CoNLL-2005	Greedy F1	Viterbi F1	Δ F1
LISA	81.99	82.24	+0.25
+D&M	83.37	83.58	+0.21
+Gold	*86.57*	*86.81*	*+0.24*

CoNLL-2012	Greedy F1	Viterbi F1	Δ F1
LISA	80.11	80.70	+0.59
+D&M	81.55	82.05	+0.50
+Gold	*85.94*	*86.43*	*+0.49*

Table 7: Comparison of development F1 scores with and without Viterbi decoding at test time.

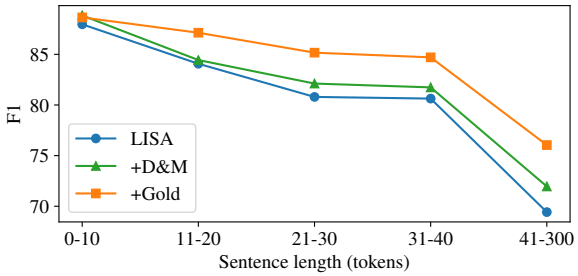

Figure 5: F1 score as a function of sentence length.

A Supplemental Material

A.1 Supplemental analysis

Here we continue the analysis from §4.3. All experiments in this section are performed on CoNLL-2005 development data unless stated otherwise.

First, we compare the impact of Viterbi decoding with LISA, D&M, and gold syntax trees (Table 7), finding the same trends across both datasets. We find that Viterbi has nearly the same impact for LISA, D&M and gold parses: Gold parses provide little improvement over predicted parses in terms of BIO label consistency.

We also assess SRL F1 as a function of sentence length and distance from span to predicate. In Figure 5 we see that providing LISA with gold parses is particularly helpful for sentences longer than 10 tokens. This likely directly follows from the tendency of syntactic parsers to perform worse on longer sentences. With respect to distance between arguments and predicates, (Figure 6), we do not observe this same trend, with all distances performing better with better parses, and especially gold.

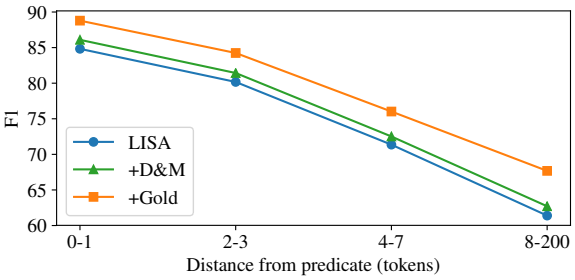

Figure 6: CoNLL-2005 F1 score as a function of the distance of the predicate from the argument span.

	L+/D+	L-/D+	L+/D-	L-/D-
Proportion	37%	10%	4%	49%
SA	76.12	75.97	82.25	65.78
LISA	76.37	72.38	85.50	65.10
+D&M	76.33	79.65	75.62	66.55
+Gold	*76.71*	*80.67*	*86.03*	*72.22*

Table 8: Average SRL F1 on CoNLL-2012 for sentences where LISA (L) and D&M (D) parses were correct (+) or incorrect (-).

A.2 Supplemental results

Due to space constraints in the main paper we list additional experimental results here. Table 9 lists development scores on the CoNLL-2005 dataset with predicted predicates, which follow the same trends as the test data.

A.3 Data and pre-processing details

We initialize word embeddings with 100d pre-trained GloVe embeddings trained on 6 billion tokens of Wikipedia and Gigaword (Pennington et al., 2014). We evaluate the SRL performance of our models using the `srl-eval.pl` script

WSJ Dev	P	R	F1
He et al. (2018)	84.2	83.7	83.9
Tan et al. (2018)	82.6	83.6	83.1
SA	83.12	82.81	82.97
LISA	83.6	83.74	83.67
+D&M	**85.04**	**85.51**	**85.27**
+Gold	*89.11*	*89.38*	*89.25*

Table 9: Precision, recall and F1 on the CoNLL-2005 development set with gold predicates.

provided by the CoNLL-2005 shared task,[7] which computes segment-level precision, recall and F1 score. We also report the predicate detection scores output by this script. We evaluate parsing using the `eval.pl` CoNLL script, which excludes punctuation.

We train distinct D&M parsers for CoNLL-2005 and CoNLL-2012. Our D&M parsers are trained and validated using the same SRL data splits, except that for CoNLL-2005 section 22 is used for development (rather than 24), as this section is typically used for validation in PTB parsing. We use Stanford dependencies v3.5 (de Marneffe and Manning, 2008) and POS tags from the Stanford CoreNLP `left3words` model (Toutanova et al., 2003). We use the pre-trained ELMo models[8] and learn task-specific combinations of the ELMo representations which are provided as input instead of GloVe embeddings to the D&M parser with otherwise default settings.

A.3.1 CoNLL-2012

We follow the CoNLL-2012 split used by He et al. (2018) to evaluate our models, which uses the annotations from here[9] but the subset of those documents from the CoNLL-2012 co-reference split described here[10] (Pradhan et al., 2006). This dataset is drawn from seven domains: newswire, web, broadcast news and conversation, magazines, telephone conversations, and text from the bible. The text is annotated with gold part-of-speech, syntactic constituencies, named entities, word sense, speaker, co-reference and semantic role labels based on the PropBank guidelines (Palmer et al., 2005). Propositions may be verbal or nominal, and there are 41 distinct semantic role labels, excluding continuation roles and including the predicate. We convert the semantic proposition and role segmentations to BIO boundary-encoded tags, resulting in 129 distinct BIO-encoded tags (including continuation roles).

A.3.2 CoNLL-2005

The CoNLL-2005 data (Carreras and Màrquez, 2005) is based on the original PropBank corpus (Palmer et al., 2005), which labels the Wall Street Journal portion of the Penn TreeBank corpus (PTB) (Marcus et al., 1993) with predicate-argument structures, plus a challenging out-of-domain test set derived from the Brown corpus (Francis and Kučera, 1964). This dataset contains only verbal predicates, though some are multi-word verbs, and 28 distinct role label types. We obtain 105 SRL labels including continuations after encoding predicate argument segment boundaries with BIO tags.

A.4 Optimization and hyperparameters

We train the model using the Nadam (Dozat, 2016) algorithm for adaptive stochastic gradient descent (SGD), which combines Adam (Kingma and Ba, 2015) SGD with Nesterov momentum (Nesterov, 1983). We additionally vary the learning rate lr as a function of an initial learning rate lr_0 and the current training step $step$ as described in Vaswani et al. (2017) using the following function:

$$lr = lr_0 \cdot \min(step^{-0.5}, step \cdot warm^{-1.5}) \quad (8)$$

which increases the learning rate linearly for the first $warm$ training steps, then decays it proportionally to the inverse square root of the step number. We found this learning rate schedule essential for training the self-attention model. We only update optimization moving-average accumulators for parameters which receive gradient updates at a given step.[11]

In all of our experiments we used initial learning rate 0.04, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-12}$ and dropout rates of 0.1 everywhere. We use 10 or 12 self-attention layers made up of 8 attention heads each with embedding dimension 25, with 800d feed-forward projections. In the syntactically-informed attention head, Q_{parse} has dimension 500 and K_{parse} has dimension 100. The size of $predicate$ and $role$ representations and the representation used for joint part-of-speech/predicate classification is 200. We train with $warm = 8000$ warmup steps and clip gradient norms to 1. We use batches of approximately 5000 tokens.

[7] http://www.lsi.upc.es/~srlconll/srl-eval.pl

[8] https://github.com/allenai/bilm-tf

[9] http://cemantix.org/data/ontonotes.html

[10] http://conll.cemantix.org/2012/data.html

[11] Also known as *lazy* or *sparse* optimizer updates.