Identifying Speakers and Addressees in Dialogues Extracted from Literary Fiction

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam,
Gintarė Grigonytė & Sofia Gustafson Capková
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
{adam.ek, mats.wiren, robert, kristina.nilsson, gintare, sofia}@ling.su.se

Abstract

This paper describes an approach to identifying speakers and addressees in dialogues extracted from literary fiction, along with a dataset annotated for speaker and addressee. The overall purpose of this is to provide annotation of dialogue interaction between characters in literary corpora in order to allow for enriched search facilities and construction of social networks from the corpora. To predict speakers and addressees in a dialogue, we use a sequence labeling approach applied to a given set of characters. We use features relating to the current dialogue, the preceding narrative, and the complete preceding context. The results indicate that even with a small amount of training data, it is possible to build a fairly accurate classifier for speaker and addressee identification across different authors, though the identification of addressees is the more difficult task.

Keywords: literary corpora, speaker identification, addressee identification, quote attribution

1. Introduction

During the last few years, quantitative approaches to literary analysis have increasingly progressed from stylistic problems to higher-level phenomena such as plot, community structure and interaction between protagonists. One example of this is the recent interest in constructing social networks from literary fiction, either manually (Moretti, 2011; Agarwal et al., 2012; Yeung and Lee, 2016; Vala et al., 2016) or using automatic methods (Newman and Girvan, 2004; Elson et al., 2010; Rydberg-Cox, 2011). Typically, the goal has been to mirror relations between entities or events extracted from the text as a whole. Arguably, however, a more fine-grained perspective can be obtained by studying the direct speech between characters separately from the narratives in which the speech is embedded. Direct speech, in literary fiction usually framed by devices such as dashes, quotation marks and paragraphs, can be seen as the lowest level of narrative transmission (Koivisto and Nykänen, 2016). In this sense, it provides an independent level in which the relations between characters can be studied, including phenomena such as stance and sentiment as expressed through (the rendering of) the characters themselves.

To properly analyse dialogue interactions, we need to identify both the speakers and the addressees in occurrences of direct speech. The former problem, also known as quote attribution, has been explored in literary fiction by, among others, Elson et al. (2010), O'Keefe et al. (2012), He et al. (2013) and Muzny et al. (2017). As far as we know, however, the problem of identifying addressees (liisteneners) in literary fiction has previously only been dealt with by Yeung and Lee (2017).

For the purpose of specifying our method, we make the following assumptions, aimed at covering differing author styles. We refer to a sequence of direct speech interactions as a *dialogue*; this consists of one or more *turns*, each of which we assume is associated with one speaker and one or more addressees. A turn consists of one or more *lines* (framed by dashes in the example in Figure 1), and a line consists of one or more utterances. Literary fiction consists of alternating dialogues and instances of narrative structure, the latter of which we refer to as *narratives*. In addition, we refer to the entire text before a dialogue (narratives as well as other dialogues) and back to the beginning of a chapter as the *global context*.

> Olle very skilfully made a bag of one of the sheets and stuffed everything into it, while Lundell went on eagerly protesting.
>
> When the parcel was made, Olle took it under his arm, buttoned his ragged coat so as to hide the absence of a waistcoat, and set out on his way to the town.
>
> – He looks like a thief, said Sellén, watching him from the window with a sly smile. – I hope the police won't interfere with him! – Hurry up, Olle! he shouted after the retreating figure. Buy six French rolls and two half-pints of beer if there's anything left after you've bought the paint.

Figure 1: Example narrative and dialogue turn translated from our Swedish data (from Chapter 6 of August Strindberg, *The Red Room*, 1879). The first two paragraphs constitute a narrative. The third paragraph is a turn consisting of three lines, each of which is marked by a dash, with Sellén as speaker. In the first line, the speaker is explicitly tagged ("said Sellén"); in the second, the speaker is implicit; and in the third line, the speaker is anaphoric ("he shouted"). The three lines have two distinct addressees (Lundell, Lundell, and Olle, respectively). A sequence of turns uninterrupted by narratives constitutes a dialogue.

This is exemplified in Figure 1, which shows a narrative with a subsequent dialogue turn. Figure 1 also illustrates the three ways of signalling the identity of a speaker that we distinguish:

1. Explicit speaker: a speech tag consisting of a speech verb and an explicit name ("said **Sellén**").
2. Anaphoric speaker: a speech verb and an anaphoric expression in the form of a pronoun ("**he** shouted") or a definite description ("said **the angry man**").
3. Implicit speaker: none of the above; the speaker must be inferred from the previous lines, preceding dialogue, preceding narrative and/or global context.

Analogously, we distinguish three ways in which the identity of an addressee can be signalled:

1. Explicit addressee: the name of the addressee is mentioned explicitly ("Hurry up, **Olle**!").
2. Anaphoric addressee: the addressee is referred to with a pronoun ("...after **you**'ve bought the paint") or a definite description (...he shouted after "**the retreating figure**").
3. Implicit addressee: none of the above; the addressee must be inferred from the previous lines, preceding dialogue, preceding narrative and/or global context.

As mentioned above, we assume that a turn has a single speaker. As illustrated by the example in Figure 1, however, different lines within a turn may have different addressees. Also, a speaker may address more than one person simultaneoulsy, which means that one line may have several addressees.

2. Previous Work

Among the first papers to consider quote attribution applied to literary fiction is Elson et al. (2010). They use a variety of supervised machine learning approaches (JRip, J48 and Logistic Regression) to assign a speaker to quotes. The system extracts candidates from the surrounding text and for each quote the system selects the most likely speaker. The quotes are divided into seven syntactical categories corresponding to different manners in which speakers are indicated.

An SVM-ranking approach to the problem was used in He et al. (2013). Unlike Elson et al. (2010), the candidates were extracted during the preprocessing step, and for each quote the system selects the most likely speaker from the set of candidates. Each candidate is assigned a set of features capturing the turn-taking, dependency relations, name and gender matching, character frequencies, distances to the utterance and mentions in the quote. Also, an unsupervised topic-actor model was used as a feature.

Recently, a sieve approach was applied to the problem by Muzny et al. (2017). The approach determines the speakers in two steps, first candidate speakers of each quote is identified in the text. Secondly, from the candidate speakers the most likely speaker is selected. The sieves used in determining candidates capture dependency relations, mention recency, the turn-taking heuristic and mentions in and around the quote. To selecting a speaker from the candidates, co-reference resolution, name matching, the turn-taking heuristic and mentions in the quote are used.

The only work aimed at identifying addressees that we are aware of is Yeung and Lee (2017).[1] This uses a CRF sequence labeling algorithm such that for each quote, the two surrounding sentences are extracted. Each word in the extracted sentences is then assigned a feature set containing the part-of-speech tag, dependency relations, distances to the quote, and matches in the line. Each word is then classified as "speaker", "listener" or "neither" by the system.

3. Data

In this section, we describe the data set used and how the data set was annotated.

3.1. Overview

The data used in the experiments reported here consists of parts of four novels by different authors: August Strindberg, *The Red Room* (1879; obtained from the National Edition of August Strindberg's Collected Works, published in 1981); Hjalmar Söderberg, *The Serious Game* (1912); Birger Sjöberg, *The Quartet That Split Up*, part I (1924); and Karin Boye, *Kallocain* (1940). Table 1 specifies the total number of dialogues and lines which have been annotated, and how they make up the training and test set, and the development set. The development set consists of Chapters 1 and 21 from *The Red Room* by August Strindberg, whereas all the remaining chapters are included in the training and test set. The distribution of chapters, dialogues and lines in the training and test set across the four novels is shown in Table 2.

In total, the test and training corpus consists of 822 lines distributed over 268 dialogues. Specifically, for each turn we annotated each line with its speaker, its addressee or addressees (compare Section 3.2.), and an indicator for the ways in which the identity of the speaker and the addressee were signaled as described in the previous section (explicit, anaphoric or implicit). Table 3 shows the variation of indicators for speakers across the authors. Furthermore, the variation of indicators for addressees is shown in Table 4.

CORPUS	DIALOGUES	LINES
Training and test	268	822
Development	23	75

Table 1: Number of dialogues and lines in the annotated corpus.

As shown in Tables 3 and 4, both speakers and addressees are mostly referred to implicitly in our data. When this is not the case, however, speakers are more commonly referred to explicitly, whereas addressees are more commonly referred to anaphorically, mostly with pronouns.

[1] Strictly speaking, Yeung and Lee (2017) take the goal to be to identify listeners. As exemplified by the last line in Figure 1, however, the addressees (the intended recipient or recipients of an utterance) may be a subset of the listeners (the people overhearing the utterance).

Corpus	Chapters	Dialogues	Lines
Strindberg	4	93	393
Sjöberg	10	82	216
Söderberg	2	37	93
Boye	5	56	121
All	21	268	822

Table 2: Number of dialogues and lines in the training and test data.

Author	Exp	Imp	Ana-P	Ana-D
Strindberg	86	285	20	2
Sjöberg	117	67	10	21
Söderberg	26	52	15	0
Boye	21	44	56	0
All	250	448	101	23

Table 3: Indicators for speaker identity across the authors. Exp = explicit; Imp = implicit; Ana-P = anaphoric, pronoun; Ana-D = anaphoric, definite description.

Author	Exp	Imp	Ana-P	Ana-D
Strindberg	31	192	133	37
Sjöberg	34	160	5	16
Söderberg	9	67	16	1
Boye	9	74	29	9
All	83	493	183	63

Table 4: Indicators for addressee identity across the authors. Exp = explicit; Imp = implicit; Ana-P = anaphoric, pronoun; Ana-D = anaphoric, definite description.

Different authors and print editions use different conventions for framing turns and lines in dialogues, such as dashes, quotation marks or angle brackets, thus delimiting the speech in different ways. For example, the first line in Figure 1 might have been rendered as

> "He looks like a thief", said Sellén, watching him from the window with a sly smile.

We use a script to to normalize these different conventions into a format using dashes as shown in Figure 1 .

3.2. Annotation

The data was annotated by two of the authors. The data consists of raw text with the annotations being inserted tags indicating where a line ends, containing who the speaker is and who the addressee is, followed by in which way these are indicated. The components of the annotation tag are the following:

1. <speaker--addressee>
2. <type_speaker--type_adressee>

Where 1 is always followed by 2. Using Figure 1 as an example, the annotations are the following:

> – He looks like a thief, said Sellén, watching him from the window with a sly smile. <Sellén--Lundell><EXP--IMP>
> – I hope the police won't interfere with him! <Sellén--Lundell><IMP--IMP>
> – Hurry up, Olle! he shouted after the retreating figure. Buy six French rolls and two half-pints of beer if there's anything left after you've bought the paint. <Sellén--Olle><ANA--EXP>

The start of each line in a turn is indicated by a dash and the annotation is inserted at the end of the line. We have only annotated lines where there are a clear speaker and addressee. Cases in which the same character is both the speaker and addressee have not been annotated. If the addressee is a group of people it is annotated as "SEVERAL". For addressees, there may be conflicts between different tags as in the last line of Figure 1, where Olle may be annotated as explicit or as a definite description. In these cases explicit mentions supersede anaphoric mentions, for anaphoric mentions pronouns supersede definite descriptions.

4. Method

In this section, we describe the task to be performed, how we will perform the task and which features we have extracted from the text.

4.1. Task

Identification of speakers and addressees is realized as a sequence labeling task. For each chapter, a precompiled list of the characters appearing as speakers or addressees, along with their known aliases, is provided to the system. For each line in a dialogue, the system selects the most likely character from the character list.

A text is considered as a sequence of paragraphs and dialogues. A dialogue consists of n turns, each of which contains one or more lines. We consider each line as an independent unit with a speaker and an addressee label assigned to it. The task is to find the sequences of speakers and addressees that are most likely given the dialogue.

A variety of algorithms have been applied to sequence labeling tasks. For the current task the averaged perceptron (Collins, 2002) has been selected, due to its good performance and the efficient implementation it permits.[2]

4.2. Features

The features are based on information from the dialogue, the preceding narrative and the global context. Since we consider the task as a sequence labeling task, the previously selected speakers and addressees are also considered as features. The features are presented in Table 7, where each feature is binary.

Mention in Line: If a character is mentioned in a line, that character is likely relevant to the current line is some manner. The character mentions are captured for the current line by feature 1, and for the two preceding lines by features 2 and 3.

With Speech Verb in Line: Authors may indicate the speaker of a line explicitly by using their name with a

[2] Our implementation is freely available at https://github.com/adamlek/dialogue-fiction.

Id	Feature	Information Source
1	c_i mentioned in l_j	Dialogue
2	c_i mentioned in l_{j-1}	Dialogue
3	c_i mentioned in l_{j-2}	Dialogue
4	c_i + speech verb in l_j	Dialogue
5	c_i + speech verb in l_{j-1}	Dialogue
6	c_i + speech verb in l_{j-2}	Dialogue
7	c_i + speech verb/mentioned in l_j	Dialogue
8	c_i + speech verb/mentioned in l_{j-1}	Dialogue
9	c_i + speech verb/mentioned in l_{j-2}	Dialogue
10	$c_i = x_{i-k}$ $(k = 1 \ldots 6)$	Dialogue
11	c_i mentioned in narrative	Narrative
12	$count(c_i) = 0$ in narrative	Narrative
13	$0 < count(c_i) \leq 2$ in narrative	Narrative
14	$count(c_i) > 2$ in narrative	Narrative
15	$count(c_i) \geq 5$ in global context	Global context
16	$count(c_i) \geq 15$ in global context	Global context
17	c_i is n:th most recent mention	Global context
18	c_i is 0:th mention*	Global context
19	c_i is 0:th mention+speech verb in l_j*	Global context

Table 5: Feature templates used in the identification of speakers. l_j indicates the current line and features marked with an asterisk (*) are only used if the line contains an anaphoric pronoun.

speech verb. In these cases it is certain that character c is the speaker of the current line. Characters which occur with a speech verb in the current line is captured by feature 4, occurrences with speech verbs in two preceding lines are captured by feature 5 and 6.

Mention and Speech Verb in Line: Features 1 to 6 are usually strong indicators of participation in the dialogue. Given this, we combine the mention and speech verb features into one. Mentions and speech verbs for the current line is captured in feature 7. Features 8 and 9 capture mentions and speech verbs for the two preceding lines.

Hypothesis: A common heuristic applied to dialogues is that of turn-taking. The turn-taking heuristic states that in a dialogue between two characters, one character will occupy lines $l_j, l_{j-2} \ldots$ and the other character $l_{j-1}, l_{j-3} \ldots$. This pattern is captured by feature 10, which matches previously selected characters to the current character. The turn-taking is violated from time to time in the data, as such this heuristic is not implemented as a hard constraint.

Mention in Narrative: If a character is mentioned in the narrative of a dialogue, the character most likely has some relevance for the dialogue. We capture mentions in the narrative with feature 11.

Frequency in Narrative: The raw frequency of the characters in the narrative is captured by feature 12, 13 and 14. These features check (1) if the character is mentioned zero times, (2) if the character is mentioned one or two times and (3) if the character is mentioned more than two times.

Frequency in Global Context: In addition to the character frequency in the narrative, two features (15 and 16) capture the raw frequency of the character in the chapter at the current dialogue. We consider two thresholds, determined from the development set, (1) if the character occurs five or more times and (2) if the character occurs more than 15 times. The intuition behind these features is to capture import characters from a larger context.

Mention Order: The order in which the characters are mention is considered an import factor, where recently mentioned characters are likely participants of the current dialogue. A list is compiled from the order in which the characters are mentioned, where the first character is the most recent mention. The index of the current character is captured by feature 17.

Pronoun: Anaphoric pronouns appear in the lines, both with and without speech verbs. Two features (18 and 19) are designed to deal with them. Feature 18 is true if the current character is the most recent mention character and there is a pronoun in the line. Feature 19 is true if the current character is the most recent mention and the pronoun occurs with a speech verb.

For speaker identification, when a line contains a character c with a speech verb, we constrain the search to only consider hypotheses where the speaker of that line is c. This amounts to treating feature 4 as a hard constraint.

Our features are modeled in such a way to only capture information which has been given previously, e.g. we capture no information that appear after the current line.

4.3. Training and Evaluation

As mentioned in Section 4., training is performed using an averaged structured perceptron (Collins, 2002). A beam search with beam size 10 is used to keep several possible character sequences as the hypothesis. The hypothesis with the highest score is selected as the character sequence for the current dialogue.

The model's performance is estimated using cross-validation with the authors as folds, since we are primarily interested in the extent to which the features generalize across different author styles. The results are compared against three baselines for speakers and addressees, respectively:

1. Random baseline: For each dialogue, two characters are selected randomly, and are distributed in an alternating pattern across the lines.

2. Latest mentions with speech verb: The two latest characters that occurred with a speech verb are distributed over the lines in an alternating pattern (compare below). This corresponds to the baseline used by O'Keefe et al. (2012).

3. Latest mentions: The two latest mentioned characters are distributed in an alternating pattern across the lines.

For the second and third baselines, if there are less than two characters satisfying the conditions, the remaining characters are generated randomly. Also, the latest character to satisfy the conditions is designated as the first speaker and the second addressee, and the second character is assigned as the second speaker and first addressee.

5. Results

In this section, we report the results from our speaker and addresee identification experiments along with the results from our feature ablation.

5.1. Sequence Labeling

To test how well the model generalizes to other authors we test the authors against each other. Cross-validation was performed where each author represents a fold, e.g. one author is selected as the test set and the remaining authors as the training set. The results are presented in table 6.

TEST	RANDOM	LATEST	LATEST-VB	SEQ
		SPEAKER		
Strindberg	20.3	45.5	29.7	68.9
Sjöberg	26.8	44.7	27.8	73.4
Söderberg	29.0	53.7	33.3	70.9
Boye	32.0	34.7	36.3	41.3
Average	27.0	44.6	29.2	63.7
		ADDRESSEE		
Strindberg	17.5	46.3	35.6	42.4
Sjöberg	24.5	38.2	23.9	65.1
Söderberg	22.5	46.2	25.8	44.0
Boye	30.0	28.9	27.3	32.2
Average	23.6	39.9	28.1	46.0

Table 6: Accuracy of cross-author speaker and addressee identification compared with three baselines. RANDOM = random baseline; LATEST-VB = latest mentions with speech verb; LATEST = latest mentions; SEQ = accuracy of sequence labelling.

For speaker identification, the performance is around 70% for all corpora except for Boye, where the accuracy is at 41.3%. For addressee identification, the results tend to vary more. Two of the corpora, Söderberg and Strindberg, have an accuracy of 44% and 42.4%, while Sjöberg's accuracy is 65.1%. Again, Boye's accuracy is the lowest with only 32.2%.

Comparing speaker and addressee identification we see that the results are similar between them for Sjöberg and Boye, while the difference between speakers and addressees is larger for Söderberg and Strindberg.

5.2. Feature Ablation

To evaluate the impact of each feature, we performed a feature ablation experiment, the results of which can be found in Table 7.

For the feature ablation, we perform cross-validation using the authors as folds with each feature removed once. The results presented in Table 7 is the average accuracy change of the author cross-validation with the feature removed.

Generally, for speaker identification we see that the accuracy changes are rather small, with both positive and negative changes. For addressee identification, most of the accuracy changes higher than speaker identification and most changes are negative. The most important features appear to be feature 7 for speaker identification and features 8 and 10 for addressee identification.

ID	FEATURE REMOVED	SPK	ADD
1	c_i mentioned in l_j	$+0.2$	-0.6
2	c_i mentioned in l_{j-1}	-1.6	-2.8
3	c_i mentioned in l_{j-2}	$+0.6$	-1.5
4	c_i with speech verb in l_j	$+2.2$	-1.5
5	c_i with speech verb in l_{j-1}	-0.4	-2.9
6	c_i with speech verb in l_{j-2}	± 0.0	-4.1
7	c_i with speech verb/mentioned in l_j	-9.9	-2.3
8	c_i with speech verb/mentioned in l_{j-1}	$+3.0$	-6.6
9	c_i with speech verb/mentioned in l_{j-2}	$+0.7$	$+0.7$
10	$c_i = x_{i-k}$ $(k = 1, \ldots, 6)$	-1.3	-9.1
11	c_i mentioned in narrative	$+0.2$	-5.0
12	$count(c_i) = 0$ in narrative	$+2.3$	-2.9
13	$0 < count(c_i) \leq 2$ in narrative	$+0.5$	-1.3
14	$count(c_i) > 2$ in narrative	$+0.9$	-2.5
15	$count(c_i) \geq 5$ in global context	-2.6	-3.2
16	$count(c_i) \geq 15$ in global context	$+0.4$	-3.6
17	c_i is n:th most recent mention	$+0.1$	-6.0
18	c_i is 0:th mention	-0.1	-4.7
19	c_i is 0:th mention + speech verb in l_j	-0.3	$+0.5$

Table 7: Feature ablation results for speakers (SPK) and addressees (ADD). Differences are given in percentage points relative to the baseline results.

6. Discussion

In this section, we discuss the results from the author cross-validation and from the feature ablation. Additionally, we analyze a set of errors generated in the experiments manually.

6.1. Baselines

The results from Table 6 show that the latest mentions baselines (third baseline) seem to provide a better indicator for both speakers and addressees than the latest occurrence with speech verb (second baseline).

It is interesting to note that for the latter, the overall difference between speakers and addressees is only 1.1 percentage points. In contrast, the overall difference between speakers and addressees for the latest mentions baseline is 4.7 percentage points. This is quite surprising, as one would expect the latest mentions with speech verb baseline to perform better for speakers than for addressees, and the latest mentions baseline to be more balanced.

6.2. Model Accuracies

Comparing the results against the baselines we see that all models perform better than the random and latest occurrence with speech verb baseline. The system also generally performs better than the latest mentions baseline, but there are two exceptions. When testing against Strindberg and Söderberg and training on the other authors, the addressee identification model performance is 3.9 and 2.2 percentage points lower than the latest mention baseline.

The difference between speaker and addressee identification for the sequence model is 17.7 percentage points, compared to the difference between the latest mentions baseline which is 4.7 percentage points and the latest mentions with speech verb baseline where the difference is 1.1 percentage points.

The difference appears when the number of information sources increases from one to 19, that is, when more textual information is captured the difference in correct predictions between speaker and addressee identification grows. This would indicate that the speakers of a dialogue are signaled more explicitly than the addressees in the text, which would correspond to the notion that given the speakers of a dialogue, the addressees should be possible to infer from them.

Interestingly, observing Table 7 we can see that for speakers, 12 out of the 19 features produced accuracy changes of less than 1 percentage point, while for addressees only 3 features resulted in changes lower than 1 percentage point. This indicates that the features in general provide more information useful to the system regarding addressees than the speakers.

The notion that addressees are indicated to a lesser degree in the text thus seems implausible, given that the features applied to addressees show a larger change compared to speakers. However, if we consider the difference in the distribution of types presented in Table 3 for speakers and Table 4 for addressees, we see that there are significant differences for explicit, anaphoric and definite description indicators. Firstly, explicit speaker mentions are certain indicators that the character is the speaker, the indicator is less strong for addressees. An explicit mention in a line without a speech verb may be the addressee, but can also be a passer-by or a person not currently present. Thus, to predict addressees, the system must rely on other information sources than explicit mentions, which results in higher accuracy changes for more features.

Furthermore, definite descriptions are not treated in any special manner. This means that these indicators will be treated in the same manner implicit mentions. Anaphoric pronouns have some special treatment in the form of two heuristic features, 18 and 19. However, anaphora resolution is a complex problem and the current system will need a more sophisticated method for these in the future. We can see this clearly in the accuracy drop for speaker identification in the Boye corpus. In comparison to the other authors, the accuracy of Boye is roughly 30 percentage points lower. Observing Table 3 we see that approximately 50% of the speakers are indicated by anaphoric pronouns in Boye.

Given these differences of types and their impact, the differences in the feature ablation are not too surprising and the difference in accuracy between speakers and addressees makes sense.

6.3. Feature Ablation

6.3.1. Speakers

A tendency in dialogues is that there may be many different characters in a dialogue and that they may occupy lines in an irregular pattern. Another conflicting tendency is turn-taking, that states that there should be a regular alternating pattern between dialogue participants. These tendencies would primarily be caught by feature 8, which looks at the previous line. For speakers this features shows a gain of 3 percentage points while showing a loss of 6.6 percentage points for addressees. To some degree for speakers, feature 8 seem to encode both of these tendencies, which results in a feature which introduces uncertainty. For addressees on the other hand, the turn-taking pattern seems more consistent, where occurrence with a speech verb in the previous line is a good indicator of the current addressee.

Whereas mention and occurrence with speech verb is a strong indicator for addressees in the previous line, the feature (7) which encode this for the current line appears to be a strong indicator for speakers. The removal of this feature resulted in a loss of 9.9 percentage points.

That feature 7 and 8 show prominent losses for a particular role agrees with the notion that given a dialogue, the speakers and addressees tend to alternate. That is after one character passes the turn the other character responds to the first one, this scenario would be captured by observing that A speaks first, thus A is not the addressee, in the next line A was the previous speaker which indicates that A is currently being addressed.

6.3.2. Addressees

For addressee identification, there are quite a few features that show prominent accuracy losses.

The feature with the highest accuracy loss is the hypothesis feature (10), which captures the previously selected addressees, the removal of this features resulted in a loss of 9.1 percentage points. This would indicate that the structure of a dialogue seems to be an important factor, especially in comparison to speakers where the feature only shows a loss of 1.3 percentage points.

Another feature which resulted in a high loss for addressee identification is feature 11, which captures if the character is mentioned in the narrative or not. Removing this feature resulted in a loss of 5 percentage points, which indicates that presence in the narrative, i.e. the running text immediately preceding the dialogue, is an important aspect in addressee identification. For speaker identification, removing this feature only resulted in a marginal change of accuracy. Similarly, the feature that captures the relative order in which the characters are mentioned shows a loss of 6 percentage points for addressees, while for speakers only resulting in a marginal change in accuracy. This would indicate that the relative recency of a character is an important factor in addressee identification.

6.4. Error Analysis

Error analysis has been done by selecting 30 errors randomly from both speaker and addressee identification.

6.4.1. Sequence Hypothesis

A frequent problem that appears for both speakers and addressees is that the hypothesis feature (10) may reinforce an incorrect character sequence. In the model, this feature will favor an alternating pattern between two characters. If an incorrect character is selected early in the dialogue, the hypothesis feature will tend to reinforce the incorrect choice simply because it fits better with the current character sequence.

For example, consider the correct sequence ABABAB, where the current predictions are ACA. When considering the fourth prediction, the feature which captures the turn-taking heuristic (10) will favor C over B since C has occurred previously in the prediction while B has not. Given that C is predicted, followed by A, for the sixth predictions the turn-taking will be reinforced even further because C now appears twice in the predicted sequence. In this manner early incorrect predictions may reinforce or exclude the correct hypothesis.

Relatedly, characters talking about themselves, typically during introductions, may throw the speaker identification off track since characters are rarely mentioned in their own utterances. Lacking other strong evidence, this can cause a sequence like ABAB to be mistaken for BABA, if A introduces him/herself in the first line thereby making any hypothesis starting with A seem unlikely.

One particular problem found during the analysis is that our model assumes that all characters have been identified by name or alias(es), which makes dialogues involving unnamed participants (e.g., passers-by) challenging since the most useful features rely on mentions of names in the text. This error can often be seen for addressee identification where there are cases when one person is speaking to a group. A group is annotated as "SEVERAL" in the data, and thus will not be covered by most of the features. A more efficient way of handling these cases is to develop a set of features to determine if the content of a line is intended for a group of people or an individual.

In some cases, the hypothesis feature presents a problem because the dialogue does not follow a regular structure. These are dialogues where one speaker occupies two or more lines in a row, or dialogues in which there are more than two speakers.

6.4.2. Mention Order

Another problem that appears for both speakers and addressees is the mention order. In some cases, particularly in the beginning of a dialogue, the feature set of different characters tend not to show many differences. One feature however, will never be equal between the characters, namely feature 17 which capture the order in which the characters are mentioned in. This feature tends to work best in conjunction with other features, and not as the deciding factor. Resolving this is rather hard as there is not much that can be done to the feature itself. To avoid this problem, the other features used must show a larger variety to prevent one feature from being the deciding factor.

6.4.3. Narrative

A problem that appears for addressee identification is when the addressee is not mentioned in the narrative. Generally, mentions in the narrative are seen as a positive indicator for addressees, however many of the narratives are rather short and may not contain many mentions. Two possible improvements to remedy these situations is to capture mentions in previous narratives and to capture membership of previous dialogues.

A common narrative encountered in the data is the following: "...and then he said:", where the speaker is indicated with a speech verb, but the addressee is not present. These types of narratives present another problem for our system, namely that occurrences with speech verbs are not captured for narratives, and that certain narrative mentions are only relevant for speaker identification. Implementing a feature to capture this would work both as a positive indicator for speakers and a negative indicator for addressees.

6.4.4. Chapter

In some of the chapters there is a great deal of running text which includes many mentions while other chapters may primarily consist of dialogue. This presented a problem for feature 15 and 16 which capture how many times the character has been mentioned thus far. Given the different structures, different chapters in the training data were observed to assign quite different weights to this feature, which resulted in it being positive for some training sets and negative for others. To try and resolve this inconsistency there are some possibilities, one solution would be to convert the raw frequency into a relative frequency. We may also try to capture the structure of the chapters in other ways, where the total number of dialogues in comparison to paragraphs of running text is used as a feature, or as a factor in the frequency measurement.

7. Conclusions and Future Work

We have described a general method for identifying both speakers and addressees in dialogues extracted from literary fiction. The dataset we have used is small, but given that our results are based on out-of-domain training data we regard the approach as promising. This is an important aspect as it removes the need of having annotated data for each author investigated.

Direct comparison with previous work is difficult, primarily because of differences in the data-size and the experimental setup. However, our results on speaker identification are relatively similar to previously obtained results (Elson et al., 2010; He et al., 2013; Muzny et al., 2017).

For addressee identification, the only other results are from Yeung and Lee (2017). Using out-of-domain data they report a high loss of accuracy. Our proposed method seems to handle out-of-domain data efficiently.

For future work, definite descriptions and pronouns will have to be handled by a co-reference system. Based on the feature ablation and the error analysis there are several improvements that can be made to the current system. Furthermore, currently the characters are given to the system. However, it would be much more interesting to also extract these automatically. This would benefit the system as all the candidates considered will be somewhat relevant, meanwhile this need not be true for the current implementation.

Another interesting addition would be to apply a social network analysis on the chapter and use the relationships between the current character and the previously selected characters as a feature.

8. Acknowledgements

The work presented here has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003). We gratefully thank Litteraturbanken at University of Gothenburg and the Strindberg Project at Stockholm University for providing us with the electronic version of the book material.

9. Bibliographical References

Agarwal, A., Corvalan, A., Jensen, J., and Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. In *Proceedings of the Workshop on Computational Linguistics for Literature, CLfL@NAACL-HLT 2012, June 8, 2012, Montréal, Canada*, pages 88–96.

Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

Elson, D. K., Dames, N., and McKeown, K. R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 138–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria, August. Association for Computational Linguistics.

Koivisto, A. and Nykänen, E. (2016). Introduction: Approaches to fictional dialogue. *International Journal of Literary Linguistics*, 5.

Moretti, F. (2011). Network Theory, Plot Analysis. *Literary Lab, Pamphlet 2*, May.

Muzny, G., Fang, M., Chang, A., and Jurafsky, D. (2017). A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain, April. Association for Computational Linguistics.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. , 69(2):026113, February.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799. Association for Computational Linguistics.

Rydberg-Cox, J. (2011). Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(3).

Vala, H., Dimitrov, S., Jurgens, D., Piper, A., and Ruths, D. (2016). Annotating characters in literary corpora: A scheme, the charles tool, and an annotated novel. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Yeung, C. Y. and Lee, J. (2016). An annotated corpus of direct speech. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Yeung, C. Y. and Lee, J. (2017). Identifying speakers and listeners of quoted speech in literary works. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 325–329. Asian Federation of Natural Language Processing.