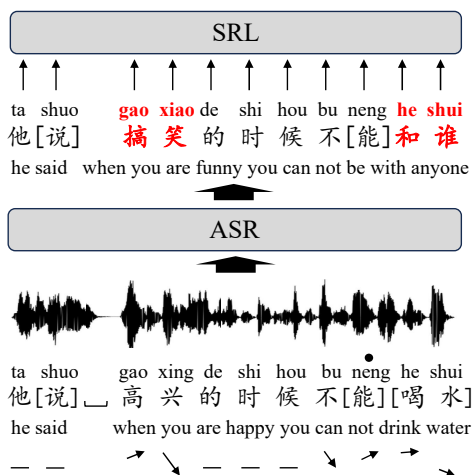


Figure 1: Pipeline of SRL from Chinese speech. ●: stresses. Space: pauses. →: undulations. Under the pipeline architecture, error propagation can influence SRL greatly (e.g., predicate “喝水(drink water)” would be lost) and acoustic features such as stresses, pauses and undulations are unable to be employed by SRL.

Zhang et al., 2020), and robot command parsing (Thomason et al., 2020; Vanzo et al., 2020).

Previous SRL typically assumes input as text (Pradhan et al., 2005; FitzGerald et al., 2015; Fei et al., 2020; Campagnano et al., 2022). However, in many real-world scenarios, such as for the elderly who may find typing challenging, or in situations requiring remote control of machines, speech often emerges as the more direct medium for conveying information (Fowler, 1986; Bingol and Aydogmus, 2020). As voice assistants (Ahmed et al., 2022), social bots (Mahdi et al., 2022), and intelligent home devices (Torad et al., 2022) become increasingly prevalent, the demand for language understanding capabilities that can process speech inputs grows correspondingly. Thus, SRL from speech directly would exhibit increasing interest.

SRL from speech can be accomplished via a two-step pipeline directly: (1) automatic speech recognition (ASR) that transcribes the speech into text

first, and then (2) text-based SRL. However, such a pipeline could be insufficient for speech-based SRL (Ghannay et al., 2018), as illustrated in Figure 1. On the one hand, the sequential pipeline might suffer from the commonly-recognized error propagation problem, i.e., the ASR errors would propagate to SRL (Hatmi et al., 2013; Simonnet et al., 2018), since traditional text-based SRL always assume well-written text as input. On the other hand, after ASR the acoustic features are completely isolated to the second step (Jannet et al., 2017), while part of them might be useful for SRL, e.g., the phonological features such as stress, pauses and undulations. To address the above issues, end-to-end learning of the two tasks is one natural solution.

In this work, we present the first work to tackle SRL from speech in an end-to-end way, uniting ASR and SRL in a joint-learning paradigm. We focus on the Chinese language as the pioneer study here. We start from the state-of-the-art ASR and SRL models, and then connect the two models by a Stright-Through Gumbel-Softmax (GS) module (Jang et al., 2017). The GS enables gradient back-propagation over the intermediate discrete transcribed texts, therefore a joint optimization towards the ASR and SRL objectives can be achieved, and the end-to-end learning is formed naturally. Additionally, as the transcribed text might be deformed into a noisy one, which makes gold-standard annotations difficult to be explored during SRL learning, we design an alignment mechanism to reach a reasonable SRL objective. This alignment also enables more robust evaluation for speech-based SRL and akin.

We conduct experiments on Chinese Proposition Bank 1.0 (CPB1.0) (Xue and Palmer, 2003) mainly to verify the effectiveness of our end-to-end model. Specifically, we recruit Mandarin-qualified annotators to obtain authentic speech for the test set for evaluation, and the speech for the remaining part is annotated by a text-to-speech (TTS) tool,¹ since the manual annotation is of high cost. In addition, we also build another evaluation dataset AS-SRL based on the open-source Mandarin speech corpus AISHELL-1 (Bu et al., 2017), annotating it with SRL by experts following the guideline of CPB1.0. Results show that our end-to-end method is highly effective, achieving averaged improvements of 0.47% and 1.20% in the F_1 score com-

pared with the standard pipeline counterpart. We also design a series of other strong baselines for comparisons, and the results show that our final model outperform all these systems. Further, we perform in-depth analysis to provide a comprehensive understanding of our methodology.

In summary, our contributions can be summarized as follows:

- We present the first study of SRL from Chinese speech and construct the first real-world Chinese speech-based SRL dataset.
- We propose an end-to-end approach for speech-based SRL and utilize alignment tools to achieve robustness to transcription noise.
- We conduct extensive experiments to compare our method with several pipeline methods, and results show that our method can advance the speech-based SRL further.

Our code and datasets will be publicly available at [github/DreamHigh/SpeechSRL](https://github.com/DreamHigh/SpeechSRL) to facilitate future research.

2 Method

A straightforward solution for SRL from speech is through a pipeline strategy, first executing ASR and then performing SRL. Since the pipeline architecture induces error propagation inevitably and meanwhile cannot use acoustic clues for SRL, we suggest end-to-end learning to unite the two tasks, as shown in Figure 2. This section first introduces a state-of-the-art pipeline, and then describes our end-to-end learning approach.

2.1 Pipeline

We use the transformer-based sequence-to-sequence Whisper (Radford et al., 2023) model for ASR and a word-based graph parsing model introduced by Zhou et al. (2022) for SRL. Both models are representative systems with top performance for ASR and SRL, respectively.

Automatic Speech Recognition. Given a piece of raw speech, ASR is to convert it into the corresponding transcription $X = (c_1, \dots, c_n)$, where c denotes the character. Specifically, the Whisper model first uses a classical feature extractor to process the raw input speech, obtaining a sequence of speech features $\mathbf{S} = (s_1, \dots, s_N)$ (commonly, $N \gg n$). Then, \mathbf{S} is fed into a transformer encoder, reaching high-level hidden states $\mathbf{H}^{\text{asr}} = (\mathbf{h}_1^{\text{asr}}, \dots, \mathbf{h}_N^{\text{asr}})$. Subsequently, a

¹<https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech>

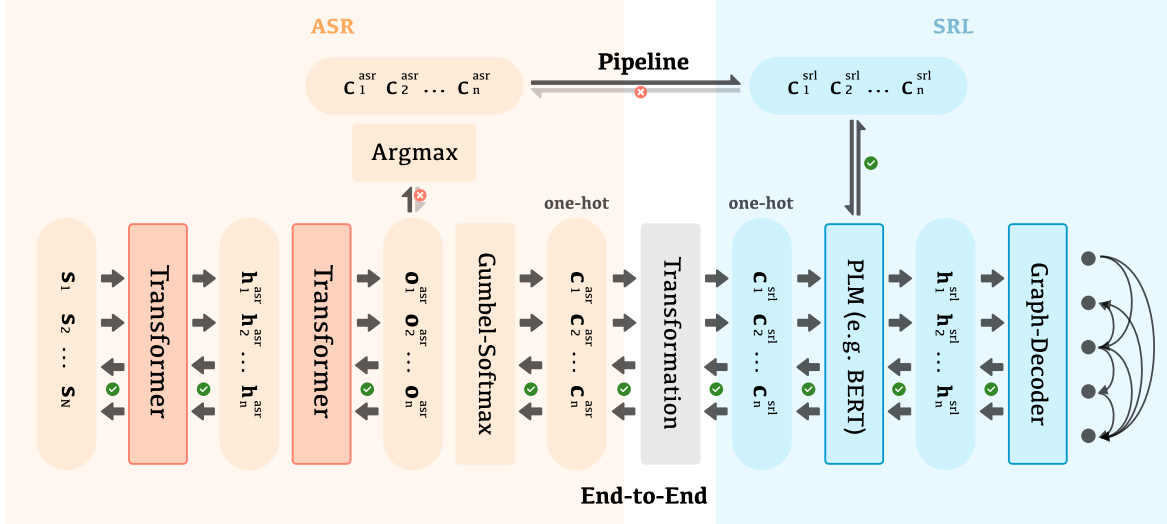


Figure 2: The overall architecture of our pipeline and proposed end-to-end models, where the green \checkmark and red \times indicates derivable and non-derivable, respectively.

transformer-based decoder is used to score the output tokens sequentially based on previous predications and the hidden states \mathbf{H}^{asr} , finally resulting in $\mathbf{O}^{\text{asr}} = (o_1^{\text{asr}}, \dots, o_n^{\text{asr}})$. The whole process can be formalized as follows:

$$\begin{cases} \mathbf{H}^{\text{asr}} = \text{Transformer-Encoder}(\mathbf{S}), \\ \mathbf{O}^{\text{asr}} = \text{Transformer-Decoder}(\mathbf{H}^{\text{asr}}). \end{cases} \quad (1)$$

During training, the cross-entropy objective is used for the ASR model parameter optimization:

$$\begin{cases} \mathbf{P}^{\text{asr}} : \mathbf{p}_1^{\text{asr}}, \dots, \mathbf{p}_n^{\text{asr}} = \text{softmax}(\mathbf{O}^{\text{asr}}), \\ \mathcal{L}_{\text{asr}} = -\sum_{i=1}^n c_i \log p_i^{\text{asr}}[c_i], \end{cases} \quad (2)$$

where $p_i^{\text{asr}}[c_i]$ indicates the probability of i th token being c_i and c_i is the one-hot vector form of c_i . During inference, we output the max-scored characters based on \mathbf{O}^{asr} .

Note that the Whisper model follows a standard transformer-based sequence-to-sequence neural architecture. Its strong capability of ASR is mainly sourced from the large-scale training corpus collected by the original work. Here we directly exploit the well-trained parameters of Whisper for ASR. For more training details, one can refer to their paper.

Semantic Role Labeling. Zhou et al. (2022) treat SRL as graph parsing, first converting predict-argument structures into a dependency graph, and then exploiting a well-designed parser to achieve the goal. The graph is built simply by connecting the boundaries of semantic roles with their predicates, and predicates are associated with a special root symbol. In our pipeline, the transcription X is

a character sequence in Chinese, thus we need to convert the original word-based dependency graph into characters. The conversion method is simple and direct, with details provided in A.1.

The overall character-based SRL parsing model can be also illustrated as an encoder-decoder architecture. First, the sentential characters $X = (c_1, \dots, c_n)$ go through a pretrained language model (PLM), obtaining the well-represented encoder output $\mathbf{H}^{\text{srl}} = (h_1^{\text{srl}}, \dots, h_n^{\text{srl}})$. Then a graph decoder is executed. For each candidate dependency edge $\langle i, j \rangle$ (i is the head and j is the modifier), a biaffine operation is adopted for scoring. After the edges are ready, a set of label-aware biaffine operations can be applied to calculate the label scores. The full process can be formalized as follows:

$$\begin{cases} \text{Encoding: } \mathbf{H}^{\text{srl}} = \widetilde{\text{PLM}}(X), \\ \text{Edge } \langle i, j \rangle : \\ \quad \mathbf{h}_i^h, \mathbf{h}_j^m = \text{MLP}^h(\mathbf{h}_i^{\text{srl}}), \text{MLP}^m(\mathbf{h}_j^{\text{srl}}), \\ \quad o_{i,j}^{\text{edge}} = [\mathbf{h}_i^h]^T \mathbf{W}_E \mathbf{h}_j^m + [\mathbf{b}_E]^T \mathbf{h}_j^m \\ \text{Label } \langle l | i, j \rangle : \\ \quad \bar{\mathbf{h}}_i^h, \bar{\mathbf{h}}_j^m = \overline{\text{MLP}}^h(\mathbf{h}_i^{\text{srl}}), \overline{\text{MLP}}^m(\mathbf{h}_j^{\text{srl}}), \\ \quad o_{l|i,j}^{\text{label}} = [\bar{\mathbf{h}}_i^h]^T \mathbf{W}_{L,l} \bar{\mathbf{h}}_j^m + [\mathbf{b}_{L,l}]^T \bar{\mathbf{h}}_j^m. \end{cases} \quad (3)$$

where $[\cdot]^T$ indicates matrix transposition.

To train the above model, we also use cross-entropy loss as objective. Essentially, the edge prediction is a binary classification of whether an edge being retained in graph, while the label prediction belongs to multi-label classification. As a

result, our SRL objective is defined as:

$$\left\{ \begin{array}{l} p_{i,j}^{\text{edge}} = \text{sigmoid}(o_{i,j}^{\text{edge}}), \\ p_{l|i,j}^{\text{label}} = \text{softmax}(o_{*|i,j}^{\text{label}})[l], \\ \mathcal{L}_{\text{srl}} = - \sum_{\langle l|i,j \rangle \in G} \log(p_{i,j}^{\text{edge}} \times p_{l|i,j}^{\text{label}}) \\ \quad - \sum_{\langle i,j \rangle \notin G} \log(1 - p_{i,j}^{\text{edge}}), \end{array} \right. \quad (4)$$

where G is the gold-standard graph of SRL. During inference, the edges with probability larger than 0.5 are kept, and the max-scored label is assigned to a given dependency edge. The decoding graph might be incompatible with SRL, which is addressed by constrained decoding. For the details, one can refer to the original paper.²

2.2 End-to-End Learning

The above pipeline isolates ASR and SRL completely, making the two components incapable of interaction. Here we introduce a simple end-to-end learning framework to connect the two tasks. The key idea is to propagate the objective loss of SRL directly into ASR, not only letting SRL robust to the ASR errors but also enabling SRL to use acoustic features directly. The major difficulty lies in that the discrete ASR outputs disable the back-propagation of SRL training gradients into the ASR component. To address this problem, Straight-Through Gumbel-Softmax is exploited here (Jang et al., 2017).

Gumbel-Softmax. As mentioned in Sec 2.1, we use the widely-accepted argmax operation for next-character prediction to obtain the transcribed characters, which are then delivered to SRL as input for training. This operation leads to non-differentiable variables, which is the source of the isolation between ASR and SRL. Different from argmax where the resulted one-hot vector is deterministic, GS introduces randomness into the prediction, resulting in a sampled vector. More importantly, GS exploits a reparameterization trick to make the sampled vector differentiable. Since the added noise follows a Gumbel distribution, the resulting next-token prediction distribution is essentially consistent with the original distribution.

Concretely, given a speech input S , we feed it into the same Whisper encoder-decoder module, obtaining the output scores O^{asr} of the transformer-decoder as shown in Equation 1. During training,

²Here we only take the first-order scoring to introduce the SRL model for brevity. The higher-order scoring can be achieved by TriAffine as mentioned by their original paper.

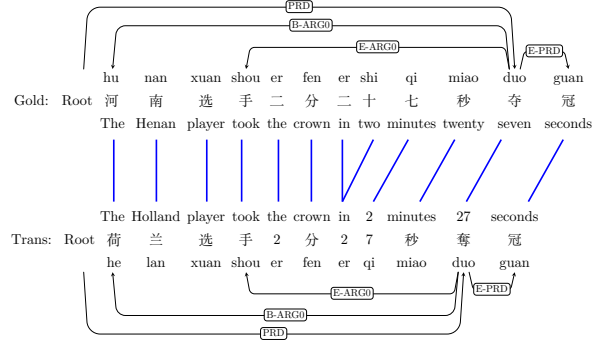


Figure 3: An example of the alignment result after using the alignment tool for a gold standard text and the corresponding transcription.

we exploit the GS operation to calculate next-token distributions instead of softmax:

$$p_{i,j}^{\text{gs}} = \frac{\exp((o_{i,j}^{\text{asr}} + g_{i,j})/\tau)}{\sum_{k=1}^{C_{\text{asr}}} \exp((o_{i,k}^{\text{asr}} + g_{i,k})/\tau)}, \quad (5)$$

where $g_{i,*}$ are i.i.d samples drawn from Gumbel(0, 1)³, τ is the temperature factor and C_{asr} denotes the output vocabulary size of the ASR model. Then, gumbel sampling is used instead of the argmax operation to obtain the one-hot output at each next-token prediction, and finally we aggregate the ASR outputs, resulting in $\mathbf{X}^{\text{asr}} = (c_1, \dots, c_n)$.

Note that the sampled one-hot outputs $\mathbf{X}^{\text{asr}} = (c_1, \dots, c_n)$ should not be converted into characters by the ASR output vocabulary, otherwise the differentiability would be lost. Therefore we need to feed them into the SRL encoder directly. However, there is a mismatch between the character vocabularies of ASR output and SRL input. To solve this issue, we maintain a matrix M to implement the transformation, which is defined as follows:

$$M_{i,j} = \begin{cases} 1, & \text{if } V_{\text{asr}}[i] = V_{\text{srl}}[j], \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where V_{asr} and V_{srl} denote the character vocabularies of ASR output and SRL input, respectively. Using M , we can get the input one-hot vectors of SRL by $\mathbf{X}^{\text{srl}} = \mathbf{X}^{\text{asr}} M$. After that, we feed \mathbf{X}^{srl} into the encoding part PLM as shown in Equation 3, and then ASR and SRL connect naturally.

SRL Alignment. Through the GS module, we connect ASR and SRL together for training. However, there still exists one big problem. The sam-

³The Gumbel(0, 1) distribution can be sampled using inverse transform sampling by drawing $\mu \sim \text{Uniform}(0, 1)$ and computing $g = -\log(-\log(\mu))$.

pled vector of ASR can often differ from the well-written gold-standard text for SRL training, as shown in Figure 3. When the phenomenon happens, how to construct the oracle character-based dependency graph in terms of the original gold-standard graph?

To solve the problem, we suggest a character-level alignment tool to project the original gold-standard graph into the sampled text. The idea is borrowed from the line of annotation projection in cross-lingual dependency parsing (Schuster et al., 2019; Zhang et al., 2019; Ozaki et al., 2021; Zhang et al., 2023). We can use a similar projection strategy since SRL is modeled as dependency parsing here. High-quality projections would be expected as the noisy text by GS is mostly similar to the original one.

Concretely, there are two steps. First, we obtain a high-performance character alignment tool. by using the unsupervised fastAlign⁴. A large-scale parallel corpus is constructed automatically with no laborious cost by mocking the GS sampling, which is used to train the fastAlign. Second, we project the character-level SRL dependencies from the source text into the target noisy text one by one. The process is very easy since accurate alignments can be obtained in most cases because very few ambiguities exist in our scenario. In cases of exceptions the dependencies are discarded directly.

Training and Inference. For training, we adopt the SRL objective as the training loss named $\mathcal{L}_{\text{joint}}$, which can jointly update the ASR module with the SRL module at the same time. For inference, there is no difference to the pipeline.

3 Experiments

3.1 Settings

Dataset. Our experiments are evaluated based on two Chinese datasets: CPB1.0 (Xue and Palmer, 2003) and AISHELL-1 (Bu et al., 2017). CPB1.0 is a widely used Chinese SRL dataset and contains gold-standard text and corresponding SRL labels. For the test set, we recruit Mandarin-qualified annotators from colleague to obtain authentic speech for evaluation. The training and development sets are annotated by Microsoft Azure TTS tool because of the high cost of manual annotation. AISHELL-1 is a widely used open-source Mandarin speech corpus, containing authentic speech and gold-standard

Dataset		#Sent	#PA-tuple
CPB1.0	Train	8,665	65,809
	Devel	549	4,445
	Test	983	8,001
AS-SRL	Train	7,500	22,536
	Devel	500	1,471
	Test	1,000	3,106

Table 1: Overview of datasets, where #PA-tuple denotes the number of automatic predicate-argument tuples: <predicate, argument, relation>.

text. Due to the extensive size of AISHELL-1, we randomly select 9,000 speech-text pairs and invite experts with a proficient understanding of the SRL task to annotate them with SRL following the guideline of CPB1.0. The statistics about the two datasets are illustrated in Table 1. The specific annotation process details are placed in A.2 for limited space.

Evaluation Metrics. Consistent with prior research of Chinese ASR, we employ the character error rate (CER) to evaluate the performance of ASR. For SRL, we measure the performance by the atomic predicate-argument structure, i.e., a tuple of <predicate, argument, relation>. We regard a tuple as correct only when the predicate, argument and relation are all exactly matched with the gold-standard answers. The precision, recall as well as their F_1 score are calculated.

Noticeably, the first-step ASR errors can make the SRL metric calculation problematical, because we cannot determine the matching status by the difference between the ASR outputs and the desired gold-standard text, as shown in Figure 3. To solve the issue, we perform a character-level alignment among the two texts before the metric computation. We achieve this goal by our well-trained fastAlign tool followed with a human checking. In fact, we find that the human checking can be almost omitted as the fastAlign can have an accuracy higher than 99% during this particular alignment.

Model Details. Regarding Whisper, we select large-v2⁵ (Radford et al., 2023), which performs well on ASR. Particularly, we performs a continual pretraining on the training set to adapt Whisper for our benchmark datasets. By default, the Whisper parameter updating is achieved using LORA (Hu et al., 2022) for efficient learning. We utilize bert-

⁴https://github.com/clab/fast_align

⁵<https://huggingface.co/openai/whisper-large-v2>

Method	CPB1.0				AS-SRL			
	CER	P	R	F ₁	CER	P	R	F ₁
Traditional Text-based SRL								
Gold Text + SRLGParser (oracle)	0.00	86.28	78.05	81.96	0.00	81.29	78.49	79.87
Pipeline								
Whisper + SRLGParser	3.16	81.25	72.53	76.64	4.45	74.21	71.44	72.80
Whisper + SRLGParser (+GS AUG)	3.16	80.95	72.93	76.73	4.45	77.49	70.70	73.94
End-to-End								
Whisper ◦ SRLGParser (gold SRL)	3.16	81.21	73.15	76.92	4.48	75.35	70.86	73.04
Whisper ◦ SRLGParser	3.16	80.27	74.19	77.11	4.47	75.84	72.25	74.00

Table 2: Main results on CPB1.0 and AS-SRL. (+GS AUG) denotes the SRLGParser is trained with additional GS augmented training instances. (gold SRL) indicates that the SRL objective is computed only when the sampled GS text equals the gold-standard text.

base-chinese⁶ (Devlin et al., 2019) to obtain text embeddings for our SRL model. For convenience, we denote the baseline SRL model as SRLGParser all through this work.

Hyperparameters. All experiments are conducted using a single NVIDIA A100 Tensor Core GPU (40G) card. We run each setting by 5 times with different random seeds, and the median evaluation scores are reported. For the setting of the SRL model parameters, we follow the settings of (Zhou et al., 2022). The temperature factor in Equation 5 is set as $\tau = e^{-5}$. We finetune the model for 20 epochs using learning rate of $3e-5$, and select the best iteration number according to the development results.

3.2 Main Results

Table 2 shows the main results on CPB1.0 and AS-SRL. It can be observed that our proposed end-to-end method achieves better results than the standard pipeline across almost all settings. Our final model, i.e., Whisper ◦ SRLGParser that optimizes with both ASR and SRL objectives in an end-to-end way, can lead to improvements of 0.47% and 1.20% on the two datasets,⁷ respectively. Following, we make comprehensive comparisons between the pipeline and end-to-end models.

First, we examine the SRL performance of SRLGParser in terms of auto ASR transcriptions and gold texts as inputs under the pipeline architecture. We can see that the CER values of Whisper are 3.16 and 4.45 for CPB1.0 and AS-SRL, respectively, which is highly competitive. Compared with

gold-text inputs, the ASR errors can bring significant decreases of $81.96 - 76.64 = 5.32$ for CPB1.0 and $79.87 - 72.80 = 7.07$ for AS-SRL. This suggests that the quality of the ASR text impacts the SRL performance much.

It is worth noting that the end-to-end learning procedure can be regarded as a special case of data augmentation. As illustrated in Sec 2.2, the gumbel sampled texts together with their oracle SRL annotations provide a new source of training instances. To test the effectiveness as data augmentation, we simply dump all augmented instances and add them into the pipeline SRL training, as denoted by Whisper + SRLGParser (+GS AUG). As shown, we can see that significant improvements $73.94 - 72.80 = 1.14$ can be achieved on AS-SRL, whereas the increase on CPB1.0 is very marginal, where the possible reason could be that the TTS-generated speech leads to little variation during GS sampling. The observation indicates that GS could be one prospective strategy for data augmentation.

Whisper ◦ SRLGParser (gold SRL) denotes that only when the ASR transcription exactly matches the gold-standard text, the SRL objective is calculated. This is an easy solution of end-to-end learning without SRL alignment. Even in this case, the joint learning can bring better performance, which further demonstrates the effectiveness of our end-to-end learning approach. Since no extra SRL instance is exploited in comparison with the pipeline, the performance gains indicate that acoustic features are useful for SRL from speech. By comparing Whisper ◦ SRLGParser (gold SRL) with the final model, we can again recognize the effectiveness of the augmented SRL instances by GS sampling. On both CPB1.0 and AS-SRL datasets,

⁶<https://huggingface.co/bert-base-chinese>

⁷Both improvements are significant with a p-value below 10^{-5} under pairwise t-test.

Statistics	CPB1.0	AS-SRL
Whole characters	41,842	14,685
-Matched characters	40,254	14,036
-Unmatched characters	1,588	649
Error characters	347	62
Accuracy	99.17%	99.58%

Table 3: Manual statistics of character alignment for the test set. Matched characters denotes the number of character that is aligned to the same character.

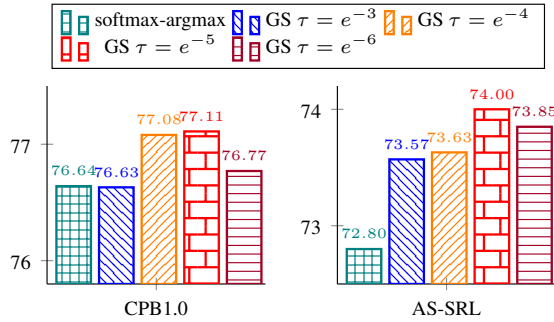


Figure 4: F_1 scores of our end-to-end method as the temperature changes.

we can observe appreciable improvements.

3.3 Analysis

The quality of alignment tool. The quality of the alignment tool plays a vital role in end-to-end SRL from speech. Therefore we conduct human evaluation to check the alignment accuracy. Specifically, during our SRL-performance evaluation, we have collected such data: first automatic alignment by our tool and then human collection. We take the collected data from the evaluation of our final model as an example (i.e., Whisper \circ SRLGParser). Table 3 reports the results. We can see that the manual correction can be almost omitted, since the proportion of corrected characters is only 0.83% for CPB1.0 and 0.42% for AS-SRL, yet most of them might be irrelevant with SRL evaluation. The high alignment accuracy also supports the high-quality of our SRL annotation projection greatly.

The sensitivity of GS temperature to our end-to-end learning. The GS temperature plays an important role in sampling. When the value is close to 1, GS would be almost equivalent to the standard softmax. We conduct experiments on the values of temperature from e^{-3} to e^{-6} to observe whether the model is sensitive to the temperature setting. The final model Whisper \circ SRLGParser is employed for verification. Figure 4 shows the

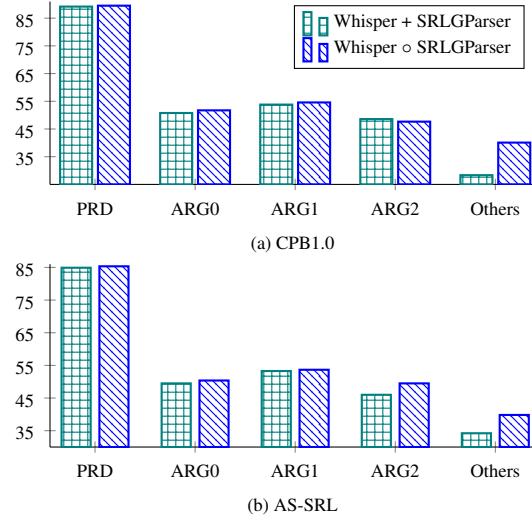


Figure 5: F_1 scores of predicate recognition and role labeling recognition with and without speech features.

results. We can see that the model can be influenced by the factor. As the value approaches to 1, the performance decreases significantly. However, the the value is lower than 10^{-5} , the performance also a downtrend. As τ nears 0, the output distribution would become a meaningless one-hot distribution.

Fined-grained performance by the concrete SRL roles.

Previous evaluation mainly focuses the overall tuple-level metrics. Here we look into the fined-grained SRL results in terms of different SRL roles, including predicates, core SRL arguments such as ARG0-2, and other SRL arguments such as TMP, LOC and etc. We compare our final end-to-end learning model with the pipeline system. Figure 5 shows the results on our investigated two datasets. As shown, the predicates and the core semantic arguments always have a better performance than others, since these roles could be fully trained because of their high occurrence. As expected, the end-to-end learning can give better performance for almost all type of SRL roles. Moreover, the scores for predicate recognition and role labeling recognition on the AS-SRL dataset improves more significantly, suggesting that the the introduction of authentic speech features is more capable of facilitating the model to recognize the arguments and its roles better than TTS speech features. On the one hand, since the predicates are stressed most of the time, the accuracy of recognizing predicates improves on the authentic speech AS-SRL dataset by 0.44%. On the other hand, the role Others in both datasets are substantially improved, as Chinese has a variation of tones and

our model becomes more sensitive to recognizing location, time, etc.

The ASR alterations after our end-to-end learning. According to our experimental results, we can see that our best model has a slightly higher CER than the pipeline model `Whisper + SRLGParser`, but the SRL results are all better than the method of pipeline. We analyze the differences between our model and the results transcribed by the pipeline model, and find that our model have fewer predicate transcription errors. For example, in the sentence “涉及人才管理(involve talent management)”, the predicate “涉及(involve: she ji)” is not transcribed correctly by Whisper, which transcribes it as “设计(design: she ji)”, whereas the Whisper after our end-to-end learning, is transcribed successfully. While we find sometimes the Whisper can produce SRL-irrelevant errors, e.g., the adjacent words such as “高新(high-tech: gao xin)” which could be correctly recognized by our baseline but confuse to “高薪(high-pay: gao xin)” by our end-to-end model, under a context “高新开发区(high-tech development zone)”.

4 Related Work

Semantic role labeling was originally conceptualized as a task for shallow natural language understanding by Gildea and Jurafsky (2000). The conventional approach to SRL involves two subtasks: (1) predicate identification followed by (2) argument role labeling (Swanson and Gordon, 2006; Punyakanok et al., 2008; Roth and Lapata, 2016; Marcheggiani and Titov, 2017; Strubell et al., 2018; Xia et al., 2019). In some studies, SRL task is decomposed with the goal of improving argument role labeling performance when predicates are provided as gold (He et al., 2018; Tan et al., 2018; Ouchi et al., 2018; Li et al., 2018; Fei et al., 2021). Recently, there has been an increasing interest in consolidating the two subtasks by a single model (Zhou et al., 2022). The line of work completes SRL fully with SOTA performance, and meanwhile exhibits significant efficiency advantages. Therefore, we follow this line of work, selecting (Zhang et al., 2022; Zhou et al., 2022) as the SRL baseline.

ASR has been concerned since very early (Lee et al., 1990; Huang et al., 1993; Chou et al., 1994). Deep learning has advanced this task greatly, with models like LAS employing RNNs to capture the contextual nuances of speech signals, leading to

notable enhancements (Chan et al., 2015; Amodei et al., 2016; Sriram et al., 2018). Moreover, the breakthrough success of the Transformer architecture in various NLP tasks has inspired the integration of attention mechanisms into ASR, further advancing its capabilities (Tang, 2023; Dong et al., 2023a). Whisper employs a Transformer-based architecture, trained on large-scale corpora (Radford et al., 2023). It has essentially achieved SOTA performance on ASR tasks. Thus, Whisper is selected as the model for our exploration of ASR.

The advance ASR technologies (Cheng et al., 2023; Radford et al., 2023) and the increasing availability of large-scale speech datasets (Bu et al., 2017; Zhu et al., 2019; Wang et al., 2021) have prompted a transition towards integrating speech directly into NLP tasks. Many studies, addressing fundamental NLP tasks such as sentiment analysis (Shon et al., 2021; Aicher et al., 2022), machine translation (Ren et al., 2020; Fang and Feng, 2023b,a), and spoken language understanding (Seo et al., 2022; Dong et al., 2023b), have adeptly leveraged speech, employing end-to-end approaches to achieve remarkable performance by exploiting intrinsic features of speech. However, for more intricate NLP tasks like parsing (Caines et al., 2017; Jamshid Lou et al., 2019; Pupier et al., 2022; Lai et al., 2023; Tseng et al., 2023; Shao et al., 2023), pipeline models that convert speech to text before applying conventional NLP strategies are commonly used. Our study pioneers SRL from speech, proposing an end-to-end framework, marking a novel approach in this field.

5 Conclusion

In this paper, we proposed a novel approach to SRL from speech through an end-to-end methodology, integrating ASR and SRL into a unified learning framework. This work, pioneering in its focus on the Chinese language, leverages state-of-the-art ASR and SRL models, interconnected via a Stright-Through Gumbel-Softmax module. The GS module facilitates gradient back-propagation across intermediate, discrete transcriptions, enabling joint optimization of ASR and SRL objectives and the natural formation of end-to-end learning. To address the challenge of transcriptions potentially becoming noisy, an alignment tool is devised, ensuring that SRL learning can effectively utilize gold-standard annotations and allowing for a more robust evaluation of speech-based SRL.

We constructed two Chinese speech-based SRL datasets and performed experiments on them to evaluate our end-to-end method. Our method not only demonstrates significant performance improvements over traditional pipeline methods but also outperforms a series of strong baseline models. Further in-depth analysis provides a comprehensive understanding of the approach, illustrating its potential for broader application and setting a new benchmark in the field.

Limitations

Our work has four major limitations. First, we only conducted our experiments on Chinese datasets. While our approach is applicable to other languages, there is a lack of experimental validation in these cases due to the time-consuming and high cost of data annotation. In the future, we will validate the effectiveness of our method on other languages. The second is that our experiments were confined to traditional classification-based SOTA SRL methods. Large language models (LLMs) excel in numerous NLP tasks. However, they currently fall short in SRL tasks and present optimization challenges due to their high computational costs. Nonetheless, investigating the integration of LLMs into speech-based SRL remains a promising research direction. We plan to thoroughly investigate the effectiveness of end-to-end SRL utilizing LLMs in future. Moreover, the training data of the ASR model we used (Whisper) we used is not publicly known, so it is uncertain there are no data leakage issues affecting ASR performance in the experiments. We will experiment on open source ASR models in the future. The last one is that ASR tasks in experiments (laboratory speech) may be easier than in the real world. Real-world speech contains more noise, and we plan to validate the effectiveness of our method on real-world speech.

Ethical Statement

We construct two datasets for Chinese speech-based SRL. All instances in the experiment are from two existing datasets: CPB1.0 and AISHELL-1, which are publicly available and do not contain sensitive and harmful information. We have obtained the license for these datasets. During recruiting annotators, we claim that all potential participants have the freedom to choose whether to partake in the study, and they may withdraw at any time without facing any negative consequences. All

annotators are properly compensated under open market competition. Moreover, the entire annotation task is conducted anonymously, with no connection to any private information of the annotators. We have not made any changes to the annotation results. Overall, the establishment of our datasets is compliant with ethics.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work is supported by the National Natural Science Foundation of China (NSFC) Grant Nos.U23B2055 and Nos.62176180.

References

- Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. 2022. Towards more robust keyword spotting for voice assistants. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2655–2672.
- Annalena Aicher, Alisa Gazizullina, Aleksei Gusev, Yuri Matveev, and Wolfgang Minker. 2022. [Towards speech-only opinion-level sentiment analysis](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2000–2006, Marseille, France. European Language Resources Association.
- Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse H. Engel, Linxi Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Han, Patrick LeGresley, Xiangang Li, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Sheng Qian, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Chong Wang, Yi Wang, Zhiqian Wang, Bo Xiao, Yan Xie, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2016. [Deep speech 2 : End-to-end speech recognition in english and mandarin](#). In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 173–182. JMLR.org.
- Mustafa Can Bingol and Omur Aydogmus. 2020. [Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot](#). *Engineering Applications of Artificial Intelligence*, 95:103903.
- Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. [AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline](#). In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*,

- O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017, pages 1–5. IEEE.
- Andrew Caines, Michael McCarthy, and Paula Buttery. 2017. [Parsing transcripts of speech](#). In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 27–36, Copenhagen, Denmark. Association for Computational Linguistics.
- Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. [SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.
- William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. [Listen, attend and spell](#). *CoRR*, abs/1508.01211.
- Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. 2023. [OpenSR: Open-modality speech recognition via maintaining multi-modality alignment](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6592–6607, Toronto, Canada. Association for Computational Linguistics.
- Wu Chou, Chin-Hui Lee, Biing-Hwang Juang, and Frank K. Soong. 1994. [A minimum error rate pattern recognition approach to speech recognition](#). *Int. J. Pattern Recognit. Artif. Intell.*, 8(1):5–31.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Fang Dong, Yiyang Qian, Tianlei Wang, Peng Liu, and Jiuwen Cao. 2023a. [A transformer-based end-to-end automatic speech recognition algorithm](#). *IEEE Signal Process. Lett.*, 30:1592–1596.
- Linhao Dong, Zhecheng An, Peihao Wu, Jun Zhang, Lu Lu, and Ma Zejun. 2023b. [CIF-PT: Bridging speech and text representations for spoken language understanding via continuous integrate-and-fire pre-training](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8894–8907, Toronto, Canada. Association for Computational Linguistics.
- Qingkai Fang and Yang Feng. 2023a. [Back translation for speech-to-text translation without transcripts](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4567–4587, Toronto, Canada. Association for Computational Linguistics.
- Qingkai Fang and Yang Feng. 2023b. [Understanding and bridging the modality gap for speech translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15864–15881, Toronto, Canada. Association for Computational Linguistics.
- Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021. [Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, Online. Association for Computational Linguistics.
- Hao Fei, Meishan Zhang, and Donghong Ji. 2020. [Cross-lingual semantic role labeling with high-quality translated training corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.
- Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. [Semantic role labeling with neural network factors](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal. Association for Computational Linguistics.
- Carol A. Fowler. 1986. [An event approach to the study of speech perception from a direct–realist perspective](#). *Journal of Phonetics*, 14(1):3–28.
- S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin. 2018. [End-to-end named entity and semantic concept extraction from speech](#). In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699.
- Daniel Gildea and Daniel Jurafsky. 2000. [Automatic labeling of semantic roles](#). In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.
- Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meignier. 2013. [Named entity recognition in speech transcripts following an extended taxonomy](#). In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22-23, 2013*, volume 1012 of *CEUR Workshop Proceedings*, pages 61–65. CEUR-WS.org.
- Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. [Jointly predicting predicates and arguments in neural semantic role labeling](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Melbourne, Australia. Association for Computational Linguistics.
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. [Question-answer driven semantic role labeling: Using natural language to annotate natural language](#).

- In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Xuedong Huang, Fileno A. Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld. 1993. [The SPHINX-II speech recognition system: an overview](#). *Comput. Speech Lang.*, 7(2):137–148.
- Paria Jamshid Lou, Yufei Wang, and Mark Johnson. 2019. [Neural constituency parsing of speech transcripts](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2756–2765, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Mohamed Ameer Ben Jannet, Olivier Galibert, Martine Adda-Decker, and Sophie Rosset. 2017. [Investigating the effect of ASR tuning on named entity recognition](#). In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2486–2490. ISCA.
- Cheng-I Jeff Lai, Freda Shi, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David Cox, David Harwath, et al. 2023. [Audio-visual neural syntax acquisition](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- Kai-Fu Lee, Hsiao-Wuen Hon, and Raj Reddy. 1990. [An overview of the SPHINX speech recognition system](#). *IEEE Trans. Acoust. Speech Signal Process.*, 38(1):35–45.
- Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. [A unified syntax-aware framework for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411, Brussels, Belgium. Association for Computational Linguistics.
- Hamza Mahdi, Sami Alperen Akgun, Shahed Saleh, and Kerstin Dautenhahn. 2022. [A survey on the design and evolution of social robots—past, present and future](#). *Robotics and Autonomous Systems*, page 104193.
- Diego Marcheggiani and Ivan Titov. 2017. [Encoding sentences with graph convolutional networks for semantic role labeling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.
- Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. [Special issue introduction: Semantic role labeling: An introduction to the special issue](#). *Computational Linguistics*, 34(2):145–159.
- Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. [A survey on open information extraction](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. [A span selection model for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642, Brussels, Belgium. Association for Computational Linguistics.
- Hiroaki Ozaki, Gaku Morio, Terufumi Morishita, and Toshinori Miyoshi. 2021. [Project-then-transfer: Effective two-stage cross-lingual transfer for semantic dependency parsing](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2586–2594, Online. Association for Computational Linguistics.
- Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Daniel Jurafsky. 2005. [Semantic role labeling using different syntactic views](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 581–588, Ann Arbor, Michigan. Association for Computational Linguistics.
- Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. [The importance of syntactic parsing and inference in semantic role labeling](#). *Computational Linguistics*, 34(2):257–287.
- Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jerome Goullan. 2022. [End-to-End Dependency Parsing of Spoken French](#). In *Proc. Interspeech 2022*, pages 1816–1820.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. [SimulSpeech: End-to-end simultaneous speech to text translation](#).

- In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online. Association for Computational Linguistics.
- Michael Roth and Mirella Lapata. 2016. [Neural semantic role labeling with dependency path embeddings](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.
- Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. [Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.
- Seunghyun Seo, Donghyun Kwak, and Bowon Lee. 2022. [Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7152–7156. IEEE.
- Yutong Shao, Arun Kumar, and Ndapa Nakashole. 2023. [Database-aware ASR error correction for speech-to-sql parsing](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.
- Suwon Shon, Pablo Brusco, Jing Pan, Kyu Jeong Han, and Shinji Watanabe. 2021. [Leveraging pre-trained language model for speech sentiment analysis](#). In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3420–3424. ISCA.
- Edwin Simonet, Sahar Ghannay, Nathalie Camelin, and Yannick Estève. 2018. [Simulating ASR errors for training SLU systems](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. [Cold fusion: Training seq2seq models together with language models](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.
- Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. [Linguistically-informed self-attention for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. [Using predicate-argument structures for information extraction](#). In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan. Association for Computational Linguistics.
- Reid Swanson and Andrew S. Gordon. 2006. [A comparison of alternative parse tree paths for labeling semantic roles](#). In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 811–818, Sydney, Australia. Association for Computational Linguistics.
- Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. [Deep semantic role labeling with self-attention](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Lina Tang. 2023. [A transformer-based network for speech recognition](#). *Int. J. Speech Technol.*, 26(on):531–539.
- Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidion, Justin Hart, Peter Stone, and Raymond Mooney. 2020. [Jointly improving parsing and perception for natural language commands through human-robot dialog](#). *Journal of Artificial Intelligence Research*, 67:327–374.
- Mohamed A Torad, Belgacem Bouallegue, and Abdelmoty M Ahmed. 2022. [A voice controlled smart home automation system using artificial intelligent and internet of things](#). *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 20(4):808–816.
- Yuan Tseng, Cheng-I Jeff Lai, and Hung-yi Lee. 2023. [Cascading and direct approaches to unsupervised constituency parsing on spoken sentences](#). In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi. 2020. [Grounded language interpretation of robotic commands through structured learning](#). *Artif. Intell.*, 278.
- Changan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. [VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Qingrong Xia, Zhenghua Li, Min Zhang, Meishan Zhang, Guohong Fu, Rui Wang, and Luo Si. 2019. Syntax-aware neural semantic role labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7305–7313.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese treebank. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 47–54, Sapporo, Japan. Association for Computational Linguistics.

Meishan Zhang, Huiyao Chen, Xin Zhang, Jing Chen, and Min Zhang. 2023. Faking a teacher works! dependency scoring learning and corpus boosting for translation-based cross-lingual dependency parsing. *Dependency Scoring Learning and Corpus Boosting for Translation-Based Cross-Lingual Dependency Parsing*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 997–1006, Hong Kong, China. Association for Computational Linguistics.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4212–4227, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

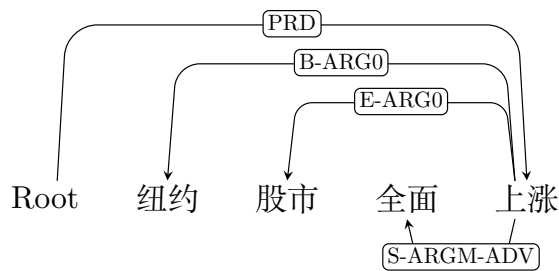
Shilin Zhou, Qingrong Xia, Zhenghua Li, Yu Zhang, Yu Hong, and Min Zhang. 2022. Fast and accurate end-to-end span-based semantic role labeling as word-based graph parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4160–4171, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Su Zhu, Zijian Zhao, Tiejun Zhao, Chengqing Zong, and Kai Yu. 2019. CATSLU: the 1st chinese audio-textual spoken language understanding challenge. In *International Conference on Multimodal Interaction, ICMi 2019, Suzhou, China, October 14-18, 2019*, pages 521–525. ACM.

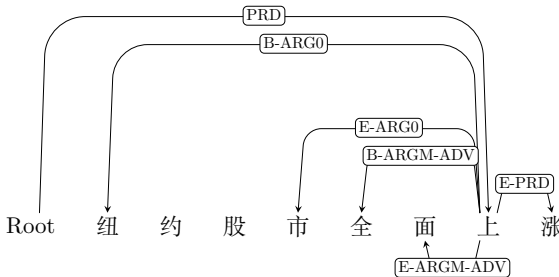
A Appendix

A.1 Character-based Dependency Graph

SRLGParser is a word-based graph, which also needs to take into account its positional information



(a) An example for word-based dependency graph.



(b) An example for character-based dependency graph.

Figure 6: Results of word-based dependency graph and character-based graph for the same sentence. In character-based graph, the recognition that the predicate span is achieved by adding the label **E-PRD** to the dependency graph.

when connecting arguments to obtain a complete argument span. It proposes four attach schemata to build the graph, and we choose the BES schema by its default. The BES schema uses "B/E/S-r" as the label, indicating the beginning, end, and single. Given a sentence that made up of words, first add a "Root" node to the front of the sentence, and it points to all predicates. The relation between a predicate and an argument is represented by the edge. If the argument is just a word, the predicate is directly points to the argument, and its label is "S-r". Otherwise, the predicate points to the head and tail of the argument span, and its labels are "B-r" and "E-r" respectively. This can form a word-based dependency graph as in Figure 6 (a).

It is worth noting that the predicate for word-based SRL only consists of a single word, however, for character-based SRL, the predicate may consist of multiple characters. In order to solve the problem that the original SRLGParser cannot recognize the predicate span, we add a new label "E-PRD" to indicate the tail of predicate. Unlike constructing word-based graph, in character-based graph, when the predicate is composed of multiple characters, "Root" points to the head of the predicate span, and the relation between the predicate and the argument is reflected by the connection between the head of

Table 4: The semantic role label distribution of AS-SRL.

Role	Num	Role	Num
ARG0	7,469	ARG1	10,209
ARG2	1,052	ARG3	36
ARG4	1	ARGM-ADV	5,060
ARGM-TMP	1,156	ARGM-MNR	286
ARGM-LOC	696	ARGM-EXT	114
ARGM-PRP	95	ARGM-DIS	754
ARGM-DIR	91	ARGM-TPC	30
ARGM-CND	2	ARGM-BNF	63
ARGM-FRQ	9	-	-

the predicate and the argument, and additionally, the connection to the tail of the predicate. The result is shown in Figure 6 (b).

A.2 Annotation Process

In this work, we annotate two speech-based Chinese SRL datasets. To ensure that we obtain high-quality labeled datasets, we undertake the following procedures.

For CPB1.0, we recruit three college students with Mandarin proficiency (level 2-A or above) to serve as annotators. The test set is evenly divided among them for reading and speech recording. While each annotator reads the text, the other two annotators act as listener to check whether the current speech can be understood correctly. Otherwise, the speaker will be asked to read the text again until it could be understood correctly.

For AS-SRL, the entire annotation workflow is divided into three stages: (i) developing a user-friendly online tool for annotators; (ii) establishing standard annotation guidelines based on the CPB1.0 rules; (iii) recruiting some experts to annotate the dataset, ensuring that each instance receives three annotations.

First, we use the label-studio platform to design an annotation platform for the SRL task. Then, we organize the CPB1.0 label definitions and selected 100 pairs of CPB1.0 data as examples to develop annotation guidelines.

Next, three university students specializing in Chinese linguistics are recruited for annotation. Before commencing the task, annotators are required to thoroughly review the annotation process guide and attempt annotating some data. Official annotation begins once their annotation results meet the

required standards. We ask two of them to annotate the data. Upon completion of annotation, a proof-reading process is conducted and the consistency rate between the two annotators is 86%. Then, the third annotator check the different labeling results to decide which one to use. If all three have different opinions, the results will be sent to experts for correction.

Finally, we obtain AS-SRL, with the semantic role label distribution as shown in in Table 4. Note that in the annotation process, we follow the annotation method of CPB1.0 span-based SRL, i.e., we first annotate the spans corresponding to predicates and arguments, then annotate the corresponding relations, and finally converting these annotations into the dependency graph format.