Morpheme-level Coreference Annotations for Pro-dropped Languages

Tuğba Pamay Arslan and **Gülşen Eryiğit**
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 (PDMRLs) due to dropped pronouns (e.g., null subjects and omitted possessive pronouns). Although there exist representations relying on the insertion and annotation of artificial tokens, the absence of any standard for where these tokens should be placed leads to various shortcomings. In this paper, we introduce an alternative representation at the morphology-level incorporating dropped pronouns into coreference resolution. The scheme is language-independent and may be applied to several PDMRLs, which we believe will promote diversity and universality in multilingual language technologies. The proposal is directly related to WG3 since it will serve as an example for the data conversion phase of the forthcoming WG3 evaluation campaign on morpho-syntactic analysis-parsing.

Relevant UniDive working groups: WG3, WG1

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 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. The mention clustering stage collects mentions referring to the same real-world entity under the same cluster, resolving which extracted mentions are coreferential. These steps could be trained sequentially or jointly.

Although CR is an NLP subject that has been studied for quite a long time (Ng and Cardie, 2002; Straka and Straková, 2022), studies on PDMRLs are still in their infancy. 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 is provided below[1], where verbal agreement and possessive suffixes allow the drop of personal and possessive pronouns. Morphemes emphasized with bold font describe the dropped pronouns: '-m' holds for the pronoun 'benim' (*my*) and '-n' holds for 'sen' (*you*). Moreover, the morphological richness in these languages may reveal the appearance of multiple coreference relations on a single token which is also illustrated below. The word 'annemin' (*of my mother*) in the below example carries multiple coreferential relations[2] to different people: *me* and *my mother*.

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?

In the CR literature, various annotation schemes exist to represent coreferential dropped-pronouns, and may be listed under two categories: 1) annotation on other tokens i.e., artificially inserted (Pradhan et al., 2012; Nedoluzhko et al., 2022) or 2) annotation on existing tokens (Rodrıguez et al., 2010; Klemen and Žitnik, 2021) other than the dropped pronouns, such as verbs carrying personal suffixes. The second approach is constrained, permitting only one coreferential annotation per token, that is, it does not facilitate the annotation of the word 'annemin' in the provided example. Moreover, representations relying on artificially inserted tokens have their deficiencies, although eliminating the multiple coreference issue, such as extra coding of the already available information easily deducible from morphology, difficulty in determining the most accurate and natural position of the artificial token in the sentence, and corruption of the original sentence flow.

This paper describes a recently published work and aims to introduce the morpheme-level repre-

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

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

#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

sentation scheme for coreference resolution task to the UniDive community as a means of supporting inter-language diversity. The published paper showed the validation of the proposed scheme on Turkish coreference resolution, and introduced an updated version of the CoNLL evaluator (Pradhan et al., 2014) supporting the scheme.

2 Morpheme-level Annotation

Morphologically rich languages allow nouns and verbs to contain pronominal markers in their morphological analyses: possessive markers for nouns, and personal markers 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. Figure 1 shows how coreferential relations between pronominal markers and mentions are annotated on top of the base CoNLL format for a sample Turkish sentence with its English translation. In the base CoNLL format, coreference annotations are given in the last column. 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: 'yeni okulunda' and 'Okulunu', both are annotated with the same cluster number, 17.

The base CoNLL format assumes and describes one coreference annotation per token; however, as described in the previous section, 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 the figure, 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 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.

The proposed representation scheme benefits from morphemes carrying pronominal meaning and can be adapt to other PDMRLs. Since, the scheme is language-independent, it is regarded as an effective method for promoting diversity across languages.

References

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

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 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.

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. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

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.

Milan Straka and Jana Straková. 2022. Úfal corpipe at crac 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37.