BERT-PersNER: a New Model for Persian Named Entity Recognition

Farane Jalali Farahani and **Gholamreza Ghassem-Sani**

Presenter: Farane Jalali Farahani

Sharif University of Technology

Tehran, Iran

fjalali@ce.sharif.edu

September 1, 2021

Named Entity Recognition (NER)
Task Definition

- **Named entities** are terms representing real-world objects like person, location, organization, drug, date, etc.
- Example:

  Marta [**Person**] joined google [**Organization**] as a data scientist in Zurich [**Location**].
- **Named entity recognition (NER)** is the task of identifying named entities
- Applications of NER: question answering, information retrieval, machine translation, text summarization, etc.

Named Entity Recognition
The problem with Persian NER

- Persian is an under-resource language, although it is spoken by more than 110 million people worldwide
- The available training data is scarce
- There is not any clue for proper nouns in Persian; whereas in English, proper nouns begin with a capital letter

Named Entity Recognition
Active learning as a possible solution

- Active learning looks for the most informative data, instead of training the model on the whole dataset (Settles, 2010).
- There are three main forms of Active learning:
  1. Membership query synthesis (Angluin, 1988)
  2. Stream-based selective sampling (Cohn et al., 1990)
  3. Pool-based sampling (Lewis and Gale, 1994)

- Comparison: the pool-based sampling provides the possibility of running a comparison among all instances, while the stream-based approach makes query decisions individually

Our Contribution

Our Contribution

- BERT-PersNER (BERT based Persian Named Entity Recognizer): a new model, which employs **active learning** and **transfer learning** approaches for NER in Persian
- We show an effective way of using limited labeled data in Persian NER (i.e., for the first time)
- Experimentally, we show an advantage in the model performance by maximizing the knowledge gain of the model during querying from a pool of unlabeled data

Background Concepts
Transfer learning (model-based approach)

- This approach tries to transfer knowledge through the shared parameters (Pan and Yang, 2010)
- A well-trained model on the source domain has learned a well-defined structure
- This structure can be transferred to the target model
- We used BERT (Devlin et al., 2019) as our source

Background Concepts
Active learning (pool-based sampling)

The AL framework consists of:

1. Training the **initial model** with a small part of labeled data (i.e., 1% in our work)
2. Sorting the instances based on **selection strategies** and selecting top-ranked instances (i.e., 10% in our work)
3. Training the model again using the newly compiled data (L) and update the parameters
4. Stages 2 and 3 are repeated until that the instances of the pool (U) are exhausted

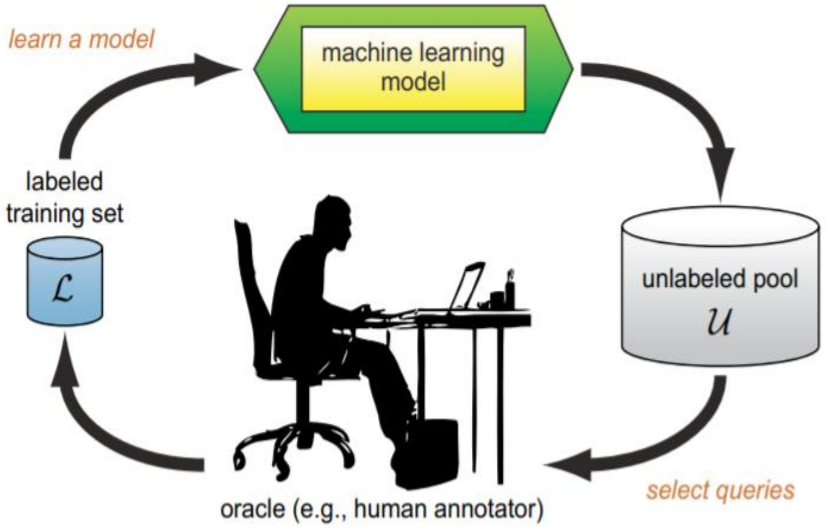


Figure 1 - The pool-based active learning (Settles, 2010)

Background Concepts
Selection strategy

- A selection strategy is considered as a core part of an AL approach
- The informativeness of each instance is determined by the selection strategy
- We employed the following selection strategies:
  - Normalized Least Confidence (NLC) (Lewis and Gale, 1994)
  - Margin (M) (Scheffer et al., 2001)
  - Sequence Entropy (SE) (Settles and Craven, 2008)

$$\phi^{NLC}(x) = 1 - \frac{1}{N} P(y^*|x;\theta)$$

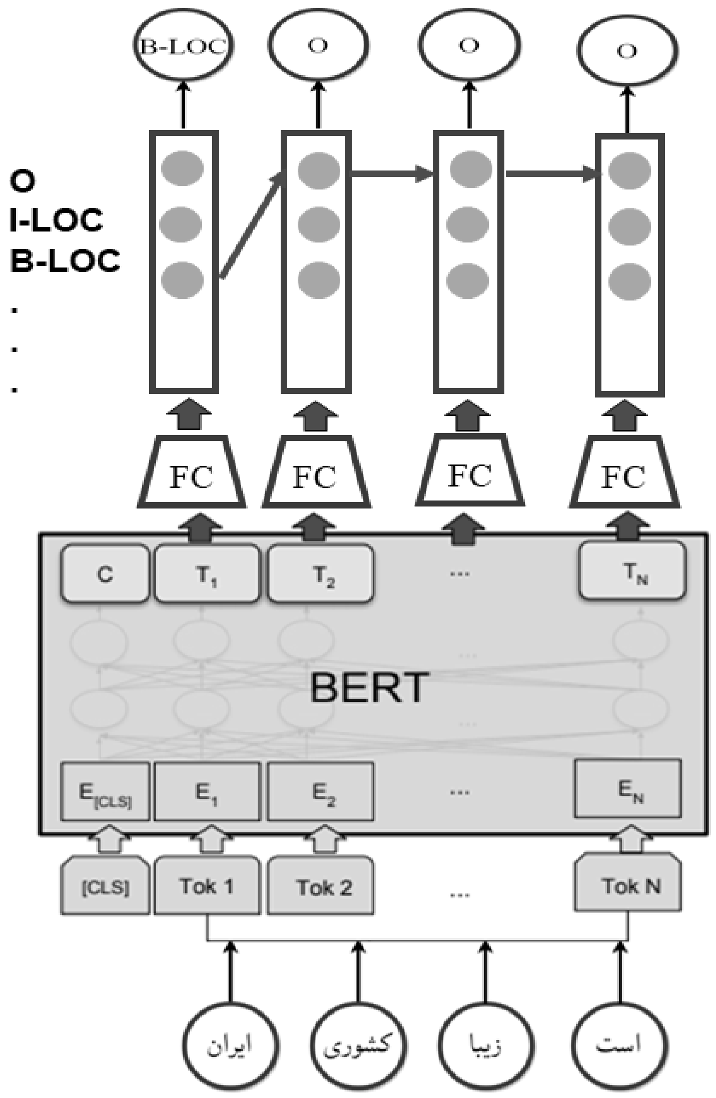
$$\phi^{M}(x) = -(P(y_1^*|x;\theta) - P(y_2^*|x;\theta))$$

$$\phi^{SE}(x) = -\sum_{y'} P(y'|x;\theta) log P(y'|x;\theta)$$

Model Architecture

Model Architecture

- Input Representation
- Context Encoder
- Tag Decoder



Experiments

Experiments
Settings

- Datasets:
  1. **Arman**: 250k tokens and six classes (location, organization, person, facility, product, and event)

  2. **Peyma**: 300k tokens and seven classes (location, organization, person, time, data, money, and percent)
- Pre-trained model: BERT-base-multilingual-cased
- Supervised learning settings: word-level, phrase-level
- Active learning evaluation: word-level
- Results have been averaged over 3-fold cross-validation

Experiments
Results (supervised learning)

Dataset	Entities	Word-level B-	Word-level I-	Phrase-level
Arman	Person	92.26	93.59	90.52
	Organization	81.61	87.97	79.43
	Location	82.08	78.67	81.83
	Facility	75.62	80.78	69.82
	Product	70.95	75.30	65.89
	Event	70.95	78.83	60.44
	All classes	84.23		80.80
Peyma	Location	86.78	76.02	84.89
	Person	86.88	91.19	84.10
	Organization	83.09	87.23	78.29
	Time	75.90	82.72	70.96
	Date	84.23	86.91	81.38
	Money	92.52	92.01	81.72
	Percent	91.64	94.48	89.32
	All classes	86.14		82.05

Table 1- F1 scores (in percentage) of running our model on Arman and Peyma.

Work	Arman Word-level	Arman Phrase-level	Peyma Word-level	Peyma Phrase-level
Deep-CRF (Bokaei and Mahmoudi, 2018)	81.50	76.79	N/A	N/A
(Shahshahani et al., 2019)	N/A	N/A	**87**	80
BERT-PersNER	**84.23**	**80.80**	86.14	**82.05**

Table 2-comparison between F1 scores (in percentage) of BERT-PersNER and our baselines.

Experiments
Results (Active learning)

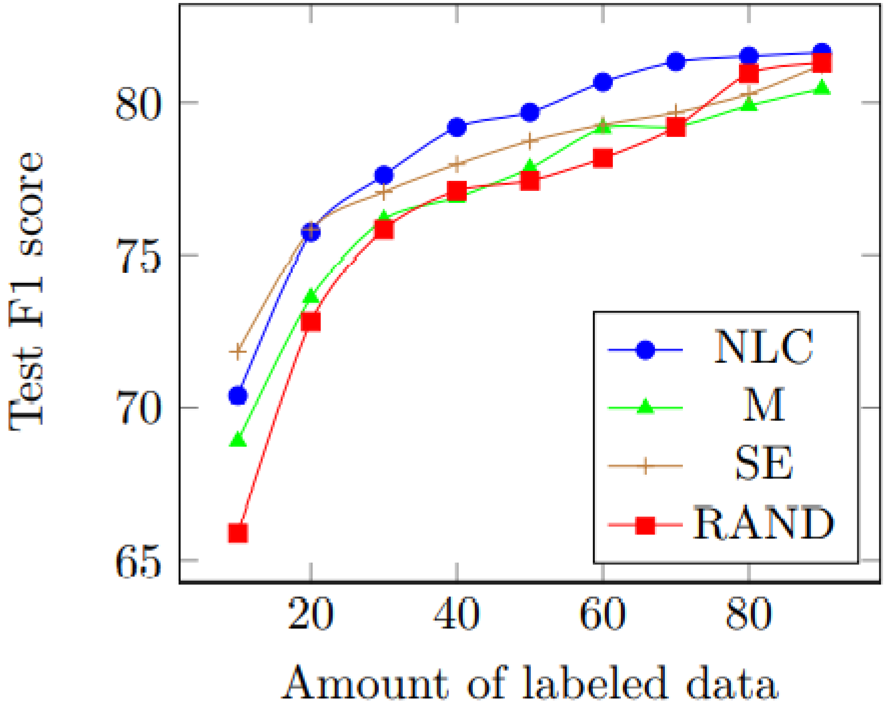


Figure 2: BERT-PersNER performance on Arman, using different selection strategies

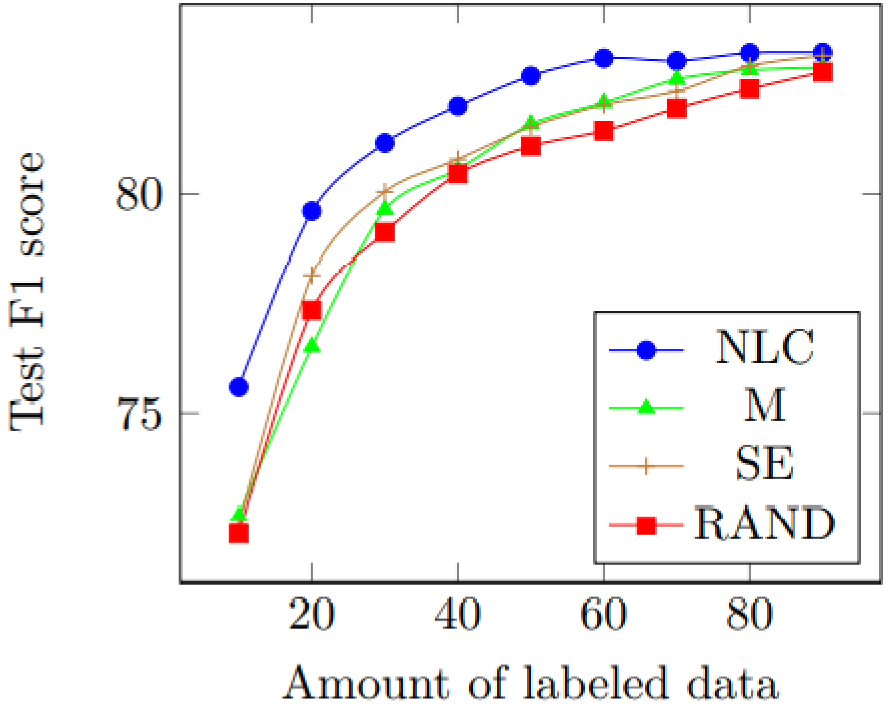


Figure 3: BERT-PersNER performance on Peyma, using different selection strategies

Conclusion & Future Work

Conclusion and future work

- With Bert-PersNER , we can choose unlabeled data for annotation in a way that maximizes the knowledge gain for the model fine-tuning process
- Using only 30% of Arman, we achieved 92.15% performance of the supervised learning method
- In the case of Peyma, using 20% of data, Bert-PersNER reached 92.41% performance of the supervised learning approach
- For future:
  - We intend to investigate the impact of other selection strategies on BERT-PersNER
  - We also plan to evaluate the proposed approach using other newly published pre-trained models

Thank you!
Question?

References

- Burr Settles. 2010. Active learning literature survey. Technical report, University of Wisconsin, Madison
- D. Angluin. Queries and concept learning. Machine Learning, 2:319–342, 1988
- D. Cohn, L. Atlas, R. Ladner, M. El-Sharkawi, R. Marks II, M. Aggoune, and D. Park. Training connectionist networks with queries and selective sampling. In Advances in Neural Information Processing Systems (NIPS). Morgan Kaufmann, 1990.
- D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994.
- Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. IEEE Educational Activities Department.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics.

References (Cont.)

- David Lewis and William Gale. 1994. A sequential algorithm for training text classifiers. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Springer-Verlag
- Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis. Springer-Verlag
- Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics

Appendix

$$P(y|x) = \frac{e^{Score(x,y)}}{\sum_{y' \in Y(x)} e^{Score(x,y')}}, \quad (1)$$

$$Score(x,y) = \sum_{i=0}^{N} T_{y_i,y_{i+1}} + \sum_{i=1}^{N} P_{i,y_i} \quad (2)$$

$$y^* = \underset{y' \in Y(x)}{\operatorname{argmax}}\, logP(y'|x) \quad (3)$$

$$\phi^{NLC}(x) = 1 - \frac{1}{N} P(y^*|x;\theta), \quad (4)$$

$$\phi^{M}(x) = -(P(y_1^*|x;\theta) - P(y_2^*|x;\theta)), \quad (5)$$

$$\phi^{SE}(x) = -\sum_{y'} P(y'|x;\theta) logP(y'|x;\theta) \quad (6)$$