Supplementary Material

Variables Summary

Variable	Domain	Description
n	$\mathbb{N}_+$	sentence length
r	$\mathbb{N}_+$	number of semantic roles
m	$\mathbb{N}_+$	number of senses
T	$\mathbb{N}_+$	number of refinement iterations
j	$\mathbb{N}_+$	position of the given predicate
R	Δ_{r-1}^{n}	semantic roles probability
P	Δ_{m-1}	senses probability
Π	$\mathbb{R}^{m \times d_\pi}$	predicate senses embedding
x^{w}	$\mathbb{R}^{n \times d_w}$	sentence tokens embeddings
x^{dep}	$\mathbb{R}^{n \times d_\delta}$	dependency labels embeddings
x^{pos}	$\mathbb{R}^{n \times d_p}$	part-of-speech tags embeddings
x	$\mathbb{R}^{n \times (d_w + d_\pi + d_p)}$	concatenated representation
Baseline Model		
h	$\mathbb{R}^{n \times d_h}$	encoded sentence representation
h^{ρ_0}	$\mathbb{R}^{n \times d_{\rho_0}}$	argument feature for null role logits
h^{ρ_1}	$\mathbb{R}^{n \times d_{\rho_1}}$	argument feature for other roles logits
h^{ϱ_0}	$\mathbb{R}^{d_{\rho_0}}$	predicate feature for null role logits
h^{ϱ_1}	$\mathbb{R}^{d_{\rho_1}}$	predicate feature for other roles logits
I^{ρ_0}	$\mathbb{R}^{n}$	null role logits
I^{ρ_1}	$\mathbb{R}^{n \times (r-1)}$	other roles logits
I^{α}	$\mathbb{R}^{n \times r}$	roles logits
h^{π}	$\mathbb{R}^{d_\pi}$	predicate feature for sense disambiguation
I^{π}	$\mathbb{R}^{m}$	sense logits
Refinement Network		
g	$\mathbb{R}^{n \times d_g}$	encoded sentence representation
g^{α}	$\mathbb{R}^{n \times d_\alpha}$	argument feature
g^{π}	$\mathbb{R}^{d_\pi}$	predicate feature
o_i	$\mathbb{R}^{r-1}$	sum of other roles
z_i^{α}	$\mathbb{R}^{2r-1+2d_g+d_\pi}$	input to role refinement network
M^{α}	$\mathbb{R}^{n \times r}$	refinement role logits to be added with I^{α}
r_{π}	$\mathbb{R}^{r-1}$	sum of all roles
z^{π}	$\mathbb{R}^{2r-1+2d_g+d_\pi}$	input to sense refinement network
M^{π}	$\mathbb{R}^{m}$	refinement sense logits to be added with I^{π}

Table 1: Variables domain and description

Networks

Network	Description
BiLSTM	3 layers stacked highway BiLSTM
MLP	1 layer MLP with exponential linear units

Table 2: Networks for all languages at all occurence in the main text. Note that the input and output dimensions of MLP and BiLSTM can be decided by the other hyper-parameters at each occurrence.

Hyper-Parameters

Hyper-parameter	Value		Description
	English	Others	
p	0.3		dropout rate for all neural modules
p_r	0.3		recurrent dropout rate for BiLSTMs
d_w	1024	300	tokens embedding dimension
d_δ	64		dependency label embedding dimension
d_p	64		part-of-speech tags embedding dimension
d_h	500	428	BiLSTM hidden state dimension in one direction
d_{ρ_0}	300		dimension for feature for null role logits
d_{ρ_1}	128		dimension for feature for other role logits
d_g	200		dimension for feature for refinement networks
d_r	200		hidden dimension of refinement networks
λ_g^π	50		multiplier of Gumbel noise for sense logits
λ_g^α	5		multiplier of Gumbel noise for role logits

Table 3: Hyper-parameters value and description. Note that the input and output dimensions of MLP and BiLSTM can be decided by the other hyper-parameters at each occurrence.

References

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1027–1035, USA. Curran Associates Inc.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2377–2385.