Morpheme-level Coreference Annotations for Pro-dropped Languages

İTÜ Natural Language Processing Research Laboratory (İTÜ NLP)
Department of AI & Data Engineering, Istanbul Technical University, Istanbul, Turkey
Tuğba Pamay Arslan, Gülşen Eryiğit
[pamay, gulsen.cebiroglu]@itu.edu.tr

2nd General Meeting
University of Naples "L'Orientale Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

Scan to read the full work

Motivation

- *Related to WG1, WG3*
- 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) [4].
- In MRLs, most syntactic information is carried at the morphological level leading to the possibility of dropping pronouns.

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 ?

Existing Representation Schemes for Dropped Pronouns

- The pro-drop nature of such languages reveals the need for mention annotation on other tokens:

1. Artificially inserted tokens: OntoNotes, CorefUD [1]

(Cat.) El Juventus confirma que **(*zero*)** jugarà demà a Turquia .
*(En.) Juventus confirms that **(they)** will play in Turkey tomorrow.*

2. Existing tokens *other* than the dropped pronouns: Italian[2], Slovene

(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.*

- Representations relying on artificially inserted tokens have their deficiencies:

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

- No standard on where to add artificial tokens:

In Hungarian, added immediately after their syntactic head in the sentence (with some minor exceptions due to punctuations).

In Czech, *Spanish and* Catalan, no strict rule about their positions except that zeros are almost always placed before their heads.

The Proposed Representation Scheme

- Each **pronominal marker ~ coreferential mention**
- `Possessive marker' for nouns, and `Personal marker' for verbs.

1	**Ahmet**	*Ahmet*	LEMMA & POS	A3sg|Pnon|Nom	DEPS	(50)
2	bugün	today		A3sg|Pnon|Nom		_
3	**yeni**	new		_		(17
4	**okulunda**	at his school		A3sg|**P3sg**|Loc		(50{P3sg})|17)
5	öğretmenliğe	teaching		A3sg|Pnon|Dat		_
6	başladı	started		Pos|Past|**A3sg**		(50{A3sg})
7	.	.		_		_
1	**Okulunu**	his school		A3sg|**P3sg**|Acc		(50{P3sg})|(17)
2	çok	very much		_		_
3	sevmiş	liked		Pos|Narr|**A3sg**		(50{A3sg})
4	.	.		_		_

The Proposed Evaluation Scheme

- **Problem:**

Multiple coreference annotations per token is supported limitedly by CoNLL Scorer [3].

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

	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

- **Solution:**

Create temporary tokens automatically for dropped pronouns on the backstage in pre-processing. Then, remove these temporary tokens in post-processing.

1	**Okulunu**	his school	LEMMA	A3sg|P3sg|Acc	3 (17)
2	çok	very much		_	3 _
3	sevmiş	liked		Pos|Narr|**A3sg**	0 (50{A3sg})
4	.	.		_	3 _
X	**Okulunu**			A3sg|**P3sg**|Acc	**1** (50{P3sg})

Results

- The proposed scheme was validated on Turkish.
- **None**: The first neural Turkish CR results [4].
- **+feats**: The impact of the hand-crafted features on the CR models.
- *Diff*: Whether the external features impact each model positively or negatively

		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

References
[1] A. Nedoluzhko, M. Novák, M. Popel, Z. Žabokrtskỳ, A. Zeldes, and D. Zeman. Corefud 1.0: Coreference meets universal dependencies. In Proceedings of LREC, 2022.
[2] K. J. Rodrıguez, F. Delogu, Y. Versley, E. W. Stemle, and M. Poesio. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In Proceedings of LREC, pages 157–163, 2010.
[3] S. Pradhan, X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In Proceedings of the 52nd Annual Meeting of ACL, pages 30–35, 2014.
[4] Pamay Arslan, T., & Eryiğit, G. Incorporating Dropped Pronouns into Coreference Resolution: The case for Turkish. In Proceedings of the 17th Conference of EACL:SRW (pp. 14-25), 2023.

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the European Union