# Appendix

## A Examples of Task Decomposition

In Figure 1, we show an example of task decomposition for standard NER.

| Text | Brush | Wellman | comments | on | beryllium | lawsuits | . |
|---|---|---|---|---|---|---|---|
| Tag | B-ORG | I-ORG | O | O | O | O | O |
| Seg | B | I | O | O | O | O | O |
| Ent | ORG | ORG | O | O | O | O | O |

Figure 1: An example of NER decomposition.

In Figure 2, we show another example of task decomposition for target sentiment, in addition to the one in the main text.

| Text | KC | Concepcion | Rogue | Magazine | Photos | Continue | to | Get | Praised | by | Fans | on | Twitter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tag | B-pos | B-pos | B-neu | E-neu | O | O | O | O | O | O | O | O | S-neu |
| Seg | B | E | B | E | O | O | O | O | O | O | O | O | S |
| Senti | pos | pos | neu | neu | O | O | O | O | O | O | O | O | neu |

Figure 2: An extra example of target sentiment decomposition.

## B Full Experimental Results on Target Sentiment

The complete results of our experiments on the target sentiment task are summarized in Tab. 1. Our LSTM-CRF-TI(g) model outperforms all the other competing models in Precision, Recall and the F1 score.

## C Experiments on Named Entity Recognition

**NER datasets** We evaluated our models on three NER datasets, the English, Dutch and Spanish parts of the 2002 and 2003 CoNLL shared tasks (Sang and F., 2002; Sang et al., 2003). We used the original division of training, validation and test sets. The task is defined over four different entity types: *PERSON*, *LOCATION*, *ORGANIZATION*, *MISC*. We used the BIOES tagging scheme during the training, and convert them back to original tagging scheme in testing as previous studies show that using this tagging scheme instead of BIO2 can help improve performance (Ratinov and Roth, 2009; Lample et al., 2016; Ma and Hovy, 2016; Liu et al., 2018). As a result, the segmentation module had 5 output labels, and the entity module had 4. The final decision task, consisted of the Cartesian product of the segmentation set (BIES) and the entity set, plus the "O" tag, resulting in 17 labels.

**Results on NER** We compared our models with the state-of-the-art systems on English[1], Dutch and Spanish. For Dutch and Spanish, we used cross-lingual embedding as a way to exploit lexical information. The results are shown in Tab. 2 and Tab. 3[2]. Our best-performing model outperform all the competing systems.

## D Additional Experiments on Knowledge Integration

We conducted additional experiments on knowledge integration in the same setting as in the main text to investigate the properties of the modules. Figure 3 shows the results for Dutch and Spanish NER datasets, while Figure 4 shows the results for the Subjective Polarity Disambiguation Datasets using the in-domain data.

[1] Liu et al.'s results are different since their implementation did not convert the predicted BIOES tags back to BIO2 during evaluation. For fair comparison, we only report the results of the standard evaluation.

[2] We thank reviewers for pointing out a paper (Agerri and Rigau, 2016) obtains the new state-of-the-art result on Dutch with comparable results on Spanish.

| System | Architecture | English | | | Spanish | | |
|---|---|---|---|---|---|---|---|
| | | Pre | Rec | F1 | Pre | Rec | F1 |
| Zhang, Zhang and Vo (2015) | Pipeline | 43.71 | 37.12 | 40.06 | 45.99 | 40.57 | 43.04 |
| | Joint | 44.62 | 35.84 | 39.67 | 46.67 | 39.99 | 43.02 |
| | Collapsed | 46.32 | 32.84 | 38.36 | 47.69 | 34.53 | 40.00 |
| Li and Lu (2017) | SS | 44.57 | 36.48 | 40.11 | 46.06 | 39.89 | 42.75 |
| | +embeddings | 47.30 | 40.36 | 43.55 | 47.14 | 41.48 | 44.13 |
| | +POS tags | 45.96 | 39.04 | 42.21 | 45.92 | 40.25 | 42.89 |
| | +semiMarkov | 44.49 | 37.93 | 40.94 | 44.12 | 40.34 | 42.14 |
| Base Line | LSTM-CRF | 53.29 | 46.90 | 49.89 | 51.17 | 46.71 | 48.84 |
| *This work* | LSTM-CRF-T | 54.21 | 48.77 | 51.34 | 51.77 | 47.37 | 49.47 |
| | LSTM-CRF-Ti | 54.58 | 49.01 | 51.64 | 52.14 | 47.56 | 49.74 |
| | LSTM-CRF-Ti(g) | **55.31** | **49.36** | **52.15** | **52.82** | **48.41** | **50.50** |

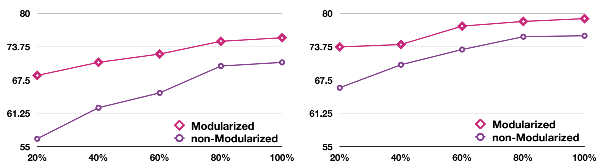
Table 1: Performance on the target sentiment task

| Model | English |
|---|---|
| LSTM-CRF (Lample et al., 2016) | 90.94 |
| LSTM-CNN-CRF (Ma and Hovy, 2016) | 91.21 |
| LM-LSTM-CRF (Liu et al., 2018) | 91.06 |
| LSTM-CRF-T | 90.8 |
| LSTM-CRF-TI | 91.16 |
| LSTM-CRF-TI(g) | **91.68** |

Table 2: Comparing our models with several state-of-the-art systems on the CoNLL 2003 English NER dataset.

| Model | Dutch | Spanish |
|---|---|---|
| Carreras et al. (2002) | 77.05 | 81.39 |
| Nothman et al. (2013) | 78.60 | N/A |
| dos Santos and Guimarães (2015) | N/A | 82.21 |
| Gillick et al. (2015) | 82.84 | 82.95 |
| Lample et al. (2016) | 81.74 | 85.75 |
| LSTM-CRF-T | 83.91 | 84.89 |
| LSTM-CRF-TI | 84.12 | 85.28 |
| LSTM-CRF-TI(g) | **84.51** | **85.92** |

Table 3: Comparing our models with recent results on the 2002 CoNLL Dutch and Spanish NER datasets.




(a) Dutch NER (b) Spanish NER

Figure 3: Experimental results on modular knowledge integration on the Dutch and Spanish NER datasets.

## E Convergence Analysis

The proposed twofold modular infusion model (with guided gating as an option) breaks the complex learning problem into several sub-problems and then integrate them using joint training. The process defined by this formulation has more parameters and requires learning multiple objectives jointly. Our convergence analysis intends to evaluate whether the added complexity leads to a harder learning problem (i.e., slower to converge) or whether the tasks constrain each other and as a result can be efficiently learned.

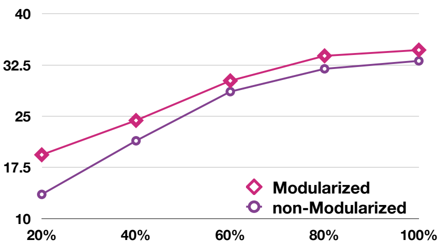



Figure 4: Experimental results on modular knowledge integration on the Subjective Polarity Disambiguation Datasets.

We compare between our LSTM-CRF-TI(g) model and recent published top models on the English NER dataset in Figure 5 and on the subjective polarity disambiguation datasets in Figure 6. The curve compares convergence speed in terms of learning epochs. Our LSTM-CRF-TI(g) model has a much faster convergence rate compared to the other models.

## References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*.

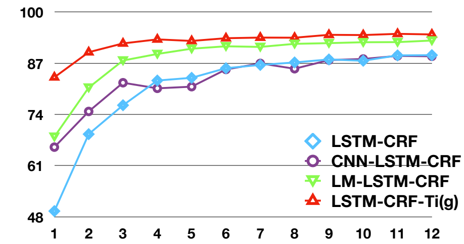



Figure 5: Comparing convergence over the development set on the English NER dataset. The x-axis is number of epochs and the y-axis is the F1-score.

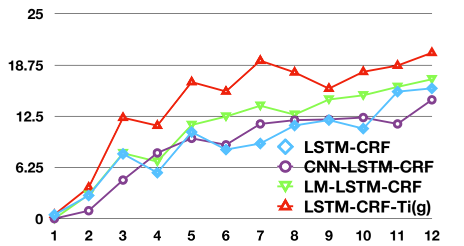



Figure 6: Comparing convergence over the development set on the subjective polarity disambiguation datasets. The x-axis is number of epochs and the y-axis is the F1-score.

Gillick, Brunk, Vinyals, and Subramanya. 2015. Multilingual Language Processing From Bytes. *ArXiv*.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*, pages 1–10.

Liyuan Liu, Jingbo Shang, Frank F. Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *AAAI*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*, pages 1064–1074, Berlin, Germany.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, 194:151–175.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Tjong Kim Sang and Erik F. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*.

Tjong Kim Sang, Erik F., and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *CoNLL*.

dos Santos and Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proc. of the Annual Meeting of the Association Computational Linguistics (ACL)*.