Multilingual Coreference Resolution in Multiparty Dialogue

Boyuan Zheng[1] **Patrick Xia**[2*] **Mahsa Yarmohammadi**[1] **Benjamin Van Durme**[1]
[1]Johns Hopkins University, USA
[2]Microsoft Semantic Machines, USA
{bzheng12, mahsa, vandurme}@jhu.edu

Abstract

Existing multiparty dialogue datasets for entity coreference resolution are nascent, and many challenges are still unaddressed. We create a large-scale dataset, *Multilingual Multiparty Coref* (MMC), for this task based on TV transcripts. Due to the availability of gold-quality subtitles in multiple languages, we propose *reusing* the annotations to create silver coreference resolution data in other languages (Chinese and Farsi) via annotation projection. On the gold (English) data, off-the-shelf models perform relatively poorly on MMC, suggesting that MMC has broader coverage of multiparty coreference than prior datasets. On the silver data, we find success both using it for data augmentation and training from scratch, which effectively simulates the zero-shot cross-lingual setting.

1 Introduction

Coreference resolution is a challenging aspect of understanding natural language dialogue (Khosla et al., 2021). Many dialogue datasets are between two participants, even though there are distinct challenges that arise in the *multiparty* setting with more than two speakers. Figure 1 shows how ‘‘you’’ could refer to any subset of the listeners of an utterance. While there are some datasets on multiparty conversations from TV transcripts (Choi and Chen, 2018), they only annotate *people*, resulting in incomplete annotations across entity types. Moreover, these datasets are only limited to English, and works in dialogue coreference resolution in other languages are rare (Muzerelle et al., 2014).

We introduce a new (entity) coreference resolution dataset focused on multiparty dialogue that supports experiments in multiple languages. We first annotate for coreference on the transcripts from two popular TV shows, in English. We then leverage existing gold subtitle translations (Creutz, 2018) in Chinese and Farsi to project our annotations, resulting in a multilingual corpus (Figure 1).

Our experiments demonstrate that coreference resolution models trained on existing datasets are not robust to a shift to this domain. Further, we demonstrate that training on our projected annotations to non-English languages leads to improvements in non-English evaluation. Finally, we lay out an evaluation for zero-shot cross-lingual coreference resolution, requiring models to test on other languages with no in-language examples. We release over 1,200 scenes from TV shows with all annotations and related metadata in English, Chinese, and Farsi, which we call MMC: Multilingual Multiparty Coreference.

2 Motivation and Related Work

Most work on coreference resolution primarily studies documents with a single author or speaker. OntoNotes (Weischedel et al., 2013) is a widely used dataset that mostly consists of single-author documents, like newswire, while other datasets like PreCo (Chen et al., 2018), LitBank (Bamman et al., 2020), and WikiCoref (Ghaddar and Langlais, 2016) also consist of documents like books. Many recent modeling contributions also focus primarily on this setting and these datasets (Lee et al., 2017, 2018; Xu and Choi, 2020; Bohnet et al., 2022) and some offload it to pretrained language models (Wu et al., 2020; Toshniwal et al., 2021) or ignore the speaker identity entirely (Xia et al., 2020) in an attempt to unify dialogue with non-dialogue domains.

The dialogue domain is less studied because we lack a suitable dataset, even though these exist for other NLP tasks (Section 2.1). In addition to filling this gap, we also present a scalable solution for dataset creation in other languages, following related work in data projection methods

*Work done at JHU/HLTCOE.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 922–940, 2023. https://doi.org/10.1162/tacl_a_00581
Action Editor: Mihai Surdeanu. Submission batch: 9/2022; Revision batch: 12/2022; Published 8/2023.

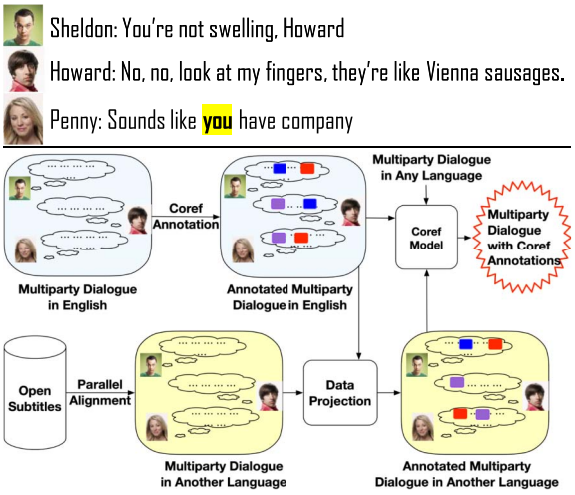

Figure 1: (*Top*) An example of ambiguous coreference due to multi-person context: Penny's *''you''* could refer to Sheldon, Howard, or both. (*Bottom*) Annotations can be projected to new languages, enabling model training beyond English.

(Section 2.2). The limitations of existing work motivate the creation of our dataset.

2.1 Multiparty Conversations

One of the focuses of this work is *multiparty* coreference resolution, which concerns coreference in conversational text with multiple participants. In particular, we are interested in conversations with more than two participants since this brings additional challenges not present in typical dialogue datasets. For example, in two-way conversations, ''you'' is typically deducible as the listener of an utterance. However, as shown in Figure 1, ''you'' in multiparty conversations with more participants could refer to any of the participants present in the conversation. Additional challenges include using a third person pronoun to refer to one of the interlocutors and plural mentions (''we'', ''you all'') that refer to a subset of the participants in the conversation (Zhou and Choi, 2018).

Multiparty conversations are ubiquitous, especially in the form of spontaneous speech and dialogue. They have been used to study tasks like discourse parsing and summarization (Afantenos et al., 2015; Liu and Chen, 2021; Manuvinakurike et al., 2021, i.a.) and coreference resolution (Walker and Reithinger, 1997; Jovanovic et al., 2005; Frampton et al., 2009; Choi and Chen, 2018). Despite the breadth of domains and formality across all datasets, each multiparty dataset itself is narrowly focused, like meetings (McCowan et al., 2005; Hsueh et al., 2006), board game play (Asher et al., 2016), fantasy storytelling (Rameshkumar and Bailey, 2020), technical or persuasive online forums (Li et al., 2020a; Wang et al., 2019), and sitcom transcripts (Choi and Chen, 2018; Sang et al., 2022).

Dialogue Coreference Resolution Coreference resolution in dialogue has recently reemerged as an area of research, with multiple datasets created and annotated for coreference resolution (Li et al., 2016; Khosla et al., 2021; more examples in Table 1) and the development of dialogue-specific models (Xu and Choi, 2021; Kobayashi et al., 2021; Kim et al., 2021). The datasets can be broadly categorized into transcripts of spoken conversations (e.g., interviews), meeting notes, online discussions, and one-on-one goal-driven genres. Table 1 shows that none of the datasets sufficiently covers spontaneous multiparty conversations. For the datasets that are multiparty, they are either incompletely annotated ($\text{Friends}_{\text{CI}}$ only annotates mentions referring to people), task-oriented (AMI), or discussion forums (ON-web, BOLT-DF). As a result, there are drawbacks to each of these datasets, like an expectation of formality (without the types of language found in spontaneous dialogue) or missing clarity on the listener or reader identities (e.g., missing usage of second person pronouns). None of these datasets aim for exhaustive annotation on multiparty dialogue in spontaneous social interactions.

$\text{Friends}_{\text{CI}}$ (Choi and Chen, 2018) is the closest dataset to the goals of this work.[1] Different from our goals, $\text{Friends}_{\text{CI}}$ is focused on *character* linking instead of general entity coreference. While pronouns like ''you'' are annotated, other entities, like objects or locations, are not. However, if we want to use coreference resolution models in downstream systems for information extraction (Li et al., 2020b) or dialogue understanding (Rolih, 2018; Liu et al., 2021), we need a dataset that aligns more closely with multiparty spontaneous conversations. We contribute a large-scale and more exhaustively annotated dataset for multiparty coreference resolution.

[1]OntoNotes is also close, but each conversational genre has its own drawbacks.

Dataset	Multiparty (> 2)	Exhaustive Entities	Spontaneous	Clear Interlocutors	Multi-lingual
TRAINS93 (Byron and Allen, 1998)		✓		✓	
$\text{Friends}_{\text{CI}}$ (Chen and Choi, 2016)	✓		✓	✓	
ON (tc) (Weischedel et al., 2013)			✓	✓	✓
ON (bc) (Weischedel et al., 2013)	✓			✓	✓
ON (wb) (Weischedel et al., 2013)	✓		✓		✓
BOLT (DF) (Li et al., 2016)	✓		✓		
BOLT (SMS, CTS) (Li et al., 2016)			✓	✓	
AMI^{C} (McCowan et al., 2005)	✓	✓		✓	
$\text{Persuasion}^{\text{C}}$ (Wang et al., 2019)		✓		✓	
$\text{Switchboard}^{\text{C}}$ (Stolcke et al., 2000)		✓	✓	✓	
LIGHT^{C} (Urbanek et al., 2019)		✓		✓	
MMC (Our work)	✓	✓	✓	✓	✓

Table 1: Examples of dialogue coreference datasets. Nothing to our knowledge satisfies our desire for modeling spontaneous multiparty conversations. Additionally, parallel data is available for MMC, which enables exploration in non-English languages. Superscript $^{\text{C}}$ indicates that they were additionally annotated by Khosla et al. (2021). OntoNotes (ON) is divided by genre.

2.2 Multilinguality

Coreference Resolution Coreference resolution models are typically developed for a single language, and although there is some prior work on cross-lingual and multilingual models (Xia and Van Durme, 2021), these methods still require some data in the desired language for best performance. While there are coreference resolution datasets in many languages (Weischedel et al., 2013; Recasens et al., 2010), they are often limited and expensive to annotate from scratch for each new language. We take a step towards a more general solution for building coreference resolution models from scratch in (almost) any language. By collecting and annotating data that already exists in a highly parallel corpus, we suggest a different approach to expensive in-language annotation: data projection.

Data Projection Using annotations in English to create data in a target language has been useful for tasks such as semantic role labeling (Akbik et al., 2015; Aminian et al., 2019), information extraction (Riloff et al., 2002), POS tagging (Yarowsky and Ngai, 2001), and dependency parsing (Ozaki et al., 2021). Previous work finds improvements when training on a mixture of gold source language data and projected silver target language data in cross-lingual tasks such as semantic role labeling (Fei et al., 2020; Daza and Frank, 2020) and information extraction (Yarmohammadi et al., 2021). The intuition of using both gold and projected silver data is to allow the model to see high-quality gold data as well as data with target language statistics. In this work, we extend projection to coreference resolution both for creating a model without in-language data and for augmenting existing annotations.

3 Multilingual Multiparty Dialogue Coreference Dataset

In this section, we present our multilingual multiparty coreference (MMC) dataset,[2] including the construction process of data alignment and filtering, annotation, and projection.[3] Core to our contribution is the choice of a multiparty dataset that *already has gold translations* and prioritizing multilinguality throughout the data collection process.

3.1 Parallel Dialogue Corpus

We construct a parallel corpus of multiparty dialogue by aligning the English transcripts from TV shows and parallel subtitles from the OpenSubtitles corpus (Tiedemann, 2012; Lison and Tiedemann, 2016), a sentence-aligned parallel corpus widely used in machine translation.[4]

TV sitcoms are an ideal target for meeting our criteria for a spontaneous multiparty genre, as

[2]This dataset is released under the Apache License.

[3]Dataset, software, and annotation infrastructure used in this work are available at: `https://github.com/boyuanzheng010/mmc`.

[4]`https://www.opensubtitles.org/`.

TV Show		Ep.	Scenes	Utter.	Speakers
TBBT	2-way	184	2,212	40,883	432
	3-way	88	1,086	19,773	249
	final	88	979	18,350	191
Friends	2-way	28	368	7,113	146
	3-way	21	270	5,226	114
	final	21	243	4,614	104

Table 2: Source data statistics (episodes, scenes, utterances, unique speakers) before and after filtering for three-way alignable episodes. *2-way* contains the union of two-way alignable episodes, while *3-way* contains the intersection, i.e., three-way alignable episodes. After a *final* filtering step, there are 1,222 scenes in total.

they contain rich multiparty dialogues, multiple references to interlocutors, and spontaneous utterances.[5] We select *Friends* and *The Big Bang Theory* (TBBT) because there is prior work in preprocessing and speaker identification for the transcripts of these shows (Roy et al., 2014; Choi and Chen, 2018; Sang et al., 2022).

We align the available data with that from two languages distant from English: Chinese and Farsi (Section 3.3). Due to missing episodes and alignments for some languages, the final three-way aligned corpus is an intersection of what is available in all three languages, and empty or clearly misaligned scenes are removed (Table 2).

3.2 English Coreference Annotation

We automatically create an initial set of proposed markable mentions, aiming for high recall. Like prior work (Pradhan et al., 2012; Poesio et al., 2018; Bamman et al., 2020), for consistent annotation, these markables are then considered for coreference linking. We mainly follow the annotation process of OntoNotes 5.0 (Weischedel et al., 2013).[6] However, we make some simplifications that are easier to understand for crowdworkers, roughly following those made by Chen et al. (2018). Unlike OntoNotes, we do not consider verbs and verb phrases as markable. Entities mentioned once (*singletons*) are annotated. Also, non-proper modifiers can be coreferred with generic mentions, and subspans can be coreferred with the whole span.

[5]While transcripts are pre-written, they are written to mimic spontaneous speech.

[6]Details are in the annotation interface instructions.

Markable Mention Proposal We ensemble predictions from the Berkeley parser with T5-Large (Kitaev and Klein, 2018; Raffel et al., 2020) and RoBERTa-based (Liu et al., 2019) spaCy[7] to detect nouns, noun phrases, and pronouns. These constitute our proposed markable mention spans.

Interface Our annotation interface (Figure 2) is derived from that of Yuan et al. (2022). The interface simplifies coreference annotation to selecting *any* antecedent for each query span (proposed markable) found by the parser. For consistency, the interface encourages users to select proposed markables, although they can also add a new antecedent mention if it is not among those proposed by the parser. They can also label a markable span as not a mention. Coreference clusters are formed by taking the transitive closure after annotation.

We make several modifications to the interface to annotate coreference more completely and in the dialogue setting. These include permitting the selection of speakers, mentions of arbitrary size for plural mentions, and an indication of uncertainty (e.g., without further context, the example in Figure 1 requires audiovisual cues). While the annotation of plural mentions and uncertainty labels are not used in this work, we hope they enable future studies.

Pilot Annotation We sampled three scenes of differing lengths from the training set for a qualification study. For these scenes, we adjudicated annotations from four experts as the gold labels. Then, we invited annotators from Amazon Mechanical Turk to participate and receive feedback on their errors. Nine high-scoring annotators on the pilot[8] (>80 MUC score) were selected for annotating the full training set. We paid US$ 7 for each pilot study, which could be completed in 25-35 minutes, including reading the instructions.

Full Annotation For the training set, the scenes were batched into roughly 250 proposed markables each. We paid $4 per batch (expected $12/hour) for each of the nine high-scoring annotators. Each of the scenes was annotated once, although we inspected these annotations to ensure they were nontrivial (i.e., not all-blank or all-singletons).

[7]We use the en_core_web_md-3.2.0 model.

[8]The lowest Cohen's Kappa score is 0.6549.

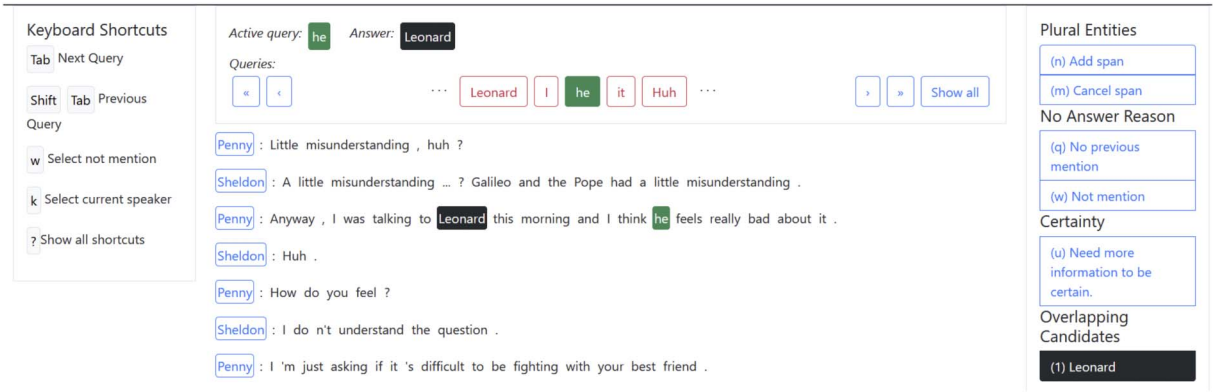

Figure 2: This figure illustrates the annotation interface. Given a set of proposed markables (‘‘queries’’), users highlight the best antecedent *or speaker* that the markable refers to or select ‘‘no previous mention’’ or ‘‘not a mention.’’ Plural entities and uncertainty due to missing context can also be annotated.

For the dev and test splits, three professional data specialists,[9] in consultation with the authors, annotated the documents with two-way redundancy. After reviewing common error and disagreement types with the authors, one of the specialists performed adjudication of the disagreements (described in Appendix B). Following several prior works (Weischedel et al., 2013; Chen et al., 2018; Toshniwal et al., 2020), we adopt MUC score as an approximate measure of agreement between annotators. The average MUC score of each annotator to the adjudicated clusters is 86.1. This agreement score is comparable to reported scores in widely used datasets: OntoNotes (89.60), PreCo (85.30). The inter-annotator MUC agreement score on this combined split is 80.3 and the inter-annotator CoNLL-F1 score is 81.55. The Cohen’s Kappa score is 0.7911, which is interpreted as ‘‘substantial agreement.’’ Note that the high agreement can be partially attributed to the agreement over non-mentions and starts of coreference chains.

In our data, 7.2%, 8.8%, and 10.0% of the clusters in training, dev, and test splits contain plural mentions. Meanwhile, 0.4%, 1.4%, and 1.6% of the mentions are marked as ‘‘uncertain.’’ The specialists working on the dev and test sets were more likely to mark an annotation as uncertain than crowdworkers.

3.3 Silver Data via Annotation Projection

Data projection transfers span-level annotations in a source language to a target language via word-to-word alignments in a fast, fully-automatic way. The projected (*silver*) target data can be used directly or combined with gold data as data augmentation.

Alignment We need to align English (source side) mention *spans* to Chinese or Farsi (target side) text spans. Our cleaned dataset contains utterance-level aligned English to Chinese and Farsi text. Using automatic tools, we obtain more fine-grained (word-level) alignments, and project source spans to target spans according to these alignments. For multi-token spans, the target is a contiguous span containing all aligned tokens from the source span.

We use awesome-align (Dou and Neubig, 2021), a contextualized embedding-based word aligner[10] that extracts word alignments based on token embedding similarities. We fine-tune the underlying XLM-R encoder on around two million parallel sentences from the OSCAR corpus (Abadji et al., 2022). We further fine-tune on Farsi-English gold alignments by Tavakoli and Faili (2014) and the GALE Chinese-English gold alignments (Li et al., 2015). See Appendix C for dataset statistics and fine-tuning hyperparameters. By projecting the annotated English mentions to the target side, the entity clusters associated with each mention are also implicitly projected.

Some coreference annotations are not transferred to the target language side either due to empty subtitles in our cleaned data or erroneous automatic word alignment and projection of the

[9]These are in-house data annotators with linguistics background who are trained by and correspond with the authors during the annotation process.

[10]https://github.com/neulab/awesome-align.

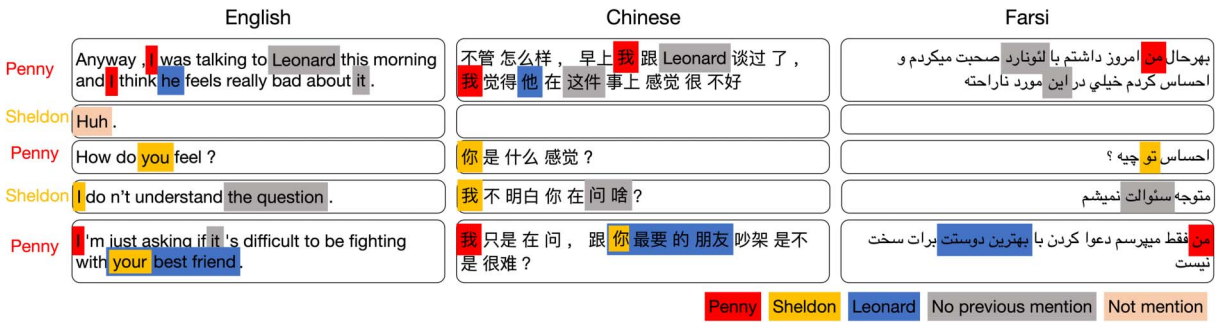

Figure 3: Example utterances with gold English and projected Chinese and Farsi coreference annotations.

Lang. (split)		Scenes	Utter.	Ment.	Clusters
En	train	955	18,477	60,997	25,349
	dev	134	2,578	10,079	4,804
	test	133	1,909	7,958	3,808
Zh	train	948	14,467	42,234	20,251
	dev	134	2,146	7,922	4,193
	$\text{test}_{\text{silver}}$	133	1,611	5,977	3,106
	$\text{test}_{\text{correct}}$	133	1,611	5,642	2,934
Fa	train	948	14,357	36,415	20,063
	dev	120	1,983	5,887	3,566
	$\text{test}_{\text{silver}}$	133	1,612	4,894	3,021
	$\text{test}_{\text{correct}}$	133	1,612	6,053	3,169

Table 3: MMC statistics. English (En) is manually annotated while Chinese (Zh) and Farsi (Fa) are projected. Since all episodes are three-way parallel, the splits for each language contain the same scenes (some empty scenes are omitted).

source text span. We refer to such cases as *null projections*.

Figure 3 shows parallel utterances with their gold English and projected Chinese and Farsi annotations. Some short English utterances do not have counterparts, such as the second utterance (‘‘Huh’’). Chinese and Farsi annotations are also a subset of English annotations due to null projections. For example, the English mention ‘‘it’’ in the last utterance is missing in the target transcripts, so this span’s annotation is missing in the projected data.

While there are the same number of episodes in the English and projected data, the number of scenes, mentions, and clusters in the projected data are smaller due to missing scenes or null projections. We see around 30% (Zh) and 40% (Fa) drop in aligned mentions (Table 3).

Alignment Correction We conducted alignment annotation for both English-Chinese and English-Farsi utterances to collect alignment corrections for the Chinese and Farsi test set with four Chinese speakers and three Farsi speakers. For each language pair, we presented the user with the utterance in each language and one of the English spans highlighted. On the target language side, the prediction by the projection model is displayed. The user makes corrections to the automatic alignments if necessary. This is conducted via the TASA interface[11] (Stengel-Eskin et al., 2019). These corrected annotations serve as the test set for both Chinese and Farsi; 1,904 (24.81%) projections are corrected in Chinese and 2,485 (32.26%) projections in Farsi. There are three types of corrections: addition, deletion, and modification, shown in Table 4. For *addition*, a mention boundary is added for a null projection. For *deletion*, the predicted projection is discarded. *Modification* is where the predicted mention boundaries are modified.

Language	Addition	Deletion	Modification
Chinese	609 (340)	441	916 (61)
Farsi	1,504 (739)	187	794 (84)

Table 4: Corrections in Chinese and Farsi test sets. The number in parentheses is the number of dropped pronouns that are recovered.

Chinese and Farsi are *pro-drop* languages. Most of the *addition* operations are related to pronouns, where the target is corrected from an empty string to the location of the trace of the pronoun (in Chinese) or the implied pronoun affix (in Farsi). In the *modification* operation, a small number of target mentions are also corrected to an empty string. This resulted in 401 and 823 additional pronoun mentions in Chinese and Farsi, respectively.

[11] https://github.com/hltcoe/tasa.

Dataset Statistics MMC contains about 101 hours of episodes, resulting in 323,627 English words, 226,045 Chinese words, and 258,244 Farsi words. Table 3 shows the final statistics of our three-way aligned, multiparty coreference corpus. This dataset is used for the remainder of the paper. To summarize, English dev and test data are two-way annotated followed by adjudication; English train is one-way annotated; and Chinese and Farsi are automatically derived via projection, but both Chinese and Farsi test alignments are corrected.

Dataset	# Speakers	Lang.	Train	Dev.	Test
ON^{En}	all	En	2,802	343	348
	≤ 1	En	2,321	260	263
	2	En	255	43	47
	> 2	En	226	40	38
$Friends_{CI}$	all	En	1,041	130	130
ON^{Zh}	all	Zh	1,810	252	218
	≤ 1	Zh	1,343	165	159
	2	Zh	173	22	18
	> 2	Zh	294	65	41

Table 5: Statistics for additional datasets used in this work. OntoNotes has a Chinese split; we are not aware of other Farsi coreference datasets.

4 Methods

4.1 Model

For all experiments, we use the higher-order inference (HOI) coreference resolution model (Xu and Choi, 2020), modified slightly to predict singleton clusters (Xu and Choi, 2021). Given a document, HOI encodes texts with an encoder and enumerates all possible spans to detect mentions. These spans are scored by a mention detector, which prunes the spans to a small list of candidate mentions. The candidate mentions are scored pairwise, corresponding to the likelihood of being coreferring, and the resulting scores are used in clustering. While mentions can be linked to their top-scoring antecedent, higher-order inference goes further and ensures high agreement between all mentions in a cluster by making additional passes. Singletons can be predicted when a high-scoring (via the mention detector) mention only has low-scoring (via the pairwise scorer) candidate antecedents. For English-only experiments, SpanBERT-large (Joshi et al., 2020) is used as the encoder while for the other experiments XLM-R-base (Conneau et al., 2020) is used. More hyperparameter details are in Appendix D.[12]

4.2 Noise-tolerant Mention Loss

The loss function used by Xu and Choi (2021) consists of a cluster loss, L_c,[13] typically used for coreference resolution (Lee et al., 2017; Joshi et al., 2020; Xu and Choi, 2020) and a binary cross-entropy mention detection loss, L_m, used to better predict singleton losses.

Compared to the two-way annotated and merged dev/test set, the one-way annotated train set is more likely to be subject to annotator biases, leading to noise in the train set. These inconsistencies are further exacerbated when projected to silver data, leading to a low recall of mentions in training, as evidenced by the number of ‘‘additions’’ in Table 4.

To address this noise, we propose a modification of L_m to downweight negative labels. Following the notation from Xu and Choi (2021), let Ψ_i^+ be the set of gold candidate mentions and Ψ_i^- be the remainder of the candidate spans. Applying a hyperparameter $\tau \in [0, 1]$, we can rewrite binary cross-entropy loss, L_m^τ, as

$$\sum_{x_i \in \Psi^+} \log(P(x_i)) + \tau \sum_{x_i \in \Psi^-} \log\left(1 - P(x_i)\right)$$

where x_i is a candidate span and $P(x_i)$ is the output of the mention scorer. Following Xu and Choi (2021), the mention loss is also weighted in the final loss, $L = L_c + \alpha_m L_m^\tau$.

4.3 Data

We evaluate the performance of models across three datasets: MMC, OntoNotes (Pradhan et al., 2012), and $Friends_{CI}$ (Choi and Chen, 2018). OntoNotes is a collection of documents spanning multiple domains, some of which include multiparty conversations or dialogues, like weblogs, telephone conversations, and broadcast news (interviews). Furthermore, OntoNotes is available in English and Chinese. $Friends_{CI}$ is a collection of annotations on TV transcripts from *Friends*, including entity linking where character entities are

[12] Code to run experiments: `https://github.com/boyuanzheng010/mmc`.

[13] L_c is based on the pairwise scorer. It is the (marginal) log-likelihood of all correct antecedent mentions for a single mention, which is provided by the gold clusters. We do not modify it in this work.

Train		Test							
		ON	$\text{Friends}_{\text{CI}}$	MMC	$\text{ON}_{\leq 1}$	ON_2	$\text{ON}_{>2}$	$\text{MMC}_{\text{Friends}}$	MMC_{TBBT}
En	Xu and Choi (2020)	**79.96**	54.41	50.97	**81.56**	**78.16**	**75.67**	48.49	51.60
	ON^{En}	78.80	54.60	53.41	80.41	76.90	74.59	49.30	50.14
	$\text{Friends}_{\text{CI}}$	30.91	**71.19**	46.64	24.52	43.32	36.67	49.99	45.76
	MMC	46.53	48.10	70.71	42.97	54.45	48.62	67.02	71.61
	MMC (XLM-R)	35.08	44.52	69.63	32.76	41.17	35.45	67.08	70.25
	ON^{En}+MMC	73.62	48.42	72.01	74.94	72.58	69.66	69.26	72.67
	$\text{ON}^{\text{En}}\rightarrow$MMC	65.72	47.40	**75.87**	67.00	65.06	61.29	**73.01**	**76.58**
Zh	ON^{Zh}	**65.52**	–	36.56	**66.59**	65.62	60.89	36.58	36.56
	MMC	23.71	–	47.89	14.72	47.12	38.92	**43.33**	49.01
	ON^{Zh}+MMC	64.19	–	47.37	63.79	**68.87**	**63.24**	41.79	48.70
	$\text{ON}^{\text{Zh}}\rightarrow$MMC	37.52	–	**48.65**	33.32	51.51	44.74	42.56	**50.01**

Table 6: F1(%) scores of models trained on a combination of different datasets for English and Chinese. All English models except MMC (XLM-R) use SpanBERT-Large as the encoder, while MMC (XLM-R) and Chinese models use XLM-R-base as the encoder.

linked. As the focus of this work is on multiparty conversations, we further separate OntoNotes into documents with 0 or 1 ($\text{ON}_{\leq 1}$), 2 (ON_2), or more than two ($\text{ON}_{>2}$) speakers/authors for evaluation. We didn't include split antecedents and drop-pronouns in the experiment, since the baseline model doesn't support predicting them. The statistics of datasets used in our experiments are in Table 5.

4.4 Evaluation

We use the average of MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF_{ϕ_4} (Luo, 2005), which is also used for OntoNotes. Furthermore, each model is trained three times and the average test score (CoNLL_{F1}) is reported.

5 Experiments and Results

First, we highlight the differences of MMC in contrast to $\text{Friends}_{\text{CI}}$ and OntoNotes. To do so, we train an off-the-shelf model on the three datasets. Additionally, we establish monolingual baselines for all three languages. Finally, we explore the cross- and multi-lingual settings to validate the recipe of using data projection for coreference resolution.

5.1 Monolingual Models

Table 6 shows the performance of several monolingual models. They highlight that models trained on other datasets (ON, $\text{Friends}_{\text{CI}}$) perform substantially worse than models trained in-domain (on MMC). Additionally, we find that both combining datasets and using continued training from OntoNotes (Xia and Van Durme, 2021) can be effective for further improving model performance: For English, this leads to gains of 2.7 F1 points (combining) and 5.2 F1 points (continued training), and continued training is also effective in Chinese.

Notably, combining the Chinese datasets yields the best scores on dialogues (ON_2, $\text{ON}_{>2}$) in OntoNotes. This highlights the utility of the *silver* MMC data as a resource for augmenting preexisting in-domain data. Combining data is less helpful for English than Chinese possibly because there is more training data in ON^{En} than ON^{Zh}, making the Chinese data augmentation more useful. The baselines for ON^{Zh} may also be less optimized by prior work than models for ON^{En}.

5.2 Cross-lingual and Multilingual Models

Next, we demonstrate the ability of the silver data in Chinese and Farsi to contribute towards creating a model with no in-language coreference resolution data. While Chinese and Farsi are the two languages we choose to study in this work, parallel subtitles for the TV Shows in MMC are available in at least 60 other languages and can be used similarly, given a projection model.[14]

[14]We assume that a language contains parallel subtitles if we find alignable episodes within the bitext between English and that language. This count is approximate and is not exhaustive.

Train	Test		
	MMC-En	MMC-Fa	MMC-Zh
Head Lemma	17.30	–	12.68
MMC-En	**69.63**	22.91	37.18
MMC-Fa	55.99	**35.00**	36.85
MMC-Zh	45.91	17.09	**47.89**
MMC-En-Zh-Fa	69.57	33.49	45.54

Table 7: Performance of models trained on datasets of different languages (English, Farsi, and Chinese) and the combination of all three of them. All four models use XLM-R-base as the encoder.

L_m	Language		
	MMC-En	MMC-Fa	MMC-Zh
$\tau = 1$	69.63	35.00	47.76
$\tau = 0$	59.49	31.32	37.30
weighted	**72.58**	**37.05**	**49.56**

Table 8: Test set performance of models trained with different τ. $\tau = 1$ is the regular binary cross-entropy mention loss reported earlier in the paper. τ is chosen according to a grid search (Section 4.2).

Simple Baseline We adopt a simple head lemma match baseline to determine a lower bound for each language if we did not have any training data. We first find the NP constituencies as candidate mentions derived from off-the-shelf constituency parsers. We adopt the Berkeley parser with T5-Large (Kitaev and Klein, 2018; Raffel et al., 2020) for English and multilingual self-attention parser (Kitaev et al., 2019) with Chinese ELECTRA-180G-large (Cui et al., 2020) for Chinese. For Farsi, we adopted the constituency parser in DadmaTools (Etezadi et al., 2022).[15] However, we were not confident in the Farsi parser quality (under 5 CoNLL_{F1} when evaluated on Farsi MMC), and could not find another widely used constituency parser for Farsi, so we omit Farsi in our results. To predict the clusters, we extract and lemmatize the head word for each mention. We link any two mentions that have the same head word lemma.

Cross-lingual Transfer We evaluate the monolingual XLM-R models for English, Chinese, and Farsi on each of the languages, i.e., ''test'' for English, ''$\text{test}_{\text{correct}}$'' for Chinese, and ''$\text{test}_{\text{silver}}$'' for Farsi. This effectively evaluates the zero-shot ability for the other two languages.

Table 7 shows that models trained on English data or silver projected data in Farsi and Chinese can achieve reasonable performance on the test set of its own language. Models trained on projected silver data in Farsi and Chinese achieve the best performance among their own test set compared with zero-shot performance of models trained in another language. Consequently, this implies that a recipe of projecting a coreference resolution dataset to another language and using that data to train from scratch outperforms naive zero-shot transfer via multilingual encoders.

Multilingual Models We combine the training data of three languages and train multilingual models. Table 7 shows that these multilingual models achieve slightly to moderately worse performance on each test set compared to their monolingual counterparts. This contrasts with prior work (Fei et al., 2020; Daza and Frank, 2020; Yarmohammadi et al., 2021) that finds benefits to using silver data. The poorer performance of the multilingual model could be due to using the same set of hyperparameters for all three languages. While it does not surpass the monolingual models, it enjoys the benefits of being more parameter efficient.

5.3 Noise-tolerant Loss Results

Table 8 shows model performance using our modified loss. We find some benefits to down-weighting negatively labeled spans, obtaining 1-3 points improvement compared to the original loss across all three languages.[16] Thus, MMC could also enable exploration into additional *modeling* questions around the use of projected and noisy annotations.

6 Analysis

We analyze our modeling results in relation to our original motivation. First, we explore differences between datasets (Section 6.1), the number

[15] We use spaCy for English (en_core_web_sm-3.4.0) and Chinese (zh_core_web_sm-3.4.0), and Hazm toolkit (`https://www.roshan-ai.ir/hazm/`) for Farsi.

[16] With the weighting, XLM-R-base can outperform SpanBERT large in English.

of speakers (Section 6.2), and overfitting (Section 6.3). For the data construction, we analyze the alignment corrections process (Section 6.4) and compare recipes for annotation projection (Section 6.5).

6.1 Comparison of Datasets

Since $\text{Friends}_{\text{CI}}$ is also based on TV Shows (*Friends*) and its dataset overlaps with MMC, we would expect a model trained on $\text{Friends}_{\text{CI}}$ to perform well on MMC. Instead, we find that its performance is over 23 F1 points worse. The main difference between $\text{Friends}_{\text{CI}}$ and MMC is that $\text{Friends}_{\text{CI}}$ only annotates characters instead of all possible mentions, and therefore there are fewer mentions per document in $\text{Friends}_{\text{CI}}$ than in MMC. A closer inspection of the precision and recall appears to validate this hypothesis, as the macro precision (across the three metrics) is 65.8% compared to a recall of 37.5%. This is also evident in the mention span precision and recall, where a model trained on $\text{Friends}_{\text{CI}}$ scores 91.5% precision but only 50.3% on recall. We see the same trend for OntoNotes: high precision and low recall both on the coreference metrics and on mention boundaries.

6.2 Number of Speakers

Table 6 also shows that in OntoNotes, models perform more poorly on documents with more speakers. However, this is not the case with both $\text{Friends}_{\text{CI}}$ and MMC, which perform best on two-person dialogues.[17] Nonetheless, the drop in performance from ON_2 to $\text{ON}_{>2}$ highlights the additional difficulty of multiparty dialogue documents (in OntoNotes). These trends are similar for both English and Chinese.

6.3 Overfitting to Specific Shows

As one of our goals is a dataset enabling a better understanding of multiparty conversations, a concern is that models may overfit to the limited (two) TV shows and the subset of characters (and names) in the training set. While the test set contains our target domain (multiparty conversation), it also shares characters and themes with the training set.

[17] Many documents in $\text{ON}_{\leq 1}$ are written documents, not dialogues, and therefore out-of-domain for $\text{Friends}_{\text{CI}}$ and MMC.

Train	**Test**		
	MMC	MMC-Name	Δ
ON	53.41	52.92	−0.49
$\text{Friends}_{\text{CI}}$	46.64	34.53	−12.11
MMC	70.70	62.38	−8.32
ON+MMC	72.01	61.19	−10.82
ON→MMC	75.87	73.60	−2.27
MMC-Name	68.91	70.88	**+1.97**

Table 9: F1(%) of models evaluated on original MMC test set and a version with character names randomly replaced per scene (MMC-Name).

Names We test whether models are sensitive to speaker names, perhaps overfitting to the character names and their traits. We replace speaker names in the original MMC dataset with random names. First, we assume the self-identified genders of the speakers through their pronoun usage. Next, for each scene, we replace the name of a character with a randomly sampled name of the same gender.[18]

The results in Table 9 show that models do overfit to character names: For models trained on MMC, $\text{Friends}_{\text{CI}}$, and ON+MMC, performance on MMC test sets drops after replacing names, thereby showing that they are sensitive to names seen in training. On the other hand, both ON and ON→MMC show more robustness to changes in speaker name. This is likely because ON does not have a persistent set of characters for the entire training set. It is less clear why ON→MMC experiences only a small drop; robustness through continued training can be investigated further in future work.

We create a *training* set (MMC-Name) without a persistent set of characters or speakers by randomly replacing the character names. While MMC performance drops slightly compared to a model trained with the original data, it outperforms on the name-replaced test set. Since we have the {original, replaced} name mapping, we can convert predictions from MMC-Name to MMC, resulting in an F1 on MMC competitive with the baseline, after post-processing. These findings support the hypothesis that models that see names used in a ''generic fashion'' are more robust towards name changes (Shwartz et al., 2020).

[18] We use the top 100 names by frequency for each gender according to `https://namecensus.com/`, which is based on the 1990 US Census.

Train	Test		
	MMC	$MMC_{Friends}$	MMC_{TBBT}
MMC	70.70	67.02	71.61
$MMC_{Friends}$	61.95	61.03	62.28
MMC_{TBBT}	71.22	69.64	71.62

Table 10: F1(%) of models trained and tested on each of the TV shows, to measure potential overfitting to the training show.

TV Series To determine overfitting to a specific TV show, we split MMC (English) into the two components: $MMC_{Friends}$ and MMC_{TBBT}, shown in Table 10. In this analysis, we find that the variance due to random seed is high, which might explain why training with MMC_{TBBT} appears to be the best model. The results suggest both models find MMC_{TBBT} easier to predict. Furthermore, training with MMC_{TBBT} outperforms $MMC_{Friends}$ when evaluated on $MMC_{Friends}$, suggesting that the substantially larger size of the MMC_{TBBT} portion beats any in-domain advantages $MMC_{Friends}$ may have.

6.4 Alignment Correction

To identify the types of systematic errors made by automatic projection, we analyzed the corrected Chinese alignments. Table 11 shows the difference in model performance between the corrected and the silver test set. Performance drops a few F1 points on the corrected set, which is caused by the distribution shift from (uncorrected, silver) training data. Naturally, MMC-Zh suffers the largest drop because it is closest in the domain to $test_{silver}$. However, it is still one of the best performing models.

The performance drop of the ON-only trained model is only 0.85 points, possibly because this model is trained on the cleaner (gold) training labels. These observations suggest that while the alignment correction yields a cleaner test set, the automatic silver data is still a good substitute for model development when no gold data is available.

There is a similar pattern in Farsi. Most of the drop is in recall, since many new mentions are added via alignment correction. These new additions are mostly words that contain compound possessive pronouns or verbal inflectional suffixes that align to a source English word, which are not often captured by automatic word alignment methods. For example, the word ''نمیشم'', is a verb with the inflectional suffix ''ـم'' aligning to the source mention ''I''. Another example is ''دوستت'', composed of the noun ''دوست'' plus the possessive pronoun ''ـت'' aligning to the source span ''your'' in Figure 3.

	Train	Test		
		corrected	silver	Δ
Zh	ON	36.56	37.41	0.85
	ON+MMC-Zh	47.37	49.38	2.01
	ON→MMC-Zh	48.65	51.09	2.44
	MMC-Zh	47.89	51.87	3.98
	MMC-En	37.18	39.23	2.05
	MMC-Fa	37.01	39.54	2.53
	MMC-En-Zh-Fa	45.54	47.30	1.76
Fa	MMC-Fa	35.00	39.76	4.76
	MMC-En	22.91	25.07	2.16
	MMC-Zh	17.09	20.06	2.97
	MMC-En-Zh-Fa	33.49	38.30	4.81

Table 11: F1 of models on Chinese and Farsi test set before and after correction.

6.5 Annotation from Scratch

Instead of relying on noisy (but free) projections of parallel English data, one could directly annotate coreference in the target language with native speakers. To investigate the quality of $test_{silver}$ and $test_{correct}$, we perform an analysis study on three randomly sampled scenes from the Chinese test set and ask an annotator to complete the full coreference annotation task. We also obtain oracle word alignments to explore the effect of alignment errors in our data projection framework.

We find MUC score (agreement) rates of 71.84, 78.23, and 87.25 using $test_{silver}$, $test_{correct}$, and oracle projections, respectively. This suggests that the corrected test set has a comparable agreement rate to that of the gold data, while the gold projections are also within the range of inter-annotator agreement. As automatic alignment methods improve, our recipe for creating multilingual coreference data will also benefit. Nonetheless, one of the limitations of MMC is that quality of the Chinese and Farsi test sets could still be higher.

Advantages Despite lower quality, the data projection method still has several advantages over from-scratch annotation as it is faster and there is less demand for an in-language expert.

First, annotation from scratch requires a syntactic parser to find constituencies for mention linking (Section 3.2). The zero-shot transfer setting usually involves lower-resource languages, where parsers, if they exist at all, may not perform well. Thus, projection may be the only solution in these cases.

Second, the annotation quality depends on the guidelines. Linguistic experts in the target language will need to design annotation guidelines and experts are not always available. However, this step can be skipped with projection (since we are releasing MMC, which has parallel text in numerous languages). Not only the projection task itself is significantly simpler to explain, it is easier to understand and can be faster than annotating from scratch. In our setting, around 70% of the predicted alignments were marked as correct. One could design heuristics to only present the difficult mention pairs, which would further reduce annotation cost.

7 Conclusion

Motivated by a desire to better understand spontaneous multilingual conversations, we developed a collection of coreference annotations on the transcripts and subtitles of two popular TV sitcoms. To reduce the cost of annotating from scratch for each language, we selected our English data such that there were already existing gold human translations available in the form of subtitles, in order to automatically project our annotations from English. After manually correcting these projections, we observe a few point differences in reported values across various multilingual models.

There exist dozens of additional languages that our annotations may be projected to in the future. If automatic projection leads to only a few point variance in the estimated performance of a model, we believe this framework is sufficient for driving significant new work in coreference across many non-English languages in the future.

8 Limitations

There are several limitations in the dataset inherent to the difficulty of the task, crowdsourcing, and the use of models for candidate proposals. The inter-annotator agreement scores are not perfect. One contributing factor is that we do not post-process or provide explicit instructions for pleonastic pronouns, so annotators used their own judgment. These account for 3.15% of the mentions in the pilot annotation. There is also a distribution difference between the (noisier) train and dev/test set caused by different annotator sources, how they were paid, and whether the annotations were adjudicated. Additionally, annotation was performed without access to ground truth video, which could impede annotation or encourage guessing when situatedness may be required. Since annotation in MMC is aided by other models (parser and aligner), system errors may not necessarily be caught during annotation.

Acknowledgments

We thank Michelle Fashandi and other linguists at HLTCOE, along with Chenyu Zhang and Kate Sanders, for their annotation efforts. We thank Elias Stengel-Eskin, Kate Sanders, and Shabnam Behzad for helpful comments and feedback and Chandler May for help in building the annotation interface. The third author acknowledges support through a fellowship from JHU + Amazon Initiative for Interactive AI (AI2AI). This work was supported by DARPA AIDA (FA8750-18-2-0015) and IARPA BETTER (201919051600005). The views and conclusions contained in this work are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, or endorsements of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv e-prints*, page arXiv:2201.06642.

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1039

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2019. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 200–210, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-0417

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2022. Coreference resolution through a seq2seq transition-based system.

Donna K. Byron and James F. Allen. 1998. Resolving demonstrative anaphora in the trains93 corpus.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1016

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-3612

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1007

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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*. European Language Resources Association (ELRA).

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. *ArXiv*, abs/2004.13922. https://doi.org/10.18653/v1/2020.findings-emnlp.58

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 3904–3914, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.321

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.181

Romina Etezadi, Mohammad Karrabi, Najmeh Valizadeh Zare, Mohamad Bagher Sajadi, and Mohammad Taher Pilehvar. 2022. Dadmatools: Natural language processing toolkit for persian language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*. https://doi.org/10.18653/v1/2022.naacl-demo.13

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Matthew Frampton, R. Fernández, Patrick Ehlen, C. Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is ‘‘you’’? combining linguistic and gaze features to resolve second-person references in dialogue. In *EACL*. https://doi.org/10.3115/1609067.1609097

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Pei-Yun Hsueh, Johanna D. Moore, and Steve Renals. 2006. Automatic segmentation of multiparty dialogue. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273–280, Trento, Italy. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. https://doi.org/10.1162/tacl_a_00300

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2005. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40:5–23. https://doi.org/10.1007/s10579-006-9006-4

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongjin Kim, Damrin Kim, and Harksoo Kim. 2021. The pipeline model for resolution of anaphoric reference and resolution of entity reference. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 43–47, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1249

Hideo Kobayashi, Shengjie Li, and Vincent Ng. 2021. Neural anaphora resolution in dialogue. In *Proceedings of the CODI-CRAC*

2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020a. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020b. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.11

Xuansong Li, Stephen Grimes, Stephanie Strassel, Xiaoyi Ma, Nianwen Xue, Mitchell P. Marcus, and Ann Taylor. 2015. GALE chinese-english parallel aligned treebank – training. LDC2015T06.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. 2016. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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. *CoRR*, abs/1907.11692.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

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

Ramesh Manuvinakurike, Saurav Sahay, Wenda Chen, and Lama Nachman. 2021. Incremental temporal summarization in multi-party meetings. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 530–541, Singapore and Online. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike

Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2005. The ami meeting corpus.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: Description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hiroaki Ozaki, Gaku Morio, Terufumi Morishita, and Toshinori Miyoshi. 2021. Project-then-transfer: Effective two-stage cross-lingual transfer for semantic dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2586–2594, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.221

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

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, Jeju Island, Korea. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:(140):1–67.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.459

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/1621969.1621982

Ellen Riloff, Charles Schafer, and David Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *COLING 2002: The 19th International Conference on Computational Linguistics*. https://doi.org/10.3115/1072228.1072298

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems.

Anindya Roy, Camille Guinaudeau, Hervé Bredin, and Claude Barras. 2014. TVD: A reproducible and multiply aligned TV series dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26–31, 2014*, pages 418–425. European Language Resources Association (ELRA).

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character comprehension in stories as speaker guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.317

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. ''You are grounded!'': Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.556

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374. https://doi.org/10.1162/089120100561737

Leila Tavakoli and Heshaam Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach. *International Journal of Information and Communication Technology*, 6:63–76.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to ignore: Long document coreference with bounded memory neural networks. In *EMNLP*. https://doi.org/10.18653/v1/2020.emnlp-main.685

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6–8, 1995*. https://doi.org/10.3115/1072399.1072405

Marilyn A. Walker and Norbert Reithinger. 1997. Standards for dialogue coding in natural language processing.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1566

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0. *Linguistic Data Consortium, Philadelphia, PA*.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.622

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.425

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2021. Adapted end-to-end coreference resolution system for anaphoric identities in dialogues. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 55–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.149

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.519

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendixes

A Split Antecedent Statistics

MMC-En has a number of split antecedents; 1,156 antecedents across 2,745 spans in the training set; 255 across 717 spans in the dev set, and 178 across 444 spans in the test set.

B Merging Two-way Annotations

A third annotator adjudicates disagreement in the two-way annotations in the dev/test set. To decide whether a pair of annotations disagree, we first build common clusters between two annotations. After annotation, each query mention is annotated with two antecedents.

$$A = \{(q_1, a_1^1, a_1^2), (q_2, a_2^1, a_2^2), \ldots, (q_n, a_n^1, a_n^2)\}$$

q_i is the i^{th} query and n is the number of candidate queries. a_i^1 and a_i^2 are the antecedents linked to the i^{th} query. We build initially agreed clusters by taking the transitive closure of the subset of A where each triplet agrees exactly (i.e., for q_i, $a_i^1 = a_i^2$) between the two annotations. Note that the annotations, a_i, can be another query span, q_j, that is also annotated. This lets us connect the annotations and form clusters.

Next, we incrementally add query spans to these clusters if both annotators link them to the same cluster ($a_i^1 \neq a_i^2$ but a_i^1 and a_i^2 are in the same cluster anyway), continuing until no further pairs agree. At the end, if there exist q_i where $a_i^1 \neq a_i^2$, then each (q_i, a_i^1, a_i^2) is marked for adjudication. The adjudicator is prompted to select between a_i^1, a_i^2, or relabel q_i entirely. Their annotation is final.

C Word Alignment

Word alignments are extracted from the fine-tuned XLM-R-large model using Awesome-align. We first fine-tuned XLM-R on English-{Chinese, Farsi} parallel data that has been filtered using LASER semantic similarity scores (Schwenk and Douze, 2017; Thompson and Post, 2020). We reuse empirically chosen Awesome-align hyperparameters from prior work for a similar task (Yarmohammadi et al., 2021): softmax normalization with probability thresholding of 0.001, 4 gradient accumulation steps, 1 training epoch with a learning rate of $2 \cdot 10^{-5}$, alignment layer of 16, and masked language modeling (‘‘mlm’’), translation language modeling (‘‘tlm’’), self-training objective (‘‘so’’), and parallel sentence identification (‘‘psi’’) training objectives. We further fine-tuned the resulting model on the gold word alignments on 1500 En-Fa and 2800 En-Zh sentence pairs with the same hyperparameters, for 5 training epochs with a learning rate of 10^{-4} and only ‘‘so’’ as the training objective.

D Hyperparameters

We reuse most of the hyperparameters from Xu and Choi (2020): We enumerate spans up to a maximum span width of 30 and set the maximum speakers to 200, ‘‘top span ratio’’ to 0.4, and maximum top antecedents (beam size) to 50. For XLM-R models, we set the LM learning rate to 10^{-5} and task learning rate to $3 \cdot 10^{-4}$. For SpanBERT models, we use a LM learning rate of $2 \cdot 10^{-5}$ and task learning rate of $2 \cdot 10^{-4}$.

Following a grid search, we set the mention loss weights (α_m) for the each language and dataset: 5 for MMC-Zh and MMC-En, 6.5 for MMC-Fa, and 0 for OntoNotes. For τ we find $\tau_{\text{Fa}} = 0.55$, $\tau_{\text{Zh}} = 0.7$, and $\tau_{\text{En}} = 0.7$ performed best on dev.