Corpus	Named Entities	Sents	Tokens	Mentions
BC2GM	Gene/Protein	20,131	569,912	24,585
BC4CHEM	Chemical	87,685	2,544,305	84,312
NCBI	Disease	7,287	184,167	6,883
JNLPBA	Gene/Protein, DNA, Cell-type, Cell-line, RNA	24,806	595,994	59,965
Linnaeus	Species	23,155	539,428	4,265

Table 3: Statistics for the Training Corpora

Corpus	Named Entities	Sents	Tokens	Mentions
BC5CDR	Chemical, Disease	13,938	360,373	28,789
BioNLP13CG	Gene/Protein, Disease, Chemical, Others	1,906	52,771	6881
BioNLP11ID	Gene/Protein, Chemical, Others	5178	166416	11084

Table 4: Statistics for global evaluation corpora. "Others" denote the NEs which do not appeared in training data, thus are not evaluated.

	Articles	Sentences	Tokens
Training set	946	14,987	203.621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

Table 5: Statistics for the CoNLL 2003 NER dataset

A Appendix

A.1 Datasets

Below we introduce the datasets in the biomedicine domain and the news domain.

A.1.1 Biomedicine domain: Local training group

The training group consists of five datasets: *BC2GM*, *BC4CHEM*, *NCBI-disease*, *JNLPBA*, and *Linnaeus*. The first two datasets are from different BioCreative shared tasks (Smith et al., 2008; Krallinger et al., 2015; Wei et al., 2015). *NCBI-disease* is created by Doğan et al. (2014) for disease name recognition and normalization. *JNLPBA* comes from the 2004 shared task from joint workshop on natural language processing in biomedicine and its applications (Kim et al., 2004), and *Linnaeus* is a species corpus composed by Gerner et al. (2010). More information about the datasets can be found in Table 3.

Below are detailed descriptions of the datasets:

BC2GM is a gene/protein corpus. The annotation is Gene. It's provided by the BioCreative II Shared Task for gene mention recognition.

BC4CHEM is a chemical corpus. The annotation is Chemical. It's provided by the BioCreative IV Shared Task for chemical mention recognition.

NCBI-disease is a disease corpus. The annotation is Disease. It was introduced for disease name recognition and normalization.

JNLPBA consists of DNA, RNA, Gene/Protein, Cell line, Cell Type. The annotation is same as the NE names, except the Gene/Protein is annotated with Protein. It was provided by 2004 JNLPBA Shared Task for biomedical entity recognition.

Linnaeus is a species corpus. The annotation is Species. The original project was created for entity mention recognition.

A.1.2 Biomedicine domain: Global evaluation group

We reemphasize here that the purpose of the global evaluation is to test the model's ability to making global predictions and efficiently adapt to global corpora. While no corpus is globally annotated, we identify several existing corpora to *approximate* the global evaluation. Each test corpus is annotated with a *superset* of several training corpora to test the model's generalizability outside of the local tag spaces.

The global evaluation group contains three datasets: *BC5CDR*, *BioNLP13CG*, and *BioNLP11ID*. Each is annotated with multiple entity types. *BC5CDR* comes from the BioCreative shared tasks (Smith et al., 2008; Krallinger et al., 2015; Wei et al., 2015). *BioNLP13CG* and *BioNLP11ID* come from the BioNLP shared task (Kim et al., 2013). More information about the global evaluation datasets can be found in Table 4.

Below are detailed descriptions of the datasets:

BC5CDR is a chemical and disease corpus. The annotation is Chemical and Disease. It's provided by BioCreative V Shared Task for chemical and disease mention recognition.

BioNLP13CG consists of Gene/Protein and Related Product, Cancel, Chemical, Anatomy and Organism and others. **BioNLP11ID** consists of Gene/Protein, Chemical, and Organism. The annotation is same as the NE types but has a finer ontology scope.

Corpus	BC2GM	BC4CHM	NCBI	JNLPBA	Linnaeus
STM	79.9	88.6	84.1	72.7	87.3
MTM Crichton et al. (2017)	73.2	83.0	80.4	70.1	84.0
MTM Wang et al. (2018)	80.7	89.4	86.1	73.5	-
MTM (ours)	80.3	89.2	85.8	73.5	**88.5**
Unified-01	70.9	83.5	79.8	80.9	79.9
Unified-11	74.2	84.1	80.5	80.9	80.7
Unified-00	79.1	87.3	84.0	**83.8**	83.9

Table 6: Local evaluation (f1 scores). The best results that are significantly better than the second best are boldfaced, while those are best but not significantly better than the second best are underlined. All the significance tests are conducted using mention-level McNemar's Chi-square test, with p-value = 0.01.

There are inconsistencies between the entity type names in different datasets, mainly due to different granularities. To remove this unnecessary noise, we manually merged some entity types. For example, we unify Gene and Protein into Gene/Protein as they are commonly used interchangeably; we merge "Simple Chemical" to "Chemical" and leave the problem of entity type granularity for future work. The information in Table 3 and 4 reflects the merged types.

A.1.3 News domain: CoNLL 2003 NER dataset

We use the CoNLL 2003 NER dataset ((Sang and De Meulder, 2003)) to evaluate the models in news domain. More information about the dataset can be found in Table 5. We use synthetic data from the dataset to simulate local training and global evaluation. Specifically, the CoNLL 2003 NER dataset is annotated with four entity types: location, person, organization, and miscellaneous entities. We randomly split the training set into four portions, each contains only one entity type respectively, with other types changed to "O". The models are trained on the four training portions and we test on the original test set with all entity types annotated.

A.1.4 Data split

For the news domain, we use the default train, dev, test portion of the CoNLL 2003 NER dataset. For the biomedicine domain, we follow the data split in Crichton et al. (2017) for both the training and the evaluation groups. All datasets are divided into three portions: train, dev, and test. We train the model on the training set of the training group and tune the hyper-parameters on the corresponding development set. Global evaluations are performed on the test set of the evaluation group.

A.2 Local Evaluation

For a sanity check, we evaluate the models on the training corpora and compare the results with state-of-the-art systems. In this setting, all the models are trained on the training set of the training corpora (without fine-tuning on global evaluation corpora) and evaluated on their test set. The results are shown in Table 6. **STM** is the single-task models we implemented, following the settings in Wang et al. (2018). The SOTA is achieved by Wang et al. (2018) with multi-task model, which is shown in the table as **MTM Wang et al. (2018)**. They trained their model on *BC2GC*, *BC4CHM*, *NCBI*, *JNLPBA*, and *BC5CDR*. **MTM (ours)** is the multi-task model we trained on our five training corpora and used as a baseline in the global evaluations. It has the same architecture as Wang et al. (2018).

As we can see, **MTM Wang et al. (2018)** achieves the best results on 3 out of 4 datasets. And our MTM achieves very similar results, showing it is a strong model on training corpora. Our proposed models do not perform very well when evaluated on the training corpora. But in the global evaluation setting, they perform much better compared to our strong MTM. This demonstrates the superiority of our proposed models on task adaptation.