hw2586, xf574, kc4627, zz4323
11-11-2024

Final Project Proposal

List of Members and Roles

Kevin Chen: Programming & Theories
Tim Han Yu Wang: Writing & Analyzing
Richard (Xu) Fang: Programming & Evaluation
Flora Zhimeng Zhang: Writing & Researching

Introduction

Our project aims to improve the accuracy of Named Entity Recognition (NER) systems, particularly for handling ambiguous, multi-word entities. In BIO tagging, where entities are marked with Beginning (B), Inside (I), and Outside (O) tags, current methods often struggle with entities that overlap or have complex, nested structures. While simpler NER tasks have been explored, more advanced algorithms are needed to accurately identify and extract multi-word or ambiguous entities. We chose dates and organizations as our focus groups because they are distinct from each other, allowing us to analyze results in a more controlled and unbiased way.

To address this gap, we are developing a refined BIO tagging algorithm designed to better classify and disambiguate complex entities. Our approach builds on existing tagging methods, enhancing them with advanced feature engineering and evaluation techniques to improve accuracy in ambiguous cases. By improving the tagging process, our goal is to deliver a solution that not only identifies but also more reliably disambiguates entities than current systems.

Our solution is directly applicable to structured data extraction from financial news articles. Using data from the Wall Street Journal, our algorithm focuses on identifying dates and organizations—two often ambiguous entity types in financial text. Possible applications include information retrieval and trend analysis. Through iterative training and evaluation, our algorithm aims to meet an F-score baseline, refining its precision, recall, and reliability across development and test sets. This improvement has the potential to make NER more robust and effective, advancing its utility in real-world scenarios.

Academic Articles

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). *Part-of-speech tagging for Twitter: Annotation, features, and experiments*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 42–47). Portland, Oregon, USA: Association for Computational Linguistics.

The model mentioned in the article is applied to related social work to our project, which has plenty of similar word groups. The development and fine-tuning of the model could be referenced for the evaluation of the project.

He, S., Wang, T., Lu, Y., Lin, H., Han, X., Sun, Y., & Sun, L. (2023). *Document information extraction via global tagging*. In Lecture Notes in Computer Science (pp. 145–158). Springer. https://doi.org/10.1007/978-981-99-6207-5_9

The article presents a token classification model for unnatural visual documents. It approaches entities arranged in complex layouts, such as lists, flex boxes, etc. The tested documents in the article match our original files, which are selected from journals and newspapers that may have various layouts. The framework could inspire the development of the project.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020). *A unified MRC framework for named entity recognition*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5849–5859).

For nested noun entity recognition where a token may be assigned by several labels, also seen in our project, the article presents a unified framework for handling nested noun extraction tasks. It would be a reference for developing a more evaluated project that could extract complex noun groups in the journals.

McDonald, R., Crammer, K., & Pereira, F. (2005). *Flexible text segmentation with structured multilabel classification*. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 987–994). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

The article presents a model for multilabel classification, which could inspire us to handle the complex tags for the evaluated implementation of the project. The model mentioned not only solves the multilabel tasks problem but also holds the accuracy of the simple tagging, which is a comparable example of our project when evaluating the whole process of development.

Ye, W., Li, B., Xie, R., Sheng, Z., Chen, L., & Zhang, S. (2019). *Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1351–1360). Florence, Italy: Association for Computational Linguistics.

The article specifically introduces the addition of BIO tagging for name entity extraction. For the sentences with multiple noun groups, a strategy could catch the complex context relation matter for the accuracy of the extraction. Since the algorithm framework mentioned in the article is similar to our project, the strategy development and measurement score could be utilized to evaluate our group project.

Strategy

To address the task of BIO tagging specific word groups, “dates,” and “organizations,” an extended methodology based on Homework 5 will be employed. Initially, data preprocessing will be conducted on files within the BBN-NE.tgz archive obtained from “Data for Class.” This data will be stratified into three subsets: training, development, and testing. Specifically, 21 folders

within BBN-NE.tgz will serve as training files, while two folders each will be designated for both development and test purposes.

Following this, the training files will be merged into a singular training dataset. A custom program, *data_reformater.py*, will be developed to standardize the format of these files. This script will parse the training file, outputting each line with one word paired with its corresponding BIO tag (derived from BBN-NE.tgz labels) and its part-of-speech (POS) tag (generated using NLTK's POS tagger). The resulting file will be named *training.pos-chunk*. A similar process will be repeated for the development and test datasets, resulting in files named *dev.pos*, *dev.pos-chunk*, *test.pos*, and *test.pos-chunk*. Notably, *dev.pos* and *test.pos* files will exclude BIO tags, whereas *dev.pos-chunk* and *test.pos-chunk* files will retain them. These pos-chunk files will serve as the basis for scoring.

Subsequently, another program, *feature_adding.py*, will be created to process the *training.pos-chunk* file, generating an enriched output file, *training.feature*. This program will also extend to the *dev.pos* and *test.pos* files, generating corresponding feature files. Various features—such as the current, previous, and next BIO and POS tags—will be incorporated into the training and development feature files, though BIO tag features will be excluded from the *dev.feature* and *test.feature* files.

The training and development feature files will be further processed according to the steps outlined in Homework 5 to generate a *result.pos-chunk* file. Using the maxent-3.0.0.jar package alongside *MEtrain.java*, a model, *model.chunk*, will be trained on the feature-enriched data. The *MEtag.java* program will then be applied to tag the development file, producing *dev_result.pos-chunk*.

Evaluation

A scoring script, *scorer.py*, will assess the accuracy of *dev_result.pos-chunk* by comparing it to the correctly tagged *dev.pos-chunk* file. To enhance the model's performance, additional experiments will be conducted by iteratively adding or removing various features from the training and development feature files, with the objective of optimizing precision, recall, and F-score. Once an acceptable score is achieved using the development set, *test_result.pos-chunk* will be generated and evaluated based on the test data.

This entire procedure will be executed twice: once for the "date" group and once for the "organization" group. A target baseline F-score of 85% has been set for the BIO tagging accuracy of both "date" and "organization" groups. In an advanced version of this project, further sub-categorization within the "organization" group (e.g., corporations, government organizations, etc.) may be incorporated.

Collaboration Plan

For the project, Flora and Tim will research existing organizations and date BIO tagging research. They will focus on analyzing existing issues with BIO tagging organizations and dates.

Possible issues can be the uniqueness of organization names and the arbitrariness of formatting the dates by different authors, which can cause inaccuracies. Tim will be focusing on researching the BIO Tagging of organizations, specifically the issues that may come up due to the unique nature of organization names, as well as what other researchers have done compared to our model. Flora will research the date groups, including the tagging algorithms. Since the structure of date groups might differ, a flexible tagging system is needed for precise search.

Kevin and Richard will collaborate on programming the project. This project will include three Python programs: *data_reformater.py*, *feature_adding.py*, and *scorer.py*. They will create a simple version of the *feature_adding.py* program collaboratively and then work separately to add more different types of features. Then, they will compare and experiment with different combinations of features to create a new final *feature_adding.py* program that will result in the highest score. Kevin and Richard will continuously report the results to Tim and Flora to guide their research direction into error analysis.