Temporally-informed analysis of Named Entity Recognition

Bloomberg

Engineering

The 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)

Shruti Rijhwani - Bloomberg Data Science Ph.D. Fellow, Carnegie Mellon University

Daniel Preoțiuc-Pietro Senior Researcher, AI Group, Bloomberg

TechAtBloomberg.com

Motivation

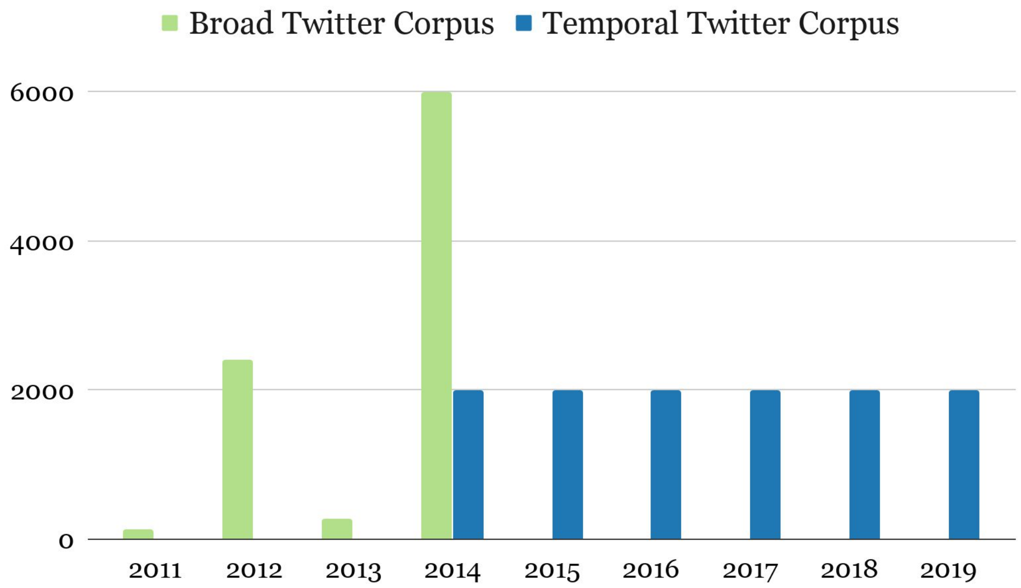
- Text data evolves over time because of **changes in language use**
- The usual setup of large-scale NLP models:
  - Trained and evaluated on **random splits of available data**
  - Make predictions on data in a **future time period**
  - Does not take temporal data drift into account
  - Lower performance on future data compared to the test set
- **Temporal information in modeling** may lead to better performance

Research Questions

- **Does temporal drift in training data affect performance?**
- **Can we leverage the temporal information of the training instances to improve performance?**
- Case study: Named Entity Recognition on English Twitter Data
  - Readily accessible timestamp information
  - Users on social media post about current events
  - Reflects changes in language use faster than other sources of data

Temporal Twitter Dataset

- Existing Twitter NER datasets do not have sufficient temporal diversity
  - Broad Twitter Corpus (2016)
    - Data from 2009-2014, unevenly distributed
- Temporal Twitter Dataset
  - **2,000 tweets from each year between 2014-2019**
  - Sampled using the same strategy as the Broad Twitter Corpus
    - Six English-speaking locales
    - *Twitterati*, i.e., individuals from array of domains including musicians, journalists and celebrities
    - Mainstream news organizations, both larger networks and local news outlets



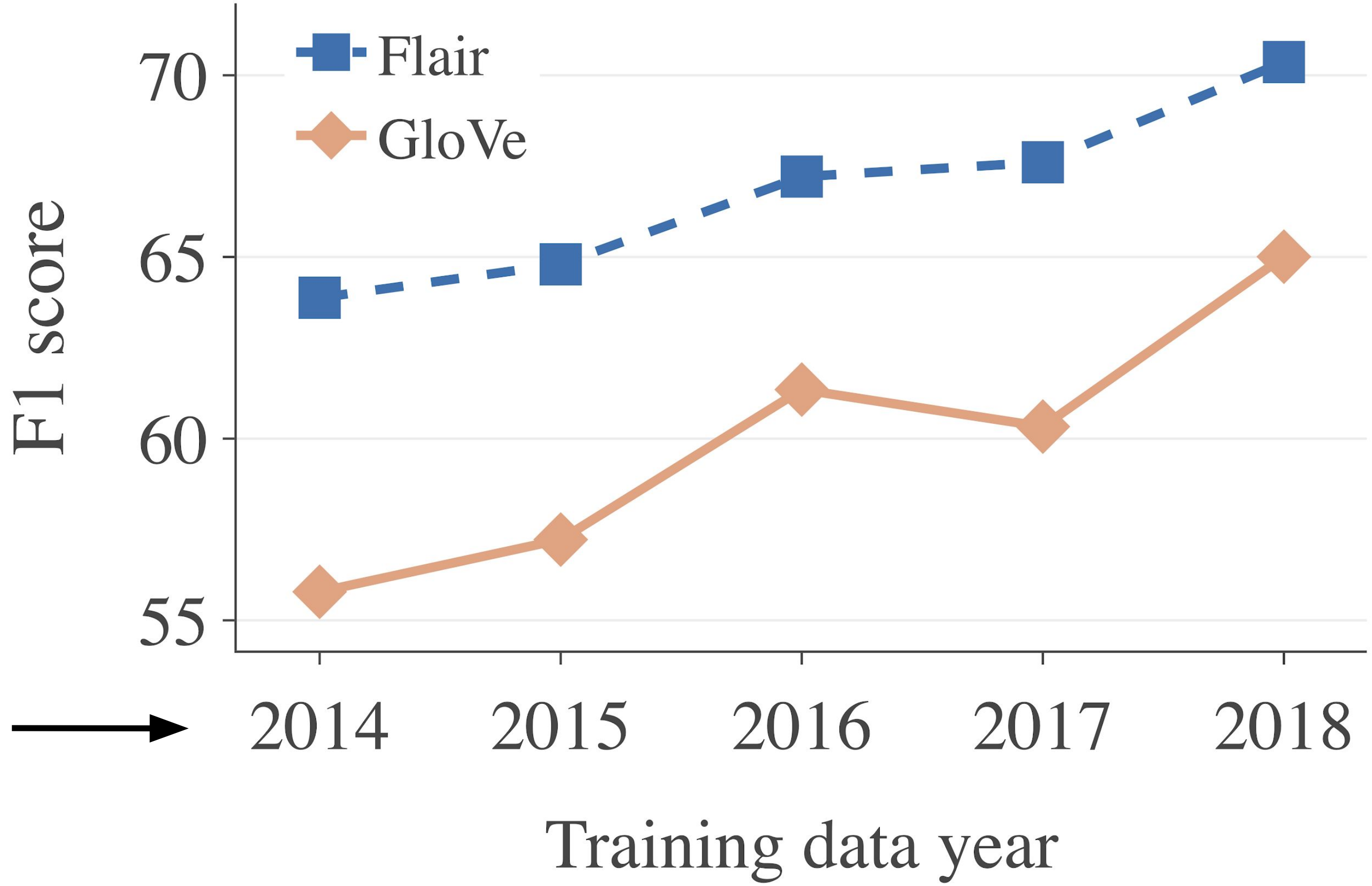
Experimental Setup

- NER Model Architecture
  - **Character and word embeddings** to represent the input text
  - **Bi-LSTM** to encode the input text
  - **CRF** to make a globally normalized prediction
- Word Embeddings
  - **GloVe:** static word embeddings
  - **Flair:** contextual word embeddings
  - All trained on data from before 2014
- Data splits
  - Train on 2014-2018 tweets
  - Validation and test on random splits of 2019 tweets
  - **Simulates a "future time period" for inference**

Does temporal drift in training data affect performance?

Data temporally closer to the target data gives better performance

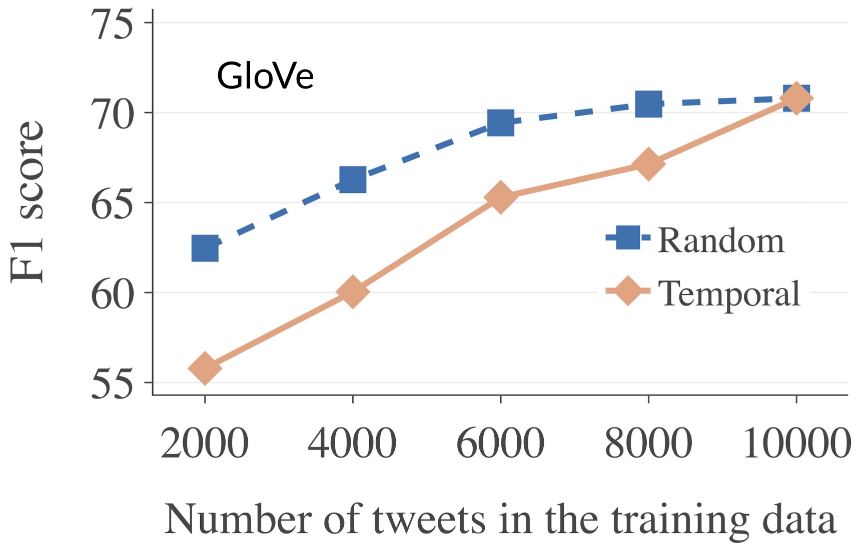
Trained on each year individually
Each model has the same number of examples



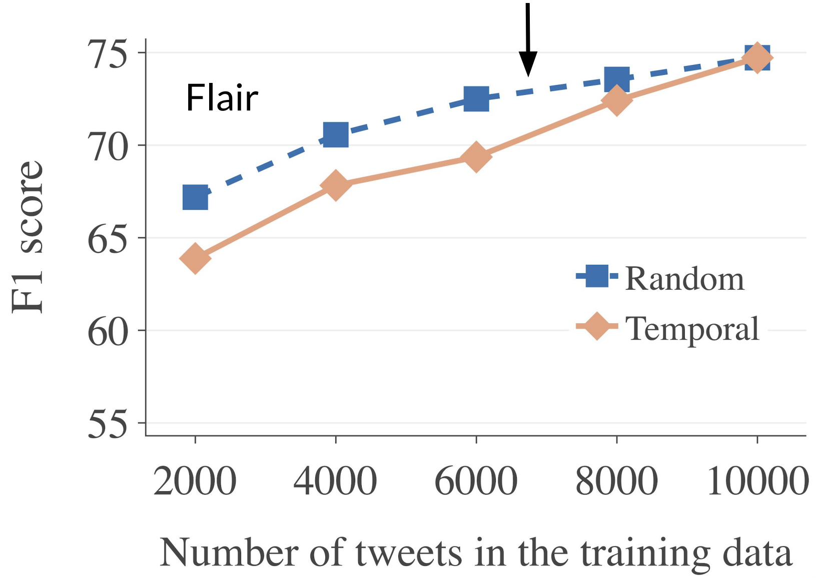
Does temporal drift in training data affect performance?

Temporal distribution of the training data impacts performance

- **Random:** Sample randomly from all years
- **Temporal:** Cumulatively add data in sequence of years

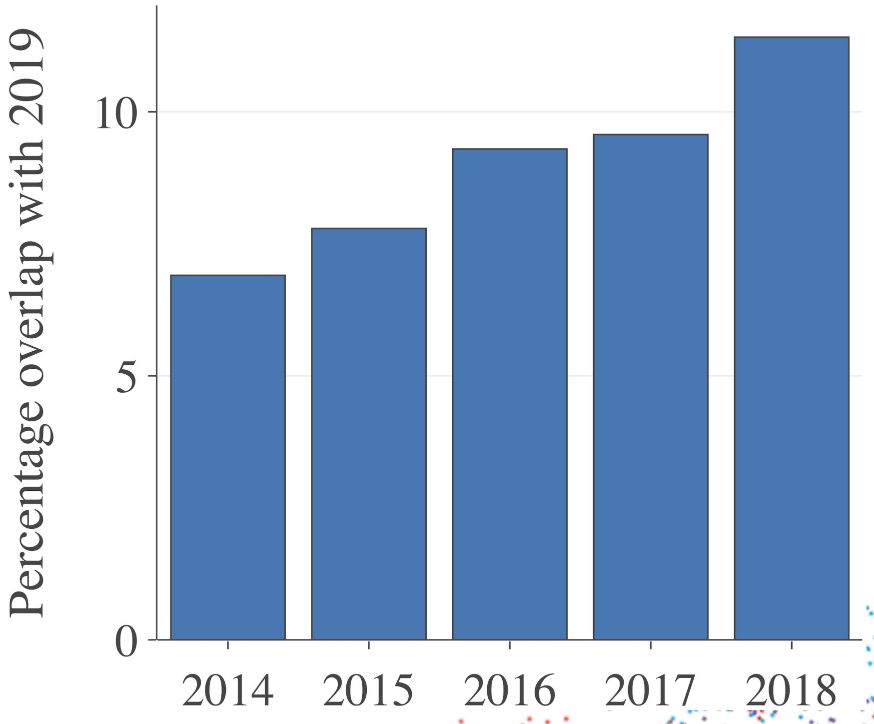


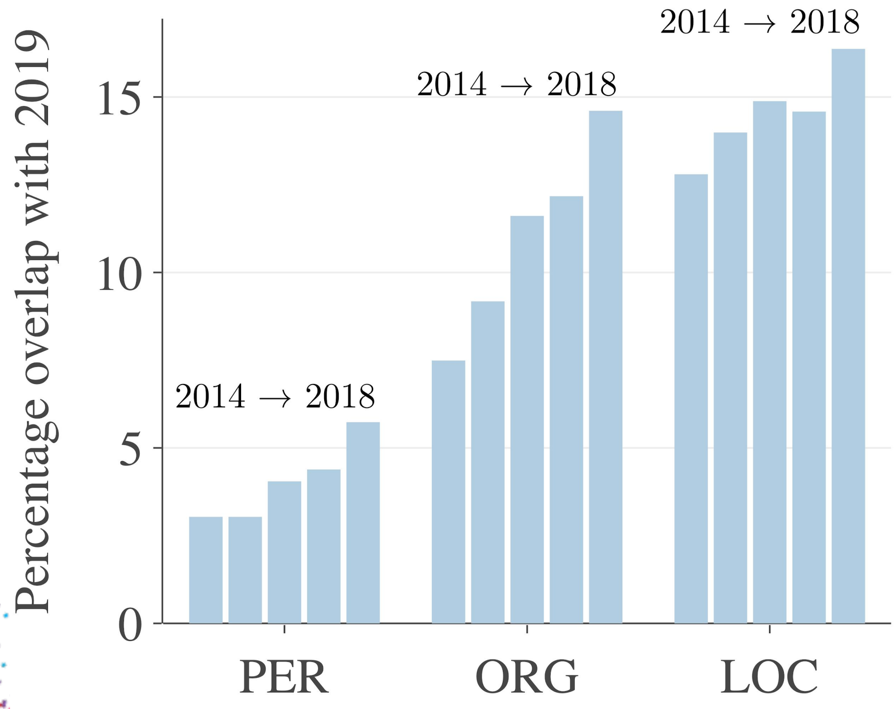
Data sampled from all years has better performance



Analysis: Entity Mention Overlap

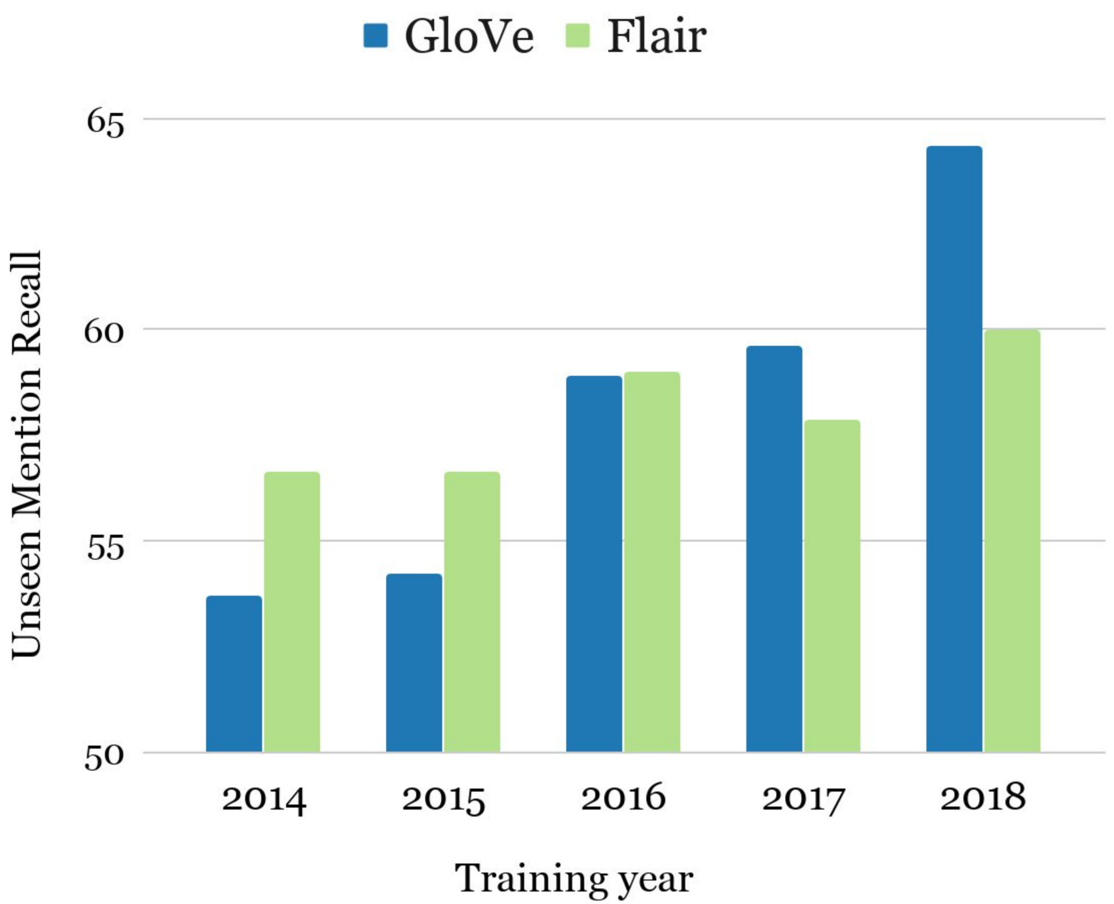
- One potential reason for better performance is the overlap of entity mentions between the training and test data
- Overlap increases as we get temporally closer to the target data: **overall and type-wise**





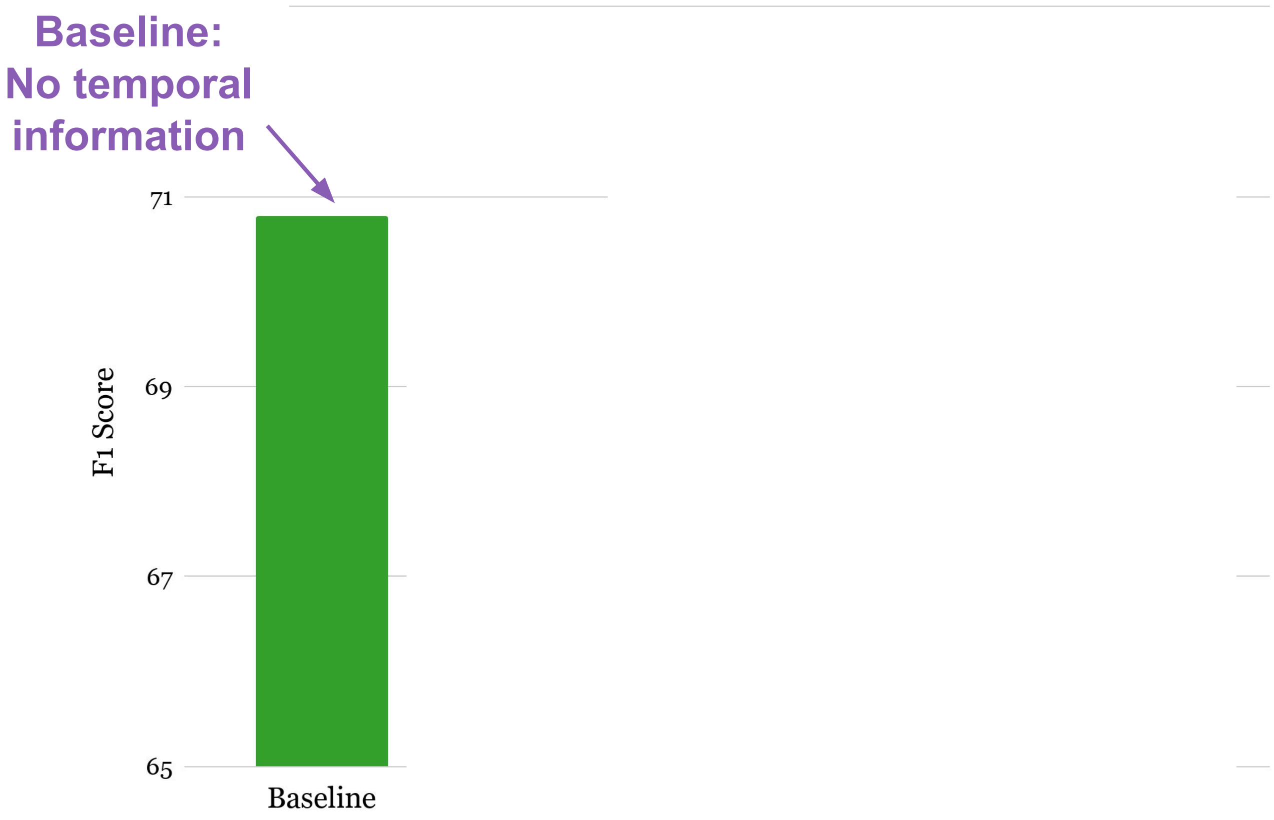
Analysis: Model Performance

- Is surface-level overlap the only factor in temporal drift?
- **Recall for mentions unseen in the training data** increases as we move temporally closer.
- The model is able to learn **more relevant context.**



Modeling Temporal Information

- Initial exploration of methods
  - Do not require significant modifications to the model
- **Sequential training:** train the model on each year sequentially
- **Temporal fine-tuning:** Train the model on data from all years and fine-tune on 2018
- **Instance weighting:** Double the weight of training data from 2018
- **Year prediction:** As an auxiliary task, use shared representations to predict the year of the instance during training



Experiments with GloVe embeddings

Takeaways

- Temporal information is useful and can be leveraged in training
  - Annotated data is collected across a time range
- In a case study on NER with Twitter data:
  - **Training on data from a closer time period to the target leads to better performance**
  - **Fine-tuning models on temporally close data improves performance** over simply combining all data for training
- Get our data set at https://zenodo.org/record/3899040

Thank you!

https://www.bloomberg.com/careers
NYC: https://careers.bloomberg.com/job/detail/80630
London: https://careers.bloomberg.com/job/detail/80683

Get our data set at https://zenodo.org/record/3899040

Shruti Rijhwani - srijhwan@andrew.cmu.edu
Daniel Preoțiuc-Pietro - dpreotiucpie@bloomberg.net

Q&A Sessions 13B & 15A
Information Extraction 8 & 11
Wednesday, 8 July 13:00 UTC & 20:00 UTC

TechAtBloomberg.com

Bloomberg Engineering