Identifying Speakers and Listeners of Quoted Speech in Literary Works

Chak Yan Yeung and **John Lee**
Department of Linguistics and Translation
City University of Hong Kong
chak.yeung@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

We present the first study that evaluates both speaker and listener identification for direct speech in literary texts. Our approach consists of two steps: identification of speakers and listeners near the quotes, and dialogue chain segmentation. Evaluation results show that this approach outperforms a rule-based approach that is state-of-the-art on a corpus of literary texts.

1 Introduction

A literary work can be analysed in terms of its conversational network, often encoded as a graph whose nodes represent characters, and whose edges indicate dialogue interactions between characters. Such a network has been drawn for *Hamlet* (Moretti, 2011), Classical Greek tragedies (Rydberg-Cox, 2011), as well as a set of British novels (Elson et al., 2010).

To automatically construct these networks, it is necessary to identify the speakers and listeners of quoted speech. Past research on quote attribution has mostly focused on speaker identification (O'Keefe et al., 2012; He et al., 2013). In the only previous study that attempts both speaker and listener identification (Elson et al., 2010), there was no formal evaluation on the listeners. Listener identification can be expected to be challenging, since they are more often implicit.

This paper presents the first evaluation on both speaker and listener identification, with two main contributions. First, we present a new model that incorporates dialogue chain segmentation. We show that it outperforms a rule-based approach that is state-of-the-art on a corpus of literary texts. Second, training data from the same author, or even from the same literary genre, cannot be assumed in a realistic scenario. We investigate the amount of training data that is required for our statistical model to outperform the rule-based approach.

2 Previous Work

Among rule-based approaches on speaker identification, most rely on speech verbs to locate the speakers (Pouliquen et al., 2007; Glass and Bangay, 2007; Liang et al., 2010; Ruppenhofer et al., 2010).

For machine learning approaches, Elson and McKeown (2010) treated the task as classification, using features such as the distance between quotes and speakers, the presence of punctuation marks, etc. O'Keefe et al. (2012) reformulated the task as a sequence labelling task. In the news domain, their statistical model outperformed a rule-basd approach; in the literary domain, however, the rule-based approach achieved the best performance. This rule-based approach will be compared with our proposed approach in our experiments. Similar to our approach, He et al. (2013) parsed the sentences near the quotation. Their method, however, includes a manual preprocessing step that extracts all name mentions in the text and clusters them into character aliases, and so cannot be directly compared with our approach, which does not require preprocessing and thus can be more easily applied to larger datasets.

3 Approach

Our task is to identify the speakers and listeners of the quotes in a text. We count the system output as correct if it identifies the specific text spans that indicate the speakers and listeners of the quote. When the text only implicitly identifies the speakers and listeners, the mentions of these characters that are closest to the quote are considered the gold answer.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 325–329,
Taipei, Taiwan, November 27 – December 1, 2017

We use a simple rule-based method to extract all direct quotes. We then perform quote attribution in a two-step process: speaker and listener identification from context, and dialogue chain segmentation.

3.1 Speaker and listener identification from context

The system first identifies the speakers and listeners in the context around the quote. We extract two sentences before and two sentences after each quote for tagging, excluding the words within the quote. We adopt a sequence labelling approach, using a CRF model (Lafferty et al., 2001) to tag each word as one of 'speaker', 'listener', or 'neither'. For each word, we extract the following features:

Word identity: The word itself.

POS: The POS tag is useful, for example, in capturing the text spans that indicate possible entities.

Head: The head of the word, extracted from the dependency tree, can captue verbs that indicate direct speech.

Dependency relation: The grammatical relation between the word and its head. As shown by He et al. (2013), dependency relations of reported speech verbs are useful in extracting speakers, with 'nsubj' usually suggesting a speaker, and 'nmod' or 'dobj' favoring a listener.

Sentence distance: The location of the sentence in relation to the quote (-1, -2, +1, or +2).

Paragraph distance: The number of paragraphs that separate the word and the quote. This captures the observation that a new paragraph usually signifies a change of speaker and listener.

Matching word in quote: A binary feature of whether the word can be found within the quote. As pointed out by He et al. (2013), the speaker name is rarely found within the quote while the reverse is true for the listener.

Initial word and POS: The first word in the sentence and its POS tag are useful in capturing the pattern of "[speech verb] [speaker]" that is often found after a quote (e.g. "..." said Peter.).

3.2 Dialogue chain segmentation

In this step, we segment the quotes in a text into dialogue chains. Each quote can be a continuation of a dialogue, i.e., its speaker (listener) is the listener (speaker) of the preceding quote; otherwise, it is the beginning of a new chain. As shown in Table 1, we label each quote either as `B`(egin) or `C`(ontinue).

Sentence	Tag
(1) *A centurion came to him, asking for help: "..."*	`B`
(2) *Jesus said to him, "..."*	`C`
(3) *But the centurion replied, "..."*	`C`
(4) *When Jesus heard this he was amazed and said to those who followed him, "..."*	`B`

Table 1: Quotes (1) to (3) form a dialogue chain; (4) starts a new one.

Similar to the first step, we use a sequence labelling approach with a CRF model (Lafferty et al., 2001). The quotes in the whole document are seen as one single sequence. For each quote, we extract the following features:

Distance: The number of sentences separating the current quote from the preceding one. The closer the two quotes are, the more likely it is for them to be in the same dialogue chain.

Speaker/Listener identity: Within a dialogue chain, the speaker and listener of the n^{th} quote are the same as those in the $(n+2)^{\text{th}}$ quote; their identities are reversed, however, in the $(n+1)^{\text{th}}$ quote. To capture these patterns, we include eight binary features that compare the predicted speakers and listeners of the current quote with those of the $(n \pm 1)^{\text{th}}$ quote and the $(n \pm 2)^{\text{th}}$ quotes.

Implicit: Two binary features — no extracted speakers and listeners — capture the observation that from the third quote in a dialogue chain, the speakers and listeners can sometimes be omitted.

Pronoun: Two binary features — the extracted speakers and listeners being pronouns — capture the observation that from the third quote in a dialogue chain, the speakers and listeners can be referred to as pronouns.

After tagging (Table 1), the system fills in any missing speakers and listeners. If two consecutive quotes belong to the same chain, the system will infer the speaker (listener) of one quote to be the listener (speaker) of the other.

4 Baselines

4.1 Speaker and listener identification

Distance baseline. We re-implemented the rule-based approach that was state-of-the-art for literary texts (O'Keefe et al., 2012), achieving the best

performance on a re-annotated version of the Elson et al. (2010) dataset. We take as entities all pronouns and all words tagged as person and organization by the Stanford NER tagger (Finkel et al., 2005). We compiled a list of quotative verbs by retrieving the verbs closest to the quotes in the training set.[1]

Dependency baseline (Dep). We parsed the sentence that contains the quote, excluding the words within the quote and replacing the trailing comma, if any, with a full-stop. If a quotative verb is modified by a word with the dependency relation 'nsubj', that word is extracted as the speaker; if it is modified by a word with 'dobj' or 'nmod', that word is extracted as the listener.

4.2 Dialogue chain segmentation

Elson et al. (2010) used the distance between two quotes to determine whether they belong to the same chain. We use the same feature to train a CRF model for chain segmentation.

5 Data

We tested our approach on two datasets: the novel *Emma* and the *New Testament*.

The *Emma* set was taken from the corpus of 19th-century British novels compiled by Elson and McKeown (2010). Since the original annotations did not annotate listeners and did not indicate the text span of the speaker that is connected with the quotation, we performed re-annotation on the *Emma* portion of the corpus.[2] This dataset contains 737 quotes; 63% of the quotes belong to dialogue chains of length two or more. We performed a four-fold cross validation on this set since each fold cannot have too few dialogue chains for meaningful evaluation on chain segmentation.

The *New Testament* (NT) set contains a total of 1628 quotes[3]; 43% of the quotes belong to dialogue chains of length two or more. We divided the text into seven folds following a natural division of the books.

[1]We require each verb to be attested at least two times in the training set.

[2]Two annotators re-annotate the first 100 quotes. The kappa was 0.89/0.83 for speakers/listeners. Most disagreements involved conversations with three or more characters, where the identity of the listener was often unclear. One of the annotators completed the rest of the re-annotation.

[3]Please see Lee and Yeung (2016) for details of this dataset. The paragraph distance feature was omitted since the corpus did not contain information on paragraphs.

6 Experimental results

We used the Stanford parser (Manning et al., 2014) for POS tagging and dependency parsing, and CRF++ (Kudo, 2005) for training CRF models.

6.1 In-domain

As shown in Tables 2 and 3, for *Emma*, our approach achieved an average accuracy of 52.46/28.46 for speakers/listeners. For the NT, the average accuracies for speakers and listeners are 66.09 and 56.97.[4] For both datasets, our approach significantly outperformed[5] both baselines.

Dataset →	***Emma***		***NT***	
↓ Model	no seg	w/ seg	no seg	w/ seg
Distance	40.06	43.73	45.39	45.61
Dep	8.98	10.20	56.63	56.63
Proposed	42.11	**52.46**	66.05	**66.09**

Table 2: Speaker identification accuracy, before and after dialogue chain segmentation (Section 3.2).

Dataset →	***Emma***		***NT***	
↓ Model	no seg	w/ seg	no seg	w/ seg
Distance	6.95	13.62	24.91	26.92
Dep	5.31	6.53	25.29	29.83
Proposed	17.99	**28.46**	52.94	**56.97**

Table 3: Listener identification accuracy, before and after dialogue chain segmentation (Section 3.2).

As shown in Table 4, our dialogue chain segmentation achieved 89.8 precision and 78.3 recall for *Emma*, and 90.2 precision and 89.5 recall for the NT. It significantly improved[6] the F-measure over the distance baseline.

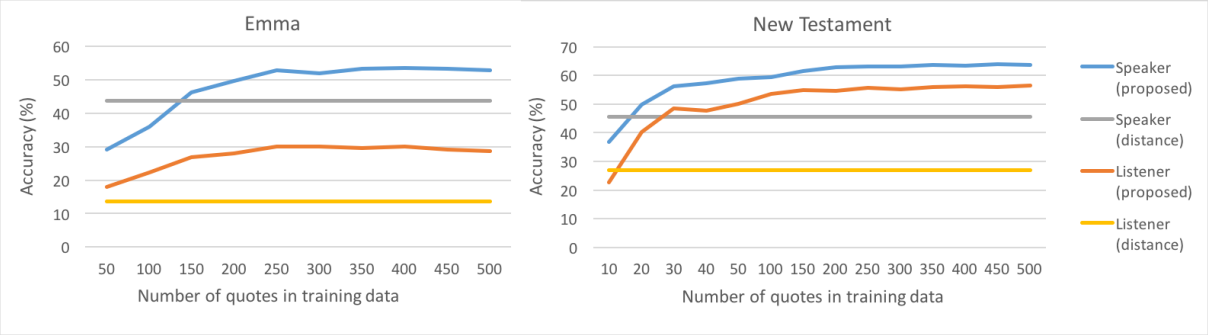
The segmentation step yielded a higher degree of improvement in accuracy for *Emma* than for the NT, due to differences in literary style. The system was often able to extract speakers and listeners in the NT from the context around the quote, and so did not benefit much from dialogue chain boundary. In novels such as *Emma*, however, the listeners were often not specified in the context around

[4]Among utterances with no speaker/listener, the system correctly output "no speaker" at 60% (out of 10) in Emma and 29% (out of 7) in NT, and correctly output "no listener" at 73% (out of 42) in Emma and 80% (out of 195) in NT.

[5]At $p \leq 0.001$ by McNemar's test for all cases.

[6]At $p \leq 0.001$ by McNemar's test.

Figure 1: Accuracy of the proposed approach and distance baseline on training sets of different sizes for *Emma* (left) and the *New Testament* (right).



the quote and the speakers were sometimes omitted. This led to lower performance in the first identification step, which only considered the context around the quote, but greater improvement with the segmentation step.

Most errors involved speakers embedded within a long description, where the act of speaking was not explicitly stated. Consider "... and Harriet then came running... which Miss Woodhouse hoped very soon to compose.", which serves as the description of a quote. Harriet's role as the speaker was only implied and could not be captured by the system. Another source of error was the verb "hear", which reversed the usual pattern of the subject being the speakers and the object being the listeners. For example, in the sentence "They were only hearing, '...'", the system tagged the word "they" as the speaker instead of as the listener. Overall, listener identification is less accurate than speaker, because listeners were less often explicitly stated. It is particularly challenging in single-quote chains, where the preceding and following quotes cannot provide hints.

Dataset → ↓ **Model**	***Emma*** P/R/F	***NT*** P/R/F
Elson et al.	91.2/54.2/67.9	90.3/66.4/76.5
Proposed	**89.8/78.3/83.7**	**90.2/89.5/89.9**

Table 4: Precision/recall/F-measure for our dialogue chain segmentation (Section 3.2), and that of Elson et al. (2010).

6.2 Out-of-domain

One advantage of the distance-based baseline is lesser reliance on in-domain training data. Indeed, our statistical approach benefits from learning the names of the frequent speakers and listeners, as well as the speech-reporting style, from the same text. In practice, however, one may not assume the availability of a large amount of training data from the same text or author, or even from the same literary genre. We therefore re-investigate the performance of our approach when it has no access to the character names, and when it has mismatched, or limited in-domain training data.

Frequent speakers/listeners. To eliminate knowledge of the frequent speakers and listeners gained by our proposed model, we replaced all words tagged as entities with "PERSON". In this setting, the speaker and listener identification accuracy of our proposed approach decreased to 51.22/27.78 for *Emma* and 65.97/57.15 for the NT. These results, however, are still significantly better than both baselines.[7]

Limited training data. The accuracy dropped to below 25% for all cases with mismatched training data. When trained on the NT, the system failed to capture speakers that appear after the quote (e.g. "..." said her father.), a pattern common in *Emma* but rare in the NT. Inversely, when trained on *Emma*, the system could not recognize the frequent NT pattern "[speaker] said to [listener]".

As an alternative solution, we investigated how much in-domain data would be needed for our statistical model to outperform the distance-based baseline. As shown in Figure 1, relatively little annotation effort would be sufficient: our model significantly outperformed all baselines with a training set of 200 quotes for *Emma*, and a training set of 20 quotes for the NT.[8]

[7] $p \leq 0.001$ by McNemar's test.
[8] $p \leq 0.001$ by McNemar's test for all cases.

7 Conclusion

We have proposed a novel approach for quote attribution that incorporates dialogue chain segmentation. We report the first evaluation on listener identification. For speaker identification, we show that our approach outperforms the state-of-the-art rule-based approach for literary texts (O'Keefe et al., 2012). Further, we show that our results can be generalized to out-of-domain literary texts with a modest amount of training data annotation.

Acknowledgments

The authors gratefully acknowledge support from the CityU Internal Funds for ITF Projects (no. 9678104).

References

D. K. Elson, N. Dames, and K. R. McKeown. 2010. Extracting social networks from literary fiction. In *Proc. Association for Computational Linguistics (ACL)*.

David Elson and Kathleen McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. In *Proc. AAAI*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 363–370.

Kevin Glass and Shaun Bangay. 2007. A Naive Salience-based Method for Speaker Identification in Fiction Books. In *Proc. 18th Annual Symposium of Pattern Recognition*.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of Speakers in Novels. In *Proc. ACL*.

Taku Kudo. 2005. *CRF++: Yet Another CRF Toolkit*. http://taku910.github.io/crfpp/.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proc. International Conference on Machine Learning*.

John Lee and Chak Yan Yeung. 2016. An annotated corpus of direct speech. In *Proc. LREC*.

J. Liang, N. Dhillon, and K. Koperski. 2010. A large-scale system for annotating and querying quotations in news feeds. In *Proceedings of the 3rd International Semantic Search Workshop*. pages 1–5.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*. pages 55–60.

F. Moretti. 2011. Network theory, plot analysis. *New Left Review* 68.

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A Sequence Labelling Approach to Quote Attribution. In *Proc. EMNLP*.

B. Pouliquen, R. Steinberger, and C. Best. 2007. Automatic detection of quotations in multilingual news. In *Proc. Recent Advances in Natural Language Processing*. pages 487–492.

Josef Ruppenhofer, Caroline Sporleder, and Fabian Shirokov. 2010. Speaker Attribution in Cabinet Protocols. In *Proc. LREC*.

J. Rydberg-Cox. 2011. Social networks and the language of greek tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(3).