出願研究科: 情報科学研究科
希望研究室: 自然言語処理研究室
氏名: 岩田　晟

固有表現認識における Distant Supervision のノイズ軽減

1　はじめに

私は，学部時代の卒業論文ではトピックモデルと点過程の一種である Hawkes 過程を Yahoo Japan の検索データに対して用いることで，検索トピック内の流行を予測する研究を行った．その過程で，形態素解析器を用いて単語分割を行ったが，「楽天市場」が「楽天」と「市場」というように分解されてしまった．「楽天」という単語には「楽天市場」以外にも「楽天イーグルス」や「楽天ポイント」などの様々な情報が含まれている．そのような分割では「楽天市場」に関する情報のみを得ることは難しい．そのため，固有表現認識の分野に興味を持った．そこで，私が奈良先端科学技術大学院大学 (NAIST) で取り組みたい研究テーマは「固有表現認識における Distant Supervision のノイズ軽減」である．本稿では，この研究テーマの研究背景，研究課題，先行研究，提案手法について述べる．

2　研究背景

固有表現認識 (Named Entity Recognition, 以下 NER) とはテキストの中から固有表現の範囲と人名や地名，組織名などの種類を認識する技術である．NER は情報抽出や対話システムなどの上流タスクとしても用いられている．近年ではバイオテキスト NER や 料理レシピにおける用語に特化した NER などの分野固有表現体系が定義されたコーパスが提案されている．これらのようなアノテーション付きコーパスを用いて教師あり学習を行うが，コーパスにアノテーションをつけるには時間や人手など多大なコストが要求される．大量のラベル付きデータが存在している分野は限られている．また，分野特有のアノテーション付与させるためには，その分野の専門知識が必要となる．

3　研究課題

このような NER のアノテーションコストを軽減する手法として，近年では，Distant Supervision による教師データの自動生成手法が注目されている．Distant Supervision とはデータベースを利用してテキストコーパスに含まれる固有表現のマッチングを行い，人手のアノテーション無しに　NER の学習データを自動生成する学習法である．

しかし，Distant Supervision により作成された擬似学習データはノイズが生じやすいという問題がある．例えば，企業に関する NER を行いたいとする．mac や iphone のメーカーである”apple”という企業が登録されているとする，一方で，データの中に”apple”(りんご) を扱う企業が出てきた場合，単純な文字列マッチングによって自動アノテーションを行うと，商品名ではなく企業名のラベルがつけられてしまう．そのため，これらのノイズにどう対処するかが課題となる．

4　先行研究

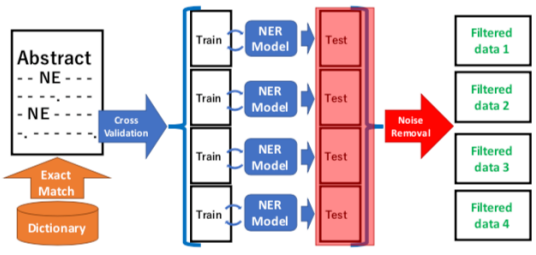

図 1　辰巳らによるノイズ除去 ([1] より引用)

先行研究として辰巳らは化学ドメインに対して Distant Supervision を用いた NER を研究している [1]．この研究にて擬似アノテーションコーパスのノイズ消去の概略は図 1 のようになる．

辰巳らは図 1 の手法が NER のノイズに対して，

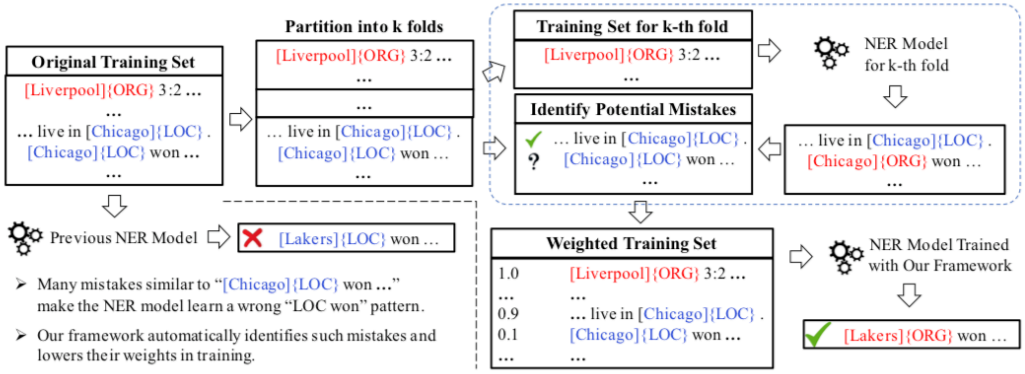

図 2　cross weigh のフレームワーク ([2] より引用)

有効であることを確認した．

5　提案手法

Wang らは NER における人手によるアノテーションのラベリングの誤りを検出し，その誤りのある文に対して重みをつけ，再度学習するようなフレームワーク CrossWeigh を提案した [2]．概略は図 2 のようになる．

誤りの推定は K 分割交差検証に似たように，K 個のサブセットに分割を行う．そのうちの一つを取り除き，残りで学習する．K 分割交差検証と異なるのは，学習データに対してテストデータに含まれるデータを削除して学習を行う．これにより，テストデータに対して堅牢になる．その後，テストデータに付与されてたアノテーションとモデルの結果が異なるものをラベル誤りであると考える．

この手法を Distant Supervision による擬似アノテーションに対して応用することでノイズに対して強くなるのではないかと考えている．辰巳ら [1] は，ノイズの原因となる文自体を学習から削除している．一方で，提案手法の場合，ノイズと判断した文も含め，重みをつけて再度学習をする．それにより，より頑健な手法になると考えられる．また，重みにより，擬似コーパス内の，どの文がノイズとなりやすいのか，どの文が識別が難しいのかといったような解釈も期待できる．

6　最後に

上記手法は既存手法と比べて，人手コストと擬似コーパスノイズを減らせることが期待できる．それにより NER の性能が上がれば，下流タスクの性能の向上に繋がると考えられる．

また，このような研究テーマに取り組むにあたって，なぜ NAIST を志望するのかというと，研究室の豊富な人材と学校の手厚いサポートである．一つの研究室に教授陣が 4 人ほどおり，博士課程の学生も他大学と比べると多い．また，授業料免除などの大学院側の学生支援も手厚いところも魅力的である．このような環境などを活用して自然言語分野に貢献していきたい．

参考文献

[1] 辰巳守祐, 後藤 啓介, 進藤裕之, 松本裕治, 辞書を用いたコーパス拡張による 化学ドメインの Distantly Supervised 固有表現認識，情報処理学会 第 241 回自然言語処理研究会，August 29， 2019.

[2] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, Jiawei Han, CrossWeigh: Training Named Entity Tagger from Imperfect Annotations，EMNLP 2019.