Incorporating Dropped Pronouns into Coreference Resolution: The case for Turkish

Tuğba Pamay Arslan and **Gülşen Eryiğit**
İTÜ NLP Research Group
Department of AI&Data Engineering
Faculty of Computer&Informatics
Istanbul Technical University
[pamay, gulsen.cebiroglu]@itu.edu.tr

Abstract

Representation of coreferential relations is a challenging and actively studied topic for pro-drop and morphologically rich languages (PD-MRLs) due to dropped pronouns (e.g., null subjects and omitted possessive pronouns). These phenomena require a representation scheme at the morphology level and enhanced evaluation methods. In this paper, we propose a representation & evaluation scheme to incorporate dropped pronouns into coreference resolution and validate it on the Turkish language. Using the scheme, we extend the annotations on the only existing Turkish coreference dataset, which originally did not contain annotations for dropped pronouns. We provide publicly available pre and post processors to enhance the prominent CoNLL coreference scorer also to cover coreferential relations arising from dropped pronouns. As a final step, the paper reports the first neural Turkish coreference resolution results in the literature. Although validated on Turkish, the proposed scheme is language-independent and may be used for other PD-MRLs.

1 Introduction

Coreference resolution (CR) is a semantic-level natural language processing (NLP) task and aims to determine sets of mentions which describe the same real-world entity (e.g., a person, a place, a thing, an event). An end-to-end CR system has two sequential steps: mention detection and mention clustering. The mention detection stage focuses on identifying all possible coreferential mentions referring to a real-world entity within a text. In the next step, the mention clustering stage collects mentions referring to the same real-world entity under the same cluster, resolving which extracted mentions are coreferential.

Although CR is an NLP subject that has been studied for quite a long time (Ng and Cardie, 2002; Sukthanker et al., 2020), studies on PD-MRLs are still in their infancy, and Turkish is one of them.

In MRLs, words may appear under different surface forms taking different types of affixes. In some languages, the richness level may be very high so that most syntactic information is carried at the morphological level leading to the possibility of dropping some functional words and pronouns. An example from the Turkish language (a highly rich MRL) is provided below[1], where verbal agreement and possessive suffixes[2] allow the drop of personal and possessive pronouns. Morphemes emphasized with bold font describe the dropped pronouns: '-im' holds for the pronoun 'benim' (*me*) and '-n' holds for 'sen' (*you*). However, the sentence is naturally made as exemplified below in the second line without personal and possessive pronouns.

Sen benim geldiğ**im**i gördü**n** mü?
~~Sen benim~~ geldiğ**im**i gördü**n** mü?
You I came see did
Did you see that I came?

The pro-drop nature of such languages reveals the need for mention annotation on other tokens (i.e., artificially inserted (Pradhan et al., 2012a; Nedoluzhko et al., 2022) or existing tokens (Rodrıguez et al., 2010; Klemen and Žitnik, 2021) other than the dropped pronouns, such as verbs carrying personal suffixes). The morphological richness in these languages may reveal the appearance of multiple coreference relations on a single token which is illustrated below. The word 'annemin' (*of my mother*) in the below example carries multiple coreferential relations[3] to different people: *me* and *my mother*.

[1] Color codes are used to indicate mentions referring to the same entity.

[2] One should note that personal and possessive suffixes differ from the phenomena called clitic pronouns in Romance languages (e.g. French, Portuguese, Italian) in two ways: 1) These suffixes always appear at the morphology level of a verb or noun although an overt pronoun depicting the same entity exist within the sentence. 2) They always appear as suffixes whereas clitic pronouns in Romance languages are written either as a separate word or as an attachment via a hyphen.

[3] '-m' holds for the pronoun 'benim' (*my*) and the word 'annem' (*my mother*) is a mention itself.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 14–25

Sen [benim] [anne[**m**]in] geldiğ**[i]**ni gördü**n** mü?
~~Sen benim~~ **annemin** geldiğ**i**ni gördü**n** mü?
You my mother came see_did
Did you see the coming of my mother?

Unfortunately, existing coreference evaluators (Pradhan et al., 2014), originally developed for non-prodrop languages (e.g., English), do not support multiple coreferential relations on a single token. On the other hand, representations relying on artificially inserted tokens have their deficiencies, although eliminating the multiple coreference issue:

1. difficulty in determining the most accurate and natural position of the artificial token in the sentence,
2. extra burden during manual annotations,
3. corruption of the original sentence flow,
4. extra coding of the already available information easily deducible from morphology.

In this paper, we propose a representation & evaluation scheme using existing tokens to incorporate dropped pronouns into coreference resolution and validate it on Turkish. Using this scheme, we extend the annotations on the only existing Turkish coreference dataset (Schüller et al., 2017; Pamay and Eryiğit, 2018), which originally did not contain annotations for dropped pronouns. We provide publicly available pre and post processors[4] to enhance the prominent CoNLL coreference scorer[5] (Pradhan et al., 2014) to also cover multiple coreferential relations arising from dropped pronouns. As a final step, the paper reports the first neural Turkish coreference resolution results in the literature providing a strong baseline for future studies in this field. The preliminary results are reported on a neural coreference resolution model with a mention-ranking approach (Klemen and Žitnik, 2021), which was introduced for Slovene, another PD-MRL. Since the coreference information is coded at the morphology level, we investigate the impact of different word embeddings (Mikolov et al., 2013a,b; Grave et al., 2018), neural language models (Peters et al., 2018; Devlin et al., 2018; Schweter, 2020), and the inclusion of hand-crafted features used in previous studies (Schüller et al., 2017; Pamay and Eryiğit, 2018) to analyze their representation power for morphological richness. Although validated on Turkish, the proposed scheme is language-independent and may be used for other PD-MRLs.

The paper is structured as follows: Section 2 gives the related work, Section 3 introduces the representation of dropped pronouns on existing data sets in the literature, Section 4 presents the proposed representation & evaluation scheme for dropped pronouns, Section 5 presents the experimental setup and results, and Section 6 gives the conclusion.

2 Related Work

Machine learning methods requiring hand-crafted features have been used in the CR literature for a long time. Generally, learning-based CR models are collected under three main categories: mention-pair (Ng, 2005; Ji et al., 2005; Nicolae and Nicolae, 2006; Yang et al., 2006; Denis and Baldridge, 2007a; Haponchyk and Moschitti, 2017), entity-mention (McCallum and Wellner, 2005; Denis and Baldridge, 2007a; Culotta et al., 2007), and ranking mechanisms (Denis and Baldridge, 2007b; Rahman and Ng, 2009, 2011). Deep neural networks have been frequently used in recent studies: mention-pair (Martschat and Strube, 2015), entity-mention (Clark and Manning, 2015), mention-ranking (Fernandes et al., 2012; Durrett and Klein, 2013; Björkelund and Kuhn, 2014; Wiseman et al., 2015, 2016). Recently, several neural end-to-end systems which focus on determining the mentions automatically before or in line with the coreference resolution stage have been also introduced (Lee et al., 2017; Joshi et al., 2020; Liu et al., 2020; Xu and Choi, 2020).

Besides the above-mentioned studies on English, the CR studies focusing on pro-drop languages have been increasing. Kong and Ng (2013) improved the CR performance reported in (Pradhan et al., 2012a) by exploiting zero-pronouns (i.e. elided pronouns) on Chinese with traditional machine learning methods. Chen and Ng (2013) enhanced the available approach (Zhao and Ng, 2007) with a richer feature set and also incorporated the dropped pronouns as a referential mention. Neural CR architectures were also employed in the Chinese CR studies (Chen and Ng, 2016; Yin et al., 2016). For Korean, Park et al. (2020) proposed a neural architecture using pointer networks to reduce the computational complexity of an available end-to-end model (Joshi et al., 2019). Guarasci et al. (2021) used ELECTRA (Clark et al., 2020) on the neural structure (Lee et al., 2018) for Italian.

[4] Available from https://github.com/TugbaP/processors-for-conll-coreference-scorer

[5] http:/github.com/conll/reference-coreference-scorers

Klemen and Žitnik (2021) proposed a neural CR model focusing on only mention clustering stage for Slovene.

Evaluation of CR systems is a challenging topic which resulted with several evaluation metrics in the literature: MUC (Vilain et al., 1995), B-Cubed (Bagga and Baldwin, 1998), mention-based & entity-based CEAF (Luo, 2005), BLANC (Recasens and Hovy, 2011), the averaged CoNLL score (Denis and Baldridge, 2009; Pradhan et al., 2014). Each metric evaluates a CR system from different perspectives and has pros and cons. A widely used evaluator (from now on referred to as the CoNLL scorer (Pradhan et al., 2014)) outputs CR performances via all these metrics.

Previous works on Turkish CR are based on traditional machine learning algorithms (Yıldırım and Kılıçaslan, 2007; Yıldırım et al., 2007; Kılıçaslan et al., 2009; Küçük and Yöndem, 2015; Schüller et al., 2017; Pamay and Eryiğit, 2018). The most recent Turkish coreference dataset (MTCC - Marmara Turkish Coreference Corpus) is from Schüller et al. (2017), and consists of a document subset extracted from METU Turkish Corpus (MTC) (Say et al., 2002). The dataset had been later extended by morpho-syntactic features by Pamay and Eryiğit (2018) using an automated Turkish NLP pipeline (Eryiğit, 2014). This dataset does not contain annotations for dropped pronouns.

3 Representation of Dropped Pronouns on Existing Data Sets

The CR literature has annotated datasets supporting various languages: MUC (Hirschman and Chinchor, 1998; Chinchor, 2001; Chinchor and Sundheim, 2003), ACE (Doddington et al., 2004), SemEval2010 (Recasens et al., 2010), OntoNotes (Pradhan et al., 2007, 2012b). The MUC covers coreference relation only for English which is not a pro-drop language. The ACE focuses on only seven pre-defined type entities, therefore, dropped pronouns were excluded in the annotation process for Arabic (a pro-drop language). Although the SemEval2010 includes pro-drop languages (e.g. Catalan, Spanish (Recasens and Martí, 2010)), dropped pronouns were not covered during the annotation. Compared with these datasets, the OntoNotes is more comprehensive and contains gold-standard coreferential relations of dropped pronouns for Chinese and Arabic.

In the OntoNotes, dropped pronouns are represented by a unique artificial token: ("`*pro*`" for Chinese and "`*`" for Arabic), which is inserted into the correct position where the subject or object is omitted in a sentence during the annotation. This token indicates that a pronoun has been dropped from this location in the sentence. Example 1 shows how a dropped pronoun is represented for Chinese.

> (Zh.) 吉林省主管经贸工作的副省长全哲洙说："**(*pro*)** 欢迎国际社会同 **(**我们**)** 一道，共同推进图门江开发事业，促进区域经济发展，造福东北亚人民。"
> (En.) *Quan Zhezhu, Vice Governor of Jinlin Province who is in charge of economics and trade, said: "**(*pro*)** Welcome international societies to join **(us)** in the development of Tumen Jiang, so as to promote regional economic development and benefit people in Northeast Asia".*

Example 1: Annotation of dropped pronouns with artificial **(*pro*)** token in Chinese (Pradhan et al., 2012b)

Chinese and English translations of the same sentence are shown in the Example 1. In Chinese, `*pro*` is inserted for an omitted subject pronoun, which is referential with another pronoun: (我们) ('us' in English).

A recent study (Nedoluzhko et al., 2022) proposed a similar representation scheme as in the OntoNotes, built on top of the CoNLL-UD framework (Nivre et al., 2016, 2017), called the CorefUD. Dropped pronouns are represented by inserted tokens, called empty nodes (i.e. zeros), and they are related to their syntactic heads (hereinafter referred to as 'owner') by dependency relations. The CorefUD introduces how the inserted tokens should be represented (with a sub-indexed token number i.e. <tokenID>.<subIndex>); however, there is no standard on where to add them across different languages. In Hungarian, they are added immediately after their owners in the sentence (with some minor exceptions due to punctuations). In Czech, Spanish and Catalan, there is no strict rule about their positions except that empty nodes are almost always placed before their owners. The decisions about their positions seem to be affected by the fact that they will have an influence on the dependency trees of the related language.

Besides the explained representation above, Iida and Poesio (2011) introduced another approach and applied it on the Italian CR dataset. Italian is a partial PD-MRL, allowing only omitted subjects, called null-subjects. In this approach, instead of an artificially inserting token, dropped subject

pronouns are directly annotated on the verbs. Example 2 shows how a dropped subject pronoun is represented for Italian.

> (It.) **(Pahor)** è nato a Trieste, allora porto principale dell'Impero Austro-Ungarico. A sette anni **(vide)** l'incendio del Narodni dom.
> (En.) ***(Pahor)*** *was born in Trieste, then the main port of the Austro-Hungarian Empire. At the age of seven* ***(he)*** *saw the fire of the Narodni dom.*

Example 2: Annotation of dropped subject pronoun in Italian (Iida and Poesio, 2011)

Italian and English translations of the same sentence are shown in Example 2. In this approach, each verb is considered as a potential coreferential mention. Mentions existing before this verbal mention in a text are considered antecedents of the verb. In the example, the predicate of the second Italian sentence, `vide`, has a coreference relation with `Pahor` in the first sentence. Since English is not a pro-drop language, the subject of the second sentence, `he`, is explicitly defined, and the coreference relation are made between `he` and `Pahor`. Slovenian CR dataset (Klemen and Žitnik, 2021) used the same representation approach for null-subjects as in Italian.

Both the OntoNotes and CorefUD approaches propose inserting new tokens to represent dropped pronouns, but from different perspectives. In the OntoNotes, all type of dropped pronouns are represented with the same artificial token which could be easily adapted to various languages. However, each dropped pronoun is represented with the same surface form creating ambiguity for automated CR systems. On the other hand, the CorefUD proposes inserting an empty token according to pronominal information at the morphology level, not a unique token for all dropped pronouns; however, it requires extra coding of the already available information easily deducible from the owner's morphology information. These approaches harm the original sentence flow, reduce human readability and also cause an extra burden to the annotation process from the perspective of determining the most accurate and natural position of the these newly inserted tokens in the sentence. However, they both allow direct use of existing evaluation tools, which may be considered as an advantage of these approaches.

Moreover, the Universal Dependencies (UD)[6] (Nivre et al., 2016, 2017) initiative suggests to reduce the use of additional artificial tokens (i.e., inflectional groups) even in case of derivational suffixes/cases requiring a new sub-token group. However, the above-mentioned approaches propose inserting extra nodes based on morphological suffixes, which may be treated as contradictory.

Using existing tokens to represent coreferential relations of dropped pronouns overcomes these drawbacks. However, in extreme PD-MRLs, dropping may occur in cases other than null-subjects; e.g., dropped possessive pronouns. The morphological richness in these languages may reveal the appearance of multiple coreference relations on a single token; e.g., a nominal as exemplified in the introduction section. This is a barrier in front of using existing evaluation tools for such kind of representations.

4 The Proposed Scheme

This section introduces our representation and evaluation scheme and its validation.

4.1 Dataset Representation

Morphologically rich languages allow nouns and verbs to contain pronominal markers in their morphological analyses. A pronominal marker may be a possessive marker for nouns or a personal marker for verbs. These markers carry information about the related person who did the action (or was affected by the action passively) or specify the properties of a pronominal possessor of a noun/noun phrase. In PD-MRLs, information about the omitted pronouns can be reached by these markers. The proposed scheme considers the pronominal markers in existing nouns/verbs as a coreferential mention and allows a coreferential relation between these markers and other mentions of the same entity. Example 3 shows coreferential relations between pronominal markers and mentions for a sample Turkish sentence with its English translation. Please refer to Figure 1 for the literal translation.

> (Tr.) **Ahmet** bugün yeni **okulunda** öğretmenliğe **başladı**. **Okulunu** çok **sevmiş**.
> (En.) *Ahmet started teaching at his new school today. He liked his school very much.*

Example 3: Representation of dropped pronouns in Turkish.

The nominal word, `okulunda`, has a morphological analysis as `okul+Noun+A3sg+P3sg+Loc` with

[6] UD aims to create a common framework for annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages.

#sntNo: 00002213_102

1	**Ahmet**	*Ahmet*	Ahmet	Noun	Prop	A3sg|Pnon|Nom	6	SUBJECT	(50)
2	bugün	*today*	bugün	Noun	Noun	A3sg|Pnon|Nom	6	MODIFIER	
3	**yeni**	*new*	yeni	Adj	Adj	_	4	MODIFIER	(17
4	**okulunda**	*at his school*	okul	Noun	Noun	A3sg|**P3sg**|Loc	6	MODIFIER	(50{P3sg}) | 17)
5	öğretmenliğe	*teaching*	öğretmenlik	Noun	Noun	A3sg|Pnon|Dat	6	MODIFIER	
6	başladı	*started*	başla	Verb	Verb	Pos|Past|**A3sg**	0	PREDICATE	(50{A3sg})
7	.		.	Punc	Punc	_	6	PUNCTUATION	

#sntNo: 00002213_103

1	**Okulunu**	*his school*	okul	Noun	Noun	A3sg|**P3sg**|Acc	3	OBJECT	(50{P3sg}) | (17)
2	çok	*very much*	çok	Adverb	Adverb	_	3	DETERMINER	
3	sevmiş	*liked*	sev	Verb	Verb	Pos|Narr|**A3sg**	0	PREDICATE	(50{A3sg})
4	.		.	Punc	Punc	_	3	PUNCTUATION	

Figure 1: Annotated CoNLL dataset sample

a possessive marker, `P3sg`. This suffix shows that a third singular person possessor, 'onun' (*his*), modifies the word. The pronoun is not explicitly defined in the context; that is a dropped pronoun. Therefore, the coreferential relation of the dropped possessive pronoun is annotated on an existing token, `okulunda`. The predicates of both sentences, 'başladı' (*start*) and 'sevmiş' (*like*) are coreferential mentions due to the personal markers deducible from their morphological analyses. These personal markers refer to the same person, 'Ahmet'. In the first sentence, the person who started teaching can be directly obtained from the syntactic analysis of the sentence. However, the second sentence does not contain an overt-subject. The predicate, 'sevmiş' (*like'*), carries a personal marker in its morphological analysis, `A3sg`. This suffix shows that a third singular person, 'o' (*he*), is the subject of this verb. The coreferential relations between the person and his personal markers are annotated on existing verbal tokens. Additionally, the word 'okulunda' (*at his school*) has two coreferential mentions: the possessive marker ('-u' holds for the pronoun 'onun' (*his*)) and the word 'okulunda', which is a mention itself. According to the proposed scheme, verbs are considered potential coreferential mentions due to the pronoun markers in their morphological analyses[7]; and possessive markers in nouns are also regarded as coreferential mentions besides the noun itself.

Figure 1 shows how coreferential relations are represented on top of the base CoNLL format for a Turkish sample. In the base CoNLL format, coreference annotations are given in the last column. Each sentence is labeled by a unique identifier containing the document and sentence number (#sntNo). Coreferential mentions are annotated by their numerical cluster identifiers, and this number is encapsulated by an opened and a closed parenthesis symbol to specify the initial and final words of a mention span. Mentions referring to the same real-world entity are labeled with the same cluster number. In Figure 1, 'Ahmet' is a coreferential mention parenthesized by the cluster number 50. Another mention 'yeni okulunda' is a bi-token mention with a cluster id 17. While the parenthesis is opened with cluster 17 for the first token, it is closed with the same number for the last token to mark the mention's border. As may be seen from the figure, relations are inter-sentential. For example, the mention 'yeni okulunda' in the first sentence has a coreference annotation with the cluster 17, and its referent, 'Okulunu', which is in the second sentence, is also annotated with the same cluster number.

The base CoNLL format assumes and describes one coreference annotation per token; however, as described in previous sections, a nominal token may contain multiple coreference relations. Therefore, in the proposed scheme, additional coreferential relations coming from dropped pronouns are annotated with the help of curly brackets including pronominal markers' information. In this way, pronominal markers existing in nominal and verbal tokens are annotated as a coreferential mention rather than adding a new token for each dropped pronoun. With this representation, the dependency tree of the actual sentence is not affected as in the newly inserted token approach.

In Figure 1, the predicate 'başladı' in the first sentence contains the third singular personal marker, `A3sg`, in its morphological analysis. This marker is annotated as a mention with cluster number 50. The marker and the person 'Ahmet' are coreferential within the same cluster. Similarly, in the second sentence, the possessive marker of the first

[7] On the contrary for null-subjects, it is not a tradition to produce morphological markers for the null-object cases (Nivre et al., 2016, 2017). However, the proposed scheme is also applicable to null-objects when needed.

token, 'Okulunu', {P3sg}, also exists in the same cluster, 50. Moreover, the first token of the second sentence contains multiple annotations separated by the pipe symbol. The first annotation stands for the coreferential relation of its possessive marker, whereas the second annotation with cluster number 17 shows the relation of the word itself.

4.2 Adaptation of Evaluators

Although in practice, the widely-used CoNLL coreference scorer accepts multiple coreference annotations per token in its input, it is reported[5] that this situation is only limitedly supported. Table 1 exemplifies this situation on some randomly selected documents having a diverse number of multiple annotations; the number of tokens having multiple coreference annotations is reported in the last column of the table. The table provides the drop in evaluation scores on gold-standard data where the key and the predicted inputs are exactly the same; in other words, we expect 100% F-Scores on all metrics. However, as it can be seen from the table, the performances are dropped as far as the number of tokens with multiple annotations increases; e.g. 9.20 percentage point drop in MUC score for the last document having 98 tokens with multiple annotations.

	MUC	B-Cubed	CEAF$_e$	#Tokens w MultAnn.
D#1	↓0.82	↓0.79	↓0.47	4
D#2	↓1.25	↓1.34	↓0.79	8
D#3	↓3.09	↓3.31	↓1.07	16
D#4	↓3.56	↓3.20	↓1.83	20
D#5	↓6.62	↓7.99	↓3.03	42
D#6	↓9.20	↓14.51	↓6.08	98

Table 1: Performance drops reported by the CoNLL scorer on documents having multiple mentions per token.

A solution to the above-described problem is to automatically create temporary tokens for dropped pronouns on the backstage, to use the scorer, and finally to remove these temporary tokens. In this manner, the scorer will not encounter problems evaluating the relations of dropped pronouns. That is, the need of artificial tokens introduced in Section 3 for dropped pronouns are handled at the software level rather than the human-annotation level, which eliminates the deficiencies listed in the same section. The proposed pre-processor[4] copies a token having multiple annotations, as many as the number of its annotations caused by pronominal markers. Then, these duplicated tokens are concatenated to the end of the sentence. After preprocessing, each copy token carries only one annotation related to a pronominal marker, whereas these relations are removed from the original token. We use the syntactic head identifier field (the seventh column in the CoNLL format in Figure 1) to keep the links between the original and temporary tokens, and use this information to aggregate everything during the post-processor stage.

4.3 Validation

We validated our proposed scheme on the Turkish language which is a strong representative of PD-MRLs. As the first step, using the proposed dataset representation, we reannotated a Turkish CR dataset (MTCC[8]) to include the dropped pronouns which were not available in the original annotations.

	MTCC	ITCC
# Documents	24	24
# Paragraphs	1564	1562
# Sentences	4744	4732
# Tokens	60788	60772
# Overt Mentions	3696	10031
# Dropped Pronouns	n/a	11584
# Total Mentions	3696	21615
# Mention Clusters	691	4065
# Multiple Annotations/doc	n/a	21.3$_{\pm 23.5}$

Table 2: Dataset statistics.

Table 2 provides statistics about the original MTCC and its extended version (referred as ITCC[4] from now on). Sentences containing only punctuations are removed from ITCC. As seen from the table, the number of dropped pronouns annotated in ITCC is 11584, which resulted in the need for the annotation of 6335 additional overt mentions and the creation of 3374 new mention clusters. An example to this may be as the following: when we annotate the dropped pronoun 'onun' (*its*) on the word 'rengi' (*its color*), we also need to annotate its referent overt mention (e.g., the cat) within the text although it had not been annotated initially due to some decisions about neglecting singletons[9]. ITCC includes 21615 mentions in total collected

[8] MTCC from Pamay and Eryiğit (2018) comes with automatically produced morphological and syntactic analyses in the CoNLL format.

[9] Singleton in CR is the situation where there appears only a single mention within a mention cluster; i.e., a mention with no coreferential antecedent.

under 4065 clusters and contains $21.3_{\pm 23.5}$ multiple coreference annotations on average with a high standard deviation. While 11 documents have less than 10 multiple annotations, this number goes up to 98 among the remaining 13 documents. Table 3 shows the distribution of referential pronominal markers in ITCC. The personal marker 'A3sg' (the third singular person, 'o' *(s/he/it)*) is the most frequent one. Similarly, possessive marker 'P3sg' (the third singular possessor, 'onun' *(his/her/its)*) has the highest distribution percentage among all types of possessive markers. One should note that there is no gender in Turkish morphology; thus, s/he/it pronouns all appear under the same surface form, which yields higher complexity in their coreference resolution.

Personal Marker					
A1sg	A2sg	A3sg	A1pl	A2pl	A3pl
815	262	3846	313	207	303
Possessive Marker					
P1sg	P2sg	P3sg	P1pl	P2pl	P3pl
499	124	4595	214	80	326

Table 3: Distribution of referential pronominal markers.

As the second step, we validate that the introduced evaluation components eliminate the errors coming from multiple annotations. The stand-alone CoNLL scorer reports the following performances on the gold-standard ITCC (as introduced in Section 4.2): MUC=96.99%, B-Cubed=96.85%, and $CEAF_e$=97.84% F-scores on average. After the introduced pre and post-processors are used together with the CoNLL scorer, the expected 100% F-Scores on all metrics are successfully obtained on the gold-standard key and predicted inputs.

5 Experiments & Results

This section introduces the first neural Turkish coreference resolution results which provides a strong baseline for future studies in the field.

5.1 Experimental setup

The neural Turkish CR performances are reported using a neural coreference resolution architecture (Klemen and Žitnik, 2021) which was introduced for Slovene, another PD-MRL. The model uses a mention-ranking approach and resolves coreferential relations on gold-standard mentions. The replicated model consists of three sequential, fully connected layers with ReLU as an activation function. The model takes a mention-pair ($mention_1$ (i.e., a head mention) and $mention_2$ (i.e., an antecedent of the head mention) as input and produces a score about how well these mentions are coreferential. Mentions and their antecedents are paired to create positive (coreferential) or negative (non-coreferential) samples. The order of mentions' occurrence in a document is also considered during pairing. A mention is paired with its antecedents that are at most 50 mention-away[10]. During inference, the model generates a score for each antecedent and the most probable one is selected as the model's prediction.

The model may utilize either word embeddings (word2vec[11] and fastText[12]) or contextual neural language models (ELMo[11] and BERT[13]). In addition to dense representations, we also extended the replicated model by including hand-crafted features used in previous Turkish studies (Schüller et al., 2017; Pamay and Eryiğit, 2018) to analyze their representation power for morphological richness. A mention is considered as a sequence of tokens so that its embedding is created from its words' embeddings. A mention embedding contains three parts: the initial token's embedding, the final token's embedding, and the weighted average of all its tokens' embeddings. The averaging step allows the model to learn the most essential token in the mention (i.e., the head token in the mention) with an intermediate fully connected layer, which may be assumed as an attention mechanism. As a result, the produced mention embedding comprises information about the head token in the mention and its right and left contexts. This paper replicates the neural CR model with the default hyper-parameters from Klemen and Žitnik (2021). Documents are split into train/validation/test parts by considering their genres. Documents having common genres are used in validation and test datasets separately. While the development set has 2 documents from news and novel, the test set contains 3 documents from news, novel, and story genres. The rest 19 documents are selected as the training dataset. The model is evaluated on four coreference metrics: MUC (Vilain et al., 1995), B-Cubed (Bagga and Baldwin, 1998), entity-based CEAF (Luo, 2005) and average CoNLL. The enhanced coreference

[10]The average mention-distance between referential mentions in a chain is 50,3 in ITCC

[11]http:/vectors.nlpl.eu/repository/

[12]http:/fasttext.cc/docs/en/crawl-vectors.html

[13]http:/huggingface.co/dbmdz/bert-base-turkish-cased

	MUC			B-Cubed			$CEAF_e$			CoNLL		
	P	R	F	P	R	F	P	R	F	P	R	F
word2vec	45.26	16.34	23.91	**80.03**	18.76	30.17	11.22	51.62	18.29	45.50	28.91	24.12
fastText	52.66	37.62	43.86	56.75	26.17	35.70	18.60	48.73	26.77	42.67	37.51	35.44
ELMo	54.19	30.76	39.02	70.29	22.50	33.84	16.31	56.06	25.08	46.93	36.44	32.71
BERT	**64.77**	**53.10**	**58.34**	56.88	**31.00**	**39.89**	**28.26**	**56.70**	**37.62**	**49.97**	**46.93**	**45.28**

Table 4: The neural CR results on ITCC with different neural language models.

		MUC	B-Cubed	$CEAF_e$	CoNLL
word2vec	None	23,91	30,17	18,29	24,12
	+feats	57,14	41,18	36,33	44,88
	Diff	↑ 33,23	↑ 11,01	↑ 18,04	↑20,76
fastText	None	43,86	35,70	26,77	35,44
	+feats	**63,80**	**45,12**	**41,15**	**50,02**
	Diff	↑ 19,94	↑ 9,42	↑ 14,38	↑ 14,58
ELMo	None	39,20	33,84	25,08	32,71
	+feats	46,80	34,56	29,37	36,91
	Diff	↑ 7,60	↑ 0,72	↑ 4,29	↑ 4,20
BERT	None	58,34	39,89	37,62	45,28
	+feats	58,22	40,06	38,56	45,61
	Diff	↓ 0,12	↑ 0,17	↑ 0,94	↑ 0,33

Table 5: The impact of the hand-crafted features on the CR models with various word embeddings.

scorer (The CoNLL-2012 Scorer[5] + our pre-post processors[4]) is used to evaluate the model.

5.2 Experimental results

The preliminary results with different word embeddings and language models are given in Table 4. We observed that the gap between precision and recall values decrease with the use of contextual models (i.e. ELMo and BERT). fastText performs better than word2vec which is expected for an MRL. The highest F-scores on all metrics are obtained with the pre-trained BERT language model. The base Turkish CR model provides a 45.28% average CoNLL F-score with BERT.

The Turkish CR model is also enhanced with hand-crafted morpho-syntactic and lexical features, and the results are presented in Table 5 in terms of F-scores for all metrics. External linguistic features predicted by Turkish NLP Pipeline[14] are integrated as a one-hot vector to mentions' representations. The table's 'Diff' raw indicates whether the external features positively or negatively impact each model. 'None' indicates that no external features are utilized, whereas the '+feats' setting benefits from features as in Pamay and Eryiğit (2018). The results show that although incorporating hand-crafted linguistic features into the CR neural model improves performances on all scenarios; its impact is higher with less-powerful word embeddings. Using the external linguistic information increases the CoNLL F-score by 20.76, 14.58, 4.2, and 0.33 percentage points for word2vec, fastText, ELMo, and BERT, respectively.

The Turkish CR model performance is increased to 50.2% average CoNLL F-score with fastText by incorporating hand-crafted linguistic features. This model performs the best over all others and provides a 5 percentage points improvement over BERT. Adding external linguistic features does not improve the performance considerably for BERT. A similar conclusion was also obtained by adding morphological information into BERT- and LSTM-based downstream tasks on several languages: Named Entity Recognition (NER) and Dependency Parsing (DP). As introduced in Klemen et al. (2022), the features help the LSTM-based models perform better on the NER and DP tasks. However, for BERT-based models, the additional morphological features only positively impact DP performance when they are gold-standard but not when they are predicted.

6 Conclusion

The paper proposed a language-independent representation and evaluation scheme to incorporate dropped pronouns into coreference resolution for pro-drop and morphologically rich languages. Pre and post-processors to enhance available CR evaluators to cover dropped pronouns (i.e., multiple annotations over a single word) are developed. The scheme was validated on the Turkish language. The study revealed the first Turkish CR dataset including the annotations for dropped pronouns and the first neural CR results for this language as a strong baseline for future studies. The impact of interaction between different text encodings and linguistic features were investigated on this task. The best performance was achieved by using fastText embeddings together with hand-crafted linguistic features with 50.2% CoNLL F-Score, which provides a 5 percentage points improvement over a BERT baseline.

[14] http:/tools.nlp.itu.edu.tr/

Limitations

The main limitation of the study is that the proposed representation scheme was validated on Turkish for now. This limitation may affect the reliability of the proposed scheme. However, it is foreseen that theoretically, the proposed representation scheme can be applied to other PD-MRLs. The proposed dataset is built on top of the widely used CoNLL format to which most datasets can be converted smoothly, and all necessary morpho-syntactic information for the conversion is already available in the CoNLL format.

A neural CR model (originally developed for Slovene) was chosen to validate the proposed scheme in Turkish due to the similar pro-dropping structure of these PD-MRLs. However, the Turkish dataset contains more complex mention spans: 1) wider types of dropped pronouns (null-subjects and also elided possessive pronouns), and 2) longer coreferential spans due to the chain of noun phrases and adjectival clauses. Even if our study introduced a strong baseline, we did not examine whether another neural, more powerful CR architecture would provide higher performance on Turkish.

Beyond the listed limitations, this paper analyzed and compared available representation schemes of dropped pronouns in the literature and introduced an easily applicable one by solving their deficiencies. The Turkish, a highly complex PD-MRL, was chosen for the validation to emphasize the importance of incorporating dropped pronouns into CR systems. Despite the increasing popularity and power of neural networks for NLP from scratch, this paper showed that employing hand-crafted linguistic features in a neural model still provides improvement for morphologically rich languages. As future work, we plan to expand the neural CR architecture with a mention prediction stage to resolve coreferential relations of automatically predicted mentions and explore the ways of improving the success on the resolution of dropped pronouns. The other future directions would be applying the proposed scheme to other pro-drop and morphologically rich languages and examining how the representation scheme affects the neural CR performance.

Ethics Statement

ITCC contains 24 documents from METU Turkish Corpus (Say et al., 2002). All necessary permissions have been obtained to use and distribute these documents.

Acknowledgements

This study was supported by İTÜ-Scientific Research Projects Coordination Unit with under project number #MDK-2022-43607#. Computing resources used in this work were provided by İTÜ Artificial Intelligence and Data Science Application and Research Center. The authors thank Prof. Deniz Zeyrek Bozşahin for allowing us to use and share subset of documents of METU Turkish Corpus (Say et al., 2002).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566. Granada, Spain.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2013. Chinese zero pronoun resolution: Some recent advances. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1360–1365.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Nancy Chinchor. 2001. Message understanding conference (muc) 7. LDC2001T02.

Nancy Chinchor and Beth Sundheim. 2003. Message understanding conference (muc) 6. LDC2003T13.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. 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 1405–1415, Beijing, China. Association for Computational Linguistics.

Aron Culotta, Michael L Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 81–88.

Pascal Denis and Jason Baldridge. 2007a. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243.

Pascal Denis and Jason Baldridge. 2007b. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1588–1593.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Gülşen Eryiğit. 2014. ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden. Association for Computational Linguistics.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Raffaele Guarasci, Aniello Minutolo, Emanuele Damiano, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Electra for neural coreference resolution in italian. *IEEE Access*, 9:115643–115654.

Iryna Haponchyk and Alessandro Moschitti. 2017. A practical perspective on latent structured prediction for coreference resolution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics(EACL)*, pages 143–149, Valencia, Spain,.

Lynette Hirschman and Nancy Chinchor. 1998. Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

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.

Heng Ji, David Westbrook, and Ralph Grishman. 2005. Using semantic relations to refine coreference decisions. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 17–24. 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.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Yılmaz Kılıçaslan, Edip Serdar Güner, and Savaş Yıldırım. 2009. Learning-based pronoun resolution for Turkish with a comparative evaluation. *Computer Speech & Language*, 23(3):311–331.

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2022. Enhancing deep neural networks with morphological information. *Natural Language Engineering*, page 1–26.

Matej Klemen and Slavko Žitnik. 2021. Neural coreference resolution for slovene language. *Computer Science and Information Systems*, (00):60–60.

Fang Kong and Hwee Tou Ng. 2013. Exploiting zero pronouns to improve chinese coreference resolution. In *Proceedings of the 2013 conference on empirical*

methods in natural language processing, pages 278–288.

Dilek Küçük and Meltem Turhan Yöndem. 2015. A knowledge-poor pronoun resolution system for Turkish. *arXiv preprint arXiv:1504.04751*.

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.

Lu Liu, Zhenqiao Song, and Xiaoqing Zheng. 2020. Improving coreference resolution by leveraging entity-centric features with graph neural networks and second-order inference. *arXiv preprint arXiv:2009.04639*.

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.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Andrew McCallum and Ben Wellner. 2005. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems*, pages 905–912.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdenek Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of LREC*.

Vincent Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 157–164. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on association for computational linguistics*, pages 104–111. Association for Computational Linguistics.

Cristina Nicolae and Gabriel Nicolae. 2006. Bestcut: A graph algorithm for coreference resolution. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 275–283. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Tuğba Pamay and Gülşen Eryiğit. 2018. Turkish coreference resolution. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7. IEEE.

Cheoneum Park, Jamin Shin, Sungjoon Park, Joonho Lim, and Changki Lee. 2020. Fast end-to-end coreference resolution for korean. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2610–2624.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012a. 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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012b. 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.

Sameer S Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526. IEEE.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pages 968–977.

Altaf Rahman and Vincent Ng. 2011. Ensemble based coreference resolution. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1884–1889.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

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.

Marta Recasens and M Antònia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language resources and evaluation*, 44(4):315–345.

Kepa Joseba Rodrıguez, 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 LREC*, pages 157–163.

Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. 2002. Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the 11th International Conference of Turkish Linguistics*, pages 183–192, Northern Cyprus.

Peter Schüller, Kübra Cıngıllı, Ferit Tunçer, Barış Gün Sürmeli, Ayşegül Pekel, Ayşe Hande Karatay, and Hacer Ezgi Karakaş. 2017. Marmara Turkish coreference corpus and coreference resolution baseline. *arXiv preprint arXiv:1706.01863*.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52. Association for Computational Linguistics.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.

Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1416–1426.

Liyan Xu and Jinho D Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. *arXiv preprint arXiv:2009.12013*.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 41–48. Association for Computational Linguistics.

Savaş Yıldırım and Yılmaz Kılıçaslan. 2007. A machine learning approach to personal pronoun resolution in Turkish. In *Proceedings of the American Association for Artificial Intelligence*, pages 269–270.

Savaş Yıldırım, Yılmaz Kılıçaslan, and Tuğba Yıldız. 2007. Pronoun resolution in Turkish using decision tree and rule-based learning algorithms. In *Language and Technology Conference*, pages 270–278. Springer.

Qingyu Yin, Weinan Zhang, Yu Zhang, and Ting Liu. 2016. A deep neural network for chinese zero pronoun resolution. *arXiv preprint arXiv:1604.05800*.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 541–550.