Open Information Extraction from Conversational Text in Movie Transcripts

Jim Chen, Rohan Thakur, Winston Lin

Final Project Report
W266 Natural Language Processing, Spring 2017

Abstract

We take advantage of the abundance and availability of movie transcripts as an opportunity to perform information extraction on conversational text. The dataset includes transcripts of popular Marvel films, such as the X-Men and Avengers franchises. Pronoun resolution and relationship extraction are performed with the help of annotations produced with the Google Cloud Natural Language API. Our model generates tuples representing binary relationships between characters and other entities in the movie. Performance is measured in terms of the precision with which correct entities are tagged to pronouns and valid relationships are created by our models.

Introduction

Most people leave superhero movies with several unanswered questions about the plot, characters, and the relationships between them. We have attempted to build some representation of the meaning for the Marvel Universe so that our model can answer simple questions about character relations. In addition, movie scripts are a good source of conversational data. We believe that this is a challenging project as it deals with messy, unlabeled data, and touches several sections of the course including information extraction, information retrieval, parsing and question answering, and possibly more. The relationships between characters are complex, and we deal with several problems such as temporal inference, and coreference resolution, optimizing around which prove critical to the accuracy of our question answering abilities. A natural extension of our project would be to perform information extraction and retrieval on real-life, conversational data, such as transcripts of trials and Congressional hearings. Such research could help public officials, researchers, journalists, and citizens better understand complex discussions about public policy and law enforcement that impact society.

Background

Whereas most traditional information extraction methods rely on manual annotation and the knowledge of relationships being searched for in advance, Open Information Extraction (Open IE) infers relationships and marks entities without the need for human intervention. Since traditional IE systems were domain specific, they were specifically designed from dependency trees and tuned NER systems to produce context specific results, but not designed to scale with corpus size and variety of articles.

Open IE is especially practical in cases where the corpus is large and relationships are indeterminable in advance, such as the web. The use of Open IE methods to extract information from Marvel scripts fits our needs nicely as we do not have access to labeled data to explore supervised learning methods, nor is it feasible to undertake an exercise of manual annotation or define generic relationships in advance.

One of the challenges of Open IE is evaluating the performance of different extraction systems. Measuring performance involves manually generating correct extractions from a corpus and comparing them to those predicted in order to calculate precision, often without addressing recall. Comparing performance across systems is also challenging since the associated corpora and tasks differ. Recent efforts have been made to address these issues by creating a benchmark corpus of extractions that can be used to evaluate a variety of systems, such as TextRunner, ReVerb, KrakeN, and Ollie.

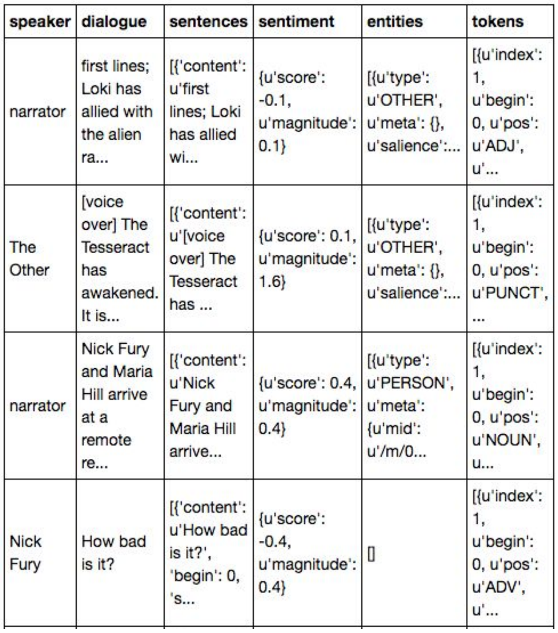
Methods

Data Collection and Preparation

Our dataset consisted of user-generated transcripts of Marvel films from Transcripts Wiki (http://transcripts.wikia.com) and user-submitted screenplays from the Internet Movie Script Database (IMSDb, http://www.imsdb.com). The transcripts were transformed into tuples of the form [speaker, dialogue]. Narrations and scene descriptions were treated as dialogue from a generic speaker, “narrator.” The results were saved to text files that were enriched using the Google Cloud Natural Language (GCNL) API. The API returned annotations containing speakers, sentences, tokens, part-of-speech tags, entities, and sentiment scores for each line of dialogue in a transcript. Table 1 shows a sample of the annotated transcripts produced by the API.

Table 1. Annotated transcripts created using Google Cloud Natural Language API

1. **Sentences:** content, sentiment score, sentiment magnitude of each sentence in the dialogue,
2. **Sentiment:** sentiment score, sentiment magnitude of the dialogue overall
3. **Entities:** name, mentions, salience, type of each entity in the dialogue
4. **Tokens:** part of speech, content, lemma, dependency index, dependency label, type of each token in the dialogue



speaker	dialogue	sentences	sentiment	entities	tokens
narrator	first lines; Loki has allied with the alien ra...	[{'content': u'first lines; Loki has allied wi...	{u'score': -0.1, u'magnitude': 0.1}	[{u'type': u'OTHER', u'meta': {}, u'salience':...	[{u'index': 1, u'begin': 0, u'pos': u'ADJ', u'...
The Other	[voice over] The Tesseract has awakened. It is...	[{'content': u'[voice over] The Tesseract has ...	{u'score': 0.1, u'magnitude': 1.6}	[{u'type': u'OTHER', u'meta': {}, u'salience':...	[{u'index': 1, u'begin': 0, u'pos': u'PUNCT', ...
narrator	Nick Fury and Maria Hill arrive at a remote re...	[{'content': u'Nick Fury and Maria Hill arrive...	{u'score': 0.4, u'magnitude': 0.4}	[{u'type': u'PERSON', u'meta': {u'mid': u'/m/0...	[{u'index': 1, u'begin': 0, u'pos': u'NOUN', u...
Nick Fury	How bad is it?	[{'content': u'How bad is it?', 'begin': 0, 's...	{u'score': -0.4, u'magnitude': 0.4}	[]	[{u'index': 1, u'begin': 0, u'pos': u'ADV', u'...

Task 1: Pronoun Resolution

The grammar and structure of dialogue allowed for simple rules-based models to resolve a significant proportion of pronouns in the corpus. Most lines of dialogue were written from a first- or second-person perspective and alternated between two or more speakers. Thus, pronouns in a dialogue often referred to the speaker of the selected line (“I”), speakers in adjacent or nearby lines (“you”). Since “I” and its forms resolved to the speaker of a line, our task was reduced to tagging references for second- (“you”) and third-person (“he/she”, “they”) pronouns.

Our baseline model resolved forms of “you” to a randomly selected speaker in the line before or after a given line and randomly associate all other pronouns with characters in the movie script. Our improved model extended the baseline model by selecting a speaker based on a probability distribution of speakers mentioned within a window of nearby lines (e.g., 10 lines before and after the selected line). The improved model also attempted to tag plural and third-person pronouns using a probabilistic sampling approach. We could not find enough leverage to distinguish second person singular forms from plural ones and treated all occurrences as singular. To improve consistency and enhance the performance of the latter relationship extraction task, we tied all appropriate entity mentionings and pronouns to standardized movie character names. The baseline and improved models were applied to five transcripts, and the resulting tags were appended to the dataset. The resulting dataset with resolved pronouns was then fed into our relationship extraction model.

To evaluate the models, we manually inspected the pronouns and the speakers tagged by the model for 20 lines randomly sampled from each transcript. Sampled lines were printed individually with context (e.g., two lines immediately before and after), along with a list of the pronouns in the line and the tagged speakers to be evaluated. For each line, we typed the count of correctly tagged pronouns after inspection, which a wrapper function then appended to the selected line in our dataset. Precision was calculated as the total number of correctly tagged pronouns divided by the total number of tagged pronouns for each sample and across all samples.

Task 2: Relationship Extraction

We used the tagged entities from the pronoun resolution model to enrich our dataset before performing relationship extraction. In order to better define the relationship extraction task, we arrived at a few categories which represent the relationships we are interested in identifying. Our model, apart from identifying relationships between two entities, also classified the relationships as being in one of our chosen relationship categories, and precision was measured by labeling a relationship as correct if it was a valid relationship between characters and was classified correctly into one of the predefined categories. We largely ignored dialogues spoken by the narrator because the dynamics of those dialogues differed from standard character lines. The five relationship categories were defined as follows:

1. **For positive, negative and mixed mentioning**: Our model looks at the sentiment score and magnitude of the line. If the sentiment score is lower than -0.4 or higher than 0.4, a negative mention or a positive mention, respectively, is created between the speaker and characters in the dialogue. If the sentiment is within that range but magnitude is high, a mixed mention is created instead.
2. **For location mention:** When an entity tagged as a place is identified, we identify whether the place is the subject or object of the sentence and accordingly classify the relation between the other entity and the place (using the verb phrase relation).
3. **For identity mention:** We leverage entity tagging provided by GCNL API and the result of our pronoun resolution model to associate characters with identities they are associated with, such as professions, family relationship, etc.

For evaluation, wrapper function samples 20 lines each from five movie scripts. Relationships and their predicted relationship category for those lines were presented to the person evaluating. For each example, the person evaluating marked an example as correct if the relation made sense, and also the category predicted was correct. At the end of the exercise, the function calculated and output a precision score.

We treated the evaluations of pronoun resolution and relationship extraction independently to better measure the performance of each task. However, pronoun resolution inevitably had an impact on relationship extraction, the downstream task. If the pronoun resolution model incorrectly tagged a character to a pronoun, the error propagated through the relationship extraction model and produced invalid relationships.

Task 3: Question Answering

We built a simple query engine on top of the two models we have described above. The engine supports two types of queries. Users can directly search for specific relationships in the form of [entity 1, entity 2, relation class] to view all lines of dialogues that contain such relations. Users can also choose to enter a free form query. The engine will then compute similarity scores of the query against relationships identified for each dialogue and return a list of dialogues that contain best matching relations. Although we were not able to address performance of the query engine due to time constraints, we acknowledge that it would be an important issue to address in future iterations of this project.

Results and Discussion

Task 1: Pronoun Resolution

The baseline model tagged forms of "I" and "you" with random nearest speakers and yielded an overall precision of 0.47. Incorrectly tagged examples consisted mostly of plural ("we", "they") and third-person pronouns ("he", "she", "they"), which were tagged with random speakers (precision of nearly 0). The improved model tagged pronouns forms of "I" and "you" with speakers based on a distribution and yielded an overall precision of 0.65, improving the baseline results by almost 50%. Again, most of the incorrect examples were due to plural pronouns, as sets of speakers were sampled from nearby speakers based on their mentions within a window, rather than actual context. This was a limitation of our method, which did not include all scene information from a given movie. Tagging of singular third-person pronouns, however, was still improved using the weighted method. Table 2 shows the performance of our models for pronoun resolution.

Table 2: Precision of Pronoun Resolution Models

Transcript	**Random Speaker (baseline)**	**Random Nearest Speakers**	**Weighted Sample Nearby Speakers**
The Avengers	0.05 (2/43)	0.33 (14/43)	0.49 (18/37)
Avengers: Age of Ultron	0.04 (3/71)	0.30 (21/71)	0.66 (42/64)
X-Men: Apocalypse	0.02 (1/55)	0.73 (40/55)	0.87 (46/53)
X-Men: Days of Future Past	0.00 (0/45)	0.53 (24/45)	0.58 (25/43)
X-Men: The Last Stand	0.00 (0/45)	0.49 (22/45)	0.59 (24/41)
Overall	**0.02 (6/259)**	**0.47 (121/259)**	**0.65 (155/238)**

Task 2. Relationship Extraction

For relationship extraction, the baseline was a naive model that traversed through all parse trees and considered each noun-verb-objective triplets as a relation. This model produced a poor precision score of 0.1. Our improved relationship extraction model performed significantly better, achieving a precision of 0.46 across scripts. The tagged entities from the pronoun resolution task added more information about the entities and enabled the relationship extraction model to produce more valid, meaningful relationships. In addition, using the sentiment score and entity tags generated by the GCNL API allowed the model to factor in sentiment between characters in order to better define relationships between them. Table 3 shows the performance of the relationship extraction models.

Table 3: Precision of Relation Extraction Model

Transcript	**Relation Extraction - no pronoun resolution input**	**Relation Extraction - with pronoun resolution input**
Captain America - Civil War	0.36 (23/64)	0.48 (29/60)
Captain America - The First Avenger	0.43 (23/53)	0.35 (19/35)
Captain America - The Winter Soldier	0.36 (24/67)	0.41 (28/68)
Iron Man 3	0.37 (22/60)	0.50 (13/26)
X-Men: The Last Stand	0.45 (29/65)	0.40 (17/42)
Overall	**0.39 (121/309)**	**0.46 (106/231)**

Conclusion

We demonstrated that a rule-based model can extract a reasonable amount of information from unlabeled movie dialogues. We believe this approach can extend to other dialogue texts and maintain comparable performance. The models, with the query engine, allowed the possibility of identifying sections of text of interest efficiently. A list of future tasks can potentially improve our model further:

- Substitute GCNL API with a language parse trained specifically on conversational data
- Gather additional information about the characters through Wikipedia or other sources to incorporate gender and other contextual information into the models
- Traverse dependency tree provided by GCNL API in order to resolve pronouns referring to characters in same dialogue.
- Attempt to split transcripts into scenes in order to limit potential references for pronouns and improve resolution of plural pronouns

We also acknowledge the inherent limitation of analyzing movie dialogue texts and document a list of obstacles that are difficult or impossible to overcome using only the text data:

- Most dialogues consist of short sentences and frequent change of speakers, making pronoun/coreference resolution more difficult
- Text data removes tone, and other voice-related information that are useful/critical in analyzing sentiments of the dialogues
- Information about movies are not always captured through conversations. Actions by the characters are important parts of movies, and not incorporating them into analysis can lead to obvious gaps.

References

1. Open Information Extraction from the Web (review article): http://allenai.org/content/team/orene/etzioni-cacm08.pdf

2. Open Information Extracton from the Web (research paper, TextRunner): http://aiweb.cs.washington.edu/research/projects/aiweb/media/papers/tmpYZBSTp.pdf

3. The Tradeoffs between Open and Traditional Relationship Extraction (pp.28-36): http://anthology.aclweb.org/P/P08/P08-1.pdf#page=72

4. Creating a Large Benchmark for Open Information Extraction: https://www.aclweb.org/anthology/D/D16/D16-1252.pdf

5. Open Language Learning for Information Extraction (Ollie): http://homes.cs.washington.edu/~mausam/papers/emnlp12a.pdf

6. Identifying relationships for Open Information Extraction (ReVerb): http://ml.cs.washington.edu/www/media/papers/reverb_emnlp2011.pdf

7. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language: https://dada.cs.washington.edu/qasrl/docs/emnlp2015_hlz.pdf

8. Google Cloud Natural Language API: https://cloud.google.com/natural-language/docs/

9. Optimizing Algorithms for Pronoun Resolution: https://pdfs.semanticscholar.org/272a/f491ddf628b827b97d6ea24be36f82b3e1d4.pdf

10. An Algorithm for Pronominal Anaphora Resolution: http://www.aclweb.org/anthology/J94-4002