Character Identification on Multiparty Conversation: Identifying Mentions of Characters in TV Shows

Henry (Yu-Hsin) Chen
Math and Computer Science
Emory University
Atlanta, GA 30322, USA
henry.chen@emory.edu

Jinho D. Choi
Math and Computer Science
Emory University
Atlanta, GA 30322, USA
jinho.choi@emory.edu

Abstract

This paper introduces a subtask of entity linking, called character identification, that maps mentions in multiparty conversation to their referent characters. Transcripts of TV shows are collected as the sources of our corpus and automatically annotated with mentions by linguistically-motivated rules. These mentions are manually linked to their referents through crowdsourcing. Our corpus comprises 543 scenes from two TV shows, and shows the inter-annotator agreement of $\kappa = 79.96$. For statistical modeling, this task is reformulated as coreference resolution, and experimented with a state-of-the-art system on our corpus. Our best model gives a purity score of 69.21 on average, which is promising given the challenging nature of this task and our corpus.

1 Introduction

Machine comprehension has recently become one of the main targeted challenges in natural language processing (Richardson et al., 2013; Hermann et al., 2015; Hixon et al., 2015). The latest approaches to machine comprehension show lots of promises; however, most of these approaches face difficulties in understanding information scattered across different parts of documents. Reading comprehension in dialogues is particularly hard because speakers take turns to form a conversation such that it often requires connecting mentions from multiple utterances together to derive meaningful inferences.

Coreference resolution is a common choice for making connections between these mentions. However, most of the state-of-the-art coreference resolution systems are not accustomed to handle dialogues well, especially when multiple participants are involved (Clark and Manning, 2015; Peng et al., 2015; Wiseman et al., 2015). Furthermore, linking mentions to one another may not be good enough for certain tasks such as question answering, which requires to know what specific entities that mentions refer to. This implies that the task needs to be approached from the side of entity linking, which maps each mention to one or more pre-determined entities.

In this paper, we introduce an entity linking task, called character identification, that maps each mention in multiparty conversation to its referent character(s). Mentions can be any nominals referring to humans. At the moment, there is no dialogue corpus available to train statistical models for entity linking using such mentions. Thus, a new corpus is created by collecting transcripts of TV shows and annotating mentions with their referent characters. Our corpus is experimented with a coreference resolution system to show the feasibility of this task by utilizing an existing technology. The contributions of this work include:[1]

- Introducing a subtask of entity linking, called character identification (Section 2).
- Creating a new corpus for character identification with thorough analysis (Section 3).
- Reformulating character identification into a coreference resolution task (Section 4).
- Evaluating our approach to character identification on our corpus (Section 5).

To the best of our knowledge, it is the first time that character identification is experimented on such a large corpus. It is worth pointing out that character identification is just the first step to a bigger task called character mining. Character mining is a task that focuses on extracting information and

[1] All our work is publicly available at:
github.com/emorynlp/character-mining

Proceedings of the SIGDIAL 2016 Conference, pages 90–100,
Los Angeles, USA, 13-15 September 2016.

constructing knowledge bases associated with particular characters in contexts. The target entities are primarily participants, either spoken or mentioned, in dialogues. The task can be subdivided into three sequential tasks, character identification, attribute extraction, and knowledge base construction. Character mining is expected to facilitate and provide entity-specific knowledge for systems like question answering and dialogue generation. We believe that these tasks altogether are beneficial for machine comprehension on multiparty conversation.

2 Task Description

Character identification is a task of mapping each mention in context to one or more characters in a knowledge base. It is a subtask of entity linking; the main difference is that mentions in character identification can be any nominals indicating characters (e.g., *you*, *mom*, *Ross* in Figure 1), whereas they are mostly related to the Wikipedia entries in entity linking (Ji et al., 2015). Furthermore, character identification allows plural or collective nouns to be mentions such that a mention can be linked to more than one character, and they can either be pre-determined, inferred, or dynamically introduced ; however, a mention is usually linked to one pre-determined entity for entity linking.

The context can be drawn from any kind of document where characters are present (e.g., dialogues, narratives, novels). This paper focuses on context extracted from multiparty conversation, especially from transcripts of TV shows. Entities, mainly the characters in the shows or the speakers in conversations, are predetermined due to the nature of the dialogue data.

Instead of grabbing transcripts from the existing corpora (Janin et al., 2003; Lowe et al., 2015), TV shows are selected because they represent everyday conversation well, nonetheless they can very well be domain-specific depending on the plots and settings. Their contents and exchanges between characters are written for ease of comprehension. Prior knowledge regarding characters is usually not required and can be learned as show proceeds. Moreover, TV shows cover a variety of topics and are carried on over a long period of time by specific groups of people.

The knowledge base can be either pre-populated or populated from the context. For the example in Figure 1, all the speakers can be introduced to the knowledge base without reading the conversation. However, certain characters, mentioned during the conversation but not the speakers, should be dynamically added to the knowledge base (e.g., Ross' mom and dad). This is also true for many real-life scenarios where the participants are known prior to a conversation, but characters outside of these participants are mentioned during the conversation.

Character identification is distinguished from coreference resolution because mentions are linked to global entities in character identification whereas they are linked to one another without considering global entities in coreference resolution. Furthermore, this task is harder than typical entity linking because contexts switch of topics more rapidly in dialogues. In this work, mentions that are either plural or collective nouns are discarded, and the knowledge base does not get populated from the context dynamically. Adding these two aspects will greatly increase the complexity of this task, which we will explore in the future.

3 Corpus

The framework introduced here aims to create a large scale dataset for character identification. This is the first work to establish a robust framework for annotating referent information of characters with a focus on TV show transcripts.

3.1 Data Collection

Transcripts of two TV shows, *Friends*[2] and *The Big Bang Theory*[3] are selected for the data collection. Both shows serve as ideal candidates due to the casual and day-to-day dialogs among their characters. Seasons 1 and 2 of *Friends* (F1 and F2), and Season 1 of *The Big Bang Theory* (B1) are collected. A total of 3 seasons, 63 episodes, and 543 scenes are collected (Table 1).

	Epi	Sce	Spk	UC	SC	WC
F1	24	229	116	5,344	9,168	76,038
F2	22	219	113	9,626	12,368	82,737
B1	17	95	31	2,425	3,302	37,154
Total	63	543	225	17,395	24,838	195,929

Table 1: Composition of our corpus. Epi/Sce/Spk: # of episodes/scenes/speakers. UC/SC/WC: # of utterances/statements/words. Redundant speakers between F1 & F2 are counted only once.

[2] friendstranscripts.tk
[3] transcripts.foreverdreaming.org

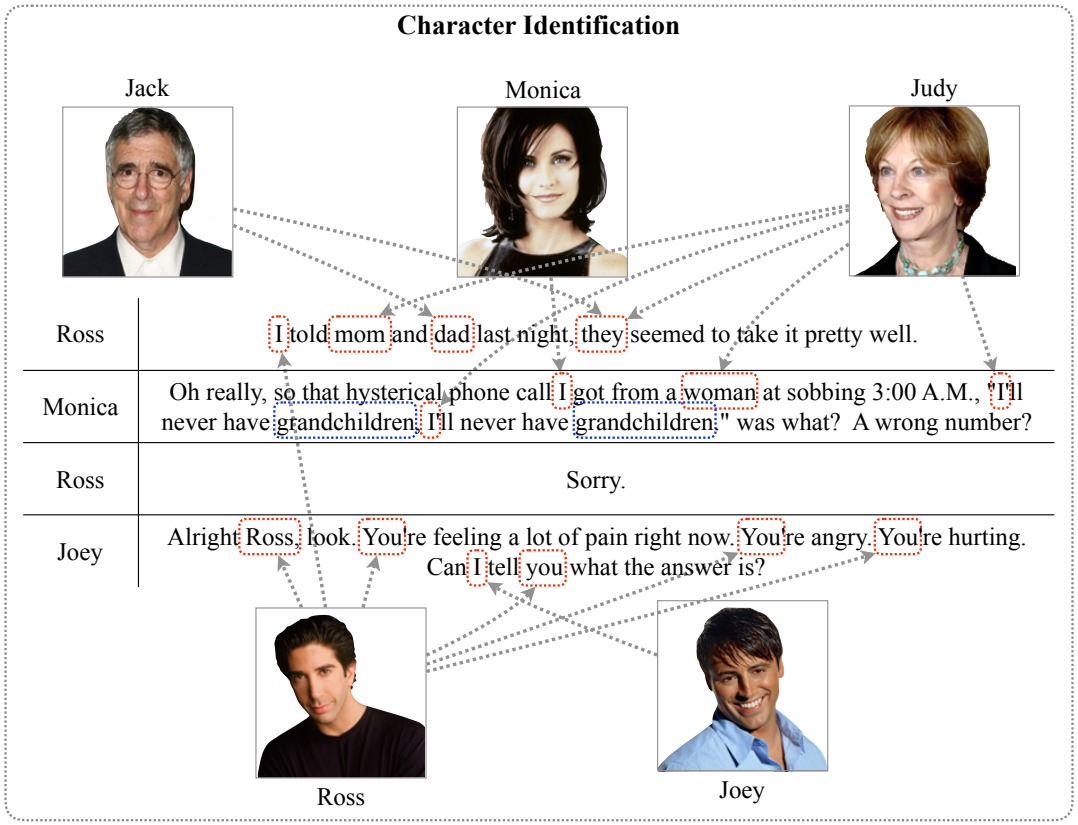

Figure 1: An example of character identification. All three speakers are introduced as characters before the conversation (Ross, Monica, and Joey), and two more characters are introduced during the conversation (Jack and Judy). The goal of this task is to identify each mention as one or more of these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

Each season is divided into episodes, and each episode is divided into scenes based on the boundary information provided by the transcripts. Each scene is divided into utterances where each utterance belongs to a speaker (e.g., the scene in Figure 1 includes four utterances). Each utterance consists of one or more sentences that may or may not contain action notes enclosed by parentheses (e.g., *Ross stares at her in surprise*). A sentence with its action note(s) removed is defined as a statement.

3.2 Mention Detection

Given the dataset in Section 3.1, mentions indicating humans are pseudo-annotated by our rule-based mention detector, which utilizes dependency relations, named entities, and a personal noun dictionary provided by the open-source toolkit, NLP4J.[4] Our rules are as follows: a word sequence is considered a mention if (1)it is a person named entity, (2)it is a pronoun or possessive pronoun excluding *it**, or (3)it is in the personal noun dictionary. The dictionary contains 603 common and singular personal nouns chosen from Freebase[5] and DBpedia.[6] Plural (e.g., *we*, *them*, *boys*) and collective (e.g., *family*, *people*) nouns are discarded but will be included in the next version of the corpus.

	NE	PRP	PNN(%)	All
F1	1,245	7,536	1,464 (24.18)	10,245
F2	1,209	7,568	1,766 (27.28)	10,543
B1	648	3,586	785 (20.05)	5,019
Total	3,102	18,690	4,015 (24.41)	25,807

Table 2: Composition of the detected mentions. NE: named entities, PRP: pronouns, PNN(%): singular personal nouns and its ratio to all nouns.

For quality assurance, 5% of the corpus is sampled and evaluated. A total of 1,584 mentions from the first episode of each season in each show are extracted. If a mention is not identified by the detector, it is considered a "miss". If a detected mention does not refer human character(s), it is considered an "error". Our evaluation shows an F1 score of 95.93, which is satisfactory (Table 3).

[4] https://github.com/emorynlp/nlp4j

[5] http://www.freebase.com

[6] http://wiki.dbpedia.org

	Miss	Error	Total	P	R	F
F1	17	19	615	96.82	94.15	94.47
F2	15	3	448	99.31	95.98	97.62
B1	19	14	475	96.93	93.05	94.95
Total	51	36	1,538	97.58	94.34	**95.93**

Table 3: Evaluation of our mention detection. P: precision, R: recall, F: F1 score (in %).

A further investigation on the causes is conducted on the misses and errors of our mention detection. Table 4 shows the proportion of each cause. The majority of them are caused by either negligence of personal common nouns or inclusion of interjection use of pronouns, which are mostly coming from the limitation of our lexicon.

1. Interjection use of pronouns (e.g., *Oh mine*).
2. Personal common nouns not included in the personal noun dictionary.
3. Non-nominals tagged as nouns.
4. Proper nouns not tagged by either the part-of-speech tagger or name entity recognizer.
5. Misspelled pronouns (e.g., *I'm → Im*).
6. Analogous phrases referring to characters (e.g, *Mr. I-know-everything*).

Causes of Error and Miss	%
Interjection use of pronouns	27%
Common noun misses	27%
Proper noun misses	18%
Non-nominals	14%
Misspelled pronouns	10%
Analogous phrases	4%

Table 4: Proportions of the misses and errors of our mention detection.

3.3 Annotation Scheme

All mentions from Section 3.2 are first double annotated with their referent characters, then adjudicated if there are disagreements between annotators. Both annotation and adjudication tasks were conducted on Amazon Mechanical Turk. Annotation and adjudication of 25,807 mentions took about 8 hours and costed about $450.

Annotation Task

Each mention is annotated with either a main character, an extra character, or one of the followings: collective, unknown, or error. *Collective* indicates the plural use of *you*/*your*, which cannot be deterministically distinguished from the singular use of those by our mention detector. *Unknown* indicates an unknown character that is not listed as an option or a filler (e.g., *you know*). *Error* indicates an incorrectly identified mention that does not refer to any human character.

Our annotation scheme is designed to provide necessary contextual information and easiness for accurate annotation. The target scene for annotation includes highlighted mentions and selection boxes with options of main characters, extra characters, collective, unknown, and error. The previous and next two scenes from the target scene are also displayed to provide additional contextual information to annotators (Table 5). We found that including these four extra scenes substantially reduced annotation ambiguity. The annotation is done by two annotators, and only scenes with 8-50 mentions detected are used for the annotation; this allows annotators to focus while filtering out the scenes that have insufficient amounts of mentions for annotation.

Adjudication Task

Any scene containing at least one annotation disagreement is put into adjudication. The same template as that for the annotation task is used for the adjudication, except that options for the mentions are modified to display options selected by the previous two annotators. Nonetheless, adjudicators still have the flexibility of choosing any option from the complete list as shown in the annotation task. This task is done by three adjudicators. The resultant annotation is determined by the majority vote of the two annotators from the annotation task and the three adjudicators from this task.

3.4 Inter-Annotator Agreement

Serval preliminary tasks were conducted on Amazon Mechanical Turk to improve the quality of our annotation using a subset of the *Friends* season 1 dataset. Though the result on annotating the subset gave reasonable agreement scores ($F1_p$ in Table 6), the percentage of mentions annotated as *unknown* was noticeably high. Such ambiguity was primarily attributed to the lack of contextual information since these tasks were conducted with a template that did not provide additional scene information other than the target scene itself. The unknown rate decreased considerably in the later tasks (F1, F2,

	Friends: Season 1, Episode 1, Scene 1	
	...	
Ross:	I_1 told mom_2 and dad_3 last night, they seemed to take it pretty well.	1. 'I_1' refers to?
Monica:	Oh really, so that hysterical phone call I got from a $woman_4$ at sobbing 3:00 A.M.,	- ...
	"I_5'll never have grandchildren, I_6'll never have grandchildren." was what?	2. 'mom_2' refers to?
Ross:	Sorry.	- ...
Joey:	Alright $Ross_7$, look. You_8're feeling a lot of pain right now. You_9're angry.	3. 'dad_3' refers to?
	You_{10}'re hurting. Can I_{11} tell you_{12} what the answer is?	- Main character$_{1..n}$
	...	- Extra character$_{1..m}$
	Friends: Season 1, Episode 1, Scene 2	- Collective
	...	- Unknown
	Friends: Season 1, Episode 1, Scene 3	- Error
	...	

Table 5: An example of our annotation task conducted. Main character$_{1..n}$ displays the names of all main characters of the show. Extra character$_{1..m}$ displays the names of high frequent, but not main, characters.

and B1) after the previous and the next two scenes were added for context. As a result, our annotation gave the absolute matching score of 82.83% and the Cohen's Kappa score of 79.96% for inter-annotator agreement, and the unknown rate of 11.87% across our corpus, which was a consistent trend across different TV shows included in our corpus.

	Match	**Kappa**	**Col**	**Unk**	**Err**
$F1_p$	83.00	79.94	13.2	33.96	3.95
F1	84.55	80.75	11.2	21.42	3.71
F2	82.22	80.42	13.13	11.69	0.63
B1	81.54	78.73	11.35	7.80	4.99
Avg.	**82.83**	**79.96**	**12.42**	**11.87**	**2.75**

Table 6: Annotation analysis. Match and Kappa show the absolute matching and Cohen's Kappa scores between two annotators (in %). Col/Unk/Err shows the percentage of mentions annotated as collective, unknown, and error, respectively.

One common disagreement in annotation is caused by the ambiguity of speakers that *you*/*your*/*yourself* might refer to. Such confusion often occurs during a multiparty conversation when one party attempts to give a general example using personal mentions that refer to no one in specific. For the following example, annotators label the *you*'s as *Rachel* although they should be labeled as *unknown* since *you* indicates a general human being.

Monica: (to Rachel) You_1 do this, and you_2 do that. You_3 still end up with nothing.

The case of *you* also results in another ambiguity when it is used as a filler:

Ross: (to Chandler and Joey) You_1 know, life is hard.

The referent of *you* here is subjective and can be interpreted differently among individuals. It can refers to Chandler and Joey collectively. It can also be unknown if it refers to a general scenario. Furthermore, it potentially can refers to either Chandler or Joey based on the context. Such use case of *you* is occasionally unclear to human annotators; thus, for the purposes of simplicity and consistency, this work treats them as *unknown* and considers that they do not refer to any speaker.

4 Approach

4.1 Coreference Resolution

Character identification is tackled as a coreference resolution task here, which takes advantage of utilizing existing state-of-the-art systems although it may not result the best for our task since it is more similar to entity linking. Most of the current entity linking systems are accustomed to find entities in Wikipedia (Mihalcea and Csomai, 2007; Ratinov et al., 2011), which are not intuitive to adapt to our task. We are currently developing our own entity linking system, which we hope to release soon.

Our corpus is first reformed into the CoNLL'12 shared task format, then experimented with two of the open source systems. The resultant coreference chains from these system are linked to a specific character by our cluster remapping algorithm.

CoNLL'12 Shared Task

Our corpus is reformatted to adapt the CoNLL'12 shared task on coreference resolution for the compatibility with the existing systems (Pradhan et al., 2012). Each statement is parsed into a constituent tree using the Berkeley Parser (Petrov et al., 2006), and tagged with named entities using the NLP4J

tagger (Choi, 2016). The CoNLL format allows speaker information for each statement, which is used by both systems we experiment with. The converted format preserves all necessary annotation for our task.

Stanford Multi-Sieve System

The Stanford multi-pass sieve system (Lee et al., 2013) is used to provide a baseline of how a coreference resolution system performs on our task. The system 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. Although no machine learning is involved, the system offers efficiency in decoding while yielding reasonable results.

Stanford Entity-Centric System

Another system used in this work is the Stanford entity-centric system (Clark and Manning, 2015). The system takes an ensemble-like statistical approach that utilizes global entity-level features to create feature clusters, and it is stacked with two models. The first model, mention pair model, consists of two tasks, classification and ranking. Logistic classifiers are trained for both tasks to assign probabilities to a mention. The former task considers the likelihood of two mentions are linked. The latter task estimates the potential antecedent of a given mention. The model makes primary suggestions of the coreference clusters and provides additional feature regarding mention pairs. The second model, entity-centric coreference model, aims to produce a final set of coreference clusters through learning from the features and scores of mentions pairs. It operates between pairs of clusters unlike the previous model. Iteratively, it builds up entity-specific mention clusters using agglomerative clustering and imitation learning.

This approach is particularly in alignment with our task, which finds groups of mentions referring to a centralized character. Furthermore, it allows new models to be trained with our corpus. This would give insight on whether our task can be learned by machines and whether a generalized model can be trained to distinguish speakers in all context.

4.2 Coreference Evaluation Metrics

All systems are evaluated with the official CoNLL scorer on three metrics concerning coreference resolution: MUC, B^3, and $CEAF_e$.

MUC

MUC (Vilain et al., 1995) concerns the number of pairwise links needed to be inserted or removed to map system responses to gold keys. The number of links the system and gold shared and minimum numbers of links needed to describe coreference chains of the system and gold are computed. Precision is calculated by dividing the former with the latter that describes the system chains, and recall is calculated by dividing the former with the later that describes the gold chains.

B^3

In stead of evaluating the coreference chains solely on their links, the B^3 (Bagga and Baldwin, 1998) metric computes precision and recall on a mention level. System performance is evaluated by the average of all mention scores. Given a set M that contains mentions denoted as m_i. Coreference chains S_{m_i} and G_{m_i} represent the chains containing mention m_i in system and gold responses. Precision(P) and recall(R) are calculated as below:

$$P(m_i) = \frac{|S_{m_i} \cap G_{m_i}|}{|S_{m_i}|}, \quad R(m_i) = \frac{|S_{m_i} \cap G_{m_i}|}{|G_{m_i}|}$$

$CEAF_e$

$CEAF_e$ (Luo, 2005) metric further points out the drawback of B^3, in which entities can be used more than once during evaluation. As result, both multiple coreference chains of the same entity and chains with mentions of multiple entities are not penalized. To cope with this problem, CEAF evaluates only on the best one-to-one mapping between the system's and gold's entities. Given a system entity S_i and gold entity G_j. An entity-based similarity metric $\phi(S_i, G_j)$ gives the count of common mentions that refer to both S_i and G_j. The alignment with the best total similarity is denoted as $\Phi(g^*)$. Thus precision(P) and recall(R) are measured as below.

$$P = \frac{\Phi(g^*)}{\sum_i \phi(S_i, S_i)}, \quad R = \frac{\Phi(g^*)}{\sum_i \phi(G_i, G_i)}$$

4.3 Cluster Remapping

Since the predicted coreference chains do not directly point to specific characters, a mapping mechanism is needed for linking those chains to certain

TRN	TST	Document: episode				Document: scene			
		MUC	B^3	$CEAF_e$	Avg	MUC	B^3	$CEAF_e$	Avg
Stanford multi-pass sieve	F1+F2+B1	80.73	44.91	27.00	50.88	79.09	62.26	50.22	63.86
Stanford entity-centric	F1+F2+B1	84.44	44.95	19.66	49.68	83.39	69.59	54.48	69.15
F1	F1	90.79	61.25	48.63	66.89	90.16	80.46	69.05	79.89
	F2	92.18	44.40	35.07	57.22	88.49	72.74	59.14	73.46
	B1	94.83	73.46	61.78	76.69	91.55	80.36	66.95	79.62
F1+F2	F1	89.83	67.18	43.98	67.00	90.02	80.48	71.44	80.65
	F2	89.27	55.94	38.55	61.25	89.61	76.76	64.34	76.90
	B1	92.94	75.26	48.61	72.27	92.87	83.55	68.09	81.50
	F1+F2	90.07	63.33	42.44	65.28	89.89	78.75	68.39	79.01
	F1+F2+B1	90.63	**65.64**	43.21	66.49	90.55	**79.84**	**68.53**	**79.64**
B1	B1	93.33	75.83	59.28	76.15	91.79	82.50	69.69	81.33
F1+F2+B1	F1	89.47	64.56	49.63	67.89	90.04	79.63	71.45	80.37
	F2	89.21	57.00	44.31	63.51	89.60	73.78	62.33	75.24
	B1	95.72	72.92	53.87	74.17	92.97	84.23	70.58	82.59
	F1+F2	89.89	62.26	47.92	66.69	89.92	76.95	67.68	78.18
	F1+F2+B1	**91.06**	64.94	**48.26**	**68.09**	**90.59**	78.53	68.37	79.16

Table 7: Coreference resolution results on our corpus. Stanford multi-pass sieve is a rule-based system. Stanford entity-centric uses its pre-trained model. Every other row shows results achieved by the entity-centric system using models trained on the indicated training sets.

characters. The resultant chains from the above systems are mapped to either a character, collective, or unknown. Each coreference chain is reassigned through voting based on the group that majority of the mentions refer to. The referent of each mention is determined by the below rules:

1. If the mention is a proper noun or a named entity that refers to a known character, it is referent to the character.
2. If the mention is a first-person pronoun or possessive pronoun, it is referent to the character of the utterance containing the mention.
3. If the mention is a collective pronoun or possessive pronoun, it is referent to the *collective* group.

If none of these rules apply to any of the mentions in a coreference chain, the chain is mapped to the *unknown* group.

5 Experiments

Both the sieve system and the entity-centric system with its pre-trained model are first evaluated on our corpus. The entity-centric system is further evaluated with new models trained on our corpus. The gold mentions are used for these experiments because we want to focus solely on the performance analysis of these existing systems on our task.

5.1 Data Splits

Our corpus is split into the training, development, and evaluation sets (Table 8). Documents are formulated into two ways, one treating each episode as a document and the other treating each scene as a document, which allows us to conduct experiments with or without the contextual information provided across the previous and next scenes.

	Epi	Sce	Spk	UC	SC	WC
TRN	51	427	189	13,681	19,575	155,789
DEV	5	46	39	1,631	2,313	17,406
TST	7	70	46	2,083	2,950	22,734
Total	63	543	225	17,395	24,838	195,929

Table 8: Data splits. TRN/DEV/TST: training, development, and evaluation sets. See Table 1 for the details about Epi/Sce/Spk/UC/SC/WC.

5.2 Analysis of Coreference Resolution

The results indicate several intriguing trends (Table 7), explained in the following observations.

5.2.1 Multi-pass sieve vs. Entity-centric

These models yield close performance when run out-of-box. It is interesting because both rule-based and statistical models give similar baseline results. This serves as an indicator of how current systems, trained on the CoNLL'12 dataset, do not work as well with day-to-day multiparty conversational data that we attend to solve in this work.

5.2.2 Cross-domain Evaluation

Before looking at the results of the models trained on F1 and F1+F2, we anticipated that these models would give undesirable performance when evaluated on B1. Those models give the average scores

TRN	TST	Document: episode					Document: scene				
		FC	EC	UC	UM	Purity	FC	EC	UC	UM	Purity
Stanford multi-pass sieve		46	53	38.64	16.33	45.97	38	60	22.15	5.97	**64.01**
Stanford entity-centric		36	60	32.59	8.41	38.78	26	60	8.85	1.49	44.12
F1	F1	19	30	30.23	4.20	61.13	21	30	4.94	1.35	54.11
	F2	12	24	40.00	3.15	42.13	17	24	17.91	4.86	51.58
	B1	9	14	0.00	0.00	75.99	14	14	6.25	1.90	70.10
F1+F2	F1	20	30	39.39	7.52	69.92	20	30	10.11	2.72	56.28
	F2	18	24	49.06	8.25	62.54	23	24	7.46	2.12	57.64
	B1	12	14	51.52	12.69	72.16	14	14	10.87	4.56	67.11
	F1+F2	30	46	42.24	7.54	66.65	26	46	9.26	1.83	45.11
	F1+F2+B1	39	60	44.22	8.44	67.67	30	60	7.76	1.35	41.79
B1	B1	11	14	25.00	1.90	80.08	12	14	14.00	5.47	72.83
F1+F2+B1	F1	25	30	21.67	4.06	73.21	20	30	9.41	3.15	51.74
	F2	25	24	29.17	3.64	64.62	25	24	5.80	1.34	58.79
	B1	9	14	20.00	1.31	71.29	15	14	6.67	1.33	69.45
	F1+F2	39	46	24.76	3.78	69.60	29	46	7.62	1.74	44.49
	F1+F2+B1	45	60	23.93	3.27	**69.21**	36	60	6.84	1.39	42.81

Table 9: Character identification results on our corpus using cluster remapping on the coreference resolution system results. FC: found clusters after remapping. EC: expected clusters from gold. UC: percentage of unknown clusters after remapping. UM: percentage of unknown mentions in the unknown clusters to all the mentions.

of 76.69 and 72.27 for B1 on the episode-level, and 79.62 and 79.01 for B1 on the scene-level, respectively. Surprisingly, the models trained on B1 do not yield a better accuracy on the episode-level (76.15), and show an improvement of 1.69 on the scene-level, which is smaller than expected. Thus, it is plausible to take models trained on one show and apply it to another for coreference resolution.

5.2.3 Cross-domain Training

When looking at the models trained on F1+F2+B1, we found that more training instances do not necessarily guarantee a continuous increase of system performance. Although more training data from a single show gives improvements in the results (F1 vs. F1+F2), a similar trend cannot be assumed for the case of the models trained on both shows (F1+F2+B1) when data of another show (B1) is added for training; in fact, most scores show decreases in performance for both episode- and scene-level evaluations. We suppose that this is caused by the introduction of noncontiguous context and content of the additional show. Thus, we deduce that models trained on data from multiple shows are not recommended for the highest performance.

5.2.4 Episode-level vs. Scene-level

We originally foresaw the models trained on the episode-level would outperform the ones trained on the scene-level because the scene-level documents would not provide enough contextual information. However such speculation is not reflected on our evaluation; the results achieved by the scene-level models consistently yield higher accuracy, which is probably because the scene-level documents are much smaller than the episode-level documents so that fewer characters appear within each document.

5.3 Analysis of Character Identification

The resultant coreference chains produced by the systems in Section 4.1 do not point to any specific characters. Thus, our cluster remapping algorithm in Section 4.3 is run on the coreference chains to group multiple chains together and assign them to individual characters. These remapped results provide a better insight of the effective system performance on our task. Table 9 shows the remapped results and the cluster purity scores.

5.3.1 Remapped Clusters

As discussed in Section 5.2.4, the scene-level models consistently outperform the episode-level models for coreference resolution. However, an opposite trend is found for character identification when the coreference chains are mapped to their referent characters. The purity scores of the overall character-mention clusters can be viewed as an effective accuracy score for character identification. The purity scores, or the percentages of recoverable character-mentions clusters, of the remapped clusters for the scene-level models are generally lower than the ones for the episode-level models. Although the percentages of unknown clusters and unknown mentions are considerably higher for the

episode-level models, we find these results more reasonable and realistic to the nature of our corpus, since the average percentages of mentions that are annotated as *unknown* are 11.87% for the entire corpus and 14.01% for the evaluation set. The primary cause of lower performance for the scene-level models is the lack of contextual information across scenes. The following example is excerpted from the first utterance in the opening scene of F1:

Monica: There's nothing to tell!
He_1's just some guy_2 I_3 work with!

As the conversation proceeds, there is no clear indication of who He_1 and guy_2 refer to until later scenes introduce the character. As a result, the coreference chains in the scene-level documents are noticeably shorter than those in the episode-level documents. When trying to determine the referent characters, fewer mentions exist in the coreference chains produced by the scene-level models such that there is a higher chance for those chains to be mapped to wrong characters. Thus, the episode-level models are recommended for better performance on character identification.

6 Related Work

There exist few corpora concerning multiparty conversational data. SwitchBoard is a telephone speech corpus with focuses on speaker authentication and recognition (Godfrey et al., 1992). The ICSI Meeting Corpus is a collection of meeting audios and transcript recordings created for research in speech recognition (Janin et al., 2003). The Ubuntu Dialogue Corpus is a recently introduced dialogue corpus that provides task-domain specific conversation with multiple turns (Lowe et al., 2015). All these corpora provide an immense amount of dialogue data. However, the primary purposes of them are aimed to tackle tasks like speaker or speech recognition and next utterance generation. Thus, mention referent information are missing for the purpose of our task.

Entity Linking is a natural language processing task of determining entities and connecting related information in context to them (Ji et al., 2015). Linking can be done on domain-specific information using extracted local context (Olieman et al., 2015). Wikification is a branch of entity linking with an aim of associating concepts to their corresponding Wikipedia pages (Mihalcea and Csomai, 2007). Ratinov et al. (2011) used linked concepts and their relevant Wikipedia articles as features on disambiguation. Kim et al. (2015) explored dialogue data in the realm of the task in attempt to improve dialogue tracking using Wikification-based information.

Similar to entity linking, coreference resolution is another NLP task that connects mentions to their antecedents (Pradhan et al., 2012). The task focuses on finding pair-wise connection between mentions and forming coreference chains of the pairs. Dialogues have been studied as a particular domain for coreference resolution (Rocha, 1999) due to the complex and context-switching nature of the data. For most of the systems presented for the task, they target on narrations or conversations between two parties, such as tutoring systems (Niraula et al., 2014). Despite their similarity, coreference resolution still differs from character identification since the resolved coreference chains do not directly refer to ant centralized characters.

7 Conclusion

This paper introduces a new task, called character identification, that is a subtask of entity linking. A new corpus is created for the evaluation of this task, which comprises multiparty conversations from TV show transcripts. Our annotation scheme allows to create a large dataset with the personal mentions and their referent characters annotated. The nature of this corpus is analyzed with potential challenges and ambiguities identified for future investigation.

Hence, this work provides baseline approaches and results using the existing coreference resolution systems. Experiments are run on combinations of our corpus in various formats to analyze the applicability of the current systems as well as the model trainability for our task. A cluster remapping algorithm is then proposed to connect the coreference chains to their reference characters or groups.

Character identification is the first step to a machine comprehension task we define as character mining. We are going to extend this task to handle plural and collective nouns, and develop an entity linking system customized for this task. Furthermore, we will explore an automatic way of building a knowledge base containing information about the characters that can be used for more specific tasks such as question answering.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Jinho D. Choi. 2016. Dynamic Feature Induction: The Last Gist to the State-of-the-Art. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16.

Kevin Clark and Christopher D. Manning. 2015. Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL'15, pages 1405–1415.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'92, pages 517–520.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Annual Conference on Neural Information Processing Systems*, NIPS'15, pages 1693–1701.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'15, pages 851–861.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP'03, pages 364–367.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In *Proceedings of Text Analysis Conference*, TAC'15.

Seokhwan Kim, Rafael E. Banchs, and Haizhou Li. 2015. Towards Improving Dialogue Topic Tracking Performances with Wikification of Concept Mentions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL'15, pages 124–128.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL'15, pages 285–294.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM'07, pages 233–242.

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*, LREC'14, pages 3199–3203.

Alex Olieman, Jaap Kamps, Maarten Marx, and Arjan Nusselder. 2015. A Hybrid Approach to Domain-Specific Entity Linking. In *Proceedings of 11th International Conference on Semantic Systems*, SEMANTiCS'15.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A Joint Framework for Coreference Resolution and Mention Head Detection. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, CoNLL'15, pages 12–21.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL'06)*, pages 433–440.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL'12, pages 1–40.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL'11, pages 1375–1384.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of

Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 193–203.

Marco Rocha. 1999. Coreference Resolution in Dialogues in English and Portuguese. In *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp'99, pages 53–60.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL'15, pages 1416–1426.