Supplementary Material

Data Collection Details We provided workers with an auto-complete drop-down menu consisting of our type vocabulary (see Figure 1). The type vocabulary was constructed by taking all the singular nouns in Wiktionary (including multiword expressions, such as "prime minister"), then pruning words that appeared less than 5 times in our training data or were not in 40,000 frequent terms in the GloVe vocabulary.

Five workers annotated each example, producing an initial set of types T_0. We expanded T_0 to T_1 by adding synonyms and hypernyms from WordNet, as well as randomly selected negative types. We then asked five different annotators which types in T_1 fit the context c. To ensure good annotation quality, we selected the most agreed-upon type in T_0 as a true-positive and one of the random negative types as a true-negative, and prevented annotators who misclassified them from completing the task.

Each pair of annotators agreed on 85% of the binary decisions (i.e. whether a type is suitable or not), and 0.47 in Fleiss's κ. To further improve consistency, the final type set T contained only types selected by at least 3/5 annotators. We removed examples without any types after the pruning.

Hyperparameters We use 300 dimensional pre-trained GloVe word vectors, [1] 50 dimensions for the location vector, and set the LSTMs' dimensions to 100. For the attention mechanism, we used 100 for the hidden dimension. For the mention span representation, the character embedding dimension was 100, and the filter number for character CNN was 50. We used pytorch for implementations. We used dropout for regularization, with a probability of 0.2 for the input sequence pre-trained embeddings, and 0.5 for mention representations. The sentences are cut off after 50 tokens, mention spans are cut off after 25 characters and we ignored mentions with longer than 10 words during training. The model parameters are optimized with Adam, with an initial learning rate of 0.001, over batches of 1000 examples.

OntoNotes Fine-grained Entity Typing Results For further analysis, we divided mentions into two categories: mentions only annotated with '/other' and all other mentions (typed). We show macro-averaged precision, recall, and F1 for typed mentions, and accuracy for '/other' mentions. In Table 1, we show the ablation study of different sets of supervision and the performance breakdown between miscellaneous and typed mentions. Our data augmentation with negative examples significantly improved detecting mentions beyond the existing ontology (/other), and also improved the overall performance.

[1] http://nlp.stanford.edu/data/glove.840B.300d.zip

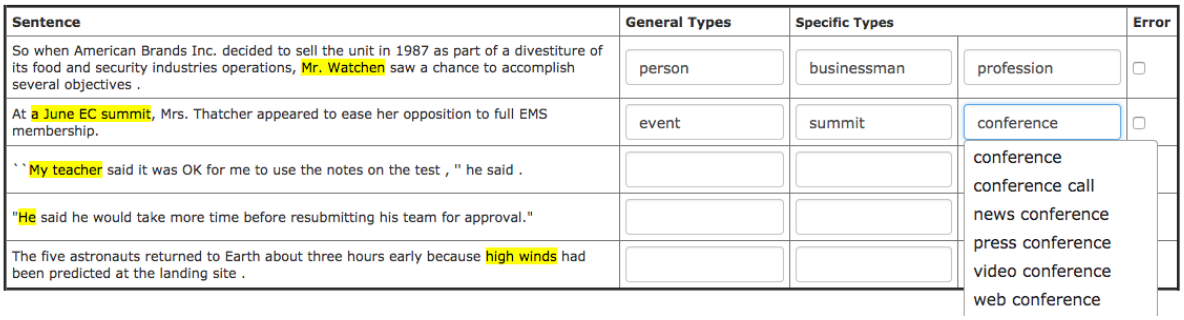

Figure 1: Data collection Framework screenshot. The crowdworkers are provided with auto-complete vocabulary which lists all nouns in the Wikitionary.

	Train Data			Total (2202)			Typed (1069)			Other (1133)
	ONTO	WIKI	HEAD	Acc.	Ma-F1	Mi-F1	P	R	Ma-F1	Accuracy
Attentive	✓			46.5	63.3	58.3	60.1	39.8	47.9	73.0
NER	✓	✓	✓	53.7	72.8	68.0	**70.2**	48.2	57.2	82.6
	✓			41.7	64.2	59.5	61.8	48.5	54.4	59.3
	✓	✓		48.5	67.6	63.6	67.1	**51.8**	58.4	70.0
Ours	✓		✓	57.9	73.0	66.9	57.9	42.3	48.9	92.6
		✓	✓	60.1	75.0	68.7	59.6	45.0	51.3	**95.7**
	✓	✓	✓	**61.6**	**77.3**	**71.8**	67.4	**51.8**	**58.6**	92.6

Table 1: Performance breakdown on the OntoNotes development set. Both new distant supervision improves the performance, both on our model and the prior model.