Movie Script Coreference Annotation Guidelines

Signal Analysis and Interpretation Laboratory (SAIL)

University of Southern California

1 Introduction

Coreference resolution is the task of *identifying* and *linking* the mentions of a specific entity or event. An **entity** can be any named object like a person, organization, country, university, etc., and an **event** is some change in the configuration of a set of such entities. Entities and events are referred to by different linguistic expressions in a text document called **mentions**. A coreference resolution system's task is to find all mentions and their referred entity or event. The sequence of all mentions of an entity or an event is called a **coreference chain**.

(1) [[Apple's]$_w$ $3 billion dollar purchase of [Beats]$_x$]$_y$ is the largest acquisition ever made by [the big tech company]$_w$. As part of [the deal]$_y$, [[Beats]$_x$ co-founders Jimmy Iovine and Dr Dre]$_z$ also joined [Apple]$_w$. "[They]$_z$'re going to be coming up with features that blow your mind," [Apple]$_w$ CEO Tim Cook tells The New York Times.

In example (1), we mark the coreference chain of three entities – *Apple*, *Beats*, and *Jimmy Iovine and Dr Dre* – and of the event, *Apple's acquisition of Beats*. Mentions of the same entity or event are bracketed and subscripted by a unique literal. Notice how mentions can be nested within each other, can refer to a group of entities (*Jimmy Iovine and Dr Dre*), or can span part of a word (*[They]'re*).

Coreference resolution has been mainly studied in news and web text, and our goal is to extend it to movie narratives. This document contains the annotation guidelines for coreference annotation in movie scripts. A **movie script** is a semi-structured text document containing the screenplay of a movie. It consists of scene headers, scene descriptions, character names, utterances, and scene transitions. Figure 1 shows an excerpt from the script of the movie 'The Shawshank Redemption', and points out the different screenplay elements. We find coreference chains of **movie characters** in this annotation task, and don't consider mentions of other named entities or events. Therefore, in the script excerpt, the only entities eligible for coreference are the characters – *Red*, *Andy*, *Heywood* and *Jigger*. Figure 1 shows the coreference chains of these four characters.

Section 2 describes how you can find character mentions. Section 3 lists the rules you should follow to handle some common scenarios. Section 4 contains instructions on how to use the annotation tool.

2 Mentions

We only consider mentions of movie characters. A mention can be a noun phrase, pronoun, or a possessive.

2.1 Noun Phrase

A **noun phrase** is a phrase with a noun or an indefinite pronoun as its head and serves the same grammatical function as a noun in the sentence. You can check if a phrase is a noun phrase if a pronoun

EXT -- EXERCISE YARD -- DAY (1947) ⟵ HEADER
Exercise period. [Red]$_w$ plays catch with [Heywood]$_x$ and
[Jigger]$_y$, lazily tossing a baseball around. [Red]$_w$ notices
[Andy]$_z$ off to the side. Nods hello. [Andy]$_z$ takes this
as a cue to amble over. [Heywood]$_x$ and [Jigger]$_y$ pause,
watching. ⟵ DESCRIPTION

[ANDY]$_z$ ⟵ CHARACTER
(offers [his]$_z$ hand)
Hello. I'm [Andy Dufresne]$_z$. ⟵ UTTERANCE
[Red]$_w$ glances at the hand, ignores it. The game continues.
[RED]$_w$
[The wife-killin'banker.]$_z$

Figure 1: Script excerpt from 'The Shawshank Redemption'

can substitute it. A typical noun phrase contains modifiers along with the noun headword. Some common types of such modifiers are (headword is in bold) –

- Determiner: *the **horse**, that **person**, a **girl***.
- Adjectives: *brown **horse**, kind **person***.
- Noun modifier: *American Quarter **horse**, business **person***.
- Prepositional phrase: ***horse** in the stable*, ***person** with the book*.
- Relative clause: *the **horse** which is grazing, that **person** who is working*.

You must mark the entire extent of the noun phrase as the mention. Mentions can be nested within each other if they refer to different entities, like in (2).

(2) [[The boy's]$_x$ mother]$_y$ scolded [him]$_x$ for getting dirty.
[She]$_y$ would have to wash [his]$_x$ clothes again.

Honorifics, for example, *[Mr.] Andy Dufresne, [Fraulein] Bridget Von Hammersmark*, etc., should also be included in the noun phrase mention.

2.2 Pronoun

Pronouns are used to substitute noun phrases and are the most common type of mentions you will find. Most pronominal mentions comprise of **personal pronouns** and their possessive and reflexive forms. Personal pronouns can be classified according to person, gender, and number. Table 1 lists all the personal pronouns under the different categories. This table will help you identify the pronouns and check if the count and gender of the pronoun matches the referred entity.

Examples (3), (4) and (5) show how pronouns can be used in coreference.

(3) [He]$_x$ destroyed [himself]$_x$.

(4) [I]$_x$ am searching for [my]$_x$ pipe.

(5) [You]$_x$ strike me as a particularly icy and remorseless man,
[Mr. Dufresne.]$_x$

While marking a pronoun mention, take care to correctly mark the extent of the pronoun in the case where it is contracted with the following verb. For example, *[I]'m, [We]'re, [She]'d,*, etc. The apostrophes and the contracted verb is not included in the span. Section 2.3 contains some more rules for verb contractions.

Person	Number/Gender	Personal Pronouns
First	**Singular**	I, me, my, mine, myself
	Plural	We, us, our, ours, ourselves
Second	**Singular/Plural**	you, your, yours, yourself, yourselves
Third	**Male**	he, him, his, himself
	Female	she, her, hers, herself
	Neutral	it, its, itself
	Plural	they, them, their, theirs, themself, themselves

Table 1: Personal Pronouns

2.3 Possessive

Possessives are words that indicate some relationship of possession. They can be possessive nouns or possessive pronouns. **Possessive nouns** are formed by adding *'s* or only the apostrophes after a noun, for example, *John's book*, *the soldier's gun*, *students' grades*, etc. **Possessive pronouns** can be either possessive determiners, which are usually followed by a noun phrase, for example, *my, your, his, her, its, our* and *their*, or independent possessive pronouns, for example, *mine, yours, his, hers, its, ours* and *theirs*. Be careful while marking the extent of a possessive noun when it is in a verb contraction form with *is*.

(6) [John]$_x$'s going to town with [Marie's]$_y$ brother. [John]$_x$ and [Marie]$_y$ are friends.

(7) [The boy's]$_x$ mother was proud of her [son's]$_x$ achievements.

In example (6), *John's* is short for *John is*. Only the word *John* refers to the person John. Therefore, the mention span will not include *'s*. Contrast this with *Marie's* which is a possessive noun and thus will include *'s*. In example (7), both the mentions are possessive nouns. The determiner will be included in the mention span, for example, *The boy's*.

2.4 Characters

We only consider **singular** and **main** characters for coreference. We do not find mentions that refer to more than one character because finding the correct referent for plural mention types (*they, them, Andy and Red*, etc.) usually requires the aid of the accompanying video clip, which will not be available to the annotator. The annotator will only be given the movie script text.

We leave it up to the annotator to decide which are the main characters. At the least, the set of main characters should contain all speaking characters: those who have some utterance in the script. If a character and its mentions can be easily identified and appear many times in the script, the annotator should consider it a main character even if they do not have any utterances. Usually, characters that appear as part of a group and are referred sparingly, often addressed by a numbered title like *MAN 1, AGENT #2*, etc., can be excluded from the set of main characters.

Figure 2 is a script excerpt from the movie 'Bourne Ultimatum'. As can be observed from the marked mentions, only singular characters, *Bourne* and the guy in the parking space are annotated for coreference. Plural entities, *occupants* and *agents*, which are underlined in the text, are not labelled.

3 Rules

The following rules govern how you should handle some common coreference scenarios.

INT./EXT. PARKING GARAGE -- TWO LEVELS BELOW ROOFTOP

[Bourne]$_x$ cuts off [a guy]$_y$ cruising for a parking space and pulls [him]$_y$ from [his]$_y$ car and races away as the agent from the roof lands hard behind [him]$_x$.

[Bourne's]$_x$ race to the exit is cut off as a 3rd CRI sedan slides into view and it's occupants open fire on Bourne in a head on charge.

The just stolen vehicle takes heavy fire as [Bourne]$_x$ reacts instantaneously; thumbing on the cruise control, shouldering [his]$_x$ door open, and slamming the gas pedal to the floor as [he]$_x$ dives out of the car.

The agents react as [Bourne's]$_x$ sedan torpedoes them head on. Hit hard they're taken out of the fight as metal collapses, glass shatters, and airbags explode.

[Bourne]$_x$ tumbles to a stop at the rear of a parked car as [his]$_x$ car implodes against the oncoming agents.

Figure 2: Script excerpt from 'Bourne Ultimatum'

3.1 Appositions

An appositive construction contains a noun phrase modified by one or more adjacent noun phrase(s), separated by a comma, colon, or parenthesis. The most specific element in the appositive construction is called the head, which refers to some character. The order of specificity is – proper noun (*Bourne*) > pronoun (*he*) > definite noun phrase (*the agent*) > indefinite noun phrase (*a agent*). The other noun phrases describe attributes of the headword. The entire appositive phrase is marked for coreference.

(8) He leaves, returning back behind the bar, with [[the YOUNG FRENCH BARMAID]$_{x\text{-HEAD}}$, the only other person in the establishment]$_x$. [She]$_x$ fills their glasses with ...

(9) [[BYRON HADLEY]$_{x\text{-HEAD}}$, captain of the guard]$_x$, slams [his]$_x$ baton into Andy's back.

(10) [[DR. HIRSCH]$_{x\text{-HEAD}}$, 70]$_x$, is, put simply, not a man to be trifled with.

(11) [You]$_x$, [John]$_x$, should behave [yourself]$_x$.

Examples (8), (9), and (10) show appositions. The head of the appositive construction is marked separately. Annotators do not have to label the head of an apposition in their annotation task. It is only shown here for explaining the appositive structure. Note that example (11) is not an apposition, despite there being two comma-separated adjacent mentions – *You* and *John*. It is not an apposition because neither of the two mentions describe some attribute of the referred character. Without the comma, both the words would be included in the same mention – [You John]$_x$.

3.2 Copula

A copular structure consists of a referent, an attribute of the referent, and a linking verb that serves to equate the referent with the attribute. Common copular verbs are *be, is, are, was, were, appear, look, seem*, etc. Only the most specific element of a copular structure should be linked to subsequent mentions. The

order of specificity is the same as described in section 3.1.

(12) Don't worry, [she]$_x$ is [a British spy], [she]$_x$'ll make the rendezvous.

In example (11), *she is a British spy* is a copular structure because of the linking verb *is*. The subject word *she* is the referent and, therefore, can be marked as a mention. The attribute *a British spy* cannot participate in any coreference.

3.3 Script Elements

Mentions of characters in the scene headers and character names should also be marked.

INT. [LANDY'S]$_x$ OFFICE. DAY
[CRONIN]$_y$
[Pam]$_x$, [You]$_x$ need to see this.
[Landy]$_x$ follows [CRONIN]$_y$ into--

Figure 3: Script excerpt from 'Bourne Ultimatum'

Figure 3 is a script excerpt from the movie 'Bourne Ultimatum', which contains a dialogue between characters Pamela Landy and Cronin. As can be seen from the labels, Landy's mention in the scene header and Cronin's mention in the character name are both marked.

3.4 Reader

The movie script will sometimes contain references to the narrator or the point of view of the reader. Such instances should be marked under a special entity name called READER. It is the only exception to rule 2.4. READER references are often first and second person personal pronouns (Table 1).

In the BACKGROUND, [WE]$_x$ SEE, [our]$_x$ three counterfeit German Officers, Hicox, Wicki, and Stiglitz, enter the basement tavern.
They obviously. see the five German soldiers, but their too far away for [us (the audience)]$_x$ to read their face.
No doubt their less then happy.

Figure 4: Script excerpt from 'Inglourious Basterds'

Figure 4 is a script excerpt from the movie 'Inglourious Basterds', where the three Basterds enter a French tavern, disguised as German Nazis. The three marked mentions are references to the reader or viewer. The third mention is also an apposition.

3.5 Generic You

Instances of the generic *you* should not be marked for coreference. Figure 5 is a script excerpt from the movie 'The Shawshank Redemption', and it contains a monologue from the character Red. None of the marked occurrences of the pronoun *you* or its derivatives are eligible for coreference.

RED (V.O.)
There's a con like me in every prison in America, I guess. I'm the guy who can get it for [you]. Cigarettes, a bag of reefer if [you]'re partial, a bottle of brandy to celebrate [your] kid's high school graduation. Damn near anything, within reason.

Figure 5: Script excerpt from 'The Shawshank Redemption'

3.6 Revelation of Identity

If the identity of a character is revealed to be identical to another as the story progresses, their coreference chains should be merged. For example, consider the script excerpt of the movie 'The Shawshank Redemption', shown in figure 6. It shows a portion of two scenes, the second scene occurring after the first. We do not know the identity of the *MAN* character in the first scene. It is later revealed to be *Mr. Quentin* in the second scene. Thus, we should merge the coreference chains of *MAN* and *Mr. Quentin*. Annotators might need a second pass over the script to resolve such cases.

```
The door bursts open.  [A MAN]x and WOMAN enter, drunk and
giggling, horny as hell.

                        ...(several scenes later)...

                              ANDY
She packed a bag and went to stay with [Mr.  Quentin.]x
```

Figure 6: Script excerpt from 'The Shawshank Redemption'

These rules should take care of most cases the annotators encounter in the script. These guidelines are largely inspired from the coreference annotation guidelines of the OntoNotes 5.0 English dataset (Pradhan et al., 2012). You can find their complete annotation codebook here [1]. For more information on coreference resolution, check out this chapter [2] from the book *Speech and Language Processing, 3rd edition draft* by Jurafsky and Martin.

4 Tool

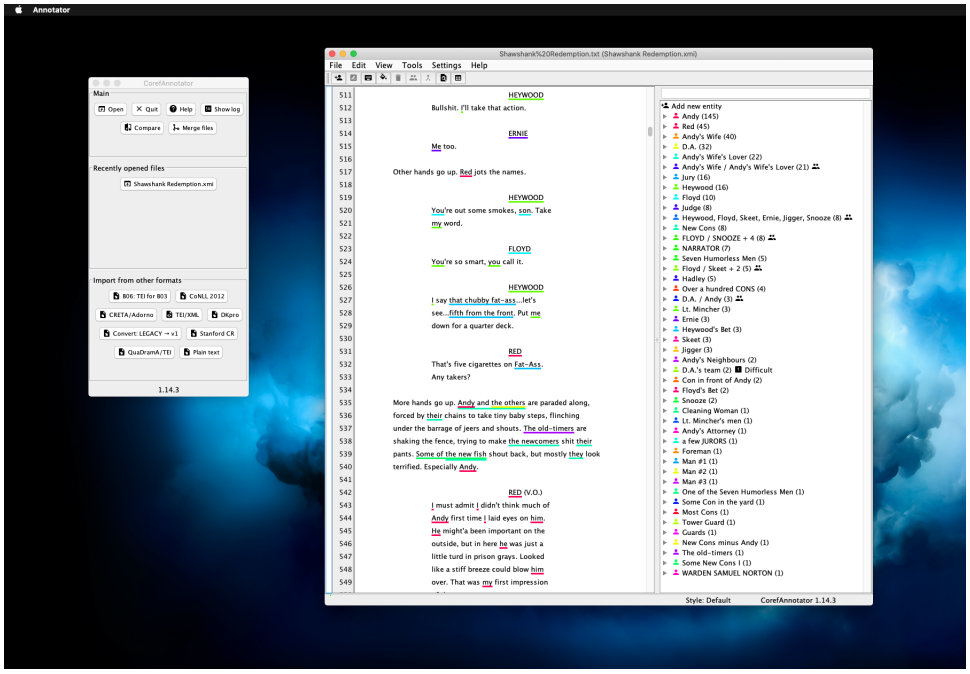

Figure 7: Main (left) and annotation window

Annotators will use the CorefAnnotator tool, developed by Neils Reiter, to annotate the coreference chains. It is easy to use and great for viewing the coreference chains as you create it. The annotation tool also allows you to tag mentions with different flags, which we shall use to distinguish named, nominal, pronominal, and appositive mentions.

[1] english-coreference-guidelines.pdf
[2] slp3/22.pdf

1. **Starting the annotation tool**

   The annotation tool can be found in the repository provided to the annotators. It is a jar file named *CorefAnnotator.jar*. You can start the annotation tool from the terminal by navigating to the repository folder and issuing the command: `java -jar CorefAnnotator.jar`. It will open the main window shown in figure 7.

2. **Opening a movie script**

   If you are opening a movie script for the first time, choose the *Plain text* option from the *Import from other formats* section of the main window (fig 7). It will open the file explorer where you can select the movie script text file that you need to annotate. Otherwise, if you have already created some annotations, you should open the saved XMI.GZ file from your previous session. You can open the XMI.GZ file using option *Open* from the *Main* section of the main window (fig 7).

   Upon opening the movie script, you will see the annotation window, as shown in figure 7. It shows the movie script on the left and coreference chains on the right. If you are opening the script for the first time, the right pane will be blank.

3. **Creating flags**

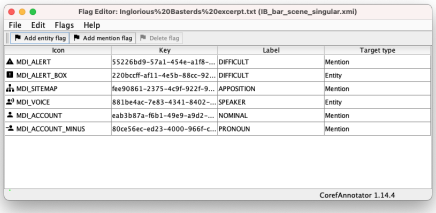

Figure 8: Flag editor window

   Before you start annotating the movie script, you need to create the PRONOUN, NOMINAL, SPEAKER, APPOSITION and DIFFICULT flags. Click on the *Tools* menu from the main annotation window and choose *Edit Flags*. It will open the *Flag Editor* window as shown in figure 8. Create the flags as shown in the figure. The icons and keys can be different.

   If you are unsure about the extent or the referred entity of a mention, or whether it is an appositive construction, etc., tag the mention with the mention-type DIFFICULT flag. If you are unsure whether an entity should be considered for coreference, or which character it corresponds to, etc., tag the mention with the entity-type DIFFICULT flag. If the character entity has some utterance, tag the entity with the SPEAKER flag. If the mention is a pronoun, tag it with the PRONOUN flag. If the mention is a noun phrase or possessive and does not contain the character's name, tag it with the NOMINAL flag. Examples of nominal mentions are *The boy*, *wife-killin' banker*, *AGENT #1*, etc. Step 7 shows how you can tag mentions with flags.

4. **Marking mentions**

   To mark a mention with its referred character, highlight the mention text and right-click on it. It will open the context menu as shown in figure 9. Select the correct character from the list. If the entity is not yet created, choose *New* and add a new entity for the character.

   The context menu only shows the most recently used characters. If the referred character does not appear in the menu, drag the highlighted mention and drop it on the corresponding entity of the character in the right pane.

5. **Editing mentions**

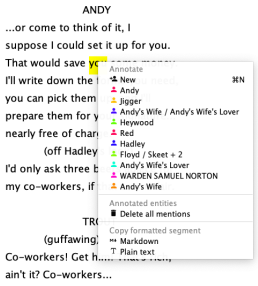

Figure 9: Labelling mentions

You can delete a mention using either the context menu or its corresponding node in the right pane. There is no option to edit the span of an existing mention. If you want to change the span of a mention, you would need to delete it and create a new mention.

6. **Renaming entities**

   You can rename an entity in the right pane by right-clicking on it and choosing the *Rename* option.

7. **Tagging flags**

   To tag a mention or entity with a flag, right-click on the mention or entity in the right pane and go to the *Flags* option. You will see the list of available flags which you created in step 3. Choose the appropriate flag from the list, according to step 3.

8. **Saving annotations**

   Click on the *File* menu from the main annotation window and choose the *Save as* option. Select the *UIMA Xmi Files* option from the *File Format* drop-down menu, and save it. You can open this next time, instead of the movie script text file, to start annotating from the point you left off.

After you have annotated the entire script, export your annotations to a CSV file using the *Export as* option from the *File* menu. You are required to submit both the XMI.GZ and CSV file to complete the task.

References

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.