Evaluating Deep Learning Approaches for Character Identification in Multiparty Dialogues Progress Report CS585: Intro to NLP

Krishna Prasad Sankaranarayanan and Sree Harsha Ramesh

Department of Computer Science
University of Massachusetts Amherst
Amherst MA, 01002, USA
{ksankaranara,shramesh}@cs.umass.edu

1 Outline

Character identification is an entity linking task that identifies each mention as a certain character in multiparty dialogue where mentions are typically nominals referring to a person and entities maybe speakers themselves or even external characters. Identifying such mentions as real characters requires cross-document entity resolution, which makes this task challenging. This task involves coreference resolution which clusters together the mentions corresponding to the same referent followed by an entity linking stage where the clusters of mentions are mapped to their corresponding entities. Historically, coreference models have been trained on the *NewsWire* dataset which is not as rich in terms of the complexity of the coreferences as those in multiparty dialogues. As has been the norm on various natural language processing tasks, deep-learning models are the state-of-the-art in coreference resolution as well. However, coreference resolution systems have been shown not to handle dialogues as well ([2],[3]). This motivates us to extend and evaluate the existing coreference systems including rule-based, statistical and deep-learning based models, for the annotated TV Show transcripts dataset released as part of SemEval 2018 Task-4 [1].

2 Related Work

Henry Y. Chen et al [4] proposes a deep learning approach to coreference resolution and entity linking for character identification. It introduces a new agglomerative convolutional neural network for returning mention and mention pair embeddings. Entity links are then mapped to each referent separately by cluster embeddings. This method takes into consideration, 20 labels viz. top 9 characters and an unknown label. It emphasizes on the intuition that the coreference resolution accuracy depends upon the size of the clusters. The combined

[1] https://competitions.codalab.org/competitions/17310

implementation of cluster and mention embeddings improved upon the singular use of mention embeddings in terms of accuracy. Although, agglomerative CNN being an incremental feature approach has its fair share of advantages in terms of word embeddings, the existing approach still lacks the handling of plurals and collective nouns. This scope for overall improvement is aimed to be leveraged by us while looking for extending and evaluating the deep learning approaches.

Clark and Manning [5] introduces an entity centric system using mention pair models as features. Agglomerative clustering is used to build coreference chains formed by merging pairs of clusters at each step. A key aspect of any incremental coreference system is its local decisions. Using this to full advantage, costs are assigned to each action which are in turn trained using a cost-sensitive classification. For the ranking model, the current mention is matched to the candidate antecedents simultaneously competing with each other. The resultant prediction model depends upon the previous actions which violate the IID(independent and identically distributed) assumption of statistical learning. Hence, imitation learning is used to classify whether a particular action is the one the expert policy would take at the current state.

Kevin Clarke [6] proposes a new approach to coreference resolution using distributed word representations. An incremental coreference system is defined which acts as a feed forward neural network for mention clusters rather than mention pairs. The usage of mention pairs does not enforce transitivity and therefore relies only on local pairwise information to make coreference decisions. Mention clusters on the other hand facilitate previous coreference decisions to inform the latest ones. This is an extension to [2], wherein features are created between mention clusters using the pairwise probabilities of the mention pair model. This is extended by consideration of all features from vector representations of mention pairs to produce cluster level features. The actual benefits of deep learning on coreference are the lack of hand engineered features. This is leveraged by Clarke in creating a simple feature set which outperforms state-of-the-art approaches.

Sam Wiseman et al [7] presents a mention ranking model for coreference resolution. It emphasizes on anaphoricity detection and antecedent ranking with respect to learning feature representations. The training model using backpropagation is preceded by a pre-training segment comprising of two tasks viz. anaphoricity detection and antecedent ranking. The mention ranking model is trained with the slack-rescaled max-margin training objective which facilitates separation between highest scoring true and false antecedents of the current mention. A major challenge of coreference systems is resolving an anaphoric mention that has no previous head term. This paper intuitively evaluates the possibility of overcoming this challenge by means of non local decision making. It provides a conclusion that pronouns may not be the only coreferent mentions causing these errors and therefore a local model can also be tweaked with respect to a loss function to achieve this.

3 Preliminary Experiments

System	Document:episode_delim			Document:scene_delim		
	MUC	$CEAF_e$	B^3	MUC	$CEAF_e$	B^3
Stanford Deterministic Coref Resolution System	5.12	2.28	1.69	8.69	7.04	3.43
Stanford Neural Coref Resolution System	4.68	2.91	1.19	10.32	5.65	2.23

Table 1. Coreference resolution results on a sub-set of the Friends dataset.

Since the precursor to character identification is identification of entity mention clusters, we experimented with two of the open-source state-of-the-art coreference systems which identify coreferant words or phrases in the discourse, that could later be linked to characters based on some characteristics of the mentions in a group.

3.1 Stanford Deterministic Coreference Resolution System

This is based on the multi-pass sieve system proposed by [8]. It is composed of multiple sieves of linguistic rules that are in the orders of high-to-low precision and low-to-high recall. Information regarding mentions, such as plurality, gender, and parse tree, is extracted during mention detection and used as global features. Pairwise links between mentions are formed based on defined linguistic rules at each sieve in order to construct coreference chains and mention clusters. We created episode-delimited and scene-delimited documents and ran the rule-based deterministic coreference resolution system[2].

3.2 Stanford Neural Coreference Resolution System

This system is based on the neural coreference system proposed in [9] and [10]. We used the pre-trained models which were trained on dataset released as part of the ConLL-2012 shared task [3]. Thus, the dense vector representations learned for mention pairs in the CoNLL dataset were used to model the entity level information on our dataset. It uses minimal number of hand-engineered features compared to the rule-based system described in Section 3.1.

3.3 Coreference Evaluation Metrics

The 4 systems trained as part of the preliminary experiments, were all evaluated with the official CoNLL scorer[4] on the three metrics for measuring coreference resolution: MUC, B^3, $CEAF_e$. The results can be seen in Table 1.

[2] https://nlp.stanford.edu/software/dcoref.html
[3] http://conll.cemantix.org/2012/data.html
[4] https://github.com/conll/reference-coreference-scorers

4 Scope and Approach

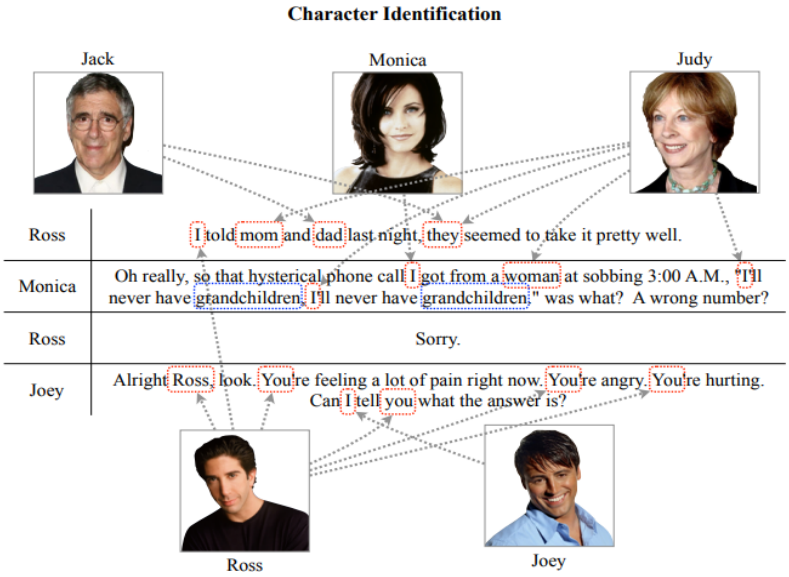

Fig. 1. An example of character identification. All three speakers are introduced as characters before the conversation (Ross, Monica, and Joey), and two more characters are introduced during the conversation (Jack and Judy). The goal of this task is to identify each mention as one or more of these characters.

The overall task of character identification in a multiparty discourse setting, could be divided into two sub-tasks –coreference resolution and entity linking. By integrating the two modules we propose to create a system which richly annotates the dialog data, by mapping mentions to their characters introduced during the discourse.

4.1 Mention Detection

To identify the mentions in a given utterance, a rule-based mention detector would be built using features such as dependency relations and named entities. As described in [1], a word sequence is considered a mention if it is a person named entity, or it is a pronoun or possessive pronoun excluding it*, or if it is in the personal noun dictionary chosen from Freebase[5] and DBpedia [6].

[5] https://developers.google.com/freebase/
[6] http://wiki.dbpedia.org/

4.2 Coreference Resolution

We would be basing our coreference resolution model on the deep learning approach introduced by [4] which involves learning mention and mention pair embeddings using convolutional neural networks. These embeddings would be used to get cluster embeddings for the subsequent stage of entity linking.

4.3 Entity Linking

In the previous stage, coreference resolution groups mentions into clusters, but it does not assign character labels to the clusters, which is required for character identification. Thus, an entity linking model is required that takes the mention embeddings and the mention-pair embeddings generated by the CNN and classifies each mention to one of the character labels. This would involve training a feed-forward neural network with back-propagation to classify each of the mentions to an entity label.

5 Dataset

The dataset released as part of SemEval-2018, would be used for training and dev-testing the system. The validation dataset would be released in January 2018, so we'd be holding back some data to report our accuracy on. The 2018 dataset comprises first two seasons of the TV show - Friends, annotated for this task. We would also be using one season of the TV show Big Bang Theory [7] released by the shared-task organizers in 2016. Each season consists of episodes, each episode comprises scenes, and each scene is segmented into sentences.

5.1 Data Format

All datasets follow the CoNLL 2012 Shared Task data format and the following are the columns for every token in an utterance.

1. Document ID: name of the show-season ID-episode ID (e.g., friends-s01e01).
2. Scene ID: the ID of the scene within the episode.
3. Token ID: the ID of the token within the sentence.
4. Word form: the tokenized word.
5. Part-of-speech tag: the part-of-speech tag of the word (auto generated).
6. Constituency tag: the Penn Treebank style constituency tag (auto generated).
7. Lemma: the lemma of the word (auto generated).
8. Frameset ID: not provided.
9. Word sense: not provided.
10. Speaker: the speaker of this sentence.
11. Named entity tag: the named entity tag of the word (auto generated).
12. Entity ID: the entity ID of the mention, that is consistent across all documents.

[7] https://github.com/emorynlp/character-mining/blob/master/md/corpus.md

6 Evaluation

As required by the shared task submission, we would be reporting the following metrics on the held-out dataset, using the provided evaluation script.

1. The label accuracy considering only 7 entities, that are the 6 main characters (Chandler, Joey, Monica, Phoebe, Rachel, and Ross) and all the others as one entity.
2. The macro average between the F1 scores of the 7 entities.
3. The label accuracy considering all entities, where characters not appearing in the tranining data are grouped as one entity, others.
4. The macro average between the F1 scores of all entities.
5. The F1 scores for 7 entities.
6. The F1 scores for all entities.

7 Proposed Tools and Resources

1. SpaCy [8] for dependency parsing
2. Pattern [9] and Stanford CoreNLP [10] for parsing, tokenization, POS tags, chunking, PNP tags and lemmata.
3. Machine Learning functions and utilities in Python from : scikit-learn, keras, pyTorch, tensorflow, numpy, matplotlib etc.

References

1. Yu-Hsin Chen and Jinho D. Choi. "Character Identification on Multiparty Conversation: Identifying Mentions of Characters in TV Shows." SIGDIAL Conference. 2016.
2. Nobal B. Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett, and Brent Morgan. 2014. The DARE Corpus: A Resource for Anaphora Resolution in Dialogue Based Intelligent Tutoring Systems. In Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC14, pages 31993203.
3. Marco Rocha. 1999. Coreference Resolution in Dialogues in English and Portuguese. In Proceedings of the Workshop on Coreference and Its Applications. CorefApp99, pages 5360.
4. Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. "Robust Coreference Resolution and Entity Linking on Dialogues: Character Identification on TV Show Transcripts." Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). 2017.
5. Kevin Clark and Chris Manning. Entity-centric coreference resolution with model stacking. In Association of Computational Linguistics (ACL), 2015.
6. Kevin Clark. "Neural coreference resolution." (2015).

[8] https://spacy.io/
[9] https://www.clips.uantwerpen.be/pages/pattern-en
[10] https://stanfordnlp.github.io/CoreNLP/

7. Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. "Learning anaphoricity and antecedent ranking features for coreference resolution." Association for Computational Linguistics, 2015.
8. Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. In Computational Linguistics 39(4)
9. Kevin Clark and Christopher D. Manning. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In Proceedings of EMNLP.
10. Kevin Clark and Christopher D. Manning. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In Proceedings of the ACL.