Using LLM for Entity Extraction

IMDB Movie Reviews Dataset

Ritesh Kumar

2024SP_MS_DSP_453-DL_SEC61: Natural Language Processing

Module 7

Assignment A.3

Nethra Sambamoorthi and Sudha BG

June 2, 2024

Introduction

This assignment aimed to compare the performance of entity and relation extraction using two different methods: SpaCy and a combination of Transformers and SpaCy. The objective was to preprocess the IMDB reviews, extract entities and relations, clean the outputs, and evaluate the performance of the two methods.

Data Extraction and Preprocessing

1. Data Extraction:
   - Extracted the first 10 rows from the IMDB reviews dataset to form the basis for the evaluation.
2. Text Cleaning:
   - Removed Extra Whitespace and Newlines: Replaced newline characters with spaces and reduced multiple spaces to a single space.
   - Handled Contractions: Expanded common English contractions to their full forms (e.g., "can't" to "cannot").
   - Removed Special Characters: Removed digits, special characters, and punctuation marks that were not part of the entities.
   - Normalized Text: Removed apostrophes, quotation marks, and extra spaces. Handled salutations (e.g., "Mr." to "Mr") and removed references to external texts.
   - Removed Paragraph Numbers and Extra Spaces: Cleaned the text by removing paragraph numbers, extra spaces, and unnecessary newlines.
3. Sentence Splitting:
   - Split the cleaned text into individual sentences to facilitate easier processing. This involved using regular expressions to split text at sentence-ending punctuation marks.

Entity and Relation Extraction

1. Entity Extraction using SpaCy:
   - Used SpaCy's pre-trained model to extract entities from each sentence. This model identified entities such as people, organizations, locations, and more.
2. Relation Extraction using SpaCy:
   - Defined custom rules to extract relations between entities based on syntactic dependency parsing. Relations were identified by examining the syntactic heads of entities and their proximity within the sentence.

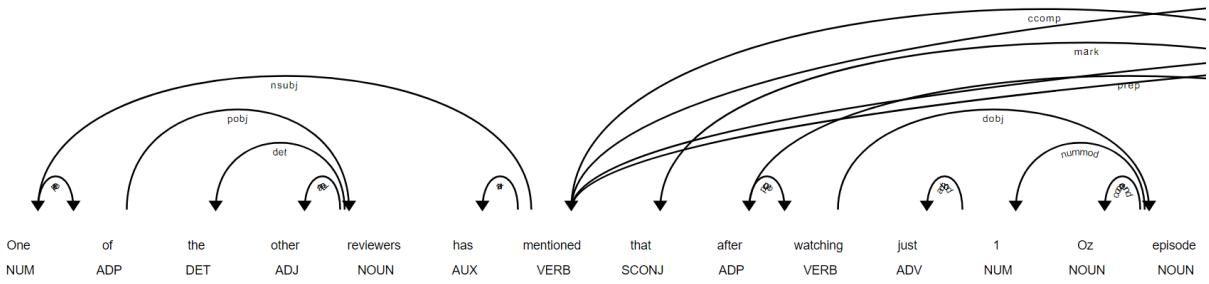
```
[259] df.spacy_entities[2] # validate output

[('summer weekend', 'DATE'),
 ('Match Point 2 Risk Addiction', 'ORG'),
 ('Woody Allen', 'PERSON'),
 ('one', 'CARDINAL'),
 ('Woody', 'NORP'),
 ('years', 'DATE'),
 ('Scarlet Johanson', 'ORG'),
 ('Devil Wears Prada', 'ORG'),
 ('Superman', 'GPE')]
```

```
[260] df.spacy_relations[2] # validate output

[('Devil Wears Prada', 'Superman'),
 ('Woody', 'years'),
 ('one', 'years'),
 ('one', 'Woody')]
```

3. Plotted Dependency Parsing for one the movies:



4. Entity Extraction using Transformers:
   - Employed a pre-trained BERT model fine-tuned for named entity recognition to extract entities from each sentence. The Hugging Face library's pipeline for NER was used to streamline this process.

4. Relation Extraction using SpaCy for Transformer Entities:

- Used the SpaCy model to extract relations from the sentences, leveraging the entities identified by the Transformer model. This involved processing the same sentences and applying similar syntactic dependency rules.

```
df['transformer_entities'][0]
```

```
[('Oz', 'I-MISC'),
 ('Oz', 'I-MISC'),
 ('O', 'I-ORG'),
 ('##Z', 'I-ORG'),
 ('Oswald', 'I-LOC'),
 ('Pen', 'I-ORG'),
 ('##ite', 'I-ORG'),
 ('##ntary', 'I-ORG'),
 ('Emerald', 'I-LOC'),
 ('City', 'I-LOC'),
 ('Em', 'I-LOC'),
 ('City', 'I-LOC'),
 ('A', 'I-MISC'),
 ('##ryan', 'I-MISC'),
 ('Muslims', 'I-MISC'),
 ('Latino', 'I-MISC'),
 ('Christians', 'I-MISC'),
 ('Italians', 'I-MISC'),
 ('Irish', 'I-MISC'),
 ('O', 'I-ORG'),
 ('##Z', 'I-ORG'),
 ('Oz', 'I-MISC'),
 ('Oz', 'I-MISC')]
```

```
df['trans_relations'][0]
```

```
[('Em City', 'be', 'Italians', 'GPE', 'NORP'),
 ('Aryans Muslims', 'gangsta', 'Latinos Christians', 'NORP', 'NORP'),
 ('Italians', 'be Italians', 'Irish', 'NORP', 'NORP')]
```

Cleaning and Aligning Outputs

The outputs, including entities and relations, from SpaCy and Transformers+SpaCy were in different formats and contain duplicates. It was crucial to clean these outputs to ensure accurate evaluation and comparison.

1. Cleaned and Aligned Entities:

- Removed Sub-token Markers: Cleaned sub-token markers (e.g., "") from the Transformer model's output.
- Merged Fragmented Tokens: Combined fragmented tokens into complete entities to match the format used by SpaCy.

- Removed Duplicates: Ensured unique entities and relations by removing duplicates from the outputs using sets to track seen entities and relations.

2. Aligned Entity Types:
   - Mapped entity types from the Transformer model to those used by SpaCy to ensure consistency in evaluation.

Evaluation

1. Defined Evaluation Metrics:
   - Established Precision, Recall, and F1 Score as the metrics for both entity and relation extraction. These metrics were used to evaluate the accuracy (precision), completeness (recall), and overall performance (F1 score) of the extractions.
2. Evaluated Performance:
   - Compared the entities and relations extracted by the Transformer model against those extracted by SpaCy, which served as the gold standard. The comparison involved calculating the number of true positives (correctly identified entities/relations), false positives (incorrectly identified entities/relations), and false negatives (missed entities/relations).

Results

1. SpaCy Extraction:
   - Extracted entities and relations using SpaCy provided the gold standard for evaluation.
2. Transformers + SpaCy Extraction:
   - Extracted entities using a Transformer model and relations using SpaCy were compared against the gold standard.
3. Evaluation Metrics:

- Calculated precision, recall, and F1 scores for both entities and relations to measure the accuracy and completeness of the Transformer-based method against the SpaCy baseline. The detailed output showed the differences in performance, highlighting areas where the Transformer model either succeeded or failed in comparison to SpaCy.
    - 'entity_precision': 0.01813186813186813,
    - 'entity_recall': 0.025757575757575757,
    - 'entity_f1': 0.020790020790020788,
    - 'relation_precision': 0.0,
    - 'relation_recall': 0.0,
    - 'relation_f1': 0.0

Conclusion

The evaluation of entity and relation extraction using a combination of Transformers and SpaCy against SpaCy as the gold standard reveals significant performance discrepancies. The entity extraction using Transformers yielded a precision of 0.0181, recall of 0.0258, and an F1 score of 0.0208. These results indicate that the Transformer model struggled to accurately identify entities, with a substantial number of false positives and false negatives. The low precision suggests that many of the entities identified by the Transformers were incorrect, while the low recall shows that the model missed a large number of entities present in the gold standard.

Furthermore, the relation extraction results were notably poor, with both precision and recall at 0.0, leading to an F1 score of 0.0. This implies that the Transformer-based method failed to correctly identify any relations between entities, reflecting a critical gap in its capability compared to SpaCy.

Overall, while the Transformer model demonstrated some ability to recognize entities, it fell short in both precision and recall, and it was ineffective in extracting relations. These findings highlight the need for further optimization and possibly more sophisticated models or hybrid approaches to improve the accuracy and completeness of entity and relation extraction tasks.

Appendix A1

Complete Output:

Text: One of the other reviewers has mentioned that after watching just 1 Oz episode you will be hooked They are right as this is exactly what happened with me The first thing that struck me about Oz was its brutality and unflinching scenes of violence which set in right from the word GO Trust me this is not a show for the faint hearted or timid This show pulls no punches with regards to drugs sex or violence Its is hardcore in the classic use of the word It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary It focuses mainly on Emerald City an experimental section of the prison where all the cells have glass fronts and face inwards so privacy is not high on the agenda Em City is home to many Aryans Muslims gangstas Latinos Christians Italians Irish and more so scuffles death stares dodgy dealings and shady agreements are never far away I would say the main appeal of the show is due to the fact that it goes where other shows would not dare Forget pretty pictures painted for mainstream audiences forget charm forget romance OZ does not mess around The first episode I ever saw struck me as so nasty it was surreal I could not say I was ready for it but as I watched more I developed a taste for Oz and got accustomed to the high levels of graphic violence Not just violence but injustice Watching Oz you may become comfortable with what is uncomfortable viewing thats if you can get in touch with your darker side

Spacy Entities: {('Emerald City', 'GPE'), ('the Oswald Maximum Security State Penitentary It', 'ORG'), ('Aryans Muslims', 'NORP'), ('first', 'ORDINAL'), ('just 1 Oz', 'PERCENT'), ('Em City', 'GPE'), ('Italians', 'NORP'), ('One', 'CARDINAL'), ('Irish', 'NORP'), ('GO Trust', 'ORG'), ('Latinos Christians', 'NORP')}

Transformer Entities: {('Oz', 'MISC'), ('ntary', 'ORG'), ('Muslims', 'MISC'), ('Emerald', 'GPE'), ('Italians', 'MISC'), ('Penite', 'ORG'), ('Oz', 'ORG'), ('City', 'GPE'), ('Em', 'GPE'), ('Christians',

'MISC'), ('City', 'MISC'), ('Oswald', 'ORG'), ('Irish', 'ORG'), ('OZ', 'ORG'), ('Aryan', 'MISC'), ('Latino', 'MISC')}

Spacy Relations: {('Em City', 'Aryans Muslims'), ('Latinos Christians', 'Italians'), ('Em City', 'Italians'), ('Aryans Muslims', 'Irish'), ('Italians', 'Irish'), ('Aryans Muslims', 'Latinos Christians'), ('the Oswald Maximum Security State Penitentary It', 'Emerald City'), ('Aryans Muslims', 'Italians'), ('Latinos Christians', 'Irish')}

Transformer Relations: {('Aryans Muslims', 'gangsta', 'Latinos Christians', 'NORP', 'NORP'), ('Italians', 'be Italians', 'Irish', 'NORP', 'NORP'), ('Em City', 'be', 'Italians', 'GPE', 'NORP')}

Text: A wonderful little production The filming technique is very unassuming very old time BBC fashion and gives a comforting and sometimes discomforting sense of realism to the entire piece The actors are extremely well chosen Michael Sheen not only has got all the polari but he has all the voices down pat too You can truly see the seamless editing guided by the references to Williams diary entries not only is it well worth the watching but it is a terrificly written and performed piece A masterful production about one of the great master of comedy and his life The realism really comes home with the little things the fantasy of the guard which rather than use the traditional dream techniques remains solid then disappears It plays on our knowledge and our senses particularly with the scenes concerning Orton and Halliwell and the sets are terribly well done

Spacy Entities: {('about one', 'CARDINAL'), ('BBC', 'ORG'), ('Williams', 'PERSON'), ('Halliwell', 'ORG'), ('Michael Sheen', 'PERSON'), ('Orton', 'ORG')}

Transformer Entities: {('BBC', 'PERSON'), ('well', 'PERSON'), ('Williams', 'PERSON'), ('Orton', 'PERSON'), ('Halli', 'PERSON'), ('Sheen', 'PERSON'), ('Michael', 'PERSON')}

Spacy Relations: {('Orton', 'Halliwell')}

Transformer Relations: {('Orton', 'concern Orton', 'Halliwell', 'ORG', 'ORG')}

Text: I thought this was a wonderful way to spend time on a too hot summer weekend sitting in the air conditioned theater and watching a light hearted comedy The plot is simplistic but the dialogue is witty and the characters are likable While some may be disappointed when they realize this is not Match Point 2 Risk Addiction I thought it was proof that Woody Allen is still fully in control of the style many of us have grown to love This was the most I would laughed at one of Woody comedies in years While I have never been impressed with Scarlet Johanson in this she managed to tone down her sexy image and jumped right into a average but spirited young woman This may not be the crown jewel of his career but it was wittier than Devil Wears Prada and more interesting than Superman a great comedy to go see with friends

Spacy Entities: {('Match Point 2 Risk Addiction', 'ORG'), ('Woody', 'NORP'), ('Scarlet Johanson', 'ORG'), ('Devil Wears Prada', 'ORG'), ('one', 'CARDINAL'), ('years', 'DATE'), ('summer weekend', 'DATE'), ('Superman', 'GPE'), ('Woody Allen', 'PERSON')}

Transformer Entities: {('Superman', 'MISC'), ('Wears', 'MISC'), ('Point', 'MISC'), ('Scarlet', 'PERSON'), ('Prada', 'MISC'), ('Allen', 'PERSON'), ('2', 'MISC'), ('Addiction', 'MISC'), ('Devil', 'MISC'), ('Woody', 'PERSON'), ('Johanson', 'PERSON'), ('Match', 'MISC'), ('Risk', 'MISC')}

Spacy Relations: {('Devil Wears Prada', 'Superman'), ('Woody', 'years'), ('one', 'years'), ('one', 'Woody')}

Transformer Relations: set()

Text: Basically there is a family where a little boy thinks there is a zombie in his closet his parents are fighting all the time This movie is slower than a soap opera and suddenly Jake decides to become Rambo and kill the zombie OK first of all when you are going to make a film you must Decide if its a thriller or a drama As a drama the movie is watchable Parents are divorcing arguing like in real life And then we have Jake with his closet which totally

ruins all the film I expected to see a BOOGEYMAN similar movie and instead i watched a drama with some meaningless thriller spots 3 out of 10 just for the well playing parents descent dialogs As for the shots with Jake just ignore them

Spacy Entities: {('BOOGEYMAN', 'ORG'), ('3', 'CARDINAL'), ('10', 'CARDINAL'), ('first', 'ORDINAL'), ('Jake', 'CARDINAL'), ('Rambo', 'PERSON')}

Transformer Entities: {('GE', 'MISC'), ('BO', 'MISC'), ('Jake', 'MISC'), ('Jake', 'PERSON'), ('MA', 'MISC'), ('Rambo', 'MISC'), ('N', 'MISC')}

Spacy Relations: {('3', '10')}

Transformer Relations: set()

Text: Petter Mattei Love in the Time of Money is a visually stunning film to watch Mr Mattei offers us a vivid portrait about human relations This is a movie that seems to be telling us what money power and success do to people in the different situations we encounter This being a variation on the Arthur Schnitzler play about the same theme the director transfers the action to the present time New York where all these different characters meet and connect Each one is connected in one way or another to the next person but no one seems to know the previous point of contact Stylishly the film has a sophisticated luxurious look We are taken to see how these people live and the world they live in their own habitat The only thing one gets out of all these souls in the picture is the different stages of loneliness each one inhabits A big city is not exactly the best place in which human relations find sincere fulfillment as one discerns is the case with most of the people we encounter The acting is good under Mr Mattei direction Steve Buscemi Rosario Dawson Carol Kane Michael Imperioli Adrian Grenier and the rest of the talented cast make these characters come alive We wish Mr Mattei good luck and await anxiously for his next work

Spacy Entities: {('Petter Mattei Love', 'ORG'), ('Carol Kane', 'PERSON'), ('Michael Imperioli', 'PERSON'), ('Arthur Schnitzler', 'PERSON'), ('one', 'CARDINAL'), ('Stylishly',

'DATE'), ('Mr Mattei', 'PERSON'), ('Steve Buscemi Rosario', 'PERSON'), ('Mr Mattei', 'ORG'), ('Mattei', 'PERSON'), ('New York', 'GPE')}

Transformer Entities: {('Gren', 'PERSON'), ('ier', 'PERSON'), ('itz', 'PERSON'), ('Busce', 'PERSON'), ('Adrian', 'PERSON'), ('New', 'GPE'), ('Michael', 'PERSON'), ('Time', 'MISC'), ('Mattei', 'PERSON'), ('ler', 'PERSON'), ('Schn', 'PERSON'), ('Kane', 'PERSON'), ('i', 'PERSON'), ('Imper', 'PERSON'), ('York', 'PERSON'), ('Love', 'MISC'), ('mi', 'PERSON'), ('of', 'PERSON'), ('in', 'MISC'), ('Petter', 'PERSON'), ('Steve', 'PERSON'), ('Arthur', 'PERSON'), ('Dawson', 'PERSON'), ('iol', 'PERSON'), ('Carol', 'PERSON'), ('Rosario', 'PERSON')}

Spacy Relations: {('Mr Mattei', 'Steve Buscemi Rosario'), ('Steve Buscemi Rosario', 'Carol Kane'), ('Steve Buscemi Rosario', 'Michael Imperioli'), ('Carol Kane', 'Michael Imperioli')}

Transformer Relations: {('Carol Kane', 'Imperioli Grenier', 'Michael Imperioli', 'PERSON', 'PERSON'), ('Steve Buscemi Rosario', 'Carol Imperioli', 'Carol Kane', 'PERSON', 'PERSON')}

Text: Probably my all time favorite movie a story of selflessness sacrifice and dedication to a noble cause but it is not preachy or boring It just never gets old despite my having seen it some 15 or more times in the last 25 years Paul Lukas performance brings tears to my eyes and Bette Davis in one of her very few truly sympathetic roles is a delight The kids are as grandma says more like dressed up midgets than children but that only makes them more fun to watch And the mother slow awakening to what happening in the world and under her own roof is believable and startling If I had a dozen thumbs they would all be up for this movie

Spacy Entities: {('the last 25 years', 'DATE'), ('one', 'CARDINAL'), ('Paul Lukas', 'PERSON'), ('Bette Davis', 'PERSON'), ('a dozen', 'CARDINAL')}

Transformer Entities: {('Lukas', 'PERSON'), ('Davis', 'PERSON'), ('Bette', 'PERSON'), ('Paul', 'PERSON')}

Spacy Relations: {('the last 25 years', 'Paul Lukas'), ('Bette Davis', 'one')}

Transformer Relations: set()

Text: I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the kid excitement in me I grew up on black and white TV and Seahunt with Gunsmoke were my hero every week You have my vote for a comeback of a new sea hunt We need a change of pace in TV and this would work for a world of under water adventure Oh by the way thank you for an outlet like this to view many viewpoints about TV and the many movies So any ole way I believe I have got what I wanna say Would be nice to read some more plus points about sea hunt If my rhymes would be 10 lines would you let me submit or leave me out to be in doubt and have me to quit If this is so then I must go so lets do it

Spacy Entities: {('Seahunt with Gunsmoke', 'ORG'), ('10', 'CARDINAL'), ('today', 'DATE')}

Transformer Entities: {('ke', 'ORG'), ('Gunsmo', 'MISC'), ('t', 'MISC'), ('Seahun', 'MISC')}

Spacy Relations: set()

Transformer Relations: set()

Text: This show was an amazing fresh innovative idea in the 70 when it first aired The first 7 or 8 years were brilliant but things dropped off after that By 1990 the show was not really funny anymore and it is continued its decline further to the complete waste of time it is today It truly disgraceful how far this show has fallen The writing is painfully bad the performances are almost as bad if not for the mildly entertaining respite of the guest hosts this show probably would not still be on the air I find it so hard to believe that the same creator that hand selected the original cast also chose the band of hacks that followed How can one recognize such brilliance and then see fit to replace it with such mediocrity I felt I must give 2 stars out of respect for the original cast that made this show such a huge success As it is now the show is just awful I cannot believe it is still on the air

Spacy Entities: {('1990', 'DATE'), ('The first 7 or 8 years', 'DATE'), ('2', 'CARDINAL'), ('one', 'CARDINAL'), ('first', 'ORDINAL'), ('today', 'DATE'), ('70', 'CARDINAL')}

Transformer Entities: set()

Spacy Relations: {('first', 'The first 7 or 8 years'), ('70', 'The first 7 or 8 years'), ('70', 'first')}

Transformer Relations: set()

Text: Encouraged by the positive comments about this film on here I was looking forward to watching this film Bad mistake I have seen 950 films and this is truly one of the worst of them it is awful in almost every way editing pacing storyline acting soundtrack The film looks cheap and nasty and is boring in the extreme Rarely have I been so happy to see the end credits of a film The only thing that prevents me giving this a 1 score is Harvey Keitel while this is far from his best performance he at least seems to be making a bit of an effort One for Keitel obsessives only

Spacy Entities: {('950', 'CARDINAL'), ('1', 'CARDINAL'), ('one', 'CARDINAL'), ('One', 'CARDINAL'), ('Harvey Keitel', 'PERSON'), ('Keitel', 'ORG')}

Transformer Entities: {('l', 'PERSON'), ('Harvey', 'PERSON'), ('Keite', 'PERSON')}

Spacy Relations: {('1', 'Harvey Keitel'), ('One', 'Keitel')}

Transformer Relations: set()

Text: If you like original gut wrenching laughter you will like this movie If you are young or old then you will love this movie hell even my mom liked it Great Camp

Spacy Entities: {('Great Camp', 'FAC')}

Transformer Entities: {('Great', 'MISC'), ('Camp', 'MISC')}

Spacy Relations: set()

Transformer Relations: set()