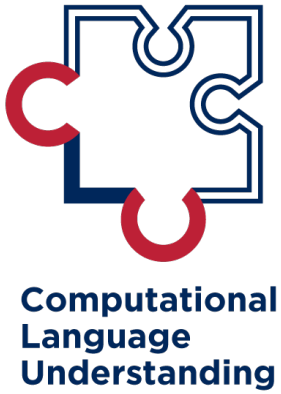
ELLEN: *Extremely Lightly Supervised Learning For Efficient Named Entity Recognition*

Haris Riaz, Razvan-Gabriel Dumitru, Mihai Surdeanu



Authors

Motivation

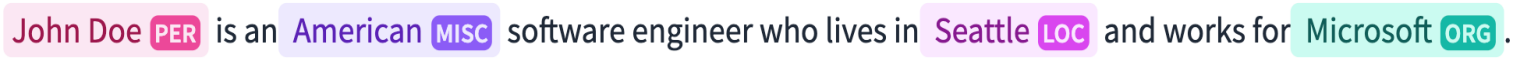
- Named Entity Recognition (NER)
  - With minimal supervision, e.g., 10 examples per class. This is motivated by the fact that in many specialized domains e.g., intelligence gathering, pandemic surveillance etc., there is an absence of labeled data.
  - Using models that can be deployed at scale, e.g., small encoders rather than LLMs.

NER Background

Input:

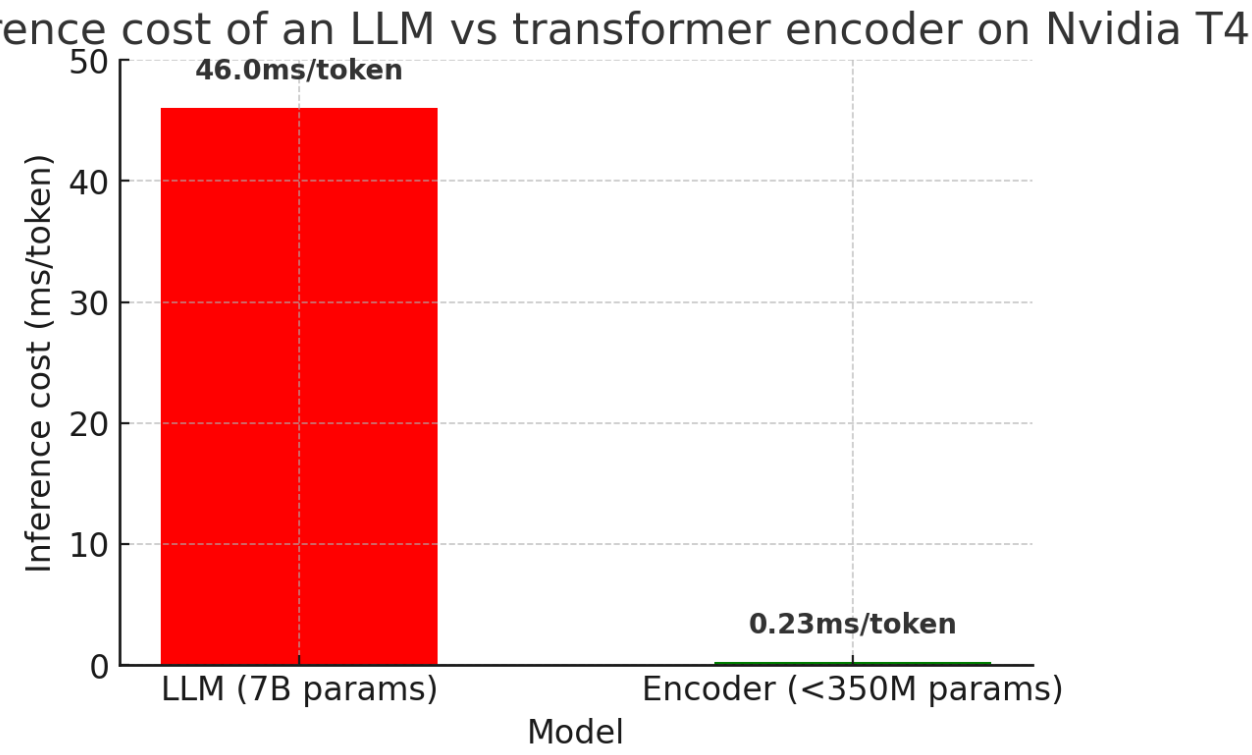
"John Doe is an American software engineer who lives in Seattle and works for Microsoft."

Expected Output:



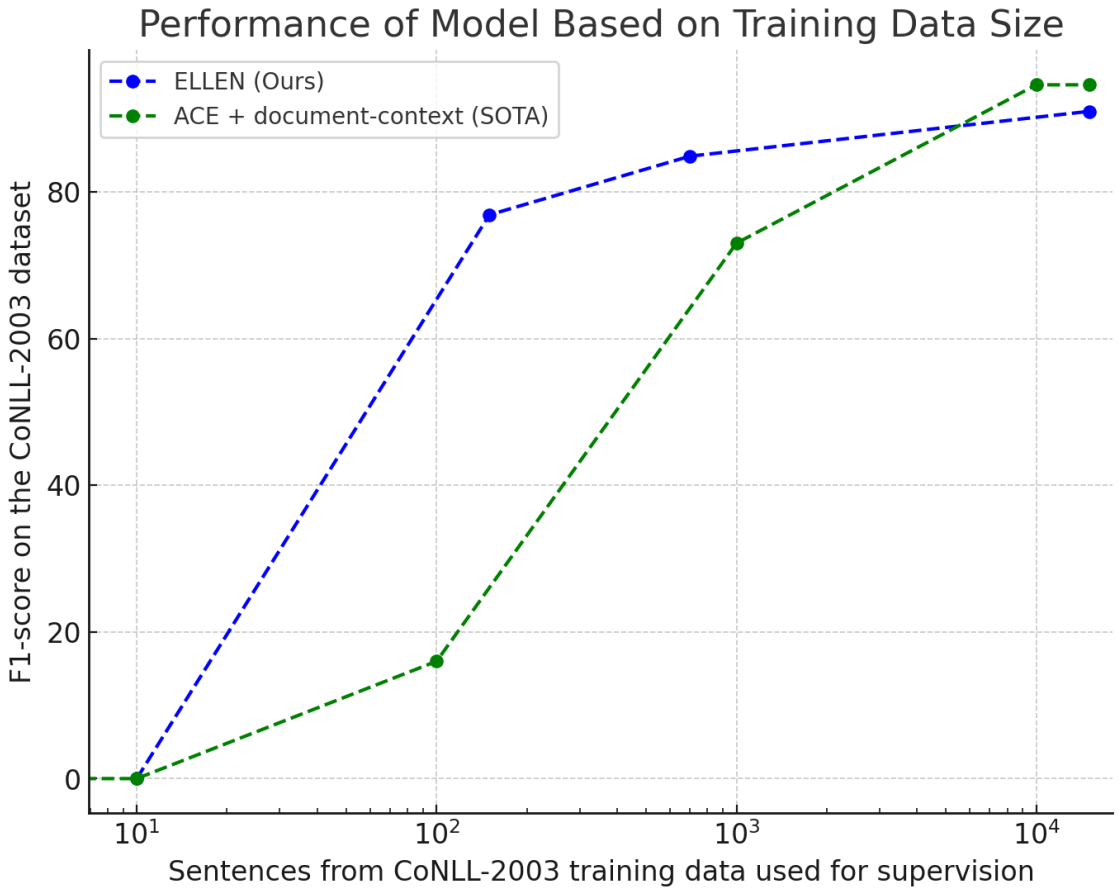
Goal

- We want high NER performance using a lightweight transformer encoder.



Goal

- We want high NER performance with minimal supervision.



NER – already solved?

- LLMs and fully supervised transformer-based methods achieve very high performance >90% on benchmark datasets (CoNLL-2003, OntoNotes, ACE2005 etc.).

English CoNLL2003 (FULL)			
Model	**Precision**	**Recall**	**F1**
Baselines (Supervised Model)			
BERT-Tagger (Devlin et al., 2018)	-	-	92.8
BERT-MRC (Li et al., 2019a)	92.33	94.61	93.04
GNN-SL (Wang et al., 2022)	93.02	93.40	93.2
ACE+document-context (Wang et al., 2020)	-	-	**94.6 (SOTA)**

Many of these methods (including LLMs) are trained on large annotated datasets

- The setting for many current NER methods is not realistic – in many domains data with gold labels is not available, or the amount of labeled data is extremely scarce.

- Even current SOTA semi-supervised NER methods use an **impractically high degree of supervision** (5% of gold data).

Contributions

We redefine the semi-supervised NER task – create a new setting of **Extremely Light Supervision** (<1% of the data)

- We propose a setting where the only source of supervision comes from a lexicon of only 10 example named entities per class provided by a domain expert.
- The domain expert **does not have access to any of the actual labels from the dataset.**
- The domain expert can create the lexicon **in less than 30 minutes.**

Contributions

Example lexicon chosen by the domain expert

Category	Entities
ORG	Reuters, PUK, NATO, Honda, Ajax Amsterdam, Motorola, PSV Eindhoven, PKK, Hansa Rostock, Commonwealth
LOC	Germany, Australia, Britain, Spain, Italy, LONDON, Russia, China, Japan, NEW YORK
MISC	Dutch, British, French, Russian, German, Iraqi, Israeli, English, Australian, American
PER	Clinton, Yeltsin, Arafat, Lebed, Wasim Akram, Waqar Younis, Mushtaq Ahmed, Netanyahu, Williams, Rubin

Contributions

1. We need more data.

2. We need to use the new data carefully, e.g., sentences may contain labeled entities as well as unlabeled ones.

3. Solution: **Combine deep learning with insights from linguistics.**

Contributions

Insight # 1: Masked Language Modeling Heuristic

- We can use a **Masked Language Model** (MLM) e.g., BERT, RoBERTA etc., to obtain "free supervision" based on just the lexicon provided by the domain expert.

- We present a novel technique to use a pre-trained Masked Language Model as a **fully unsupervised NER algorithm**, which on its own achieves an F1 score over **56%.**

Contributions

Insight # 1: Masked Language Modeling Heuristic

- We can detect spans of possible Named Entities, fully unsupervised, using a very simple linguistic pattern based on Part of Speech (POS) tags:

$$(NNP \mid NNPS) + (IN(NNP \mid NNPS)+)?$$

Contributions

Insight # 1: Masked Language Modeling Heuristic

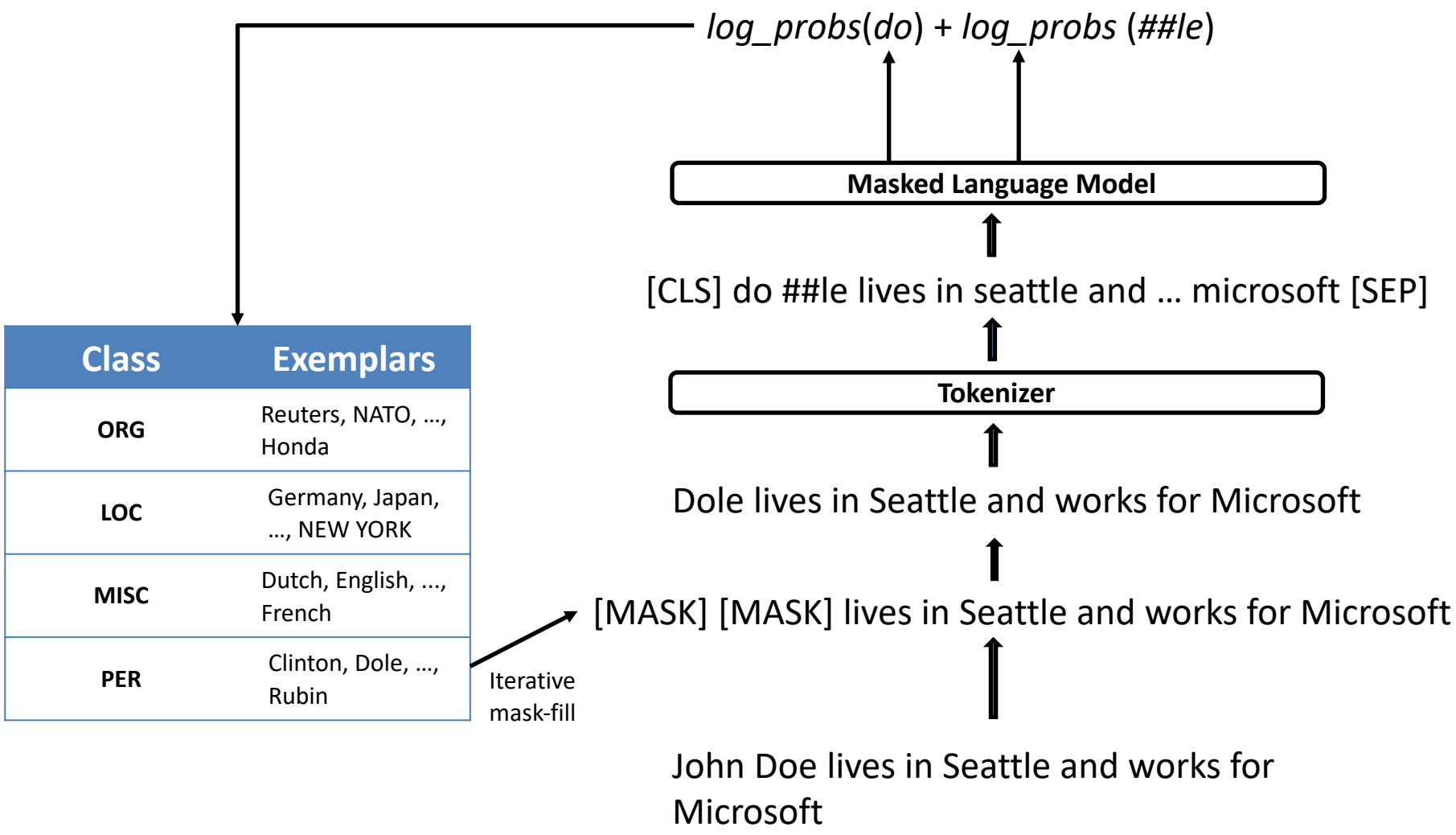
- This simple linguistic pattern can detect possible named entity boundaries/spans with high precision:

Precision	Recall	F1 Score
85.16%	90.96%	87.96%

Contributions

Insight # 1: Masked Language Modeling Heuristic

- Once candidate named entity spans in the unlabeled corpus are identified, we can estimate their labels with a simple masking heuristic:



Class	Exemplars
ORG	Reuters, NATO, …, Honda
LOC	Germany, Japan, …, NEW YORK
MISC	Dutch, English, ..., French
PER	Clinton, Dole, …, Rubin

Contributions

Insight # 1: Masked Language Modeling Heuristic

MLM heuristic on its own achieves 56% F1 on CoNLL-03 dev:

Entity Type	Precision	Recall	F1
Overall	61.78%	51.90%	56.41%
LOC	69.72%	41.53%	52.05%
MISC	45.18%	55.15%	49.67%
ORG	44.85%	40.88%	42.77%
PER	85.07%	65.02%	73.71%

Contributions

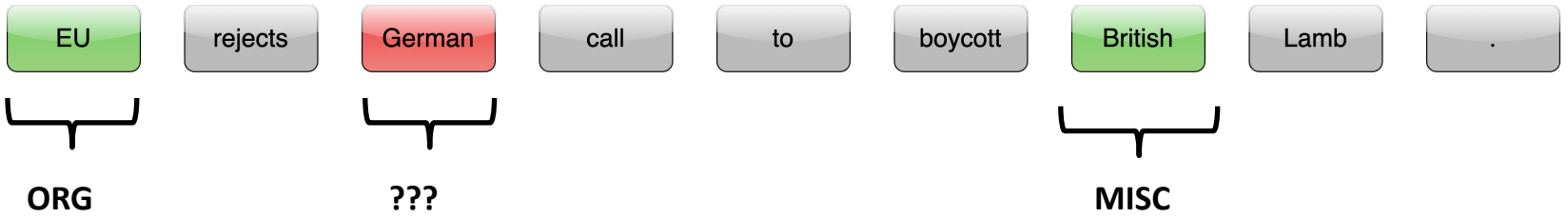
Insight # 2: We must use pseudo-labeled data very carefully

- What should we do when some tokens are labeled, and some are not?

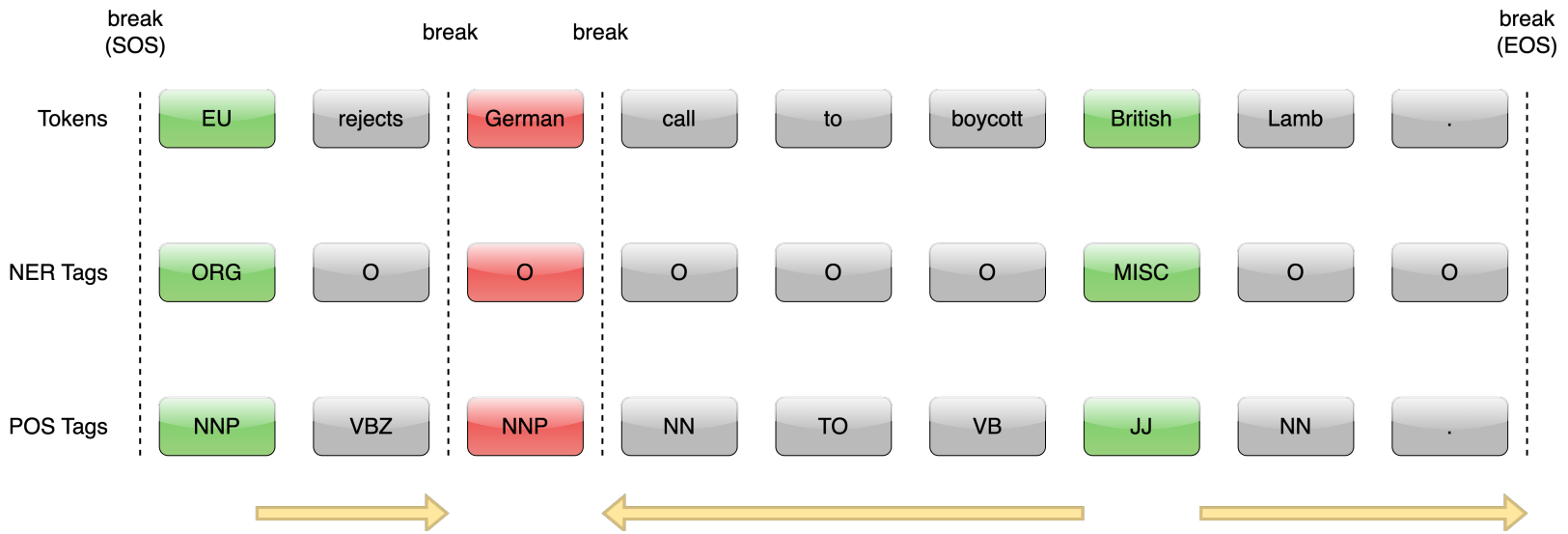
- Proposal: **Dynamic window filtering** – a simple, run-time efficient and linguistically inspired algorithm for solving the "unlabeled entity problem" i.e., eliminating false negative entities from the sparsely annotated training data, again using POS tags.

Contributions

Insight # 2: We must use pseudo-labeled data carefully



Contributions



New filtered sentence fragments:

"EU rejects"

"call to boycott British Lamb"

Contributions

Insight # 3: Combine commonsense linguistics, statistics and active learning for automatically correcting pseudo-labels

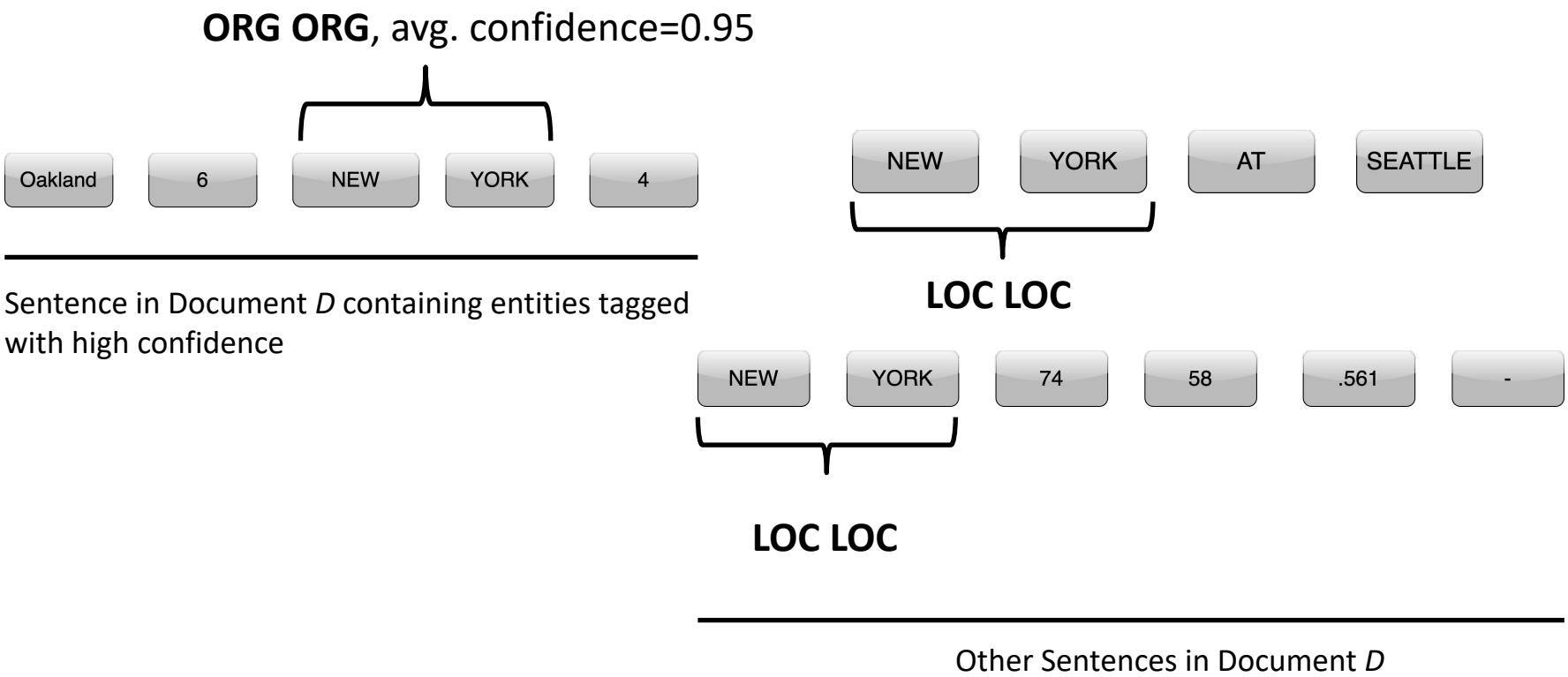
One Sense Per Discourse

William A. Gale
Kenneth W. Church
David Yarowsky

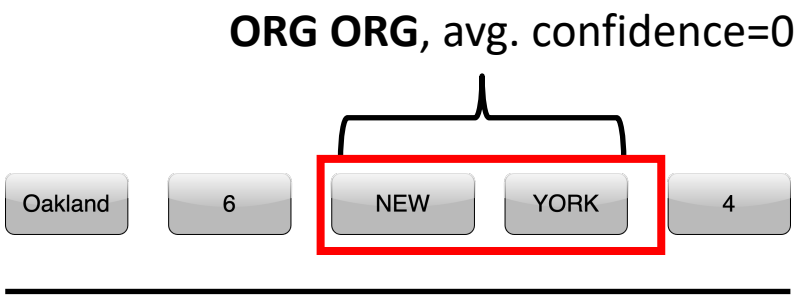
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974-0636

It is well-known that there are polysemous words like *sentence* whose "meaning" or "sense" depends on the context of use. We have recently reported on two new word-sense disambiguation systems, one trained on bilingual material (the Canadian Hansards) and the other trained on monolingual material (Roget's Thesaurus and Grolier's Encyclopedia). As this work was nearing completion, we observed a very strong discourse effect. That is, if a polysemous word such as *sentence* appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense. This paper describes an experiment which confirmed this hypothesis and found that the tendency to share sense in the same discourse is extremely strong (98%). This result can be used as an additional source of constraint for improving the performance of the word-sense disambiguation algorithm. In addition, it could also be used to help evaluate disambiguation algorithms that did not make use of the discourse constraint.

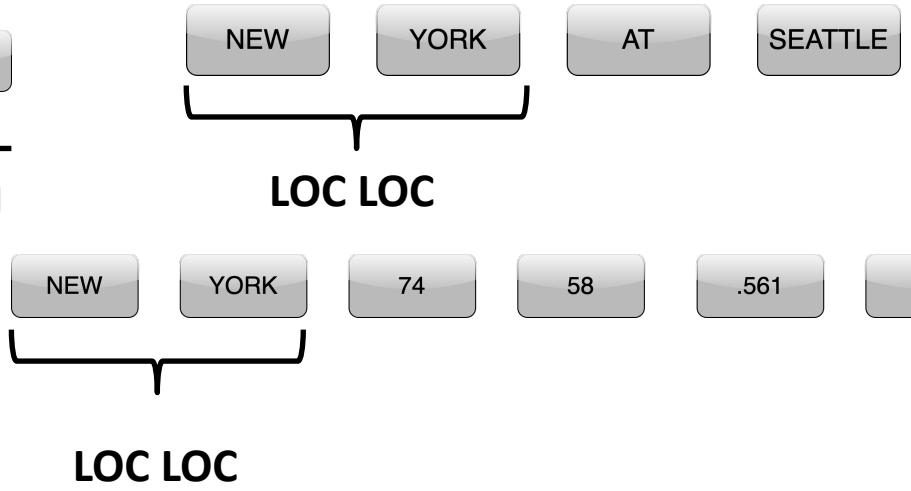
Linguistic Heuristics



Linguistic Heuristics

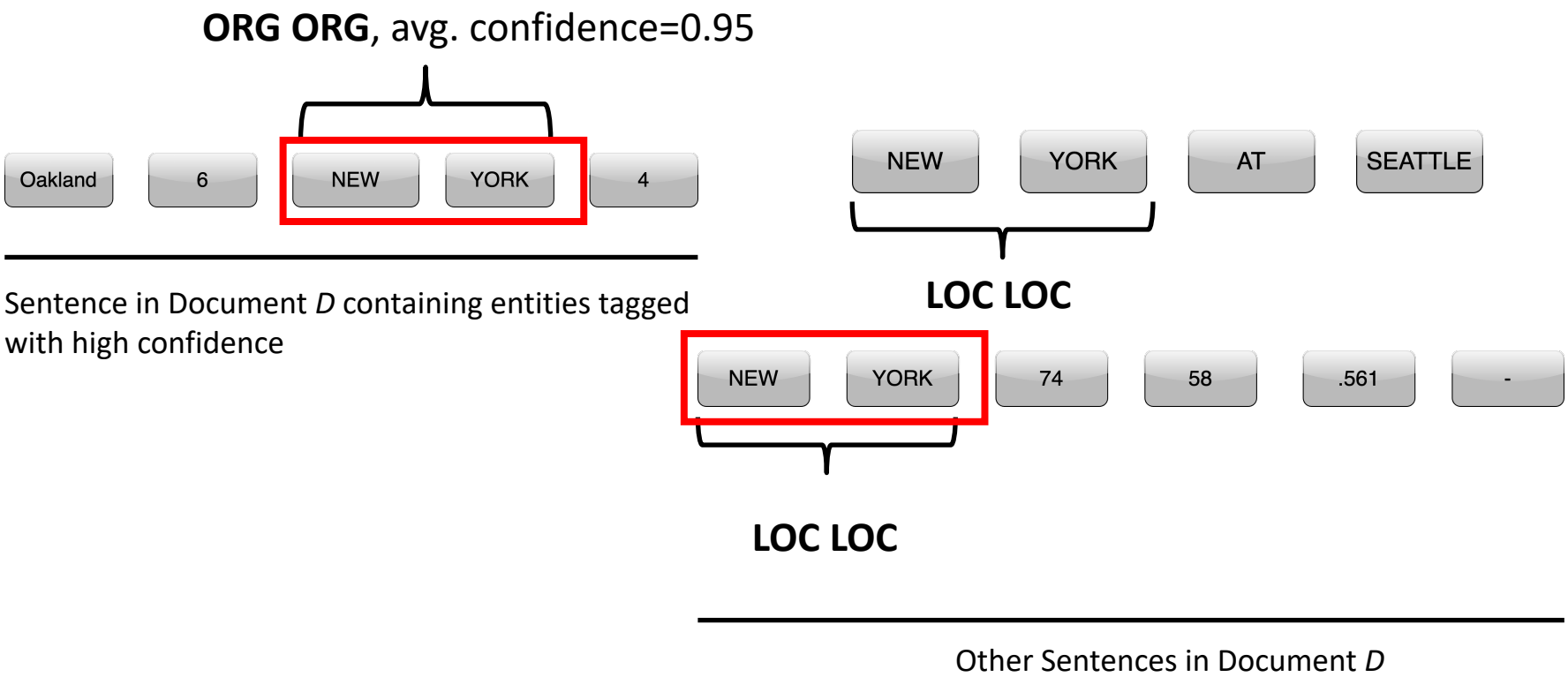


Sentence in Document *D* containing entities tagged with high confidence

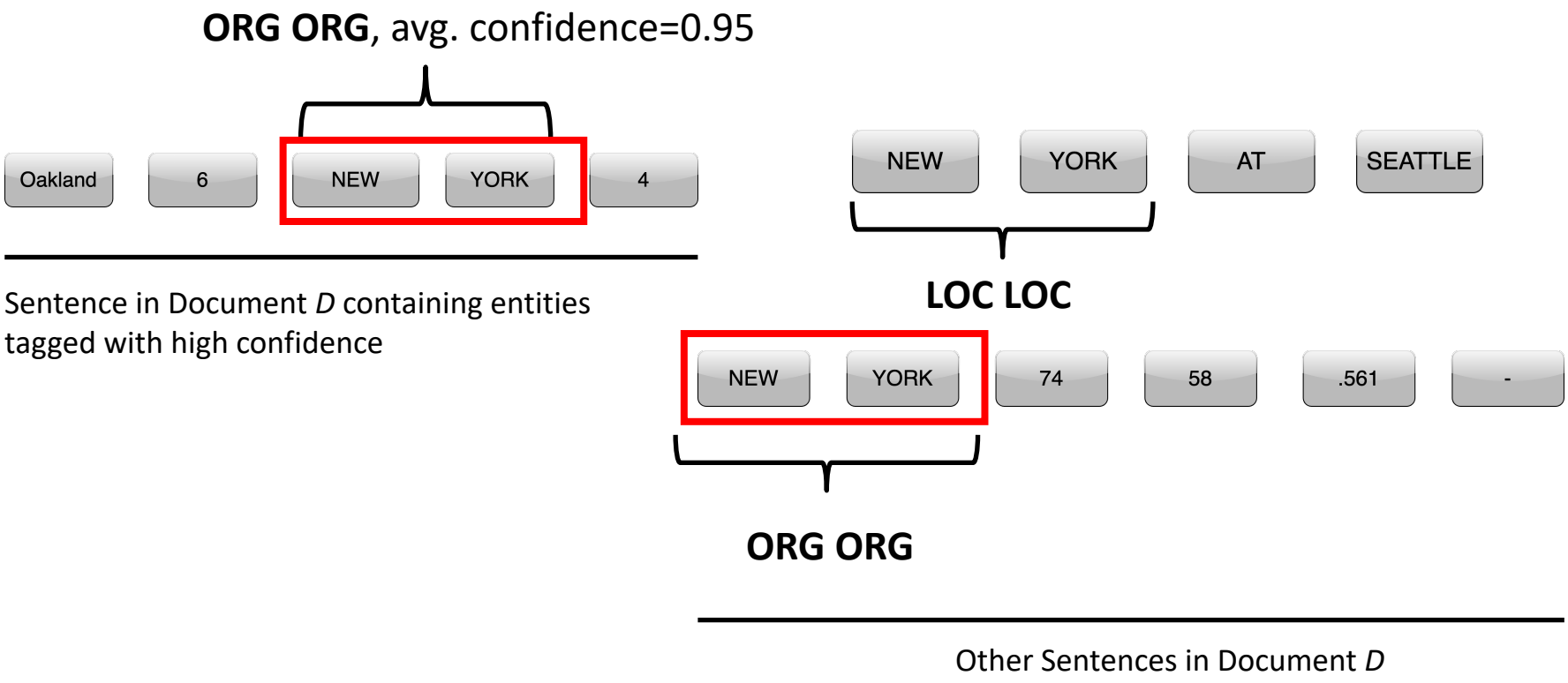


Other Sentences in Document *D*

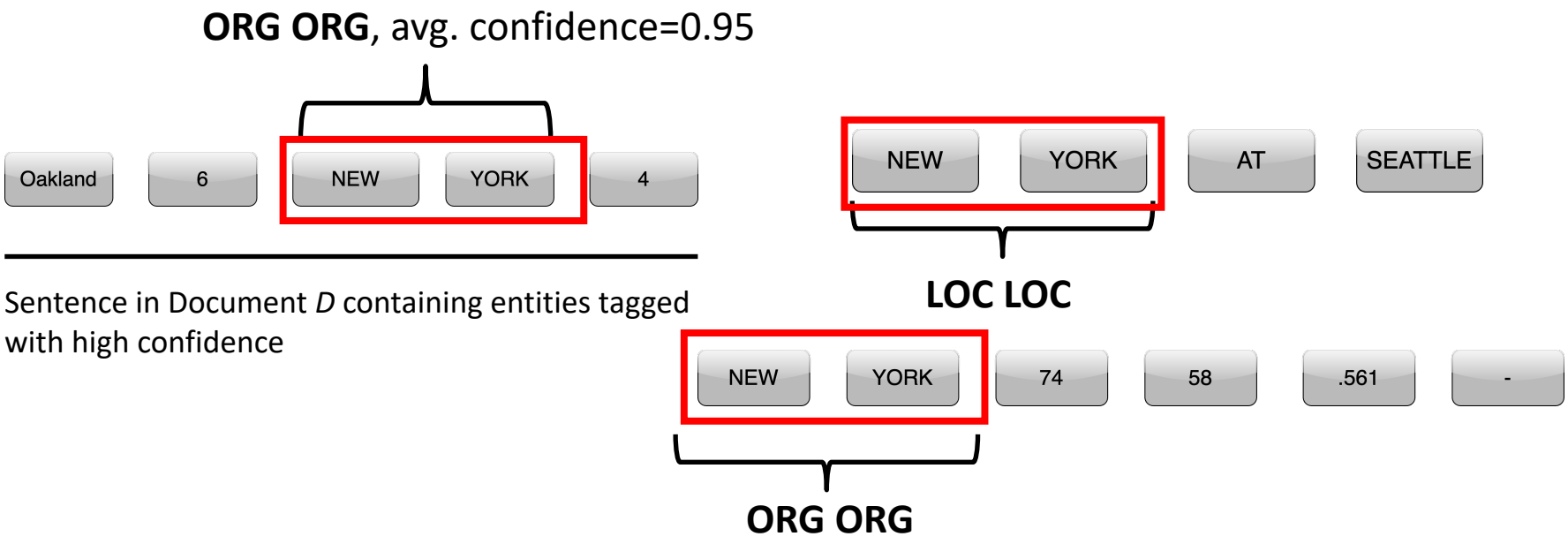
Linguistic Heuristics



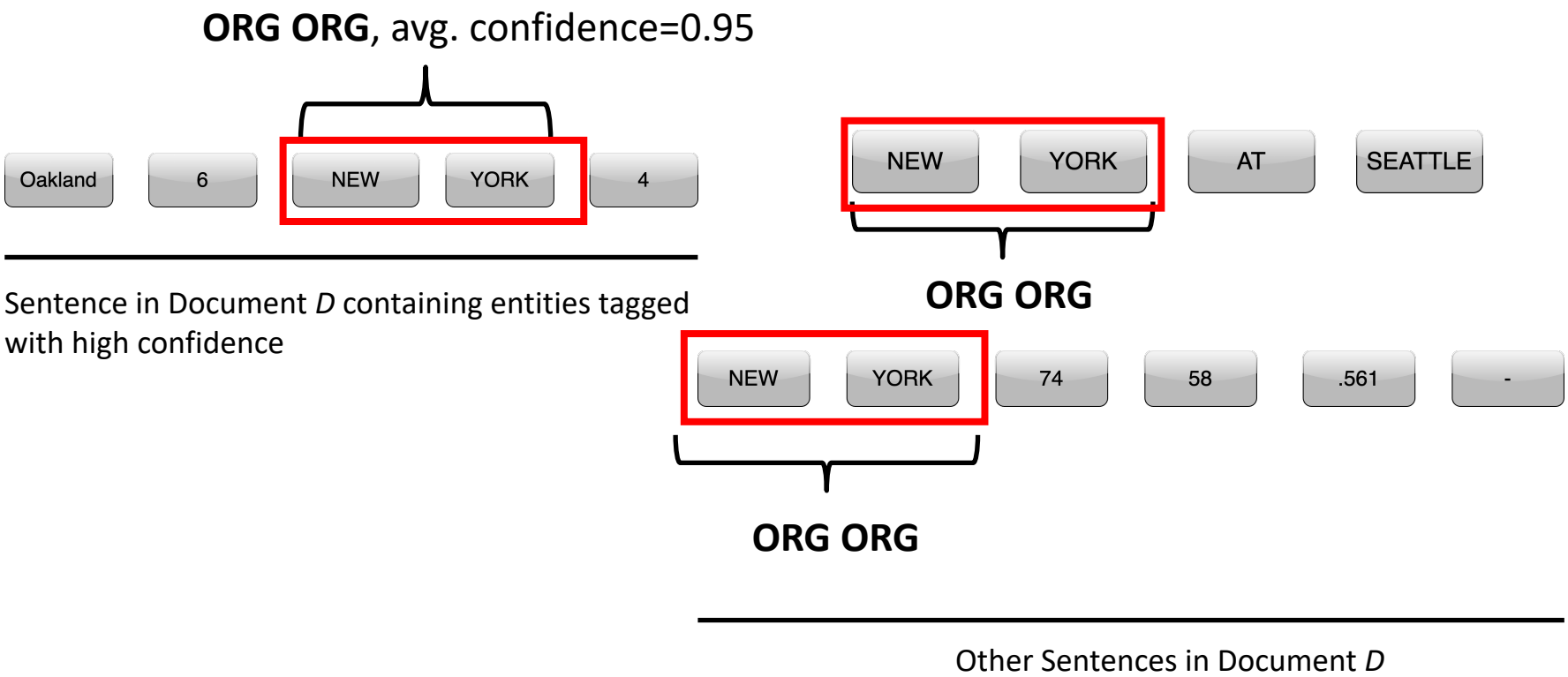
Linguistic Heuristics



Linguistic Heuristics



Linguistic Heuristics



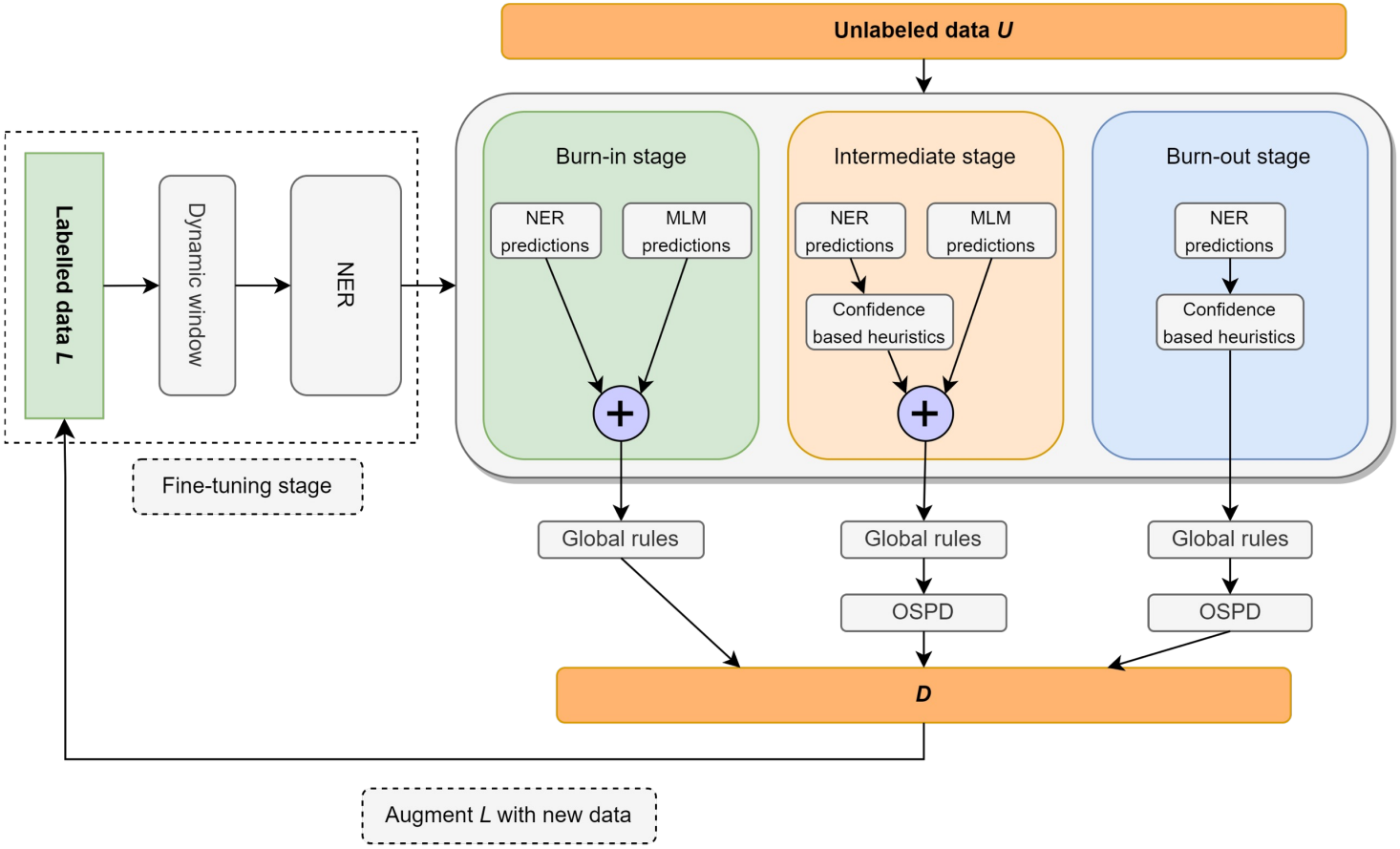
Contributions

Insight # 4: Self-training

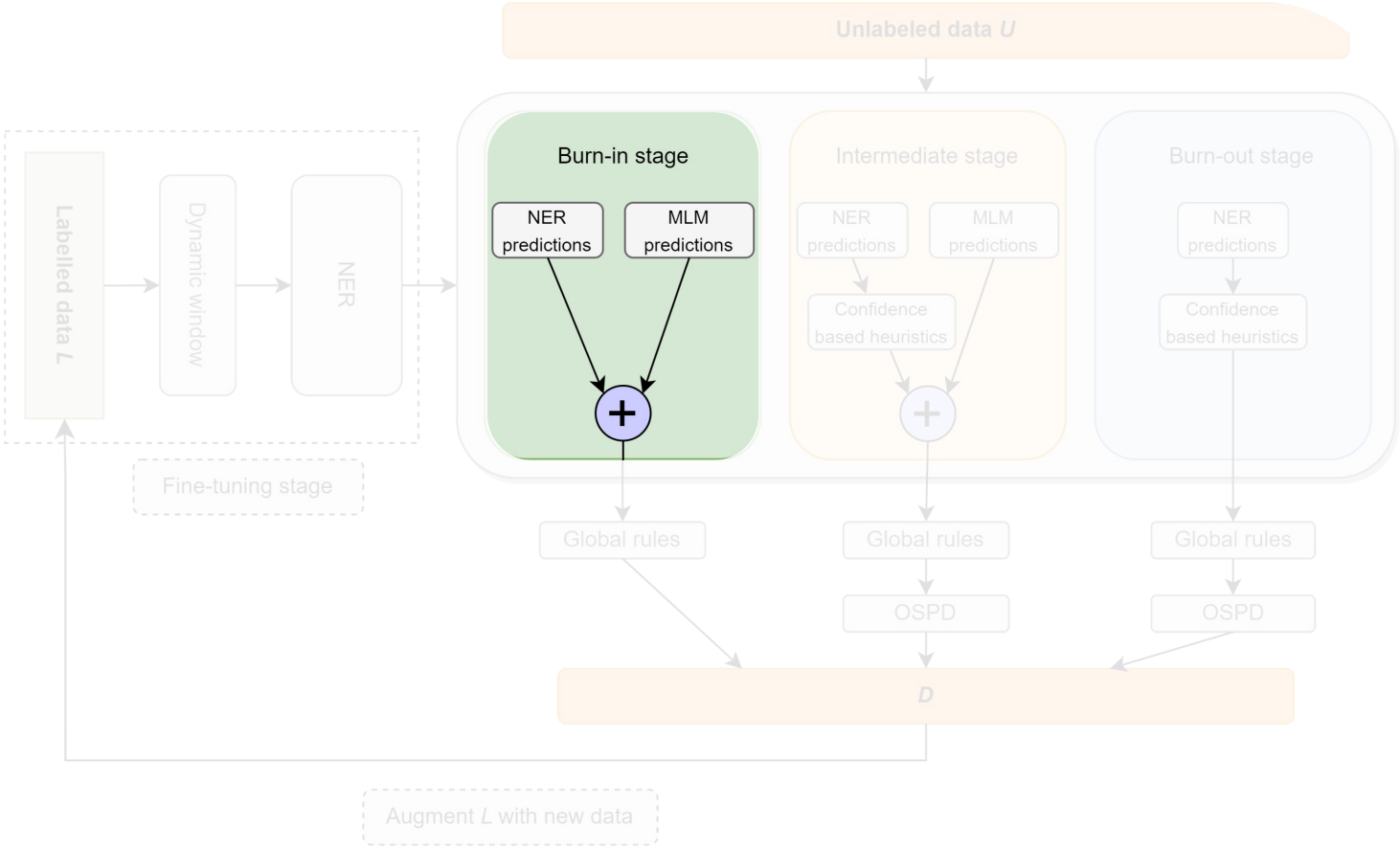
Algorithm 1 A simple NER self-training algorithm

1: **Given:**
2: L - a small set of labeled training data
3: U - unlabeled data
4: **for** k iterations **do**
5: **Step 1:** Train a NER C_k based on L
6: **Step 2:** Extract new data D based on C_k
7: **Step 3:** Add D to L
8: **end for**

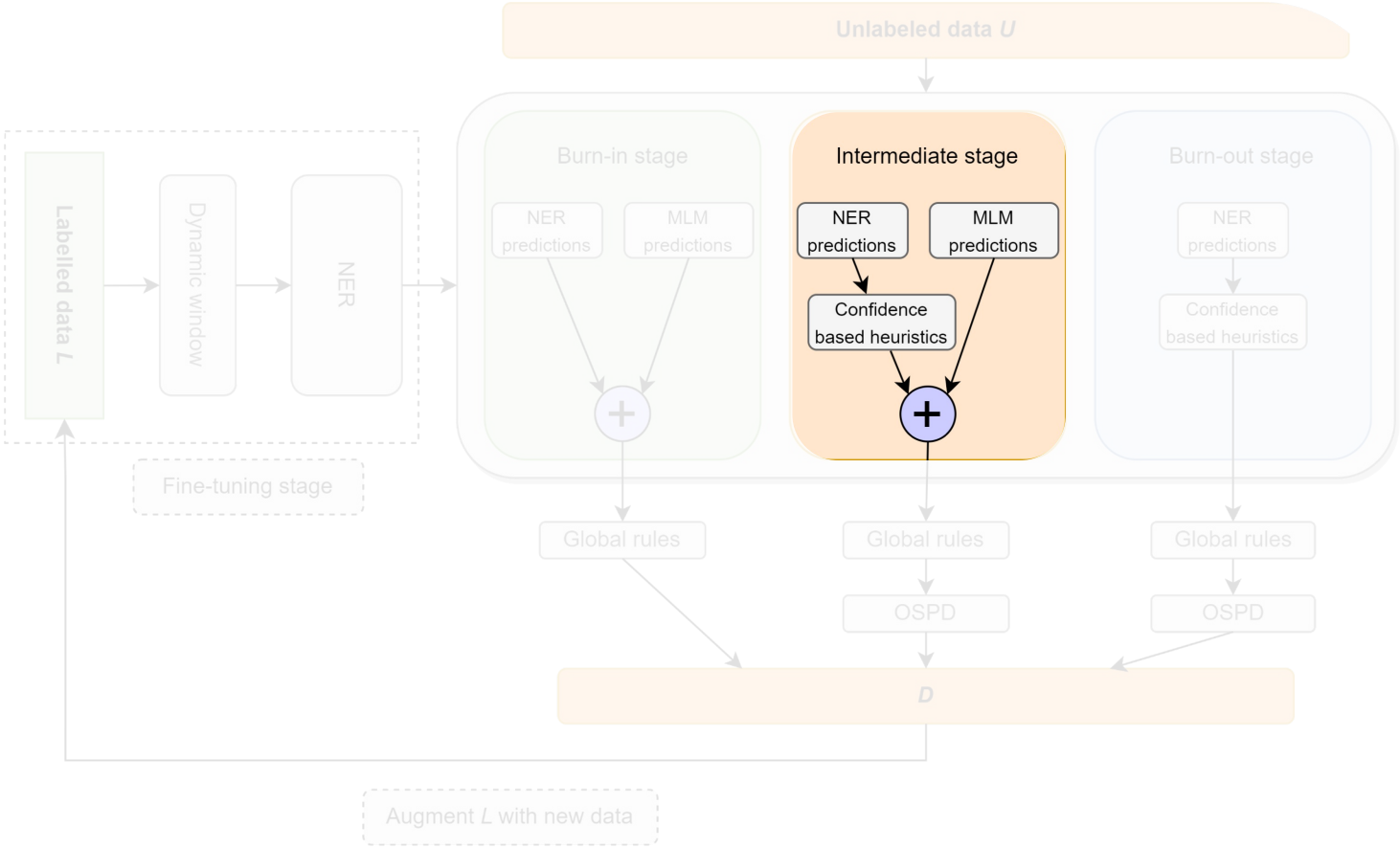
Insight # 5: Blend linguistics and deep learning in a simple, modular framework



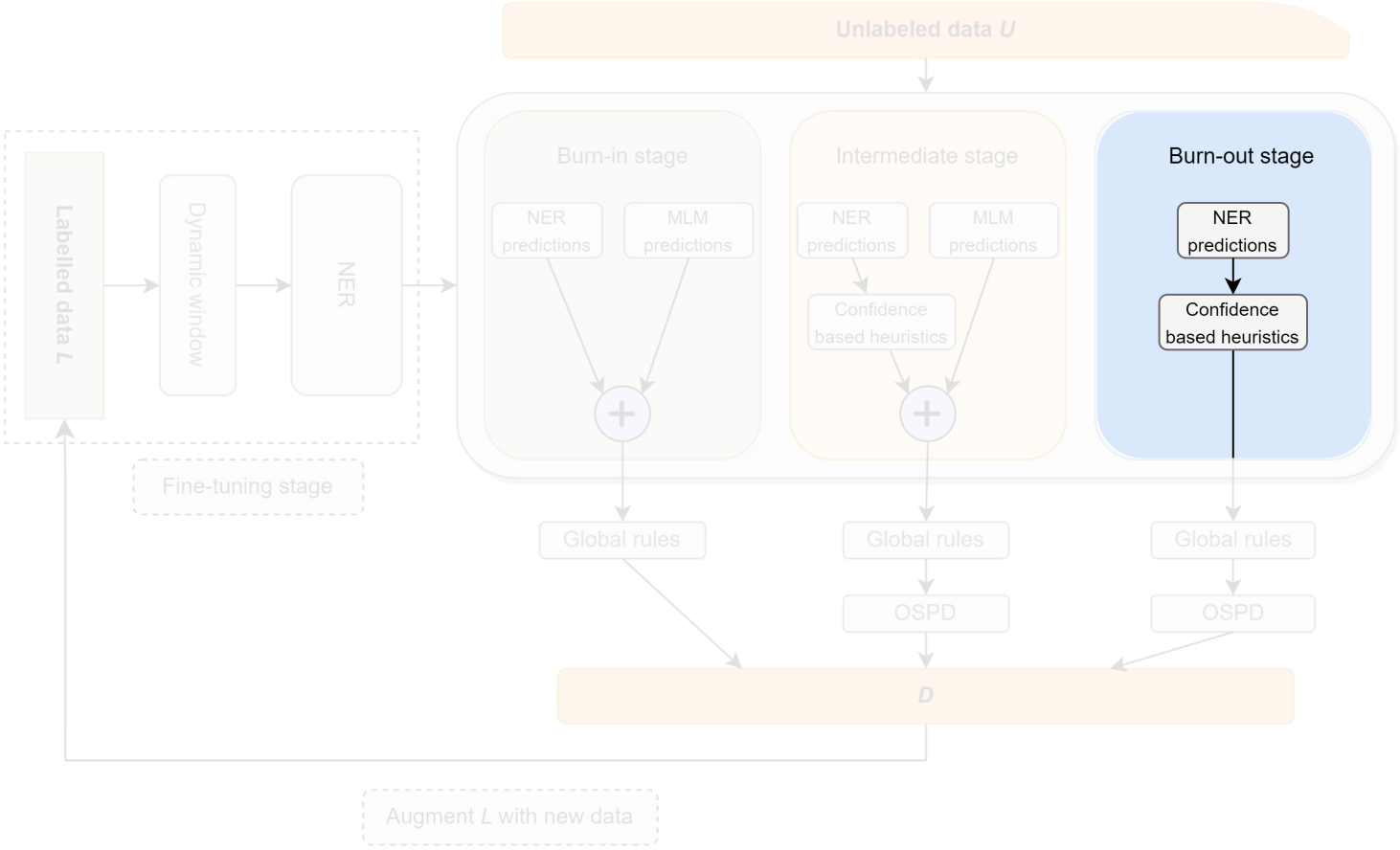
Insight # 5: Blend linguistics and deep learning in a simple, fully modular framework



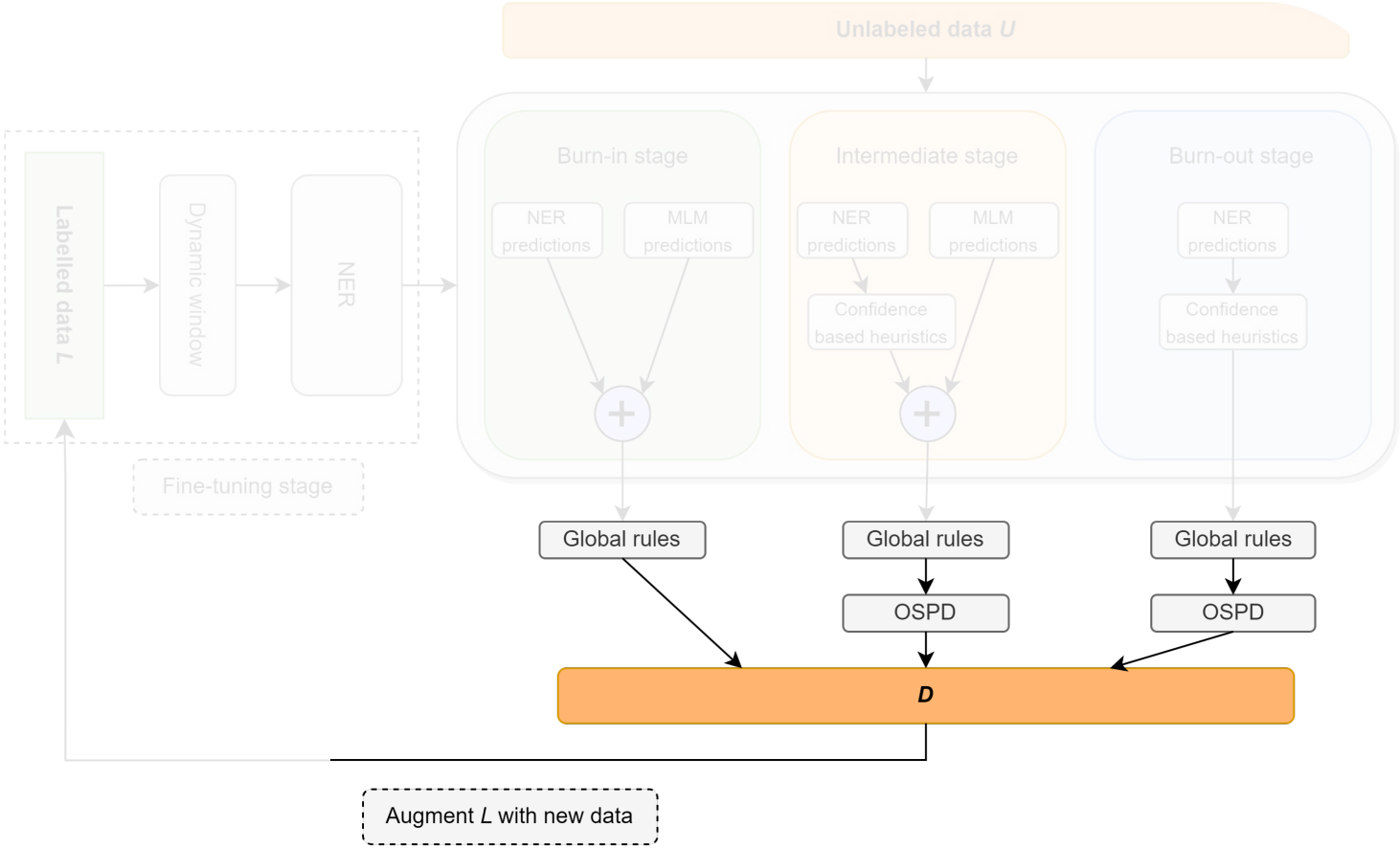
Insight # 5: Blend linguistics and deep learning in a simple, fully modular framework



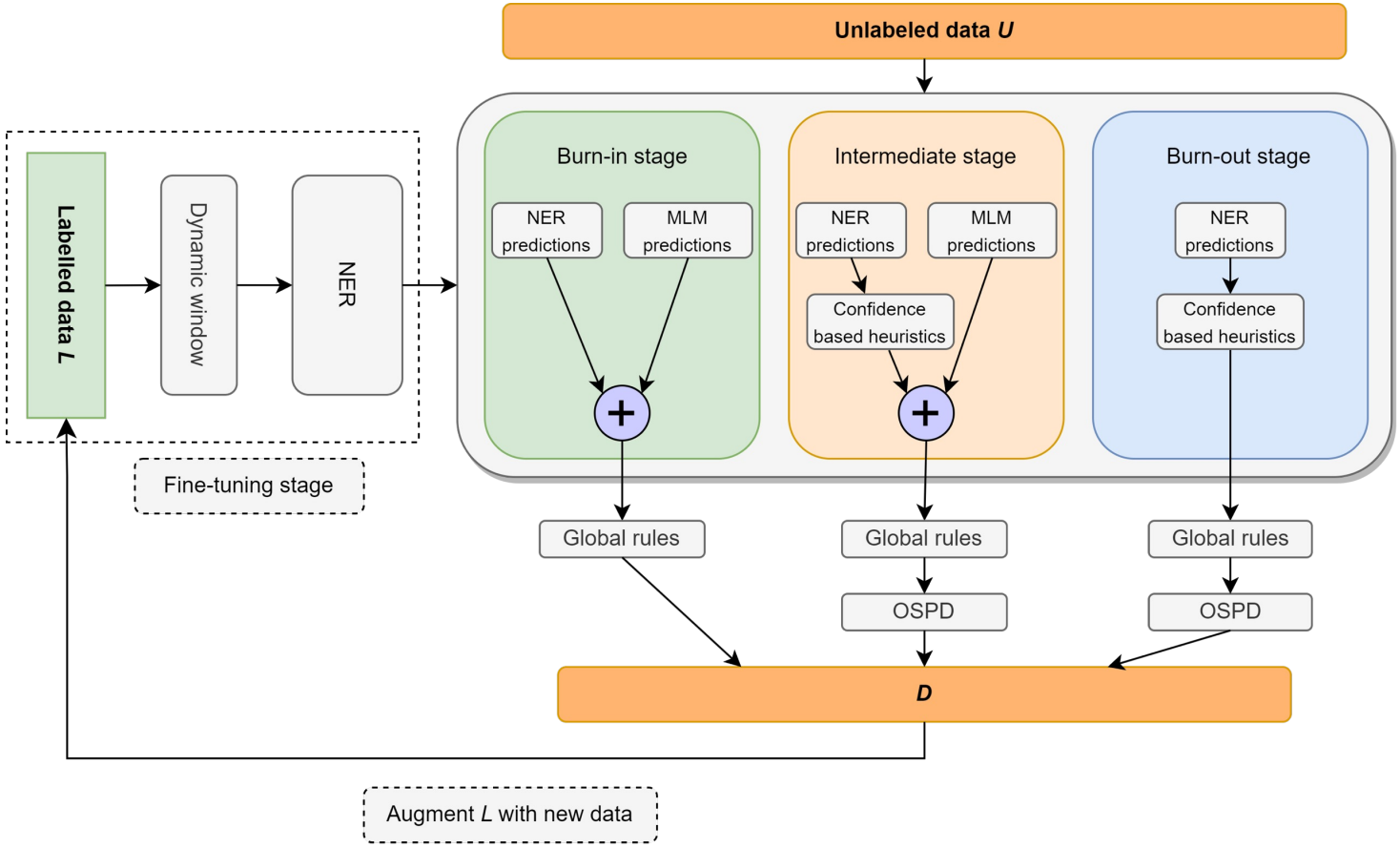
Insight # 5: Blend linguistics and deep learning in a simple, fully modular framework



Insight # 5: Blend linguistics and deep learning in a simple, fully modular framework



Insight # 5: Blend linguistics and deep learning in a simple, fully modular framework



Evaluation

CoNLL-2003, domain: news wire, classes: **O, PER, ORG, LOC, MISC.**

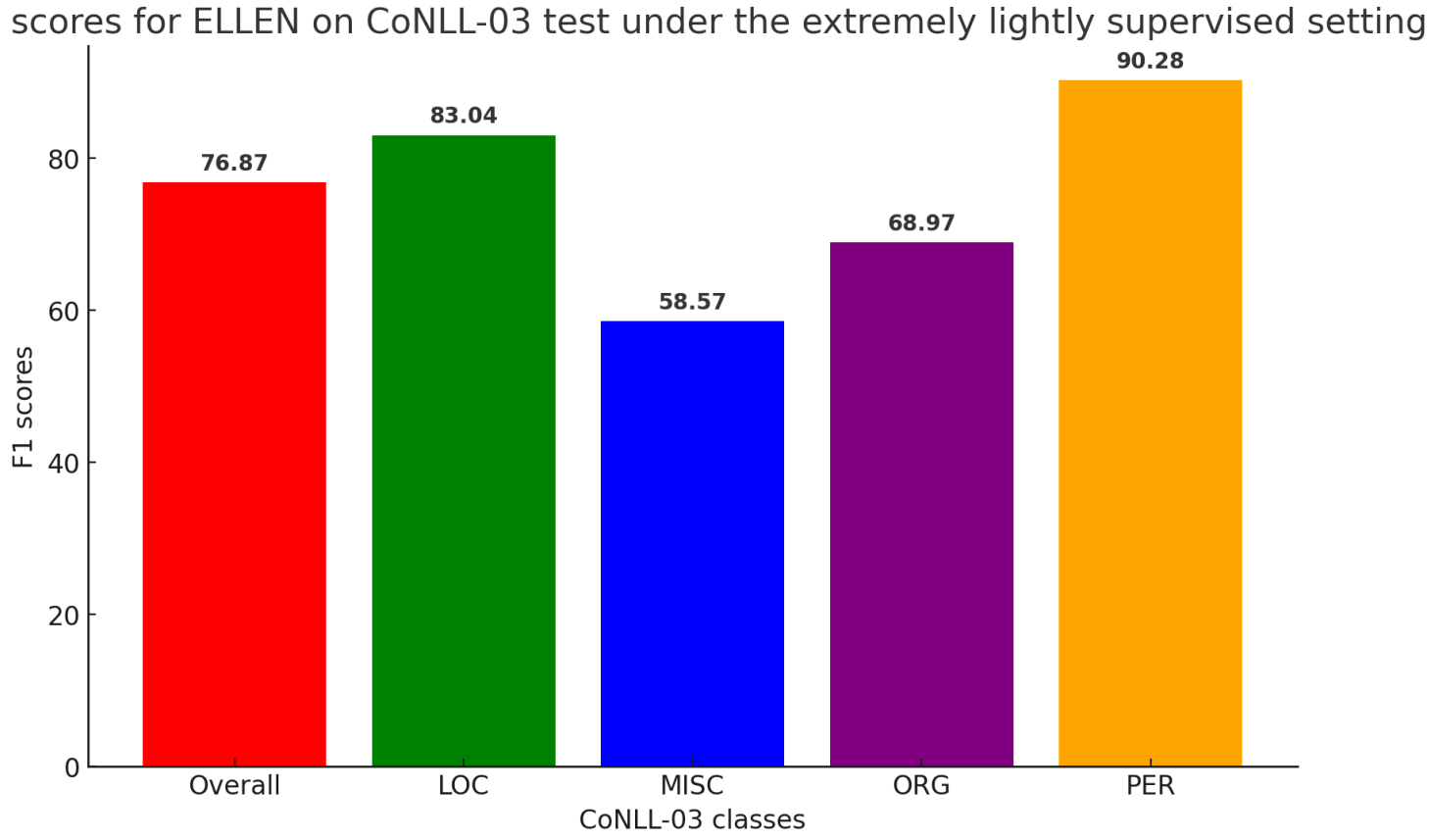
Settings: 1% supervision, 5% supervision and full supervision.

WNUT-17, domain: user-generated text, classes:

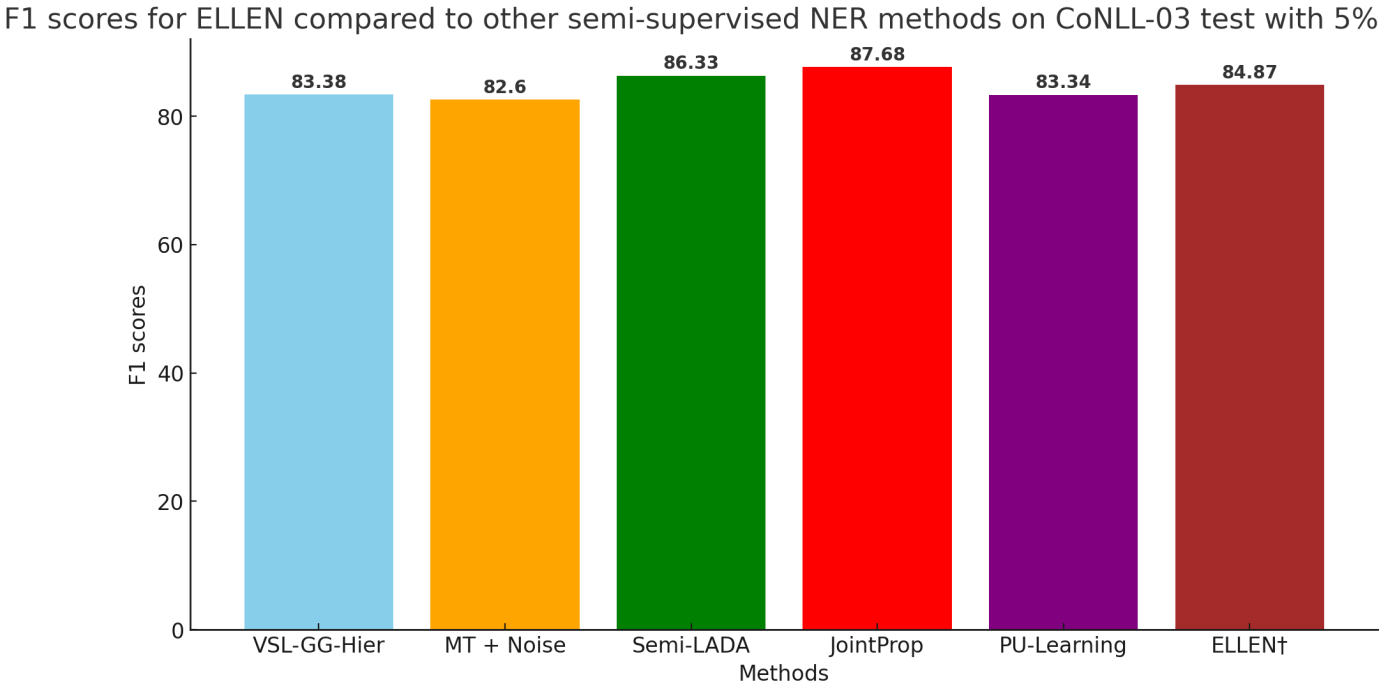
O, corporation, creative-work, group, location, person, product.

Setting: Zero-Shot.

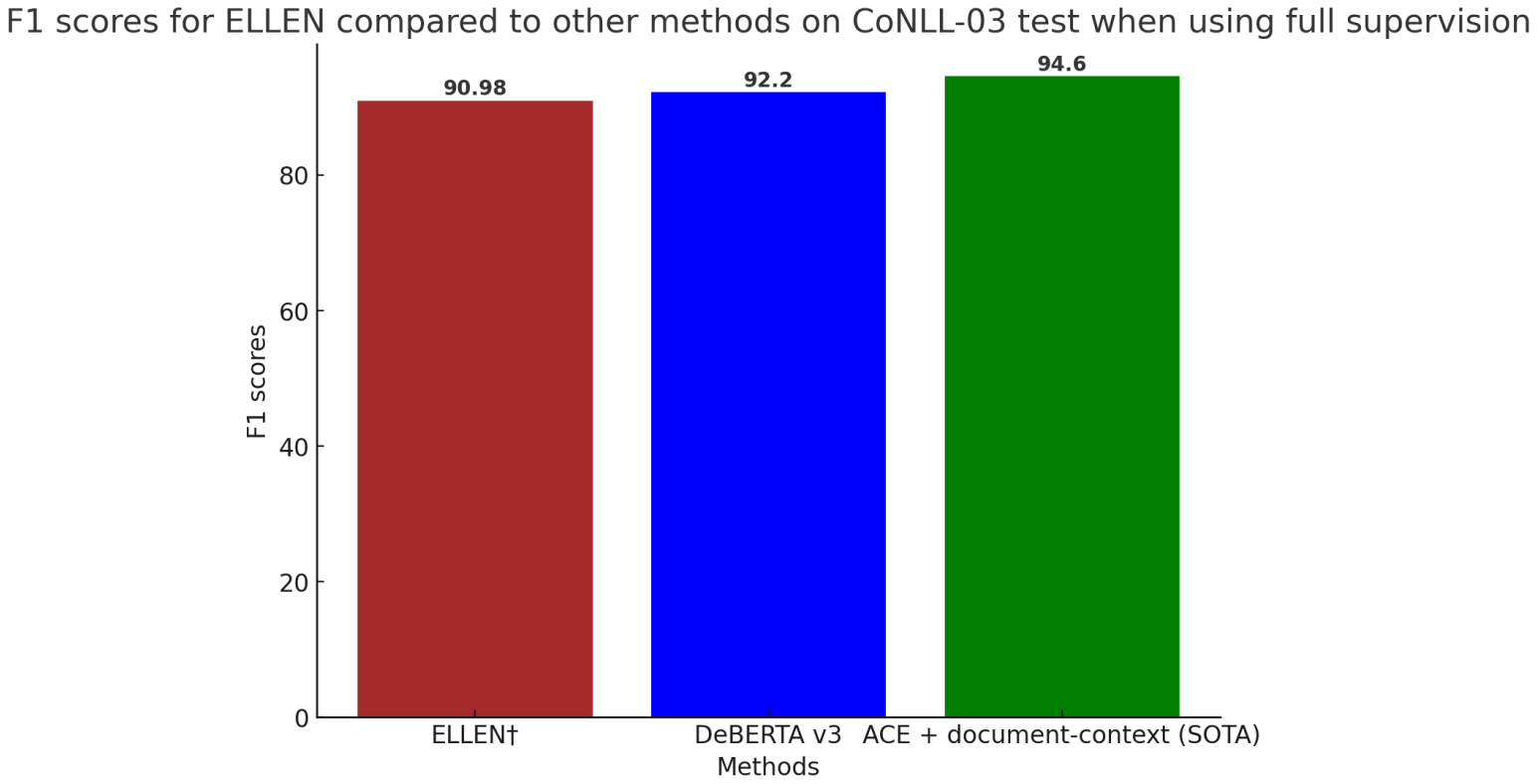
Results in the **extremely lightly supervised (1%)** data setting



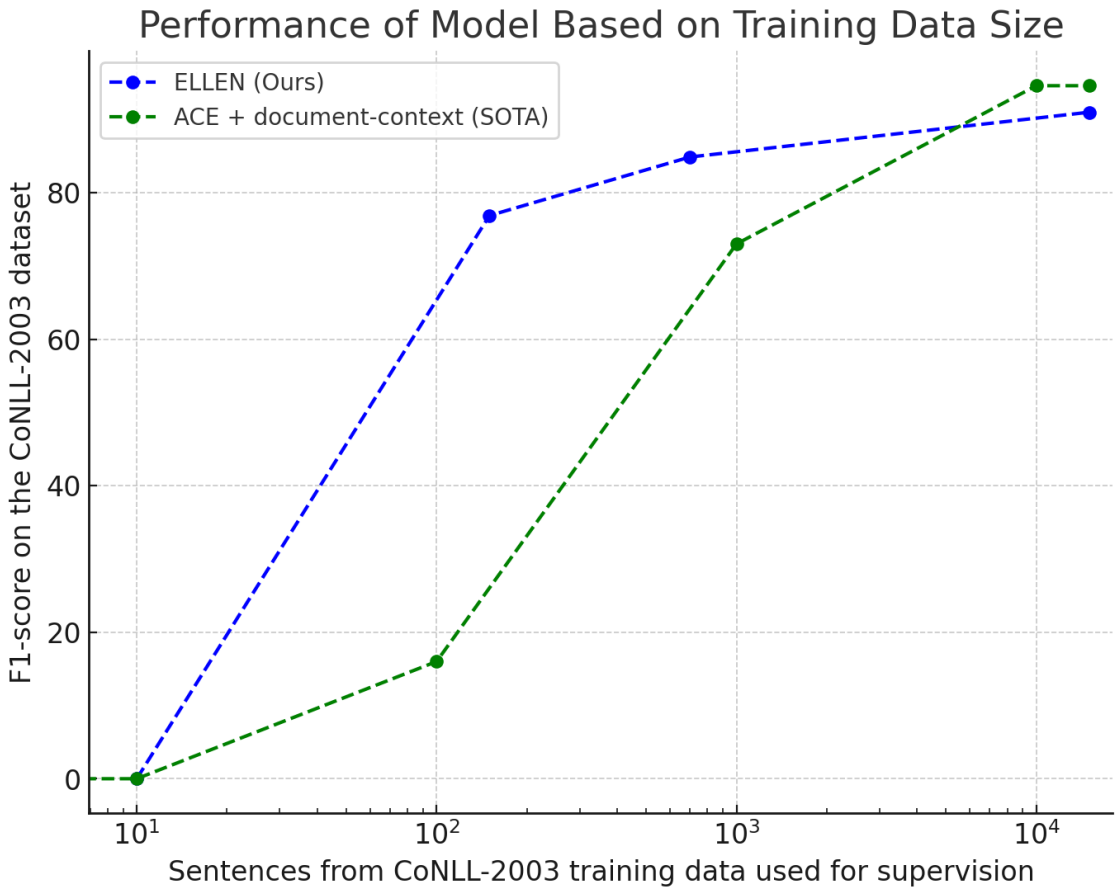
Comparison with other SOTA semi-supervised NER methods under 5% degree of supervision – Our method *scales* with increasing supervision



Our method *continues to scale* even when using full supervision

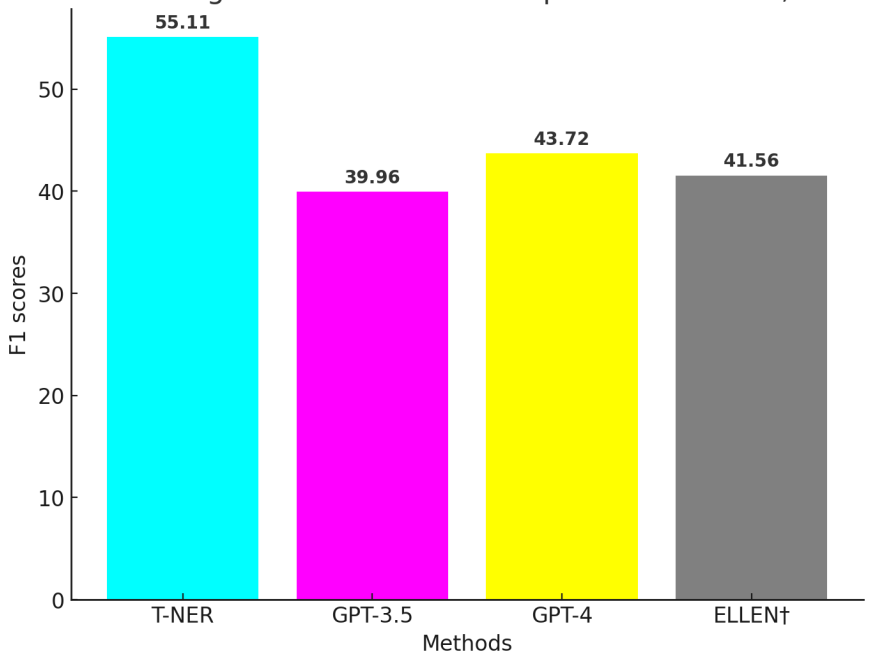


Data Efficient Learning



Zero-Shot evaluation on WNUT-17:

F1 scores for ELLEN in a zero-shot setting on WNUT-17 test compared to GPT-3.5, GPT-4 and a fully supervised model



Conclusions

- We propose a method to assemble a fast, encoder-only NER system **in less than half a day** for any specialized domain, given the availability of a domain expert/lexicon.

- We demonstrate that linguistics and deep learning can co-exist to overcome the scarcity of labeled data for NER.

THANK YOU!

Link to the paper

Link to code

Haris: hriaz@arizona.edu