Cross-Lingual Speaker Identification from Weak Local Evidence

Anonymous ACL submission

Abstract

Speaker identification, determining which character said each utterance in text, benefits many downstream tasks. Most existing approaches use expert-defined rules or rule-based features to directly approach this task, but these approaches come with significant drawbacks, such as lack of contextual reasoning and poor cross-lingual generalization. In this work, we propose a speaker identification framework that addresses these issues. We first extract large-scale distant supervision signals in English via general-purpose tools and heuristics, and then apply these weakly-labeled instances with a focus on encouraging contextual reasoning to train a cross-lingual language model. We show that our final model outperforms the previous state-of-the-art methods on two English speaker identification benchmarks by 5.4% in accuracy, as well as two Chinese speaker identification datasets by up to 4.7%.

1 Introduction

Speaker identification (also called quote attribution) is the task of deciding which character said or implied each quote/utterance in a document (Elson and McKeown, 2010). It is mostly studied in the domain of literature and novels because, unlike news, the speakers in stories are often not explicitly specified by a name. This task directly benefits many downstream applications such as character detection (Vala et al., 2015), character profiling (Kokkinakis and Malm, 2011), and text-to-speech (Iosif and Mishra, 2014). While good systems exist (e.g., Muzny et al. (2017) report >80% accuracy), speaker identification is still challenging. As speaker identification datasets are usually too small-scale to sufficiently train large models, most previous work directly rely on language-specific patterns and heuristics, which cannot sufficiently solve hard cases (e.g., those that are implicit and require contextual reasoning). This kind of knowledge also cannot be easily transferred to other languages, limiting cross-lingual performances.

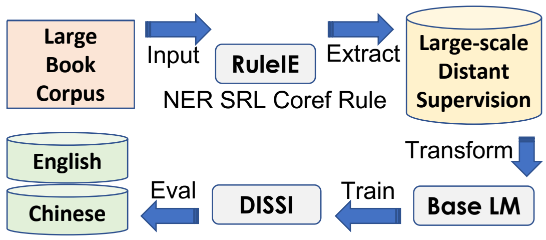

Figure 1: Overview of our framework. RULEIE extracts incidental supervisions that are used to train DISSI.

In this work, we address these issues with a novel framework for cross-lingual speaker identification *without relying* on any domain, task, or language-specific annotation. The framework, as overviewed in Fig. 1, starts with extracting large-scale distant and incidental supervision (Roth, 2017) from unstructured corpora. **We propose a rule-based system** called RULEIE to do this (§3). **We collect 100K weakly-labeled instances** with RULEIE and transform them to encourage more contextual reasoning (§4). **We train a cross-lingual language model** (LM) (Conneau et al., 2020) with the resulting data and name the resulting model DISSI (**Di**stantly-**S**upervised **S**peaker **I**dentification). We hypothesize that DISSI may improve cross-lingual performance because the speaker identification task shares many language-invariant features (§5).

Experimental results[1] show that DISSI achieves state-of-the-art English performance on the P&P dataset (He et al., 2013), improving 2.4% in the unsupervised setting, and 5.4% with full supervision. With minimum language-specific efforts, our cross-lingual model also outperforms state-of-the-art methods on two Chinese datasets WP (Chen et al., 2019, 2021) and Jinyong (Jia et al., 2020), by up to 4.7%. Comparing to the baseline LM, our distant supervision brings an improvement of more than 40% in realistic few-shot settings.

[1]We will release all code and data upon publication.

2 Related Work

Speaker Identification. Language-specific expert-designed rules, patterns, and features (Elson and McKeown, 2010; He et al., 2013; Muzny et al., 2017; Ek et al., 2018) are widely used to identify speakers. To leverage large unlabeled corpora, previous work (Pavllo et al., 2018) starts from a small number of seed patterns and obtains more lexical patterns by conducting an unsupervised bootstrapping, which however will lead to semantic drifts, and pattern-based methods usually suffer from low recall. This work studies the usage of high-precision heuristics and patterns, which fully leverage coreference resolution information, to build distant supervision data without hurting model generalization. In addition, previous cross-lingual studies in this direction mainly focus on direct speech identification (Kurfali and Wirén, 2020; Byszuk et al., 2020). To the best of our knowledge, this is the first work on cross-lingual speaker identification without the need for redesigning rules, patterns, and features for a new language.

Indirect Supervision and LM. Studies have shown that distant supervision is effective in bridging the knowledge gaps in pre-trained LMs (Zhou et al., 2020, 2021). People have also discussed the ability of LMs to learn from indirect but related supervision signals (Khashabi et al., 2020).

3 English Speaker Extraction

In this section, we introduce a rule-based information extraction system named RuleIE: it receives a long document as input and output (context, utterance, speaker) triples in the document. RuleIE can be directly applied to identify speakers in English texts in a given dataset, but we mainly use it[2] to automatically extract incidental signals that approximates the target task from unlabeled corpora, which is later used as distant supervision to train our cross-lingual system Dissi in §5.

3.1 Main Heuristics

The core of this RuleIE component follows three basic rules. Inspired by previous work (He et al., 2013; Muzny et al., 2017), we design the first two: direct speaker identification for explicit speakers and conversational pattern for implicit speakers (i.e., no speaker mentions exist in the nearby context). We introduce a novel and intuitive third rule based on local coreference to further improve the precision and recall of this component.

Direct Speaker Identification. We use semantic role labeling (SRL) to identify direct speakers (e.g., ***Mary*** *said: "..."*). We construct a list of 113 speech verbs (e.g., *"say"* and *"answer"*).[3] If an utterance is either ARG-1 or ARG-2 in a frame whose verb exists in this list, we treat the ARG-0 of that frame as the direct speaker. If the speaker mention is named (e.g.,*"Mary"* but not *"his sister"*), we assign the utterance to the corresponding character.

Conversational Pattern. Often times, the speaker names for some utterances are implicit because of ongoing dialogues between a limited amount of characters (typically two). In these cases, we, the readers, may identify the speakers by tracking the alternation. As a result, if multiple utterances are not separated by additional context, we decide that a given utterance is not from the speaker of the immediate previous or next utterance, but are likely from the same speaker of the skip-utterances.

Local Coreference Resolution. Previous work only use coreference resolution (coref) to resolve direct speaker mentions (Muzny et al., 2017). We extend the application of coref to all pronouns in the utterance, because i) any linked names of first-person pronouns (*"I"*, *"me"*, *"my"*) indicates the actual speaker and ii) those of second and third-person pronouns (*"you"*, *"she"*, and *"they"*) are excluded from the candidate speakers. We run coref on every three-sentence-windows to avoid mistakes made by trying to reduce the number of clusters. Empirically, we find that modern coref tools perform reasonably well on short literal texts, even when the texts contain dialogue alternations.

3.2 Iterations and Voting

RuleIE runs in iterations with different heuristics for best precision-recall tradeoff. In the first iteration, it extracts direct speaker mentions, collect all person names, and try to link other nominal/pronouns to a name. We do this first to introduce only high-confidence predictions to the following two iterations, which use conversational patterns (noise-sensitive) and pronoun coreference resolution. Instead of using a hard assignment that may produce conflicts, we let each rule to "vote" or "vote against" for a speaker and assign the character with the highest vote count to each utterance.

[2]This is because RuleIE is not guaranteed to produce a predicted speaker for every utterance.

[3]We will also release the speech verb list.

4 Distant Supervision Acquisition

We hypothesize that the speaker identification task shares many commonalities across languages (e.g., the patterns people use to describe explicit, implicit, and anaphoric speakers in texts). If we can do well on one language, we may improve on other languages with the help of cross-lingual language models. In this section, we describe how we use RULEIE to acquire large-scale English speaker identification instances as distant supervision.

4.1 Automatic Extraction

We use Project Gutenberg, which contains over 60,000 books, as the source corpus.[4] We identify sentences that contain at least one utterance by simply running a sentence chunker and finding quotation marks in each sentence. As a result, we collect 1.5M sentences that contain utterances and their surrounding context. For each sentence, we run named entity recognition (NER) to find person-named entities in the chapter that includes the sentence and use them as candidate characters. We then run RULEIE to try to assign characters to utterances. From the raw sentences, we extract 100K (context, utterance, speaker) triples. We view these triples as distant supervision as they are automatically collected (therefore with a certain level of noise) from external resources and do not rely on any task or domain-specific annotation.

4.2 Contextual Reasoning with Masking

As argued in §1, we need to build models that approach speaker identification with contextual understanding and reasoning. However, many of automatically extracted instances have explicit speakers (53% discussed in § 6.4) and do not contribute much to a stronger reasoning model. As an improvement, we mask explicit speaker mentions with *"someone"* with a probability of 15%, so that models are forced to use other textual clues to identify the speaker, which often times involve understanding the story and the context. In addition, to avoid the model overfitting on speaker names, which are relatively irrelevant in determining who said each utterance, we randomly assign each character a masked name *"Person [X]"* (where [X] is a letter except those meaningful letters (e.g., *"A"* and *"I"*), and we replace corresponding mentions in the input context with the masked name.

[4] https://www.gutenberg.org/ (books are not protected by copyright laws and distributed for free use).

5 Cross-Lingual Model

Given the large amount of English-based distant supervision, we explore the possibility of transferring mono-lingual signals to cross-lingual applications, under the help of pre-trained cross-lingual LMs. In this section, we propose and describe DISSI.

5.1 Model Formulation

We formulate the data into a span-selection task. We use the previous three sentences and the next two sentences, together with the sentence containing the target utterance, to form an input document. For each document, following previous work, we assume a given list of characters and their *named* aliases. For the distant supervision data, we approximate such lists via NER and span overlap.

We format the list of character names and an input document as `People:[C-1][C-2]...[C-N] [SEP] [Document]` and a corresponding question that specifies the target utterance `who said "[U]"?`. Here `[C-1]...[C-N]` are the character names in the document, `[SEP]` is a model-specific separator token, `[Document]` is the input document, and `[U]` is the target utterance, which is a sub-string of the input document. The labels are the span start and end indices of the speaker mention (one of `[C-1]...[C-N]`) in character list provided at the beginning of the input.

6 Experiments

6.1 Data and Baselines

For English, we use Pride & Prejudice (P&P) and its official splits and settings (He et al., 2013). We shorten the utterances if they are too long and replace character mentions with masked names following §4.2. We also report results on the Emma dataset (Muzny et al., 2017), but we remove 127 test instances due to conflicting aliases (dataset error), hence making the comparison on Emma with previous work indirect. For Chinese, we use two datasets, one based on *Jinyong* novels (JY) and another based on novel *World of Plainness* (WP).

We compare with published best results on each dataset, and the baseline language model in multiple settings. Emma does not provide training data, so no in-domain numbers are reported.

6.2 Implementation Details

We use AllenNLP (Gardner et al., 2017) for SRL, NER, and coref. As base LMs, we use RoBERTa-large (Liu et al., 2019) for English and XLMR-

System	Supervision	*P&P*	*Emma*
Muzny et al. (2017)	no	83.6	(75.3)
Muzny et al. (2017)	in-domain	85.2	–
RoBERTa	in-domain	71.1	–
DISSI-R w/o masking	no	85.2	79.1
DISSI-R	no	86.0	**81.2**
DISSI-R	in-domain	**90.6**	–

Table 1: Accuracy (%) on English speaker identification datasets. Supervision in *w/o masking* is not masked per §4. Numbers in parentheses are for reference only. DISSI-* are our proposed systems.

System	Supervision	*JY*	*WP*
MLP†	in-domain	95.6	70.5
CSN†	in-domain	–	82.5
XLMR	in-domain	98.3	53.4
DISSI-X	in-domain+distant	**98.4**	**87.2**
XLMR	mini	51.7	40.9
DISSI-X	mini+distant	**95.6**	**67.8**
Random†	no	33.7	37.6
DISSI-X	no	**70.7**	**50.3**

Table 2: Accuracy (%) on Chinese speaker identification datasets (†: numbers from (Jia et al., 2020) and (Chen et al., 2021)). *Mini* uses 200 in-domain instances.

large (Conneau et al., 2020) for other languages such as Chinese. Both LMs are trained on our distant supervision data for one epoch, which we denote as DISSI-R and DISSI-X respectively. We report single-run results. We use Transformers (Wolf et al., 2020) and default parameters. Both runs finish in an hour with single RTX A6000.

Inference. For English evaluation, we apply an inference process similar to §3 to both the baseline LM and our proposed LM with distant supervision. We treat any *named* mentions identified as direct speakers as final predictions. If the direct speaker mention is a pronoun that indicates genders (e.g., *he*, *she*), we remove all gender-incompliant candidates. We also apply conversational patterns onto the output probabilities to achieve maximum likelihood for any conversational sequences.

6.3 Main Results

Table 1 compares English speaker identification accuracy with state-of-the-art (SOTA) numbers (Muzny et al., 2017; Yoder et al., 2021). DISSI-R outperforms previous SOTA results by 5.4%. The masking process proposed in §4.2 evidently contributes to this gain, improving as much as 2.1%.

Table 2 shows performance on Chinese benchmarks. With full supervision, our model DISSI-X improves 2.8% and 4.9% over previous SOTA on *JY* and *WP* respectively, and it gains 34% over the XLMR baseline on *WP*. We also achieve comparable performance (+44%) on *JY* with only 200 training instances (*Mini*).

System	Explicit	Anaphoric	Implicit
XLMR	52.3	54.2	48.3
CSN†	93.2	81.3	75.9
DISSI-X	**97.7**	**84.9**	**89.7**

Table 3: Accuracy (%) by type according to the WP dataset. Results are produced with full supervision.

As Table 3 shows, we find that our method outperforms previous methods on identifying all three types of speakers by a large margin. On the WP dataset that provides ground truth type labels for instances, for the most challenging implicit category, our method obtains a 13.8% improvement compared with the state-of-the-art performance.

6.4 The Quality of Weakly-Labeled Data

Based on 100 random extractions from §4, we find that 29% require contextual reasoning as no direct evidence exists. In the following example, the speaker of the utterance *"I wasn't far...been there."* is correctly identified (Person X).

... </s> ``It is always the way," said Person X. ``If you miss a day, it is sure to be the best thing of the season. An hour and a quarter with hardly anything you could call a check! It is the only very good thing I have seen since I have been here. Person T was with them all through." </s> ``And I suppose you were with Person T. " </s> <u>``I wasn't far off. I wish you had been there."</u> ...

This, to some extent, explains the large gain achieved by our method on the implicit instances as shown in Table 3. The accuracy of RULEIE on the selected samples is 68%.

7 Conclusion and Future Work

In this work, we propose a multi-step framework for speaker identification that includes **i)** RULEIE, a ruled-based system which we use to extract **ii)** 100K distant supervision instances. We use them to train **iii)** a cross-lingual model DISSI that outperforms previous bests on English and Chinese benchmarks, by as much as 5.4%, and over 40% in low-resource settings. The limitations of our work also inspire future directions, which may include i) improving distant supervision accuracy, ii) proposing global inference for long documents that cannot fit into LMs, and iii) auto-learning and generalizing rules and heuristics such as those in §3 on the fly.

References

Joanna Byszuk, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of the LT4HALA*, pages 100–104.

Jia-Xiang Chen, Zhenhua Ling, and Lirong Dai. 2019. A chinese dataset for identifying speakers in novels. In *Proceedings of the INTERSPEECH*, pages 1561–1565.

Yue Chen, Zhen-Hua Ling, and Qing-Feng Liu. 2021. A neural-network-based approach to identifying speakers in novels. *Proceedings of the INTERSPEECH*, pages 4114–4118.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the ACL*, pages 8440–8451.

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam, Gintarė Grigonytė, and Sofia Gustafson Capková. 2018. Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the LREC*, pages 817–824.

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the AAAI*, pages 1013–1019.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. In *Proceedings of the NLP-OSS*, pages 1–6.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the ACL*, pages 1312–1320.

Elias Iosif and Taniya Mishra. 2014. From speaker identification to affective analysis: A multi-step system for analyzing children's stories. In *Proceedings of the CLFL*, pages 40–49.

Yuxiang Jia, Huayi Dou, Shuaiying Cao, and Hongying Zan. 2020. Speaker identification and its application to social network construction for chinese novels. *Proceedings of the IALP*, pages 13–18.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single QA system. In *Findings of the EMNLP*, pages 1896—-1907.

Dimitrios Kokkinakis and Mats Malm. 2011. Character profiling in 19th century fiction. In *Proceedings of the LaTeCH*, pages 70–77.

Murathan Kurfali and Mats Wirén. 2020. Zero-shot cross-lingual identification of direct speech using distant supervision. In *Proceedings of the LaTeCH-CLfL*, pages 105–111.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*, cs.CL/1907.11692v1.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the EACL*, pages 460–470.

Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. In *Proceedings of the ICWSM*, pages 231–240.

Dan Roth. 2017. Incidental supervision: Moving beyond supervised learning. In *AAAI*.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the EMNLP*, pages 769–774.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP*, pages 38–45.

Michael Miller Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Penstein Rosé. 2021. Fanfictionnlp: A text processing pipeline for fanfiction. In *Proceedings of the WNU*, pages 13–23.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the ACL*, pages 7579–7589.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the NAACL-HLT*, pages 1361–1371.