Toward Implicit Reference in Dialog: A Survey of Methods and Data

Lindsey Vanderlyn **Talita Anthonio** **Daniel Ortega** **Michael Roth** **Ngoc Thang Vu**
University of Stuttgart
Institute for Natural Language Processing
{lindsey.vanderlyn,talita.anthonio,daniel.ortega,
michael.roth,ngoc-thang.vu}@ims.uni-stuttgart.de

Abstract

Communicating efficiently in natural language requires that we often leave information implicit, especially in spontaneous speech. This frequently results in phenomena of incompleteness, such as omitted references, that pose challenges for language processing. In this survey paper, we review the state of the art in research regarding the automatic processing of such *implicit references* in dialog scenarios, discuss weaknesses with respect to inconsistencies in task definitions and terminologies, and outline directions for future work. Among others, these include a unification of existing tasks and evaluation metrics, addressing data scarcity, and taking into account model and annotator uncertainties.

1 Introduction

In natural language conversations, speakers often leave out parts of the conversation which are understood by the other party through the shared context, as exemplified in Figure 1. This can either serve as a way to add variance to a conversation, to make the dialog more efficient by not repeating information, or to accomplish a specific conversational goal, such as displaying skepticism (Carberry, 1989). These omissions can take the form of syntactically correct sentences that leave out important semantic information or even incomplete sentence fragments (Fernández et al., 2007; Raghu et al., 2015). Figure 1 shows an example of a dialog which contains both of these. Turn 2 demonstrates a syntactically correct sentence where the user asks about the capital, but leaves out which country they are referring to. Turn 3 shows an example of a syntactically incomplete sentence, where the user leaves out both the country and the verbal phrase.

In this paper we refer to these omitted entities as *implicit references* because while there is no direct reference, e.g., a pronoun, it is still understood that the user is referring to a specific entity. We propose implicit reference as unifying term encapsulating this type of phenomena and including implicit arguments, zero-anaphora, and certain types of noun ellipsis, which we expand on in section 2.

Turn		Utterance
1	**USR**	Who is the Chancellor of Germany?
	SYS	Olaf Scholz is the current Chancellor of Germany.
2	**USR**	And what is the capital _? [of Germany]
	SYS	The capital of Germany is Berlin.
3	**USR**	And _ the population _? [what is], [of Germany]

Figure 1: Dialog between a user USR and a system SYS, with examples of implicit references (in Turn 2 and 3) and another implicit element (in Turn 3) indicated by underscores in red. The correct resolution of each implicit element is shown in brackets in red.

While parsing such sentences is a simple task for humans, it poses a larger problem for automatic systems, which are often designed to only consider a single dialog turn at a time. Therefore research in this area focuses on trying to exploit the dialog context to find what, if any, information has been included only implicitly in a current dialog turn (Mittal et al., 2018; Tseng et al., 2021; Maqbool et al., 2022). This can be especially challenging, however, when such information can lie anywhere in the conversational history (Wu et al., 2021).

2 Definitions

In this section, we provide an overview of linguistic phenomena related to the concept of implicit reference and discuss overlaps and differences in definition. Our focus lies on phenomena that occur in dialogue and written text involving an omitted element referring to an entity.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 587–600
November 20–23, 2022.

Ellipsis. Ellipsis is a syntactic phenomenon in which a constituent is omitted because it can be resolved from the context. Although there are multiple types of ellipsis, we limit the scope of this survey to focus only on nominal ellipsis.

Nominal ellipsis occurs when the head noun inside a noun phrase is implicit. An example taken from the NOEL corpus (Khullar et al., 2020) is: *Let's party at Sam's* $\underline{\quad}_{\text{NP}}$ *this Friday*. In this case, Sam's location is omitted. When the noun phrase is fully omitted, noun ellipsis can also be seen as zero anaphora. Therefore, some researchers use the term noun phrase ellipsis to refer to instances for which only a part of the noun phrase got deleted (Menzel, 2016) whereas others use the term to indicate both types of cases (Khullar et al., 2020).

Implicit argument. An implicit argument is the filler of a semantic role that is not realized in the local syntactic context of its predicate. Frequent examples in English include logical subjects in passive voice sentences (*He was* ***called*** ___) and omitted arguments of nominalized predicates (*They approved the* ***use*** _). Implicit arguments are related to ellipsis in that a subset of them can be viewed as the semantic equivalent of the omission of syntactic constituent. In frame-semantic theory (Fillmore, 1977) implicit role fillers are also referred to as *null instantiations* (NI) and categorized into definite, indefinite and constructional NIs. Definite NIs refer to a definite entity in the context, whereas other NIs can have an unspecific, existential interpretation.

Zero-Anaphora. In general, *anaphora* are references to other expressions in context. Unlike explicit elements, such as pronouns, zero-anaphora are a special case in which the expression itself is omitted. The term is mostly commonly used in context of *pro-drop languages*, in which pronouns can be omitted in general or under specific circumstances. In languages such as Japanese and Chinese, such omissions of pronouns can also occur in obligatory syntactic positions. There are exceptional cases in which this is also possible in non-prodrop languages such as English. An example from a recipe is: *Bake _ for 30 minutes* (Jiang et al., 2020, p.822). Zero-anaphora are related to implicit arguments in that they fill a semantic role in addition to serving a anaphoric function.

Implicit Reference. In the remainder of this paper, we will use *implicit references* as a general term to cover all referential expressions to entities that are omitted in context. Because such expressions can typically be realized as constituents, they form a subset of nominal ellipsis. By definition, implicit references do not have to fill a semantic role or a anaphoric function. Therefore, they form a superset of implicit arguments and zero anaphora.

3 Implicit Reference Tasks in Dialog

This section introduces the most common areas of research on implicit references in dialog.

3.1 Conversational Semantic Role Labeling

Semantic Role Labeling (SRL) is a task in which the predicates in a sentence are analyzed regarding their arguments, in order to determine "*who* did *what* to *whom*". The task is also referred to as Predicate Argument Structure Analysis. Generally, SRL can be divided into three subtasks: 1) recognizing the predicates in a given sentence, 2) finding their arguments, and 3) assigning corresponding semantic labels (He et al., 2017). While much research has investigated automatically extracting such arguments in text (Carreras and Màrquez, 2005; Pradhan et al., 2013; Zhou and Xu, 2015; He et al., 2021; Tan et al., 2018), these methods can have difficulty adapting to a dialog context (Xu et al., 2021). While traditional SRL methods often consider only one sentence at a time, conversations generally contain implicit or explicit references to entities from previous utterances.

The goal of conversational Semantic Role Labeling (CSRL) is, given a dialog, to predict complete semantic-role structures for each predicate, even in the case of implicit arguments that are outside the context of a single dialog turn. Performance on this task is generally evaluated either explicitly via precision, recall, and F1-scores over (predicate, argument) tuples (Wu et al., 2021; Imamura et al., 2014; Xu et al., 2021; He et al., 2021) or implicitly via their performance on a downstream task such as conversational utterance rewriting (Xu et al., 2020).

3.2 Conversational Utterance Rewriting

This task has been referred to by many names, including: conversational query understanding (Ren et al., 2018), conversational ellipsis filling (Zhang et al., 2020), ellipsis and coreference resolution (Ni and Kong, 2021), zero-label anaphora resolution (Maqbool et al., 2022), incomplete utterance rewriting (Liu et al., 2020a) incomplete utterance restoration (Pan et al., 2019), question rewriting in

context (Elgohary et al., 2019), conversational question reformulation (Lin et al., 2020), non-sentential utterance restoration (Raghu et al., 2015). In this paper, we use *Conversational Utterance Rewriting* (CUR) as a general term to encapsulate the task.

The goal of CUR is, given a user utterance and conversational context, to rewrite the utterances such that all information needed to understand it is contained in the rewrite (Ren et al., 2018). This often implicitly or explicitly requires the use of the conversational context to reconstruct the implicit references (Vakulenko et al., 2021). However, implicit reference is never the sole consideration of this task, rather it is part of a more holistic approach including coreference and verb ellipsis resolution in order to generate a fully grammatical expanded version of the user utterance (Tseng et al., 2021).

Data may include labels for anaphora (including zero anaphora) (Regan et al., 2019; Dalton et al., 2020; Raghu et al., 2015; Zhang et al., 2020) or only dialog turns and their corresponding rewrites (Raghu et al., 2015; Elgohary et al., 2019; Pan et al., 2019; Su et al., 2019; Zhou et al., 2019). Performance is generally measured either explicitly – e.g., using metrics such as exact matches or BLEU score between suggested system rewrites for the utterance and a set of gold label annotations (Zhang et al., 2020) – or implicitly – based on performance of downstream tasks such as question answering, database querying, or dialog act classification (Guo et al., 2018; Mittal et al., 2018).

3.3 Noun Ellipsis Detection and Resolution

While Noun ellipsis resolution is often implicitly considered in CUR, we define noun ellipsis detection and resolution as a separate task in this paper. As the scope is far narrower/more precise in this task, it may attract the interest of a different group of researchers than the broader task of CUR.

Khullar et al. (2020) suggest that noun ellipsis detection can be thought of as a classification task, where given a tri-gram, the goal is to predict whether it includes evidence of ellipsis (called an ellipsis licensor). Similarly, they suggest that noun ellipsis resolution can be considered a classification problem. For a given triad of [Licensor, Antecedent, all tokens in a Sentence], the classifier must predict whether the antecedent candidate is the resolution of the ellipsis. Both tasks can then be evaluated with an F1-score, precision, and recall against gold-label annotations.

4 Data

In the following subsections, we describe datasets which have been collected for studying implicit references in dialog. The datasets are directly compared in Table 1 as well as described below.

4.1 Noun Ellipsis

NoEL is an English dataset (Khullar et al., 2020) that contains 946 annotated instances of noun ellipsis from the first 100 movies from the Cornell Movie Dialogs Dataset (Danescu-Niculescu-Mizil and Lee, 2011).

4.2 Conversational Semantic Role Labeling

CSRL The most popular dataset for conversational semantic role labeling is the CSRL dataset collected by Xu et al. (2020). The dataset is in Chinese and composed of three different subsets: 1) SRL annotations for 3,000 dialogs (33,673 predicates in 27,198 utterances) from the DuConv dataset, a knowledge-driven dialog corpus focusing on celebrities and movies. 2) 300 sessions from Personal-Dialog (1,441 predicates in 1,579 utterances), a dataset created by crawling Weibo[1] posts. 3) 200 sessions from NewsDialog (3,621 predicates in 6,037 utterances), a corpus collected by asking two participants to discuss news articles.

Other Datasets Other smaller datasets include that of Zhang et al. (2020) who annotate 1,689 user utterances from the Gunrock dataset (Chen et al., 2018) and that of Wu et al. (2022) which includes annotations for 972 user utterances from PersonaChat (Zhang et al., 2018) and CMU-DoG (Zhou et al., 2018). Both of these datasets are in English.

4.3 Utterance Rewriting

GECOR The GECOR dataset (Quan et al., 2019) is an extension of the task-oriented, English language CamRest676 dataset (Wen et al., 2016). Here, the authors added manual annotations to label sentences which contain coreference or ellipsis and provide rewritten versions of these sentences which do not. Additionally if it were possible to transform a complete sentence to contain either ellipsis or coreference, this was done. The dataset contains 2,744 user utterances of which 1,174 originally contained ellipsis and 1,331 were rewritten to include ellipsis or coreference.

[1]Weibo is a popular Chinese social media website.

Dataset	Language	Size	Annotation	Dialog Type
TREC CAsT (Dalton et al., 2020)	English	38,426,252	Sentence Rewrites	Conversational QA
CANARD (Elgohary et al., 2019)	English	40,527	Sentence Rewrites	Conversational QA
Question Completion (Raghu et al., 2015)	English	7,400	Sentence Rewrites+	Conversational QA
CQR (Regan et al., 2019)	English	*3,000	Sentence Rewrites+ Anaphora Classes	Task Oriented
GECOR (Quan et al., 2019)	English	2,744	Sentence Rewrites+	Task Oriented
Hybrid-EL-CMP (Zhang et al., 2020)	English	2,258 1,689	Sentence Rewrites+ Semantic Roles+	Chit-chat
Zero-Shot-XCSRL (Wu et al., 2022)	English	927	Semantic Roles	Chit-chat
NoEl (Khullar et al., 2020)	English	**100	Ellipsis Licensors	Movie Script
Psuedo Rewrite (Zhou et al., 2019)	Chinese	6,846,467	Sentence Rewrites	Social Media
Restoration-200K (Pan et al., 2019)	Chinese	200,000	Sentence Rewrites	Social Media
Dialog Utterance Rewrite Corpus (Su et al., 2019)	Chinese	40,000	Sentence Rewrites	Social Media
CSRL (Xu et al., 2020)	Chinese	*3,000 *300 *200	Semantic Roles	Document Based Social Media Chit-chat

Table 1: Comparison of implicit reference datasets; where possible, the dataset column acts as a link to the data itself. Unless otherwise indicated (by * or **), size refers to the number of utterances in the datset. * indicates datasets which measured size as the number of dialogs rather than turns and ** indicates NoEL which measured size as the number of movie scripts annotated. + Refers to datasets which include labels/statistics for which sentences include ellipsis, coreference, or both.

CANARD The CANARD dataset (Elgohary et al., 2019) rewrites questions from the conversational QA dataset QuAC (Choi et al., 2018) to resolve ellipsis and anaphora and to disambiguate coreferences. The dataset contains 40,527 English language questions and their rewritten versions.

Dialog Utterance Rewrite Corpus Su et al. (2019) introduce a new Chinese language dataset extracted from multi-turn dialogs from social media. The dataset contains 40,000 original utterances as well as rewritten versions of those including ellipsis, coreference, or both. While the authors do not explicitly label which sentences contain such phenomena they randomly sampled 2,000 dialogs and found roughly half needed to be rewritten.

Question Completion Raghu et al. (2015) introduce an English language dataset where crowdsourced workers were presented a question–answer pair and asked to come up with a follow-up question both in an elliptical form and in a fully resolved form. The data set contains 7,400 entries, each with a question, an answer, an elliptical follow-up question, and a resolved follow-up question.

TREC CAsT CAsT-19 (Dalton et al., 2020) is a dataset of 38,426,252 passages from the TREC Complex Answer Retrieval (Dietz et al., 2017) and Microsoft Machine Reading Comprehension datasets (Nguyen et al., 2016). The questions contain implied context, ellipsis and topic shifts. CAsT-19 provides resolved versions of each turn, including those with ellipsis as well as entity annotations.

Other Datasets Zhang et al. (2020) present an English language dataset containing 2,258 user utterances from the Gunrock dataset, among them 1,124 utterances contain ellipsis, and 204 complete utterances which were modified to include a version with ellipsis. Pan et al. (2019) introduce Restoration-200K, a Chinese language dataset containing 200,000 utterances obtained from discussions on the online community Douban Group. Dialogs contain at least six turns and were professionally annotated to resolve utterances omitting information. Zhou et al. (2019) also provide a Chinese language dataset collected by crawling Douban Group. They collected 6,844,393 entries each containing an utterance, one turn of context, an automatically generated rewrite of the utterance, and the response from the next turn. Finally Regan et al. (2019) provide an English language dataset containing approximately 3,000 dialogs over three domains including 2,287 rewrites. They provide

both rewrite annotations as well as labels for what type of anaphora are present in a sentence, i.e., zero (1,436 instances), pronomial (445 instances), locative (239 instances), nominal (184 instances).

5 Methods

In the following section we outline methods which have been used for tasks related to implicit reference in dialogs. We provide an overview (Figure 2) of both classical approaches and state of the art methods and brief description of each approach.

5.1 Noun Ellipsis Resolution and Detection

As one of the few papers focused solely on noun ellipsis detection and resolution, Khullar et al. (2020) demonstrated classical machine learning approaches can be used for both of these tasks. They compared several classifiers from the sklearn toolkit (Pedregosa et al., 2011), testing their performance on a dataset of movie scripts.

5.2 Conversational Semantic Role Labeling

5.2.1 Classical Approaches

Imamura et al. (2014) were some of the first researchers to tackle SRL in dialog. They investigated zero-anaphora cases in Japanese, first training a maximum entropy-based classifier on the NAIST (Iida et al., 2007b) newspaper corpus and then adapting it to a dialog corpus which they collected. The general approach first identified all predicates in a sentence and then generated a list of candidate arguments from the current sentence and dialog history. For each candidate, relevant features were selected to predict the most likely predicate/argument pairs. This approach significantly outperformed text-based classifiers, when tested on dialog data.

5.2.2 Neural Approaches

BERT-based Approaches Recently, SRL has gained popularity for dialog applications. Xu et al. (2020), were some of the first to approach this task. The authors adapted a RoBERTa (Liu et al., 2019) based model pre-trained for text SLR (Shi and Lin, 2019) to work in the dialog domain. This was approached in two ways. They later (Xu et al., 2021) expanded their model to include self attention and additional inputs such as a speaker indicator, a dialog turn indicator, and a predicate indicator as well as the encoded dialog text. In both cases, the authors also tested the performance on downstream tasks such as dialog query rewriting (Xu et al., 2020, 2021) and dialog generation tasks (Xu et al., 2021).

He et al. (2021) proposed improving upon the work of Xu et al. (2021) by replacing BERT with K-BERT (Liu et al., 2020b), which introduces knowledge from an external graph into BERT pre-training. The proposed model consisted of four parts: 1) the K-BERT encoder using CN-DB-Pedia (Xu et al., 2017) – a large-scale open-domain Chinese encyclopedia – as the knowledge graph, 2) a dialog turn indicator and a predicate indicator encoder, 3) K self-attention layers, and 4) a softmax prediction layer. The model was trained on DuConv-CSRL subset of the dataset from Xu et al. (2021) and demonstrated increased performance compared to a baseline of the same architecture without knowledge graph enhancement.

Graph Approaches Wu et al. (2021) proposed a different approach to graph integration, rather than seeking to encode external information, the authors used a graph structure to better model the dialog context. The model included three components: 1) A pre-trained language model able to generate local and contextual representations for tokens, similar to the model proposed by Xu et al. (2021). 2) A new attention strategy to learn predicate-aware contextual representations for tokens. And 3) a Conversational Structure Aware Graph Network (CSAGN) for learning high-level structural features to represent user utterances. The authors trained their model on the three Chinese dialog datasets annotated by Xu et al. (2021), outperforming their BERT-based baseline.

5.3 Conversational Utterance Rewriting

5.3.1 Classical approaches

In general, approaches to utterance rewriting fall into three categories: those based on semantics (Waltz, 1978), syntax (Hendrix et al., 1978), or pragmatics (Carberry, 1989). Semantics-based approaches work to reconstruct implicit references through an understanding of the meaning of the sentence and the preceding context, syntactic approaches through the structure of the sentence and its context, and pragmatics-based approaches through an understanding of a speaker's discourse goals. Early work emphasized the generation of logical rules derived from examples and case studies (Carberry, 1989), while more recent work has shifted to statistical and machine-learning approaches. Raghu et al. (2015), for example, devel-

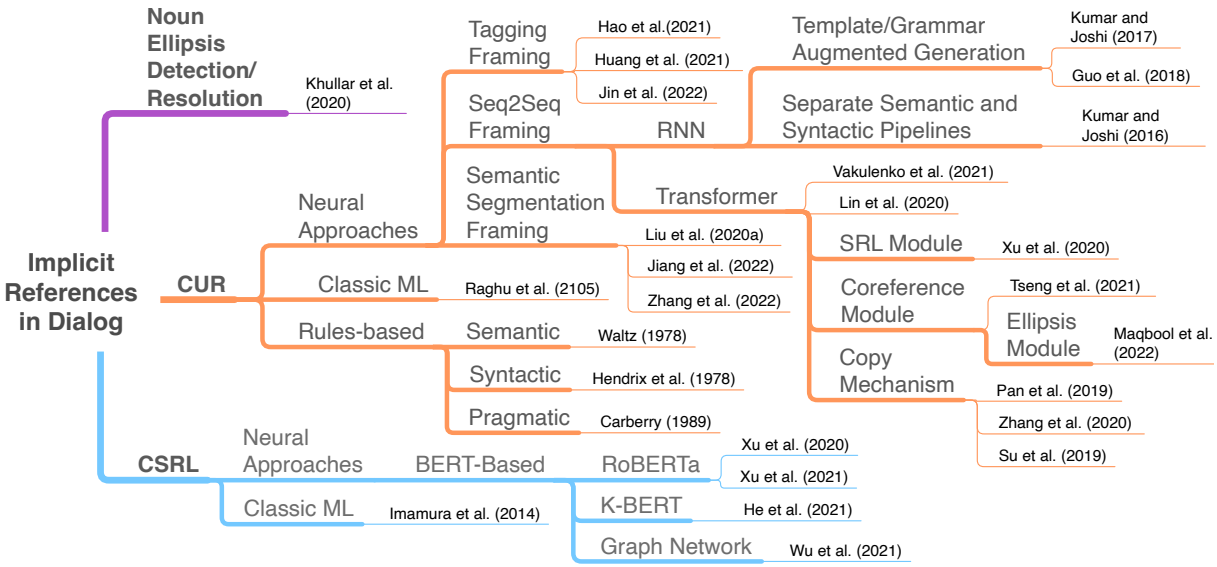

Figure 2: Overview of methods used for implicit reference in dialog.

oped a system which learned to extract keywords from incomplete utterances and expand them using delexicalized templates. Candidate rewrites were then ranked by a support vector machine based on semantic and syntactic features.

5.3.2 Neural approaches: Sequence to Sequence Framing

Due to the nature of this task, many neural approaches frame utterance rewriting as a sequence-to-sequence problem similar to machine translation: mapping an incomplete original user utterance along with its conversational context to a complete (intended) user utterance (Vakulenko et al., 2021). However, unlike machine translation there are two types of input which can be passed to the model (the current user utterance and the context) rather than only a single source (Ren et al., 2018).

Copy Mechanisms Another unique property of the rewriting task, is that most generated words come from the previous utterance or context sentences. Several approaches thus try to exploit this property to improve performance. Elgohary et al. (2019), for example, implemented a sequence to sequence model with attention and a copy mechanism (See et al., 2017). Quan et al. (2019), presented a similar approach, separately encoding the user utterance and complete dialog context before passing these inputs to a decoder which included either a copy (Gu et al., 2016) or a gated copy mechanism (modified from See et al. (2017)). In contrast, Pan et al. (2019) implemented what they refer to as a "pick and combine" model which used the pre-trained language model BERT (Devlin et al., 2019) as a classifier to select omitted words from the context to be given as input to a pointer generator network. Their model could then copy words from the input by directly taking the attention score as the prediction probability. Another approach was proposed by Su et al. (2019), who demonstrated a transformer-based rewriting architecture with a pointer network, while Ni and Kong (2021) explored implementing a speaker highlight dialogue history encoder to create a global representation of the dialogue history as well as a top-down hierarchical copy mechanism.

Handling Data Scarcity A key difficulty with the sequence to sequence approach, is the lack of large-scale parallel corpora. To tackle this, Kumar and Joshi (2016) tried to decompose the problem, proposing an RNN-based encoder/decoder ensemble model, combining a syntactic sequence model for learning linguistic patterns, and a semantic sequence model for learning semantic patterns. An alternate approach by Kumar and Joshi (2017) instead framed the problem as a retrieval problem, implementing a retrieval based sequence to sequence model. Here the authors used a set of pre-computed semantically correct question templates to guide question generation and a language model to rank candidates for syntactic correctness. Guo et al. (2018) propose a similar architecture, using a small grammar rather than template questions. To better make use of the dialog context, however, they also introduced a dialog memory module to track

entities, predicates, and actions which were mentioned in the dialog. In another approach to the data scarcity problem, Zhou et al. (2019) propose a training architecture using automatically generated rewrites for incomplete user utterances. They first train a GRU based encoder-decoder model enhanced with CopyNet (Gu et al., 2016) on the generated data. Then results are fine-tuned using reinforcement learning to correct for errors learned from the automatically generated training data.

Large Pre-trained Language Models Large pre-trained models are a powerful tool for many natural language tasks. To demonstrate their applicability to query rewriting, Lin et al. (2020) perform experiments testing multiple language models and configurations. Tseng et al. (2021) propose a more complex architecture, using GPT-2 (Radford et al., 2019) as a decoder with a coreference resolution module built on-top to act as input for their final query rewriter. Maqbool et al. (2022) incorporate both BERT and GPT-2 into their model architecture as a way to help generate and score possible rewritten utterances. Their model consisted of three stages, 1) an encoding stage with two parallel pipelines – one for handling the case of ellipsis and the other for coreference, 2) a candidate selection phase, and 3) a refinement phase using a masked language model BERT and GPT-2 to refine the output fluency.

Other Approaches Other approaches include augmenting the rewrite model with predicted semantic role information (Xu et al., 2020) or tackling downstream tasks by predicting two outputs (with rewritten or incomplete utterance as input) then using an expert knowledge-guided selector to make the final decision (Zhang et al., 2020).

5.3.3 Neural Approaches: Semantic Segmentation Framing

In contrast to framing CUR as a sequence to sequence task, recent approaches (Liu et al., 2020a; Jiang et al., 2022; Zhang et al., 2022) propose to consider it similar to semantic segmentation or object detection in computer vision. Rather than trying to generate a new utterance from scratch, this formulation, introduces the idea of edit operations being performed between word pairs of the context utterances and the incomplete utterance. Given relevant features between word pairs as an matrix (Liu et al., 2020a; Jiang et al., 2022), or the self attention weight matrix from the encoder (Zhang et al., 2022) a model can predict the edit type (substitute, insert, or none) for each word pair as a pixel-level mask. The ability to take global features into account in these approaches has shown increased performance compared to pure text generation approaches (Liu et al., 2020a; Jiang et al., 2022).

5.4 Neural Approaches: Tagging Framing

The tagging framing of CUR is very closely related to the semantic segmentation framing, however, rather than working on word pairs, edit decisions are made for single tokens. In general, the goal of this approach is to determine whether to delete, keep or change each token in a given input sentence (Huang et al., 2021), although this can also be framed as whether to delete a token or insert information from the dialog context after the token (Hao et al., 2021; jin et al., 2022). In this way, the search space is greatly reduced compared to sequence to sequence approaches, which can also make this approach more robust to changes between training and test data (Hao et al., 2021). Approaches using this framing largely distinguish themselves in the way they handle the change/insert step: choosing a single span from the context for each token (Hao et al., 2021), choosing multiple spans from the context (jin et al., 2022), or auto-regressive text generation for the inserted phrase (Huang et al., 2021).

6 Further Readings

In this section we provide an overview of related tasks, which handle (explicit) anaphora in dialogue or implicit information in written text settings and may serve as a useful reference as they aim to address similar problems.

Anaphora. Anaphora resolution is the task of identifying which parts of a text refer to the same discourse entity, which is based on the idea that different expressions can refer to the same entity. Lata et al. (2021) provide a survey of approaches to anaphora resolution in text. For dialog specific anaphora resolution, there are multiple shared tasks which have been organized, such as the CODI-CRAC 2021 shared task on anaphora resolution in (spoken) dialogues (Khosla et al., 2021), which focuses on entity coreference resolution, bridging resolution, discourse deixis/abstract phenomena as a follow-up of CRAC-18. Additionally datasets such as MuDoCo (Martin et al., 2020) provide annotations for thousands of dialogs, which contain

entity mentions and coreference links.

Ellipsis. Several studies have investigated the detection and resolution of ellipsis in written texts. For example, previous work applied classical machine learning techniques to detecting and resolving ellipsis in the British National Corpus (Nielsen, 2003a,b) and in the Penn Treebank (Nielsen, 2004). Earlier work approached ellipses with syntactic patterns (Hardt, 1992).

Most work has focused on verb ellipsis, with the first study on noun ellipsis detection in texts performed by Khullar et al. (2019), who took a small dataset from the UD treebank that did not contain a noun phrase. For detection and resolution, they used a rule-based system using syntactic constraints of licensors of ellipsis and part-of-speech similarity between the licensors of ellipsis and the modifier of the antecedent.

Zero-Anaphora. Most work on zero-anaphora has focused on pro-drop languages, in particular Asian languages such as Japanese (Konno et al., 2021; Iida et al., 2007a, 2006; J., 2013; Iida et al., 2016; Isozaki and Hirao, 2003; Sasano and Kurohashi, 2011; Sasano et al., 2008; Seki et al., 2002; Yamashiro et al., 2018; Umakoshi et al., 2021; Ueda et al., 2020) and Chinese (Converse, 2005; Chen and Ng, 2014; Kong and Zhou, 2010; Liu et al., 2017; Yin et al., 2018). Zero-anaphora has also been studied in Romance languages, including Italian (Iida and Poesio, 2011), Spanish (Palomar et al., 2001; Rodríguez et al., 2010) and Portuguese (Pereira, 2009). In English, zero-anaphora has been studied in conversation analysis (Oh, 2005) and in recipes (Jiang et al., 2020).

Implicit arguments. Implicit argument prediction in text has been modeled as a special case of anaphora resolution (Silberer and Frank, 2012), by leveraging (explicit) semantic role labeling (Schenk and Chiarcos, 2016; Chen et al., 2010; Laparra and Rigau, 2013), a combination of the two (Roth and Frank, 2013), as a cloze-task (Cheng and Erk, 2018) and as a binary classification problem (Gerber and Chai, 2010; Feizabadi and Padó, 2015). The most commonly used dataset for evaluating implicit argument prediction in texts is by Gerber and Chai (2010). A larger dataset was recently made available by Ebner et al. (2020).

7 Future directions

After presenting the current state of research on implicit reference in dialog, we propose the following future directions:

Benchmarking Resolving implicit references in dialog has primarily been explored through the tasks of conversational semantic role labeling or conversational utterance rewriting. In conversational utterance rewriting in particular, results are reported on different datasets in different languages and with various settings. Thus, it is very challenging to draw conclusions and to compare among proposed computational methods. Therefore, one of the first steps towards advancing systems for resolving implicit references in dialog is to establish a model agnostic benchmark, such as GLUE (Wang et al., 2018), to collect resources for training, evaluating, analysing such systems.

Data Explication In many cases, implicit references can be successfully resolved and clarified in the course of a dialogue. For computational models of language understanding, this is nevertheless problematic, since the relevant context can be quite broad and implicit references are by definition not explicit in the relevant position. Supervised methods in particular therefore require explicit training signals for the resolution of implicit references. Existing work on implicit arguments in text attempts to address this problem by using artificial training data based on explicit reference chains, sentence-based semantic roles, or event representations (Silberer and Frank, 2012; Schenk and Chiarcos, 2016; Cheng and Erk, 2018). Similar to Zhou et al. (2019), one research direction would be to create similar data for dialogue scenarios, for example, by collecting resolution patterns observable over multiple utterances and generalizing/applying such patterns in comparable contexts.

Modeling State-of-the-art systems to resolve implicit references in dialog are mostly based on deep learning models. One of the known weaknesses of such models is their uncertainty values. They are often overconfident (Wang et al., 2021), i.e. their certainty values are not good indicators of the actual likelihood of a correct prediction. When resolving implicit references, there may be multiple entities in the context which the reference might refer to. In such cases, estimated uncertainty values play an important role, especially in the context of

dialog systems where it is possible to gain explicit feedback from a user to resolve ambiguities. While there are already uncertainty metrics used in similar fields, e.g., reconstructing user utterances after ASR errors (Cho et al., 2021; Fan et al., 2021) these methods have not yet been integrated into work on implicit references in dialog. Additionally, these approaches focus only on implicit feedback from the user, e.g., rephrasing an initial query, and do not explore the opportunity of eliciting explicit feedback. A reliable uncertainty value would aid dialog policies in choosing a meaningful next step, e.g., whether to use a current utterance or ask for clarification. Thus, one meaningful research direction is to explore methods for estimating reliable uncertainty values for such implicit reference resolution in dialog.

Evaluation An open problem regarding phenomena of implicit language is that there may be multiple possible interpretations depending on the context. Existing work on implicit references in texts in particular has shown that, depending on the exact task, annotators themselves only exhibit low to moderate levels of agreement (Gerber and Chai, 2010). By considering uncertainty values, such disagreements can already be taken into account in modeling. In addition, however, the possibility of resolving an implicit reference in different ways is also relevant for the evaluation of corresponding models. To allow different potential assessments in context, we recommend developing evaluations that can take an interactive form, so that systems can ask clarification questions when multiple interpretations are possible for an implicit reference.

Acknowledgements

Parts of the research presented in this paper were funded by the DFG Emmy Noether program (RO 4848/2-1).

References

Sandra Carberry. 1989. A pragmatic-based approach to ellipsis resolution. *Computational Linguistics*, 15(2):75–96.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pages 152–164.

Chen Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–774, Doha, Qatar. Association for Computational Linguistics.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, et al. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa Prize Proceedings*.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden. Association for Computational Linguistics.

Pengxiang Cheng and Katrin Erk. 2018. Implicit argument prediction with event knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 831–840, New Orleans, Louisiana. Association for Computational Linguistics.

Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized search-based query rewrite system for conversational AI. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 179–188, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Susan P. Converse. 2005. Resolving pronominal references in Chinese with the Hobbs algorithm. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. *CAsT-19: A Dataset for Conversational Information Seeking*, page 1985–1988. Association for Computing Machinery, New York, NY, USA.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Xing Fan, Eunah Cho, Xiaojiang Huang, and Chenlei Guo. 2021. Search based self-learning query rewrite system in conversational ai. In *2nd International Workshop on Data-Efficient Machine Learning (DeMaL)*.

Parvin Sadat Feizabadi and Sebastian Padó. 2015. Combining seemingly incompatible corpora for implicit semantic role labeling. In *Proceedings of STARSEM*, pages 40–50, Denver, CO.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Charles J Fillmore. 1977. Scenes-and-frames semantics. zampolli, a.(ed.): Linguistic structures processing.

Matthew Gerber and Joyce Chai. 2010. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. RAST: Domain-robust dialogue rewriting as sequence tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4913–4924, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Hardt. 1992. An algorithm for vp ellipsis. In *ACL*.

Boyu He, Han Wu, Congduan Li, Linqi Song, and Weigang Chen. 2021. K-csrl: Knowledge enhanced conversational semantic role labeling. In *2021 13th International Conference on Machine Learning and Computing*, pages 530–535.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Gary G Hendrix, Earl D Sacerdoti, Daniel Sagalowicz, and Jonathan Slocum. 1978. Developing a natural language interface to complex data. *ACM Transactions on Database Systems (TODS)*, 3(2):105–147.

Mengzuo Huang, Feng Li, Wuhe Zou, and Weidong Zhang. 2021. Sarg: A novel semi autoregressive generator for multi-turn incomplete utterance restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13055–13063.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632, Sydney, Australia. Association for Computational Linguistics.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007a. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Trans. Asian Lang. Inf. Process.*, 6.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007b. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the linguistic annotation workshop*, pages 132–139.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 804–813, Portland, Oregon, USA. Association for Computational Linguistics.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using

multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1244–1254, Austin, Texas. Association for Computational Linguistics.

Kenji Imamura, Ryuichiro Higashinaka, and Tomoko Izumi. 2014. Predicate-argument structure analysis with zero-anaphora resolution for dialogue systems. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 806–815.

Hideki Isozaki and Tsutomu Hirao. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 184–191.

Antony P. J. 2013. Machine translation approaches and survey for Indian languages. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*.

Wangjie Jiang, Siheng Li, Jiayi Li, and Yujiu Yang. 2022. Multi-turn incomplete utterance restoration as object detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8052–8056. IEEE.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (RISeC): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826, Suzhou, China. Association for Computational Linguistics.

Lisa jin, Linfeng Song, Linfeng Jin, Dong Yu, and Daniel Gildea. 2022. Hierarchical context tagging for utterance rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10849–10857.

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.

Payal Khullar, Allen Antony, and Manish Shrivastava. 2019. Using syntax to resolve NPE in english. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019*, pages 534–540. INCOMA Ltd.

Payal Khullar, Kushal Majmundar, and Manish Shrivastava. 2020. NoEl: An annotated corpus for noun ellipsis in English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 34–43, Marseille, France. European Language Resources Association.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA. Association for Computational Linguistics.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution.

Vineet Kumar and Sachindra Joshi. 2016. Non-sentential question resolution using sequence to sequence learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2022–2031, Osaka, Japan. The COLING 2016 Organizing Committee.

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval based sequence to sequence learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 705–714, New York, NY, USA. Association for Computing Machinery.

Egoitz Laparra and German Rigau. 2013. ImpAr: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1180–1189, Sofia, Bulgaria. Association for Computational Linguistics.

Kusum Lata, Pardeep Singh, and Kamlesh Dutta. 2021. A comprehensive review on feature set used for anaphora resolution. *Artificial Intelligence Review*, 54(4):2917–3006.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020a. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2846–2857, Online. Association for Computational Linguistics.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for

zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111, Vancouver, Canada. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020b. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

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 arXiv:1907.11692*.

MH Maqbool, Luxun Xu, AB Siddique, Niloofar Montazeri, Vagelis Hristidis, and Hassan Foroosh. 2022. Zero-label anaphora resolution for off-script user queries in goal-oriented dialog systems. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 217–224. IEEE.

Scott Martin, Shivani Poddar, and Kartikeya Upasani. 2020. MuDoCo: Corpus for multidomain coreference resolution and referring expression generation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 104–111, Marseille, France. European Language Resources Association.

Katrin Menzel. 2016. Understanding english-german contrasts : a corpus-based comparative analysis of ellipses as cohesive devices.

Ashish Mittal, Jaydeep Sen, Diptikalyan Saha, and Karthik Sankaranarayanan. 2018. An ontology based dialog interface to database. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1749–1752.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Zixin Ni and Fang Kong. 2021. Enhancing long-distance dialogue history modeling for better dialogue ellipsis and coreference resolution. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 480–492. Springer.

Leif Arda Nielsen. 2003a. A corpus-based study of verb phrase ellipsis. In *In Proceedings of the 6th Annual CLUK Research Colloquium*, pages 109–115.

Leif Arda Nielsen. 2003b. Using machine learning techniques for vpe detection. In *In Proceedings of RANLP*, pages 339–346.

Leif Arda Nielsen. 2004. Verb phrase ellipsis detection using automatically parsed text. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, page 1093–es, USA. Association for Computational Linguistics.

Sun-Young Oh. 2005. English zero anaphora as an interactional resource. *Research on Language and Social Interaction*, 38:267–302.

Manuel Palomar, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, and Rafael Muñoz. 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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 1824–1833, Hong Kong, China. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Simone Pereira. 2009. ZAC.PB: an annotated corpus for zero anaphora resolution in portuguese. In *Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria*, pages 53–59. RANLP 2009 Organising Committee / ACL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Dinesh Raghu, Sathish Indurthi, Jitendra Ajmera, and Sachindra Joshi. 2015. A statistical approach for non-sentential utterance resolution for interactive QA system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 335–343, Prague, Czech Republic. Association for Computational Linguistics.

Michael Regan, Pushpendre Rastogi, Arpit Gupta, and Lambert Mathias. 2019. A dataset for resolving referring expressions in spoken dialogue via contextual query rewrites (cqr). *arXiv preprint arXiv:1903.11783*.

Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1715–1724, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. 2010. Anaphoric annotation of Wikipedia and blogs in the live memories corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Michael Roth and Anette Frank. 2013. Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 306–316, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 769–776, Manchester, UK. Coling 2008 Organizing Committee.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Niko Schenk and Christian Chiarcos. 2016. Unsupervised learning of prototypical fillers for implicit semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1473–1479, San Diego, California. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 1–10, Montréal, Canada. Association for Computational Linguistics.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. CREAD: combined resolution of ellipses and anaphora in dialogues. *CoRR*, abs/2105.09914.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

David L Waltz. 1978. An english language question answering system for a large relational database. *Communications of the ACM*, 21(7):526–539.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. 2021. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162, Austin, Texas. Association for Computational Linguistics.

Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu, and Linqi Song. 2022. Zero-shot cross-lingual conversational semantic role labeling.

Han Wu, Kun Xu, and Linqi Song. 2021. CSAGN: Conversational structure aware graph network for conversational semantic role labeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2317, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cndbpedia: A never-ending chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 428–438. Springer.

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. 2020. Semantic Role Labeling Guided Multi-turn Dialogue ReWriter. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6632–6639, Online. Association for Computational Linguistics.

Kun Xu, Han Wu, Linfeng Song, Haisong Zhang, Linqi Song, and Dong Yu. 2021. Conversational semantic role labeling. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2465–2475.

Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. 2018. Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu. 2020. Filling conversation ellipsis for better social dialog understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 9587–9595.

Yong Zhang, Zhitao Li, Jianzong Wang, Ning Cheng, and Jing Xiao. 2022. Self-attention for incomplete utterance rewriting. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8047–8051. IEEE.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. 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 1127–1137.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised context rewriting for open domain conversation.