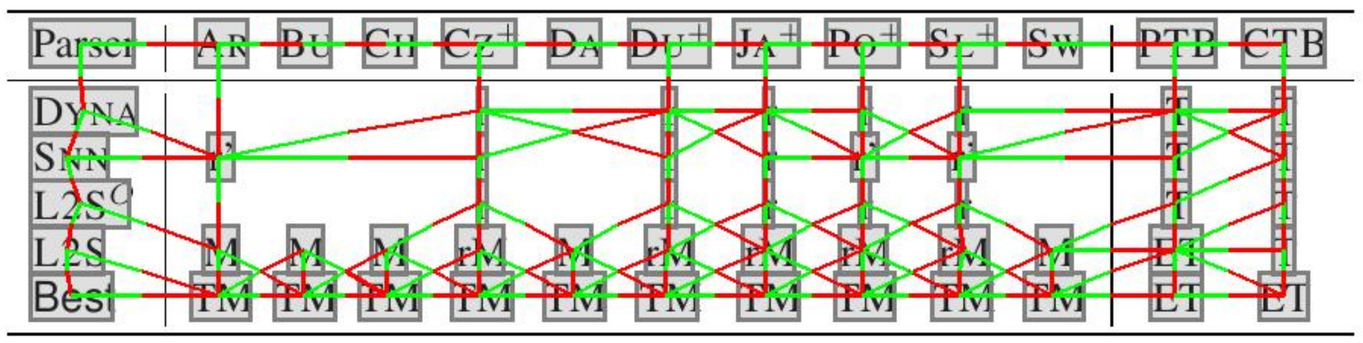

Table 3: T: tuned hyper-parameters, M: use morphological features, E: use word embedding or word clustering, r: language structure assumption that may degrades the performance (the nature of transition-based model), r': strong language structure assumption (only one head) that severely degrades the performance. Accuracy on PTB, CTB and CoNLL-X. Best: best results from the shared

	Training					Holdout		Test	
	Sents	Toks	Labels	Features	Unique Fts	Sents	Toks	Sents	Toks
POS	38k	912k	45	13,685k	629k	5.5k	132k	5.5k	130k
NER	15k	205k	7	8,592k	347k	3.5k	52k	3.6k	47k

Table 4: Basic statistics about the data sets used for part of speech (POS) tagging and named entity recognition (NER).

Part of speech tagging for English is based on the Penn Treebank tagset that includes 45 discrete labels. The accuracy reported represents number of tokens tagged correctly. This is a *pure* sequence labeling task. Named entity recognition for English is based on the CoNLL 2003 dataset that includes four entity types: Person, Organization, Location and Miscellaneous. We use the standard evaluation metric to report performance as macro-averaged F-measure. In order to cast this *chunking* task as a sequence labeling task, we use the standard Begin-In-Out (BIO) encoding, though some results suggest other encodings may be preferable [37] (we tried BILOU and our accuracies decreased). The example sentence from Figure 7 in this encoding is:

LOC ORG PER
Germany 's rep to the European Union 's committee Werner Zwingmann said ...
B-LOC O O O O B-ORG I-ORG O O B-PER I-PER O

Dependency parser is test on the English Penn Treebank (PTB) and the CoNLL-X datasets for 9 other languages, including Arabic, Bulgarian, Chinese, Danish, Dutch, Japanese, Portuguese, Slovene and Swedish. For PTB, we convert the constituency trees to dependencies by the Stanford parser 3.3.0. We follow the standard split: sections 2 to 21 for training, section 22 for development and section 23 for testing. The POS tags in the evaluation data is assigned by the Stanford POS tagger, which has an accuracy of 97.2% on the PTB test set. For CoNLL-X, we use the given train/test splits and reserve the last 10% of training data for development if needed. The gold POS tags given in the CoNLL-X datasets are used. The CTB is prepared following the instructions in [8].

D.2 Methodology

Comparing different systems is challenging because one wishes to hold constant as many variables as possible. In particular, we want to control for both **features** and **hyperparameters**. In general, if a methodological decision cannot be made "fairly," we made it in favor of competing approaches.

To control for **features**, for the two sequential tagging tasks (POS and NER), we use the built-in *feature template* approach of CRF++ (duplicated in CRF SGD) to generate features. The other ap-

POS	NNP NNP , CD NNS JJ , MD VB DT NN IN DT JJ NN Pierre Vinken , 61 years old , will join the board as a nonexecutive director ...
NER	LOC ORG PER Germany 's rep to the European Union 's committee Werner Zwingmann said ...

Figure 7: Example inputs (below, black) and desired outputs (above, blue) for part of speech tagging task, named entity recognition task, and entity-relation recognition task.