Freda Shi: Research Statement

Learning Language Structures through Grounding Signals and Beyond

Humans can learn language naturally and efficiently, as well as using natural language to interact with the world. Language structures, such as syntactic and semantic parses of sentences (Figure 1), play an important role in such processes: with awareness of structure, humans can judge whether a sentence is grammatical, compose sentence meaning, and produce grammatical sentences describing new objects and events. Even more impressively, though humans implicitly develop and use structure for language processing in their daily communication, the explicit structure of sentences is almost never given.

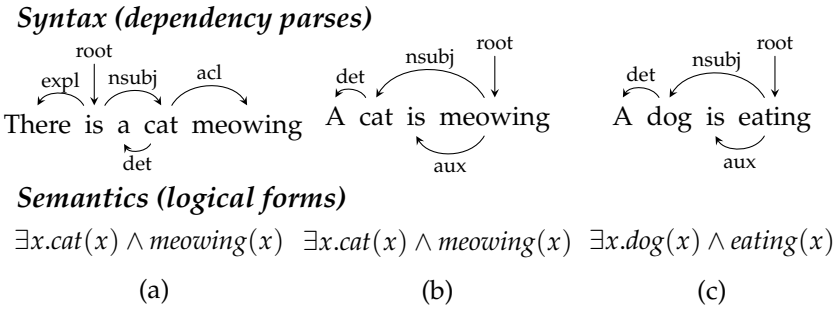
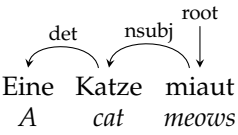


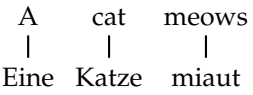
Figure 1: Examples of syntactic and semantic structures of sentences, using dependency parses and logical forms as the representatives. Sentences with different syntax may have the same semantics, i.e., express the same meaning (a and b). Sentences with different meanings may share the same syntactic structure (b and c).

My long-term goal is to build human-like systems that can learn and use language in natural settings. By natural settings, I refer to not only limited amounts of data and annotated supervision, but also situations involving cross-modal grounding signals (e.g., vision) that link language to the concrete world. I believe that learning language structures without extensive supervision is a crucial intermediate step towards my long-term goal; therefore, I have worked on learning both syntactic [**S**MGL, ACL'19; **S**LG, EMNLP'20] and semantic [**S**FGZW, EMNLP'22; M**S**WLT, NeurIPS'21] structures of sentences, through grounding signals or with very few manually annotated training examples.

In addition to human-like first language learning described above, I, as a second language learner, am interested in developing multilingual natural language processing (NLP) systems. Cross-lingual structures, such as universal dependency relations (Figure 2a) and word alignment (Figure 2b), may serve as a bridge between different languages. To this end, I have built models for cross-lingual word alignment [**S**ZW, ACL'21] and parsing [**S**LG, ACL Findings'21; **S**GL, ACL'22], through cross-lingual grounding signals such as mutually translatable sentences. Aside from structures, I have designed methods for bilingual lexicon induction [**S**ZW, ACL'21] and cross-lingual reasoning [**S**SFW+, ICLR'23].



(a) Universal dependencies



(b) Cross-lingual word alignment

Figure 2: Examples of cross-lingual structures covered in my work.

My past research has centered on addressing the following questions:

- How can we build models that learn syntactic structures of sentences through visual grounding signals?
- How can we build models that learn semantic structures by verifying them with execution?
- How can we build universal NLP systems for the diverse language families in the world?

Along the way, my work also makes general contributions to the broader NLP community by introducing new data augmentation [**S**LG, EMNLP'20; **S**LG, ACL Findings'21] and sampling [**S**FGZW, EMNLP'22] methods, as well as providing insights towards understanding pretrained large language models [**S**ZW, ACL'21; **S**GL, ACL'22; **S**FGZW, EMNLP'22; **S**SFW+, ICLR'23; T**S**SG+, RepL4NLP'20].

Below I will describe some of the approaches that I have been pursuing, as well as the future directions.

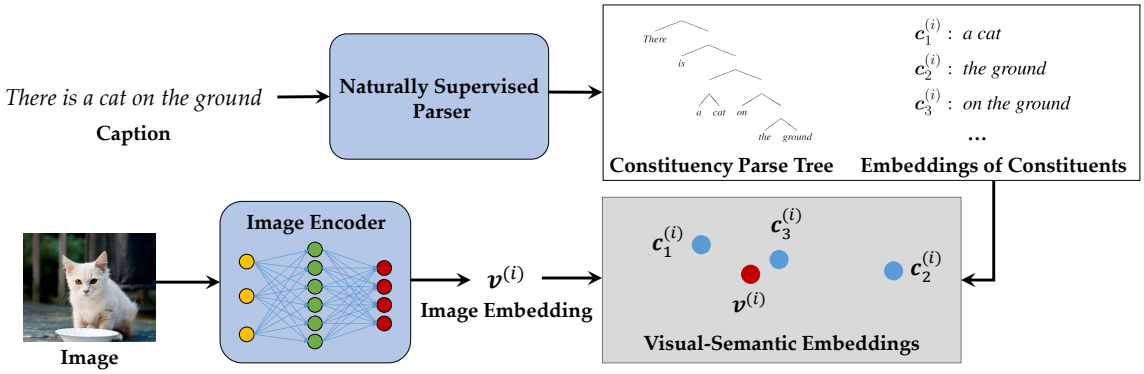


Figure 3: An overview of VG-NSL [**S**MGL, ACL'19]. The model takes image-text pairs, and parses the text based on the visual concreteness estimated within the joint visual-semantic embedding space. The model never sees explicit structure of text during the whole training procedure.

Visually grounded grammar induction. Language is rarely text in isolation: humans learn language, including syntactic structures, through interaction with others and the world. Inspired by this, I introduce the task of visually grounded grammar induction and develop the first promising system on the task [**S**MGL, ACL'19]: building on the hypothesis that more visually concrete spans of words are more likely to be phrases, I propose the Visually Grounded Neural Syntax Learner (VG-NSL; Figure 3) that induces constituency parses from paired texts and images. Such a visually grounded grammar induction model outperforms pure text-based models in terms of both parsing accuracy and stability across random seeds. VG-NSL was recognized with a nomination for the best paper award at ACL 2019, and the task of grounded grammar induction continues to attract the attention of other research groups [3; 4; 5, *inter alia*], two of which have received paper awards.

Along the same line, I have built a system that induces combinatory categorial grammars [CCG; 2], a joint formalism of both syntax and semantics, through visual question answering pairs [M**S**WLT, NeurIPS'21]. Moreover, the semantic bootstrapping hypothesis states that word meanings may assist syntax learning; as foundation for grammar induction, I have also built systems that can robustly learn lexical semantics from captioned images with contrastive adversarial examples [**S**MXJS, COLING'18].

Learning semantics with execution. Meaning expressed by natural language sentences can be represented by translating them into executable programs, and can usually be grounded into the real world: after generating the corresponding programs (i.e., semantic parses), the most intuitive way to verify and improve their quality is to execute them and check the execution results.

Therefore, I have worked on execution-supervised semantic parsing: we train a CCG induction model by comparing the execution results with the ground truth [M**S**WLT, NeurIPS'21]. Without access to any ground-truth semantic parses, our neural network–based joint syntax and semantics induction model passes the challenging generalization tests of SCAN [1], a synthetic natural language understanding benchmark, with 100% accuracy: for example, it successfully maps the command *jump twice* to the action sequence "JUMP JUMP" while only seeing the ground-truth action sequences of *jump* and *run twice* during training.

In more recent work, I present an execution-based minimum Bayes risk decoding algorithm (MBR-EXEC) for natural language to code translation [**S**FGZW, EMNLP'22]. Starting with a set of candidate programs, we define the Bayes risk of a program to estimate the discrepancy between its execution result and those of the other programs. The program with the lowest Bayes risk is chosen as the final translation. This algorithm significantly outperforms all conventional execution-unaware baselines such as maximum likelihood decoding. On the task of text-to-SQL and text-to-bash translation, with as few as 15 examples of sentence–semantic parse pairs and without access to ground-truth execution results, we reached competitive performance with prior state-of-the-art models that require thousands of examples to train.

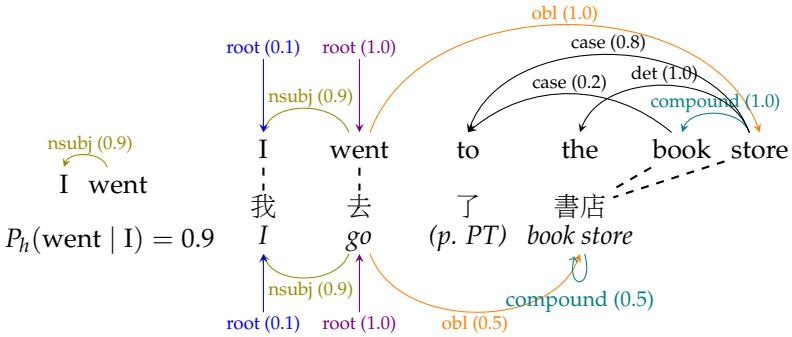
Multilingual NLP through cross-lingual grounding. There are more than 7,000 languages all over the world. While we have reached surprisingly high performance on high-resource language processing across many tasks, an ideal universal NLP system should be able to process and understand other languages as well.

To this end, I have built a unified feature-based model that extracts bilingual lexicons (Figure 4) and cross-lingual word alignment to serve as the foundation for cross-lingual transfer [**S**ZW, ACL'21]. Our induced lexicons have reached the same level of quality as the prior state of the art that requires extensive human effort to build, and our word alignment system with a similar architecture reaches a new state of the art without requiring annotated parallel sentences. This work is recognized with a nomination for the best paper award at ACL 2021 by reviewers.

English	Spanish
bank	banco *(financial institution)*
bank	orilla *(riverbank)*
shore	orilla *(riverbank)*

Figure 4: Example entries of a bilingual lexicon: each pair of words represents a mutual translation in a certain context.

In addition, I have developed two substructure-based techniques, substructure substitution [**S**LG, EMNLP'20; **S**LG, ACL Findings'21] and substructure distribution projection [Figure 5; **S**GL, ACL'22] that improve cross-lingual transfer of syntactic parsing, with zero or only a few (e.g, 50) annotated examples in the target languages. Starting from a trained English parser and using only monolingual corpora, we reached a new state of the art on zero-shot cross-lingual dependency parsing, improving over the prior best by an average absolute unlabeled attachment score of 18.9% across four distant languages.



p. PT: particle denoting past tense.

Figure 5: Overview of substructure distribution projection for zero-shot cross-lingual dependency parsing [**S**GL, ACL'22]. Let $P_h(x \mid y)$ denote the probability of x being the head of y. We project P_h predicted by an English parser (top) to Chinese (bottom) through cross-lingual word alignment. A Chinese parser is then trained from scratch to fit the projected distribution.

Future and Current Work

Building on the initial success reached by my work described above, I am enthusiastic about working on the following directions in the near future:

Using additional grounding signals. In addition to static images, introducing more forms of grounding signals such as audio or video, may model human behaviors more comprehensively and strengthen the models. Along this line, an ongoing project of mine models syntax acquisition through visually grounded speech and without a supervised speech recognition model. I am also interested in machine language understanding through other grounding signals, such as tactile and eye-tracking information, as well as through human-computer interaction. Beyond language structures, since these grounding signals provide more information about the context, I would like to explore how they can be used in modeling context-dependent understanding of language, i.e., pragmatics.

Improving efficiency and generalizability with awareness of structures. While pretrained large language models are powerful, they require a large amount of data to train, which is not always available in practice. I am interested in developing models that are similarly powerful but more efficient in terms of both data and computational resources, where structures of natural languages learned or induced from my prior work can be used to guide the learning process. Implicit or explicit awareness of structures may additionally help generalization from shorter sentences to longer ones that share similar structures: for example, I am interested

in designing models that understand complicated sentences like "*I heard that Alice said Bob said you are right*" by training with simple examples that have similar syntax, such as "*I said you are right*".

Towards universal natural language understanding. While languages all over the world share many commonalities, they are typologically diverse. Moreover, current models are arguably designed on and biased towards high-resource languages; therefore, I will seek to build systems that can work equally well for low-resource languages through developing cross-lingual transfer techniques. My future work will consider both generic modeling and language-specific parameters such as head directionality. I am particularly interested in analyzing the syntactic and semantic phenomena presented by the pretrained large language models, as well as improving the models based on my findings in a resource-efficient manner.

Scientifically, I am also interested in applying computational methods and grounding signals to facilitate linguistics and cognitive science research. For example, I am investigating whether the grammatical gender systems can be grounded to concrete visual properties in an ongoing project. I would also be happy to collaborate more with linguists and cognitive scientists on a broader range of topics.

I am excited to explore the above directions in collaboration with others.

References

[**S**MXJS, COLING'18] **Haoyue Shi**, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. "Learning Visually-Grounded Semantics from Contrastive Adversarial Samples". In: *COLING*. 2018.

[**S**MGL, ACL'19] **Haoyue Shi**, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. "Visually Grounded Neural Syntax Acquisition". In: *ACL*. 2019.

[**S**LG, EMNLP'20] **Haoyue Shi**, Karen Livescu, and Kevin Gimpel. "On the Role of Supervision in Unsupervised Constituency Parsing". In: *EMNLP*. 2020.

[**S**ZW, ACL'21] **Haoyue Shi**, Luke Zettlemoyer, and Sida I. Wang. "Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment". In: *ACL*. 2021.

[**S**LG, ACL Findings'21] **Haoyue Shi**, Karen Livescu, and Kevin Gimpel. "Substructure Substitution: Structured Data Augmentation for NLP". In: *Findings of ACL*. 2021.

[**S**GL, ACL'22] **Freda Shi**, Kevin Gimpel, and Karen Livescu. "Substructure Distribution Projection for Zero-Shot Cross-Lingual Dependency Parsing". In: *ACL*. 2022.

[**S**FGZW, EMNLP'22] **Freda Shi**, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. "Natural Language to Code Translation with Execution". In: *EMNLP*. 2022.

[**S**SFW+, ICLR'23] **Freda Shi**, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. "Language Models are Multilingual Chain-of-Thought Reasoners". In: *ICLR*. 2023.

[M**S**WLT, NeurIPS'21] Jiayuan Mao, **Haoyue Shi**, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. "Grammar-Based Grounded Lexicon Learning". In: *NeurIPS*. 2021.

[T**S**SG+, RepL4NLP'20] Shubham Toshniwal, **Haoyue Shi**, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. "A Cross-Task Analysis of Text Span Representations". In: *Proc. of RepL4NLP*. 2020.

[1] Brenden Lake and Marco Baroni. "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrentlake networks". In: *ICML*. 2018.

[2] Mark Steedman. *The Syntactic Process*. Cambridge, MA: MIT Press, 2000.

[3] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. "What is Learned in Visually Grounded Neural Syntax Acquisition". In: *ACL*. 2020.

[4] Yanpeng Zhao and Ivan Titov. "Visually Grounded Compound PCFGs". In: *EMNLP*. 2020.

[5] Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. "Video-aided Unsupervised Grammar Induction". In: *NAACL-HLT*. 2021.