Complementary Learning for NER

• For popular domains such as medical science, there are usually multiple annotated datasets for the NER task.
• However, different datasets often have different sets of annotated entity types lying in distinct label spaces.
• We propose complementary learning paradigm.
• Propose an end-to-end learning method that conducts label space fusion and proto-NER training at the same time.

![Diagram of complementary learning paradigm](image-url)
Heuristic Strategies v.s. End-to-End Training

- Heuristic Strategies
- *Tackle false negatives*: Train separate model on each dataset. Propagated labels are from separate models trained on each single dataset.
  - E.g. Train on Disease corpus → Apply to Chemical corpus to predict diseases in Chemical
- End-to-end training: Fuzzy CRF
  - Extend O to the fuzzy labels (use IOBES here)
  - Sum over all possible sequences and use DP for inference.
Extension: Self-training for complementary learning

- In complementary learning setting, some types of entities are not labeled.
- Self-training: Our algorithm should first train on the most confident predictions.
- No existing self-training algorithms are available in complementary learning setting.
Self-training layer

- Objective function: Minimize KL divergence

\[ L = KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \]

- \( p_{ij} \): ground-truth probability. When entity\_i is labeled by type T, the probability is 1 and other types 0; otherwise, it is obtained from

\[ p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}}, \quad f_j = \sum_i q_{ij} \]

Where j is not type T

- \( f_j \) is the type frequency
- This automatically select the entities with highest confidence from predictions.
- Use soft classification (prob. dist.)

- \( q_{ij} \): \( \text{Pr}(\text{entity\_i}=j|X) \), predicted probability: get from NN model

\[ \frac{e^{t_j^T v_i}}{\sum_{t_k \in L} e^{t_k^T v_i}} \]

Still working on coding part...
Misc.

• Passed the prelim exam
• TA for CS412
• 1 paper accepted by ICDM