Research Article.
Journal of Intelligent Systems: Theory and Applications 6(1) 2023: 85-95
DOI: 10.38016/jista.1225097

Neural Coreference Resolution for Turkish

Şeniz Demir[1*]

[1] MEF University, Department of Computer Engineering, İstanbul, Turkey
demirse@mef.edu.tr

Abstract
Coreference resolution deals with resolving mentions of the same underlying entity in a given text. This challenging task is an indispensable aspect of text understanding and has important applications in various language processing systems such as question answering and machine translation. Although a significant amount of studies is devoted to coreference resolution, the research on Turkish is scarce and mostly limited to pronoun resolution. To our best knowledge, this article presents the first neural Turkish coreference resolution study where two learning-based models are explored. Both models follow the mention-ranking approach while forming clusters of mentions. The first model uses a set of hand-crafted features whereas the second coreference model relies on embeddings learned from large-scale pre-trained language models for capturing similarities between a mention and its candidate antecedents. Several language models trained specifically for Turkish are used to obtain mention representations and their effectiveness is compared in conducted experiments using automatic metrics. We argue that the results of this study shed light on the possible contributions of neural architectures to Turkish coreference resolution.

Keywords: Turkish, coreference resolution, neural architectures, pre-trained language models.

Sinir Ağı ile Türkçe Eşgönderge Çözümleme

Öz
Eşgönderge çözümleme, bir metinde yer alan ve aynı temel varlığa gönderimde bulunan ifadelerin çözümlenmesiyle ilgilenir. Metin anlamanın vazgeçilmez bir unsuru olan bu zor iş, soru yanıtlama ve makine çevirisi gibi çeşitli dil işleme sistemlerinde önemli uygulamalara sahiptir. Eşgönderge çözümlemesine yönelik önemli sayıda çalışma olmasına rağmen, Türkçe üzerine yapılan araştırmalar sayıca azdır ve çoğunlukla zamir çözümlemesiyle sınırlı kalmıştır. Bildiğimiz kadarıyla, bu makale öğrenme tabanlı iki farklı modelin araştırıldığı ilk sinir ağı kullanılarak yürütülmüş Türkçe eşgönderge çözümleme çalışmasını sunmaktadır. Her iki model de ifade kümelerini oluştururken ifade sıralaması yaklaşımını takip etmektedir. İlk model, bir dizi önceden belirlenmiş özellikleri kullanırken, ikinci eşgönderge modeli, bir ifade ile onun aday öncül ifadeleri arasındaki benzerlikleri tespit için önceden eğitilmiş büyük ölçekli dil modellerinden öğrenilen kelime temsillerini kullanmaktadır. Türkçe için özel olarak eğitilmiş birçok dil modeli, kelime temsillerini elde etmek için kullanılmış ve bunları etkinlikleri yapılan deneylerde otomatik ölçütler kullanılarak karşılaştırılmıştır. Bu çalışma sonuçlarının, sinir ağı mimarilerinin Türkçe eşgönderge çözümlenmesine olası katkılarına ışık tuttuğu düşünülmektedir.

Anahtar Kelimeler: Türkçe, eşgönderge çözümleme, sinir ağı mimarileri, eğitilmiş dil modelleri.

1. Introduction

The information about real-world entities might be spread across multiple sentences in a document. It is essential to connect all mentions of the same entities and aggregate information related to these entities in order to fully understand the document. Coreference resolution is the task of identifying text spans that refer to the same entities and grouping these spans into coreference chains (clusters). The task has an impact on the performance of various natural language applications, including information extraction (Kriman and Heng, 2021), question answering (Bhattacharjee et al., 2020), and text summarization (Li et al., 2021; Steinberger et al., 2007).

Coreference resolution comprises mention detection and mention clustering (coreference resolving) subtasks that are often performed jointly. Given a document, the mention detection subtask is responsible for finding all text spans that refer to an entity and hence constitute a mention. Previous research has demonstrated that incorrect or missing identification of entity mentions negatively affected the accuracy of downstream

* Corresponding Author
E-mail: demirse@mef.edu.tr

Recieved : 27 Dec 2022
Revision : 26 Feb 2023
Accepted : 9 Mar 2023

coreference resolution. To address this task, previous research has utilized rule-based, statistical-based, and deep learning-based approaches (Lata et al., 2022). In rule-based solutions, a set of hand-crafted rules and knowledge resources are used where the rules require tremendous effort. The system developed by Sapena et al. (2010) identified noun phrases, named entities, and pronouns as mentions by relying on a set of rules that benefit from part-of-speech (POS), named entity, and syntactic information. In the work of Soraluze et al. (2012), a rule-based approach was developed for the Basque language using finite-state transducers. In statistical-based approaches, the detection task is defined as either a sequence labeling or a classification problem and the models trained on large-scale data are used for predictions. The work of Zitouni et al. (2005) addressed the task as a classification problem and used a maximum entropy Markov model classifier that utilizes various features (e.g., lexical, syntactic, and shallow parsing features) and information obtained from a gazetteer. Another detection solution based on a support vector machine classifier was implemented by Hacioglu et al. (2005). The classification was followed by a post-processing step to capture missing mentions. On the other hand, deep learning-based approaches hinder feature engineering and capture both syntactic and semantic features of candidate mentions via word embeddings. Different neural architectures were explored for mention detection such as the Bi-directional long short-term memory (BiLSTM) with conditional random field (CRF) (Park and Lee, 2015), the pointer network (Park et al., 2017), and the stacked LSTM enhanced with stack pointer (Wang et al., 2018).

An entity introduced by a mention might be referred multiple times later in a document. If the document contains multiple mentions of different entities, the mentions that are coreferent should be differentiated from those that are not and all identified mentions that refer to the same entity should be clustered in a group. Earlier studies have either performed mention detection and clustering subtasks jointly in an end-to-end fashion (Cai and Strube, 2010) or applied clustering to identified mentions (Durrett and Klein, 2013; Clark and Manning, 2015) or gold mentions provided in a dataset. Recent advancements in deep learning and transformer models motivated the development of several end-to-end solutions (Lai et al., 2022). The pioneering end-to-end solution by Lee et al. (2017) formed a contextual representation of each token in a text via a BiLSTM network that is fed by word and character embeddings. These representations were used to assign mention scores to candidate text spans and the top-scoring mentions were used as input in a feed-forward neural network (FFNN) to handle coreference resolution. Several extensions to this work have been developed since then including the use of a biaffine attention model in place of the original FFNN to compute antecedent scores (Zhang et al., 2018) and the incorporation of reinforcement learning where a reward function is used to measure the correctness of generated clusters (Fei et al., 2019).

Most coreference resolution studies have followed one of three main methods to make coreference decisions. In mention-pair approaches, the task is formulated as determining whether a given pair of mentions is coreferent or not (Soon et al., 2001; Hoste, 2016). Despite its simplicity (Uryupina and Moschitti, 2015), these approaches cannot compare all candidate antecedents for a mention at once in order to choose the most probable antecedent. On the other hand, independent decisions are eliminated and all candidate antecedents for a mention are jointly compared in mention-ranking approaches (Denis and Baldridge, 2008; Wiseman et al., 2015). The performance of these approaches was improved by the introduction of syntactic information to better prune the space of candidate antecedents and the incorporation of parse tree information (e.g., the siblings and degrees of tree nodes and the traversal node sequence) into mention span representations (Fang and Jian, 2019). The entity-mention approaches determine whether an entity represented by a possibly partially formed cluster of mentions is coreferent with a given mention (Luo et al., 2004; Yang et al., 2008). Links are built between a given mention and a discourse entity rather than a prior mention. Given that a mention might not capture adequate information about the entity it refers to, these approaches make linking decisions by utilizing the representations of clusters containing multiple mentions of the same entities. In the work of Wiseman et al. (2016), an RNN network was used to obtain a representation of a cluster (i.e., cluster level features) from the sequence of mentions that belong to it. Although cluster level features better capture entity based information, candidate clusters cannot be considered simultaneously in entity-mention approaches. Additionally, the strengths of different approaches are combined in hybrid solutions such as the cluster-ranking approaches where preceding clusters in the discourse are ranked for a given mention (Rahman and Ng, 2011). These solutions promote the expressiveness strength of entity-mention approaches and the comparison ability of mention-ranking approaches. Alternative coreference solutions have also evolved over time such as the solution where singleton clusters from mentions are initially built and larger clusters are incrementally formed by joining clusters that refer to the same entities (Bunescu, 2012; Clark and Manning, 2015), and the solution where ensemble resolvers are explored (Rahman and Ng, 2011).

There has been substantial coreference research on high resource languages. By contrast, a limited number of studies have been dedicated to low resource languages in the last decade and mainly neural network based approaches have been proposed in these studies. A mention-pair model where mention relations are learned via a convolutional neural network (CNN) was developed for Indonesian (Auliarachman and

Purwarianti, 2019). The model was equipped with a classifier in order to eliminate cases with singleton mentions being included in a cluster. Another mention-pair model that was developed for the Persian language extracted hand-crafted, embedding-based, and rich semantic features of mentions and used them as input to a fully connected neural network for coreference resolution (Sahlani et al., 2020). The adaptation of an English mention-ranking model (Lee et al., 2008) to Arabic was enhanced with performance-related improvements such as the heuristic-based preprocessing of words and the use of a separately trained mention detection approach (Aloraini et al., 2020). A Siamese network architecture and an extended feature set of mentions were used for Polish coreference resolution (Niton et al., 2018).

Coreference resolution on Turkish has received little attention. The proposed studies mostly addressed pronoun resolution such as the decision tree-based approach developed by Yıldırım et al. (2007) and the extension of that work with the use of four other learning-based approaches, namely naive Bayes, support vector machine, k-nearest neighbour, and voted perceptron (Kılıçaslan et al., 2009). In the literature, there is only one Turkish coreference resolution study (Pamay and Eryiğit, 2018). That work followed the mention-pair approach and coreference decisions were made by applying decision tree and support vector machine classifiers to some linguistic mention features.

Deep learning approaches have been shown to perform on par or considerably better than statistical-based approaches in several language processing applications. However, to our best knowledge, no deep learning-based coreference research has taken place for the morphologically-rich language Turkish. This article presents the first Turkish neural coreference resolution work where mention-ranking task is particularly addressed. Two different learning-based models are developed for ranking candidate antecedents of a mention according to a score and selecting the highest scoring antecedent. In the first model, a set of well-studied features by existing literature (Bengtson and Roth, 2008; Durrett and Klein, 2013; Wiseman et al., 2015) are extracted for a mention and its candidate antecedents and then fed to a single-layer feed-forward neural network as input. Our second model closely follows the mention ranking approach of the end-to-end coreference solution proposed by Lee et al. (2007) which was successfully applied to other languages including Arabic (Aloraini et al., 2020) and Slovenian (Klemen and Žitnik, 2022). The contextual representations of a mention and its candidate antecedent mentions are learned from pre-trained language models, and a probability distribution is obtained over all possible pairings of the mention with candidate antecedents using a two-layer feed-forward network. In order to obtain mention representations, several Turkish pre-trained language models with different architectures are explored including BERTurk and DistilBERTurk. In our work, the mention detection task is assumed to be done in advance and hence gold mentions provided by the publicly available Turkish coreference dataset are used. The model performances are evaluated using 10-fold cross-validation and compared based on automatic metrics MUC, B^3, and CEAF-e.

2. Methodology

For this study, we develop two neural models that follow the mention-ranking approach and analyze their performances on a Turkish coreference dataset. In both models, we pair an input mention ($Mention_k$) with all preceding candidate mentions in the same document ($Mention_1$, $Mention_2$, ..., $Mention_{k-1}$) and rank these identified mention pairs according to antecedent scores ($AS_{k,1}$, $AS_{k,2}$, ...$AS_{k,k-1}$). The antecedent score reflects the likelihood of two mentions being coreferent. The candidate mention in the pair with the highest score is selected as the correct antecedent of the input mention.

First Model:

In this model (Model_1), we first extract a set of features for each mention pair in order to assess the similarity between the contained mentions. Earlier approaches have utilized several features to capture similarities between mentions such as the features extracted directly from the text or the features learned from external sources (Sahlani et al., 2020). Here, we only utilized seven handcrafted features, with six of them being binary features as follows:

- *Sentence match:* whether the mentions appear in the same sentence or not
- *Number match:* whether the mentions agree in number (i.e., singular or plural) or not
- *String match:* whether mentions share the same lemma or not
- *Prefix match:* whether one mention is the prefix of the other mention or not
- *Suffix match:* whether one mention is the suffix of the other mention or not
- *Initial match:* whether one mention consists of the initials of the words in the other mention or not
- *Similarity:* the Jaro similarity of mentions

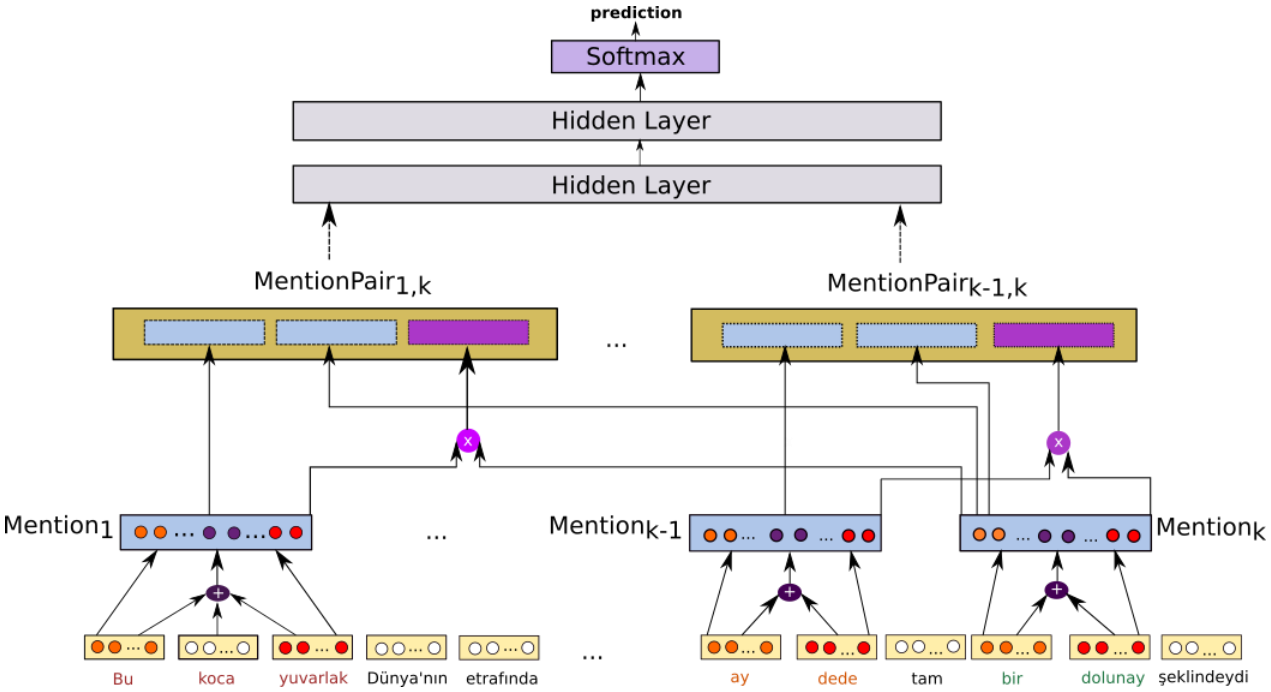


Figure 3. The architecture of the second model (Model_2)

In multi-word mentions, the number agreement is applied to the last tokens of mentions whereas the lemmas of all tokens are considered while determining the string match value. The Jaro similarity is a normalized edit distance score between 0 and 1. In our case, a Jaro score of 0 means no match between the surface forms of mentions whereas 1 means the mentions match exactly.

For each mention pair, the corresponding set of features is fed to a single-layer feed-forward neural network as input. The network applies a linear transformation to the input (x) and returns an antecedent score for candidate mention in the pair (y) by multiplying the input with a weight matrix (W) and adding a bias value (y=Wx+b). The candidate mentions are ranked according to their normalized scores and the candidate with the highest score is selected as the correct antecedent for the input mention.

Second Model:

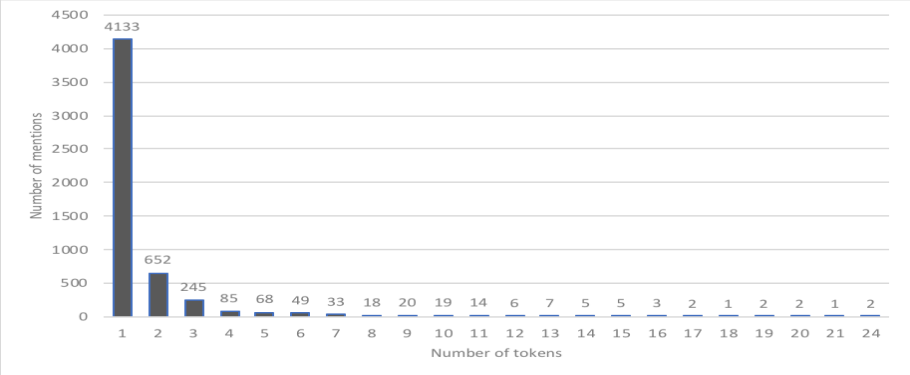
The second model (Model_2), addresses the problem by adapting a well-studied end-to-end coreference scoring model (Lee et al., 2017; Klemen and Žitnik, 2022). In this model, we obtain the representation of a mention by concatenating the embedding of the first token, the embedding of the last token, and the weighted combination of the embeddings of all tokens in the mention. The embedding of the first token is included in order to capture the left context of the mention whereas the embedding of the last token brings the right context into computation. The learned weighted combination of all token embeddings encodes the internal structure of a mention by using an attention mechanism that models its head token.

We form the representation of a mention pair by concatenating the representation of the first mention, the representation of the second mention, and element-wise multiplication of these representations. Mention pair representations are fed to a two-layer feed-forward neural network. The rectified linear activation function (RELU) is used for hidden layers. The softmax function is applied as the final activation function in order to obtain a probability distribution of all preceding candidate mentions. From among all candidates, the most probable mention is selected as the correct antecedent. The overall architecture of our second model is shown in Figure 3.

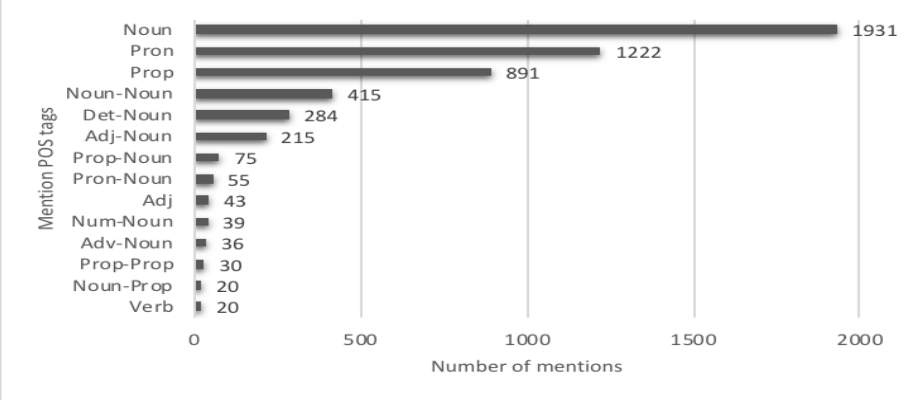
3. Experimental Setup

3.1. Dataset

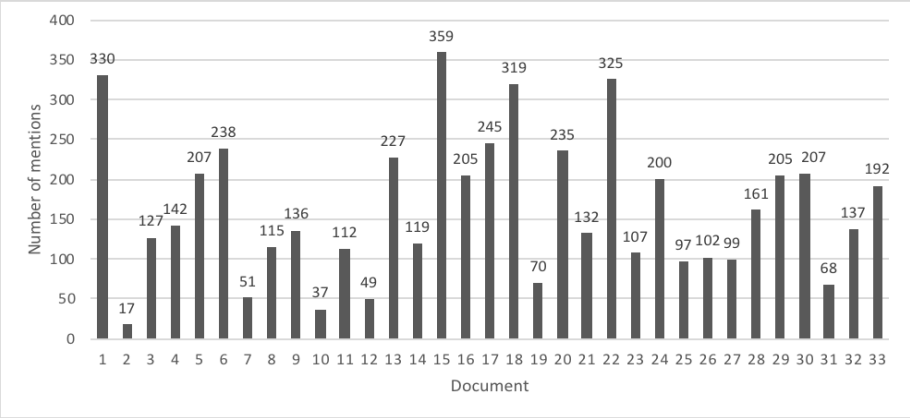
In this study, we used the publicly available Marmara Turkish Coreference Corpus (Schüller et al., 2007) as our dataset. The corpus contains 33 documents retrieved from the METU-Sabancı Turkish Treebank (Say et al., 2002) where each document consists of 26 to 424 sentences. The sentences are manually annotated with mentions and coreference chains that these mentions form. A textual expression (e.g., noun phrase, pronoun, or a nominalized adjective) that refers to an entity is considered as a mention. Each multi-word mention is annotated with the largest possible token span and no overlapping mentions are tagged. However, the remaining mentions of a document that are not part of any chain are not annotated.



a)



b)



c)

Figure 1. The statistics of annotated mentions

For this study, we also include 202 named entities (e.g., expressions that refer to a person, location, and organization) that are mentioned only once in these documents. Our dataset totally contains 5,372 mentions and 944 coreference chains capturing 5,170 mentions. As shown in Figure 1a, the majority of our mentions contain a single token (i.e., 4,133 mentions) whereas the number of tokens in a multi-word mention is between 2 and 24.

In order to answer the question of what forms a mention, we analyzed the part-of-speech (POS) tags of tokens in single-word and multi-word mentions. We used POS tags associated with the tokens in the treebank corpus and considered the POS tags of the first and last tokens for multi-word mentions. Figure 1b shows mention POS tags that appear at least 20 times in our dataset. We found that single-word expressions are often nouns (Noun), pronouns (Pron), and proper nouns (Prop). Our analysis also showed that multi-word

mentions often correspond to textual expressions that start with a noun, determiner (Det), or adjective (Adj) and end with a noun. It is particularly noteworthy that even verbs (Verb) and adverbs (Adv) are annotated as mentions. As shown in Figure 1c, the documents in our dataset contain between 17 and 359 mentions where the majority has more than 120 mentions.

In the dataset, the mentions that refer to the same real-world entity are clustered into coreference chains. Figure 2a shows the number of coreference chains that contain at most 20 mentions. In more than half of the cases, the chains contain 2 or 3 mentions and the entities with more than 20 mentions (up to 66 mentions) are observed to be rare (only 45 out of 944 chains).

We also examined how the mentions of the same entity are distributed in corresponding documents. We computed the distance between two mentions of the same chain by calculating the number of mentions that appear in between these mentions. The distance is taken as 0 if two mentions are not interleaved by other mentions. Figure 2b presents the number of coreferring mention pairs separated by at most 14 mentions. The results demonstrated that slightly more than half of the mentions that appear in the same chain have at most 3 other mentions in between.

The following is a representative chain from our dataset that consists of 8 mentions from 7 sentences. There are both single-word and multi-word mentions in the chain and the surface forms of these underlined mentions are all different. Moreover, there is more than one mention in the same sentence (i.e., the mentions numbered as fifth and sixth in the fifth sentence) and mention POS tags vary between sentences.

- Cebinde bir düğün fotoğrafı[1] durur

  He holds a wedding photo[1] in his pocket
- Kendi düğününde çekilmiş bir fotoğraf[2] , uzun yıllar önce

  A photo taken at his own wedding[2], many year ago
- Elini sol cebine sokup bir fotoğraf[3] çıkarttı , uzun_uzun baktı

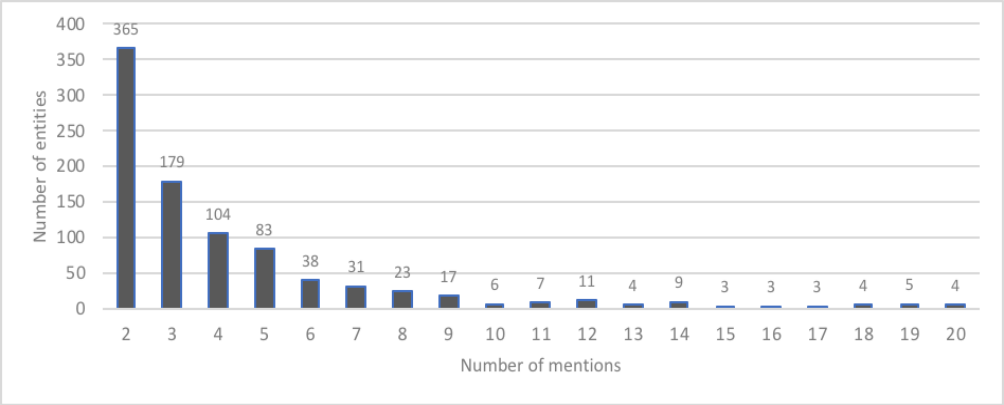
  He puts his hand in his left pocket and took out a photo[3], took a long look
- Omzunun üstünden fotoğrafı[4] görebiliyordum

  I could see the photo[4] over his shoulder
- Bir düğün fotoğrafıydı[5] bu[6]

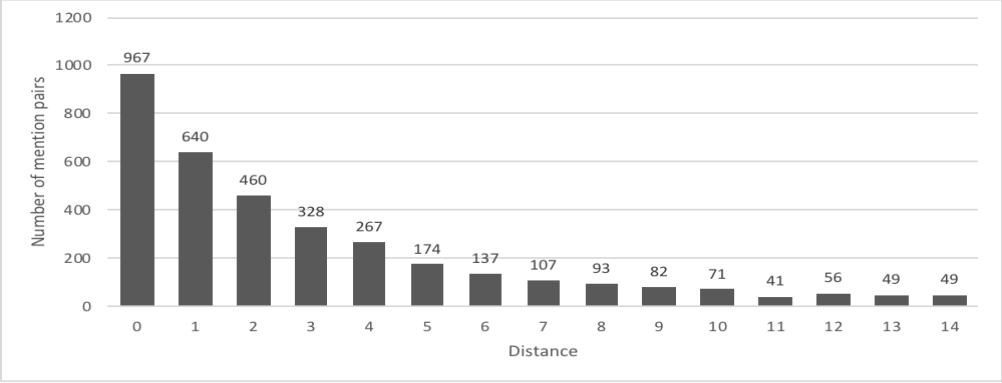
  It[6] was a wedding photo[5]
- Kerem'in sözünü_ettiği fotoğraf[7] olmalıydı

  It should have been the photo that Kerem mentioned[7]
- Lacivert takım elbiseli adam fotoğrafta[8] gençti

  The man in the navy blue suit was young in the photo[8]



a)



b)

Figure 2. The statistics of annotated coreference chains

3.2. Model parameters and evaluation metrics

For this study, we performed several experiments using both models with different settings. In Model_1, the annotations given in our dataset were used to extract the singularity and plurality information of input words and their lemmas. In Model_2, we explored the use of both context-dependent and context-independent word embeddings (Miaschi and Dell'Orletta, 2020) to assess the impact of contextual information on coreference resolution. The context-independent embedding of a word is the same regardless of the context where the word appears. Here, we used fastText embeddings trained for Turkish on Common Crawl and Wikipedia articles using character n-grams of length 5 (Grave et al., 2018). We experimented with the default embedding dimension of 300 (fastText-300) and also reduced dimensions of 100 (fastText-100) and 200 (fastText-200).

A word might have different context-dependent embeddings based on its context (surrounding words), all of which capture varying uses of the word. Literature has investigated several neural architectures while producing context-dependent embeddings (Van der Heijden et al., 2020; Mars, 2022), each with its own optimization instruments and limitations. In this work, we obtained embeddings of input tokens from Turkish language models based on BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and ELECTRA (Clark et al., 2020) architectures. In BERT architectures, some random tokens in the input are masked during training and the model learns to predict the original tokens. The Turkish BERT models (BERTurk) used in this work (Schweterr, 2021) were trained on Turkish OSCAR, Wikipedia, and OPUS corpora. We experimented with different models with small (BERTurk-32K) and large (BERTurk-128K) vocabulary sizes for both cased (BERTurk-cased) and uncased (BERTurk-uncased) training data. We also benefited from the DistilBERT architecture which is a distilled version of the BERT architecture that aims to reduce the parameter size (e.g., fewer encoder blocks) and hence increase the speed of the model. The DistilBERTurk model that we used (Schweterr, 2021) was a cased model with 32K vocabulary and trained on a portion of the original data used for training the BERTurk model. Finally, we experimented with the ELECTRA architecture where the input is corrupted by replacing some tokens with plausible incorrect alternatives during training. A generator model and a discriminator model are used to train a language model that is significantly smaller than the BERT model. The Turkish ELECTRA model used in this work (ELECTRA-base-mC4) was trained on a multilingual dataset and the cased model has a vocabulary of 32K (Schweterr, 2021).

For the experimental study, we trained Model_1 and Model_2 using 10-fold cross-validation and computed the mean scores across folds. Model_1 and Model_2 with a particular setting (i.e., with a specific pre-trained language model) was trained for 5, 10, and 15 epochs each and their performances were compared. While training Model_1, the cross entropy loss was used to adjust model weights and the stochastic gradient descent optimization with a learning rate of 0.005 was applied. In Model_2, the hidden layer size was set to 64, and the Adam optimizer with learning rate of 0.005 and dropout with a rate of 0.4 were used for model training. We reported performance scores using three automatic metrics that are often used in coreference resolution studies, namely the link-based metric MUC (Vilain et al., 1995), the mention-based metric B^3 (Bagga and Baldwin, 1998), and the entity-based metric CEAF-e (Luo, 2005).

4. Results and Discussion

Table 1 presents the evaluation scores that we obtained for Model_1 and Model_2. The results showed that Model_1 achieved the highest precision values but suffered from very low recall scores according to MUC and B^3 metrics. On the other hand, the precision of the model dropped significantly in CEAF-e metric despite an increase in its recall value. In terms of F1 scores, the performance of the model was surpassed by most of the Model_2 settings. Since Model_1 requires deep feature engineering, its performance substantially depends on the utilized features and their coverage. It is noteworthy to mention that the set of features used in this study can be extended to a larger set which would probably demonstrate a different resolution performance. Nonetheless, our scores indicate that it is yet practical to use Model_1 as a strong baseline for Turkish neural coreference studies.

In terms of all metrics, Model_2 settings where DistilBERT and ELECTRA language models are used received the lowest F1 scores among all settings with context-independent and context-dependent embeddings. The use of DistilBERT and ELECTRA embeddings resulted in very high precision values according to B^3, but as observed in other metrics, the low recall value had a negative impact on the overall performance. The setting with DistilBERT model showed poor performance as compared to the setting with ELECTRA model.

The Model_2 setting where BERTurk-uncased language model with 128K vocabulary is used obtained the highest F1 scores in MUC and CEAF-e metrics. According to the B^3 metric, its performance was the closest to the best performing model. In addition, the embeddings trained on the uncased dataset were observed to receive higher scores than those learned from the cased dataset for the same vocabulary size. When the dataset used to fine-tune a BERT model is scarce, the literature showed that the model often suffers from degraded performances (Yu et al., 2021; Zhang et al., 2021). However, here we observed that the larger model (BERTurk-128k-uncased) has a better resolution performance than the smaller model (BERTurk-32k-uncased) even though our coreference dataset is scarce.

Table 1. Performance evaluations of all models (10 epochs)

Model Name	Embedding Type	MUC			B^3			CEAF-e		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Model_1		0.958	0.353	0.501	0.983	0.401	0.554	0.245	0.681	0.347
	fastText-100	0.682	0.426	0.520	0.766	0.445	0.562	0.298	0.658	0.410
	fastText-200	0.699	0.573	0.629	0.656	0.547	0.593	0.415	0.616	0.489
	fastText-300	0.668	0.568	0.612	0.631	0.520	0.567	0.398	0.626	0.482
Model_2	*BERTurk-32k-cased*	0.699	0.631	0.660	0.604	0.553	0.560	0.437	0.549	0.472
	BERTurk-32k-uncased	0.717	0.651	0.679	0.572	0.590	0.569	0.453	0.558	0.488
	BERTurk-128k-cased	0.711	0.617	0.654	0.638	0.531	0.564	0.427	0.576	0.476
	BERTurk-128k-uncased	0.718	0.681	0.697	0.603	0.570	0.574	0.486	0.556	0.508
	DistilBERT	0.237	0.123	0.134	0.928	0.264	0.392	0.167	0.530	0.243
	ELECTRA	0.490	0.426	0.449	0.867	0.300	0.408	0.202	0.525	0.270

Using context-independent embeddings in Model_2 achieved the highest F1 score in B^3 metric, lagged behind context-dependent embeddings in MUC metric, and showed a competitive performance in CEAF-e metric. The scores achieved by the setting where 200-dimensional embeddings are used were higher than those obtained by embeddings with other dimensions. These results arguably indicate that context-independent embeddings can be used for Turkish coreference resolution in the lack of a large dataset required to fine-tune a pre-trained language model efficiently.

Figure 4 presents the behavior of all models with different settings where the training was performed for 5, 10, and 15 epochs. Increasing the number of epochs positively impacted the overall F1 performance of Model_1 (i.e., feature_based model) and Model_2 settings that utilize context-independent fastText embeddings. Model_1 achieved an improvement of 0.002 in MUC, and 0.001 in B^3 and CEAF-e metrics once the number of epochs is increased from 5 to 15. The improvements seen in Model_2 settings with fastText embeddings were in the range of 0.084-0.114 for MUC, 0.021-0.045 for B^3, and 0.064-0.118 for CEAF-e metrics, respectively. The performance of Model_2 with DistilBERT and ELECTRA embeddings were observed to decline once the epoch number is increased from 5 to 15. However, the use of BERT embeddings did not show any emergent pattern with respect to increased training epochs.

5. Conclusions

In this article, we present the first Turkish neural coreference resolution study where two different models that follow the mention-ranking approach are developed. The models cluster given mentions by ranking candidate antecedents of each mention and selecting the candidate with the highest score. The first model utilizes a set of hand-crafted features of mentions and uses them as input in a feed-forward neural network. On the other hand, the second model first obtains contextual representations of mentions using pre-trained language models and feeds representation pairs of mentions and candidate antecedents to a feed-forward neural network. We experiment with different context-dependent and context-independent language models and report their overall performances in resolving Turkish coreference mentions. Our results demonstrate that utilizing pre-trained language models (in particular BERT base models) for this task is beneficial and promising results can be obtained despite the lack of large-scale Turkish coreference datasets.

Given these encouraging results, we plan to extend this work in several directions. We first plan to collect a large-scale Turkish coreference dataset in order to fine-tune pre-trained models more efficiently and better learn the coreference relations. We argue that an extensive dataset would help us to obtain more accurate insights about the impact of Turkish language models on this task. Although this study is a significant advance in the

state-of-the-art, there is still room for future improvements. One important future work to investigate is developing a mention detection approach for Turkish and assessing its performance in an end-to-end coreference resolution system. Finally, we plan to explore new ways of achieving higher resolution performances by studying mention-pair and entity-mention approaches for Turkish using neural network architectures.

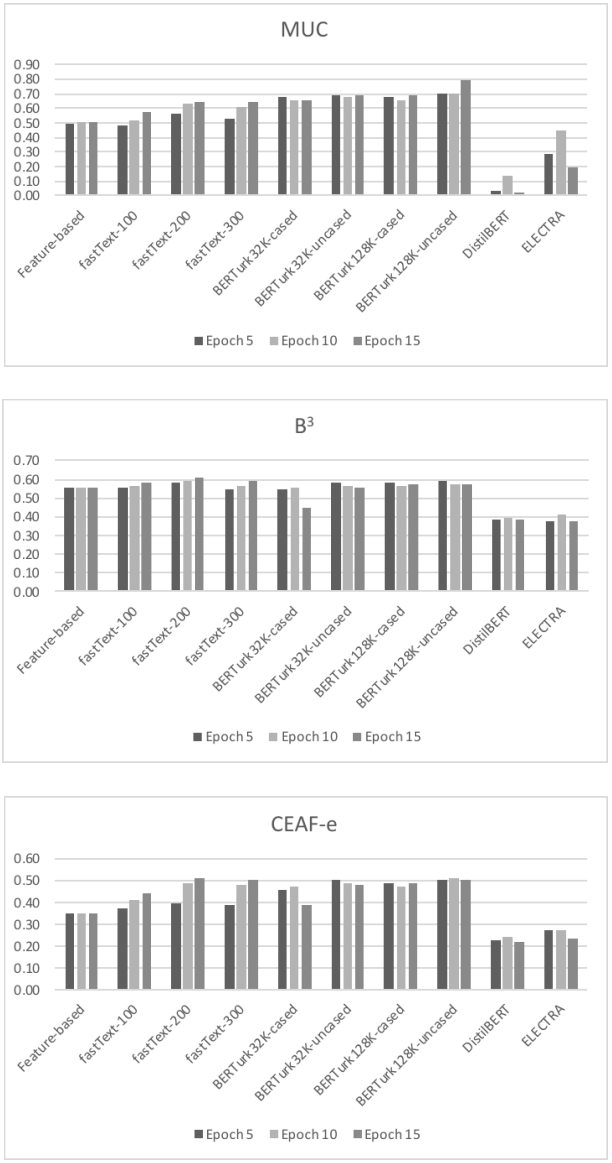


Figure 4. F1 scores of models for different epoch numbers

References

Aloraini, A., Yu, J., Poesio, M., 2020. Neural Coreference Resolution for Arabic. The Third Workshop on Computational Models of Reference, Anaphora and Coreference, pp. 99-110.

Auliarachman, T., Purwarianti, A., 2019. Coreference Resolution System for Indonesian Text with Mention Pair Method and Singleton Exclusion using Convolutional Neural Network. ICAICTA2019, The International Conference of Advanced Informatics: Concepts, Theory and Applications, pp. 1-5.

Bagga, A., Baldwin, B., 1998. Algorithms for scoring coreference chains. LREC 1998, The 1st International Conference on Language Resources and Evaluation, pp. 563-566.

Bengtson, E., Roth, D., 2008. Understanding the Value of Features for Coreference Resolution. EMNLP 2008, The Conference on Empirical Methods in Natural Language Processing, pp. 294-303.

Bhattacharjee, S., Haque, R., de Buy Wenniger, G.M., Way, A., 2020. Investigating Query Expansion and Coreference Resolution in Question Answering on BERT. In: E. Métais, F. Meziane, H. Horacek, P. Cimiano (Eds.), Natural Language Processing and Information Systems, NLDB 2020, Lecture Notes in Computer Science, 12089 pp. 47-59. Springer, Cham. doi:10.1007/978-3-030-51310-8_5

Bunescu, R., 2012. Adaptive Clustering for Coreference Resolution with Deterministic Rules and Web-Based Language Models. Seml 2012, The First Joint Conference on Lexical and Computational Semantics, pp. 11–19.

Cai, J., Strube, M., 2010. End-to-End Coreference Resolution via Hypergraph Partitioning. Coling 2010, The 23rd International Conference on Computational Linguistics, pp. 143-151.

Clark, K. Manning, C.D., 2015. Entity-Centric Coreference Resolution with Model Stacking. ACL-IJCNLP 2015, The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp.1405-1415.

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. The ACL Workshop on Computational Approaches to Semitic Languages, pp. 63-70.

Denis, P., Baldridge, J., 2008. Specialized Models and Ranking for Coreference Resolution. The 2008 Conference on Empirical Methods in Natural Language Processing, pp. 660-669.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019, The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186.

Durrett, G., Klein, D., 2013. Easy Victories and Uphill Battles in Coreference Resolution. EMNLP 2013, the 2013 Conference on Empirical Methods in Natural Language Processing, pp.1971-1982.

Fang, K., Jian, F., 2019. Incorporating Structural Information for Better Coreference Resolution. IJCAI 2019, The Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 5039-5045.

Fei, H., Li, X., Li, D., Li., P., 2019. End-to-end Deep Reinforcement Learning Based Coreference Resolution. The 57th Annual Meeting of the Association for Computational Linguistics the ACL Workshop on Computational Approaches to Semitic Languages, pp. 660-665.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T., 2018. Learning Word Vectors for 157 Languages. LREC 2018, The International Conference on Language Resources and Evaluation.

Hacioglu, K., Douglas, B., Chen, Y., 2005. Detection of entity mentions occurring in English and Chinese text. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 379-386.

Hoste, V., 2016. The Mention-Pair Model. In: M. Poesio, R. Stuckardt, Y. Versley (Eds.) Anaphora Resolution. Theory and Applications of Natural Language Processing pp. 269-282 Springer, Berlin, Heidelberg. doi: 10.1007/978-3-662-47909-4_9

Kılıçaslan, Y., Güner, E.S., Yıldırım, S., 2009. Learning-based pronoun resolution for Turkish with a comparative evaluation, Computer Speech Language, 23(3), 311-331. doi: 10.1016/j.csl.2008.09.001

Klemen, M., Žitnik, S., 2022. Neural Coreference Resolution for Slovene Language. Computer Science and Information Systems, 19(2), 495–521. doi: 10.2298/CSIS201120060K

Kriman, S., Heng, J., 2021. Joint Detection and Coreference Resolution of Entities and Events with Document-level Context Aggregation. ACL-IJCNLP 2021, The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pp. 174-179.

Lai, T.M., Bui, T., Kim, D.S., 2022. End-To-End Neural Coreference Resolution Revisited: A Simple Yet Effective Baseline. ICASSP 2022, The IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8147-8151.

Lata, K., Singh, P., Dutta, K., 2022. Mention detection in coreference resolution: survey. Applied Intelligence, 52, 9816-9860. doi: 10.1007/s10489-021-02878-2

Lee, K., He, L., Lewis, M., Zettlemoyer, L., 2017. End-to-End Neural Coreference Resolution. EMNLP 2017, The 2017 Conference on Empirical Methods in Natural Language Processing, pp. 188-197.

Lee, K., He, L., Zettlemoyer, L., 2018. Higher-order coreference resolution with coarse-to-fine inference. NAACL 2018, The Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 687-692.

Li, Z., Shi, K., Chen N.F., 2021. Coreference-Aware Dialogue Summarization. The 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 509-519.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos. S., 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. ACL 2004, The 42nd Annual Meeting on Association for Computational Linguistics, pp. 135-142.

Luo, X., 2005. On Coreference Resolution Performance Metrics. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 25-32.

Mars, M., 2022. From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough. Applied Sciences. 12(17). doi: 10.3390/app12178805

Miaschi, A., Dell'Orletta, F., 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. The 5th Workshop on Representation Learning for NLP, pp. 110-119.

Niton, B., Morawiecki, P., Ogrodniczuk, M., 2018. Deep Neural Networks for Coreference Resolution for Polish. LREC 2018, The International Conference on Language Resources and Evaluation, pp. 395-400.

Pamay, T., Eryiğit, G., 2018. Turkish Coreference Resolution. INISTA 2018, The Innovations in Intelligent Systems and Applications, pp. 1-7.

Park, C., Lee, C., 2015. Mention Detection using Bidirectional LSTM-CRF Model. The Annual Conference on Human and Language Technology, pp. 224-227.

Park, C., Lee, C., Lim, S., 2017. Mention detection using pointer networks for coreference resolution. ETRI Journal, 39(5), 652-661. doi: 10.4218/etrij.17.0117.0140

Rahman, A., Ng, V., 2011. Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution. Journal of Artificial Intelligence Research, 40, 469-521. doi: 10.1613/jair.3120

Rahman, A., Ng, V., 2011. Ensemble-based coreference resolution. IJCAI 2011, The Twenty-Second international joint conference on Artificial Intelligence, pp. 1884-1889.

Sahlani, H., Hourali, M., Minaei-Bidgoli, B., 2020. Coreference Resolution Using Semantic Features and Fully Connected Neural Network in the Persian Language. International Journal of Computational Intelligence Systems, 13(1), 1002-1013. doi: 10.2991/ijcis.d.200706.002

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv, doi: 10.48550/ARXIV.1910.01108 10.48550/arXiv.1706.01863

Sapena, E., Padró, L., Turmo, J., 2010. Relaxcor: A global relaxation labeling approach to coreference resolution. The 5th International Workshop on Semantic Evaluation, pp. 88-91.

Say, B., Zeyrek, D., Oflazer, K., Özge, U., 2002. Development of a corpus and a treebank for present-day written Turkish. ICTIL 2002, The 11th International Conference of Turkish Linguistics , pp. 183-192.

Schüller, P., Cıngıllı, K., Tunçer, F., Sürmeli, B. G., Pekel, A., Karatay, A. H., Karakaş, H. E., 2007. Marmara Turkish coreference corpus and coreference resolution baseline, arXiv, doi: 10.48550/arXiv.1706.01863

Schweterr, S., 2021. BERTurk. https://github.com/stefan-it/turkish-bert

Soon, W.M., Ng, H.T., Lim, D.C.Y., 2001. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4), 521–544. doi: 10.1162/089120101753342653

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., De Ilarraza, A.D., 2012. Mention detection: First steps in the development of a Basque coreference resolution system. KONVENS 2012, The 11th Conference on Natural Language Processing, pp.128-136.

Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K., 2007. Two uses of anaphora resolution in summarization. Information Processing Management, 43(6), 1663-1680. doi: 10.1016/j.ipm.2007.01.010

Uryupina, O., Moschitti, A., 2015. A State-of-the-Art Mention-Pair Model for Coreference Resolution. The Fourth Joint Conference on Lexical and Computational Semantics, pp. 289-298.

Van der Heijden, N., Abnar, S., Shutova, E., 2020. A Comparison of Architectures and Pretraining Methods for Contextualized Multilingual Word Embedding, pp. 9090-9097.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L., 1995. A model-theoretic coreference scoring scheme. The 6th Message Understanding Conference (MUC-6), pp. 45-52.

Wang, B., Lu, W., Wang, Y., Jin, H., 2018. A Neural Transition-based Model for Nested Mention Recognition. The 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1011-1017.

Wiseman, S., Rush, A.M., Shieber, S., Wetson, J., 2015. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. ACL-IJCNLP 2015, The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing , pp. 1416-1426.

Wiseman, S., Rush, A. M., Shieber, S. M., 2016. Learning Global Features for Coreference Resolution. The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 994-1004.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., Li., S., 2008. An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming. ACL08-HLT, The 46th Annual Meeting of the Association for Computational Linguistics and the Human Language Technology Conference , pp. 843-851.

Yıldırım, S., Kılıçaslan, Y., Yıldız, T., 2009. Pronoun Resolution in Turkish Using Decision Tree and Rule-Based Learning Algorithms. In: Z. Vetulani, H. Uszkoreit (Eds.), Human Language Technology, Challenges of the Information Society, LTC 2007, Lecture Notes in Computer Science, 5603, pp. 270-278) Springer. doi: 10.1007/978-3-642-04235-5_23

Yu, Y., Zuo, S., Jiang, H., Ren, W., Zhao, T., Zhang, C., 2021. Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach, The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1063-1077.

Zhang, R., Nogueira dos Santos, C., Yasunaga, M., Xiang, B., Radev, D., 2018. Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering. The 56th Annual Meeting of the Association for Computational Linguistics, pp. 102-107.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., Artzi, Y., 2021. Revisiting Few-sample BERT Fine-tuning. The 9th International Conference on Learning Representations.

Zitouni, I., Sorensen, J., Luo, X., Florian, R., 2005. The impact of morphological stemming on Arabic mention detection and coreference resolution. The ACL Workshop on Computational Approaches to Semitic Languages, pp. 63-70.