Annotation task

Amir Ali
Dawid Przybyliński
Aleksander Podsiad

January 2023

1 Data Annotation in NER

Data annotation for food reviews typically involves identifying and labeling different entities within the text of the review.
Similarly, data annotation for ingredients typically involves identifying and labeling the different ingredients that are mentioned in a recipe or food-related text. This include the name of the ingredient (e.g. "flour", "salt"), which in our case.

2 Difficult Issues in NER

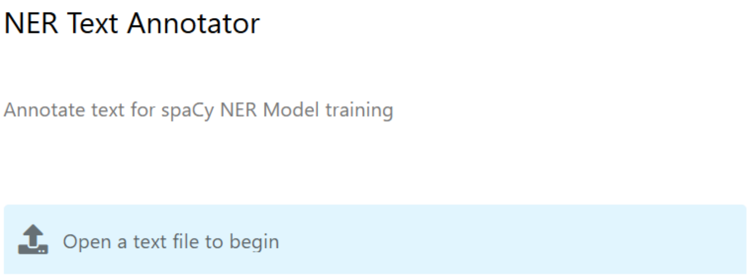
The difficult issue is the cost and time associated with manual annotation. It is expensive and time-consuming to manually annotate large amounts of text, especially for tasks that require a high level of expertise such as ingradients data annotation. This can make it difficult to scale up the annotation process to meet the needs of large datasets.
Another issue is the complexity of the text, certain task complex and require a high level of expertise. For example, ingradients text annotation requires knowledge of ingredients to identify entities correctly. Finally, the annotation process can be affected by the variability of the language used in the text, such as the use of idioms, colloquial language or acronyms, which can make it difficult to understand the text and label entities correctly. For The Future, I believe: To overcome these issues, some solutions include using active learning methods, which allow the model to ask for human feedback on uncertain examples, this can help to improve the quality of the annotation and reduce the human cost.
Another solution is to use pre-trained models, which can help to reduce the amount of manual annotation needed. Additionally, inter-annotator agreement studies, where multiple annotators label the same text, can be used to measure and improve the consistency of the annotations.

3 Instruction

For the annotation task we are using the web-interface from GitHub site: https://tecoholic.github.io/ner-annotator/. This allows for fast annotation with mostly automated workflow on the food review texts.

The steps for annotation are as follows:

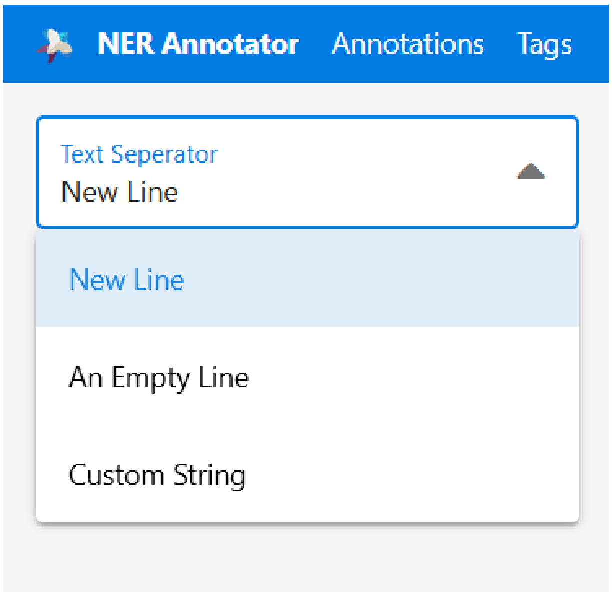
- prepare the dataset by extracting the chosen review strings and putting them into the .txt file,
- uploading the file to the previously mentioned site,
- selecting the appropriate separator,
- creating the 'INGREDIENTS' tag,
- annotating all the documents using the web-interface



NER Text Annotator

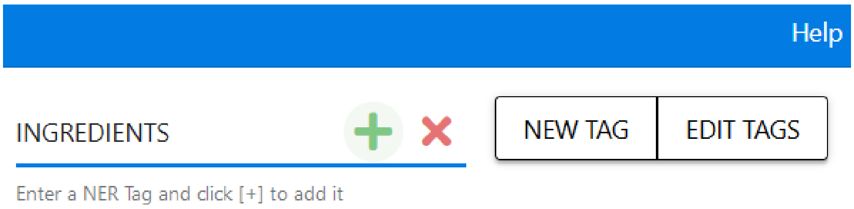
Annotate text for spaCy NER Model training

Figure 1: File input into annotator



Text Seperator
New Line

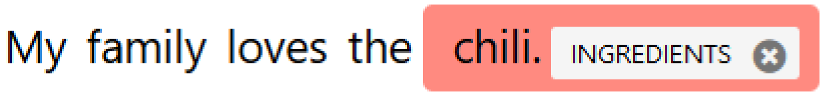
Figure 2: Choosing the separator for the file with documents



INGREDIENTS

Enter a NER Tag and click [+] to add it

Figure 3: Adding new NER tag



My family loves the chili. INGREDIENTS

Figure 4: Annotation example

4 Inter-Annotator Agreement and KAPPA

4.1 Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) is a measure of the consistency of annotations across different annotators. It is commonly used in data annotation tasks, such as named entity recognition (NER), to evaluate the reliability and quality of the annotated data.

When performing IAA, multiple annotators are given the same text to annotate, and their annotations are then compared to determine the level of agreement between them. The agreement can be measured in different ways, depending on the task, such as the percentage of entities that are labeled the same by all annotators, or the average pairwise agreement between annotators.

For NER of ingredients, IAA can be used to evaluate the consistency of the annotations of different annotators in identifying the ingredients mention in a recipe or food-related text. For example, multiple annotators can be given the same recipe and asked to identify and label the ingredients used. The IAA can then be calculated by comparing the labels assigned by different annotators to the same text.

IAA is important for ensuring the quality of the annotated data, which is a critical factor for training machine learning models. High IAA indicates that the annotations are consistent and reliable, whereas low IAA indicates that the annotations may be inconsistent and unreliable.

IAA can be improved by providing clear guidelines and instructions for the annotators, and by providing training and examples to the annotators. Another way to improve IAA is to use a consensus approach, where the annotators come to an agreement on the labels for the text after discussing any discrepancies.

4.2 KAPPA

Kappa is a statistical measure of inter-annotator agreement (IAA) that is commonly used in data annotation tasks, such as named entity recognition (NER). It is used to quantify the level of agreement between two or more annotators on a categorical task, such as labeling entities in a text. Kappa is a more robust measure of agreement than simple percentage agreement, as it takes into account the possibility of agreement by chance.

Kappa is a value between -1 and 1, where 1 indicates perfect agreement, 0 indicates agreement that is no better than chance, and negative values indicate less agreement than would be expected by chance. It is calculated by comparing the observed agreement between annotators to the agreement that would be expected by chance, taking into account the marginal frequencies of the categories in the data.

Kappa is a popular measure of IAA in data annotation tasks because it provides a more accurate representation of the level of agreement than simple percentage agreement, which can be misleading when the marginal frequencies of the categories are not evenly distributed.

It's important to note that Kappa is sensitive to the number of categories, the more categories the lower the kappa value, so it's important to keep the number of categories low if possible. Also, Kappa is sensitive to the distribution of the categories, a skewed category distribution can result in a lower Kappa value.

4.3 F1-score

In plenty of corpora for Named Entity Recognition, number of tokens not belonging to any of the analysed categories (usually tagged as "O") is substantially larger than the number of tokens that should correctly be assigned a tag. Therefore datasets are often unbalanced and Kappa score is inflated. On the other hand, the approach of ignoring the "O" label yields low Kappa scores. Because of those reasons to evaluate the similarity of tagging performed by annotators, we consider not only the Kappa score after ignoring the "O" label but also F1-score is taken into account.

5 Comparison of results

We have compared the results of annotating the set of 50 documents by two different annotators and checked the IAA for these tasks with the Kappa measure. The resulting score is 0.40 which means the moderate agreement of the annotators. The problems with manual annotations include the differences in what can be considered an ingredient and which adjectives to include when tagging different ingredients. For example in the same sentence one tagger can choose "minced garlic" as the ingredient and the other can just tag the "garlic" which already lowers the agreement score of their annotations. The analysis of the F1 score comparing the result of taggings performed by two annotators yielded the result of 0.85, which confirms the moderate agreement of the annotators.

6 Work breakdown

Team member	Tasks and time
Amir Ali	Data Collection, 50 annotations and annotation method proposal; report prepartion (12h)
Dawid Przybyliński	50 annotations and results comparison; report prepartion (8h)
Aleksander Podsiad	50 annotations and results comparison; report prepartion (8h)