IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

ELECTRA for Neural Coreference Resolution in Italian

RAFFAELE GUARASCI[1], ANIELLO MINUTOLO[1], EMANUELE DAMIANO[1],GIUSEPPE DE PIETRO[1],HAMIDO FUJITA[234],MASSIMO ESPOSITO[1]

[1]Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Naples, Italy
[2]Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam
[3]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain
[4]Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan

Corresponding author: Aniello Minutolo (e-mail: aniello.minutolo@cnr.it).

ABSTRACT

In recent years, the impact of Neural Language Models has changed every field of Natural Language Processing. In this scenario, coreference resolution has been among the least considered task, especially in language other than English. This work proposes a coreference resolution system for Italian, based on a neural end-to-end architecture integrating ELECTRA language model and trained on OntoCorefIT, a novel Italian dataset built starting from OntoNotes. Even if some approaches for Italian have been proposed in the last decade, to the best of our knowledge, this is the first neural coreference resolver aimed specifically to Italian. The performance of the system is evaluated with respect to three different metrics and also assessed by replacing ELECTRA with the widely-used BERT language model, since its usage has proven to be effective in the coreference resolution task in English. A qualitative analysis has also been conducted, showing how different grammatical categories affect performance in an inflectional and morphological-rich language like Italian. The overall results have shown the effectiveness of the proposed solution, providing a baseline for future developments of this line of research in Italian.

INDEX TERMS

Coreference resolution, ELECTRA, Italian dataset, Deep learning, Natural Language Processing

I. INTRODUCTION

Coreference resolution is one of the tasks that has always fascinated scholars of Natural Language Processing (NLP). Nevertheless, it has a troubled history in the field and it still cannot be considered as a solved task [1]–[4]. Many aspects concerning the nature itself of the coreference are still debated both in Computational [5] and Theoretical Linguistics [6].

A large number of proposed approaches have followed one another over the years, from early rule-based and statistical [7] to machine and deep learning ones [5]. Recent years with the rise of neural language models (NLM) have led to a change. Approaches based on the latest language models like ELMo (*Embeddings from Language Models*) [8], [9] and BERT (*Bidirectional Encoder Representations from Transformers*) [10]–[12] have been proposed and they have brought significant performance gains over all previous approaches [13].

This has opened the way to interesting research perspectives, but this new line of research has so far been almost exclusively the domain of the English language. The rapid growth of approaches available for English does not have an equivalent in other languages. This is partly due to the scarcity of existing resources. English, indeed, can rely on a large number of resources of various domains and sizes, the same cannot be said for the other languages.

Trying to address this shortcoming, this paper proposes a NLM-based system addressing the coreference resolution task in the Italian language, using a novel dataset built starting from OntoNotes [14].

The system end-to-end architecture is derived from [8], but, differently, it exploits a novel language model named ELECTRA (*Efficiently Learning an Encoder that Classifies Token Replacements Accurately*) [15]. ELECTRA has shown a better ability to capture contextual word representations, substantially outperforming, in its downstream performance, other models, like BERT, given the same model size, data, and compute [16]. Moreover, for ELECTRA, a pre-trained

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

model also exists for the Italian language, that is trained on a standard-Italian written corpora and, thus, results adequate for the coreference purposes.

Concering the coreference dataset used in this work, the obligatory choice has fallen on OntoCorefIT [1], a novel corpus collecting almost 55k utterances in Italian created with a cross-lingual approach starting from OntoNotes. This corpus ensures a good coverage due to its size and domain-independence, in addition collected utterances are compliant with Italian grammar being subjected to linguistic refinement step and to quantitative and qualitative evaluation.

Very few coreference resources have been developed for Italian, the majority of which are outdated and small in size. This is one of the main causes of very slow progress in the field in recent years. Moreover, currently, there are no studies that make a comprehensive analysis of different models to evaluate performances in specific NLP tasks, like coreference resolution, in Italian language. As far as known, this is the first neural coreference resolution system specifically focused on Italian.

Summarizing, this work focuses on these main objectives: i) exploiting a very promising NLM, i.e. ELECTRA, and a novel Italian corpus to make a robust and efficient coreference resolution system for the Italian language, ii) replacing ELECTRA with BERT in order to compare their performance and assess the validity of the proposed system, iii) providing a baseline for future developments of this line of research.

The contribution is structured as follows. Section 2 reviews the state-of-the art of approaches for coreference resolution starting from early rule-based approaches to modern deep learning ones. In Sections 3 and 4 the experimental assessment is outlined, including the details of the neural language model and the dataset used. The Section 5 focuses on the presentation and the discussion of the obtained results, from both a quantitative and qualitative point of view. Finally, Section 6 reports conclusion and future works.

II. BACKGROUND AND RELATED WORK

A. RULE-BASED APPROACHES

Reference resolution task in NLP has been widely considered as a task which inevitably depends on some hand-crafted rules. These rules are based on syntactic and semantic features of the text under consideration. First algorithms [17], [18] were based on syntax or discourse structure, they used rule-based approaches and a manual evaluation. Another similar work [19] added to previous ones the use of knowledge resources and morpho-semantic information. The main limitation of these early approaches was the total lack of a standard for the evaluation phase [20], especially with the advent of shared evaluation campaigns starting from MUC [21], ACE [22] and CoNLL [23], [24].

[1]OntoCorefIT is made available upon request at: https://nlpit.na.icar.cnr.it/nlp4it/#/datasets/coref

A rule-based approach with a more robust evaluation was proposed by [25]. This simple modular approach relied only on syntax and was evaluated on multiple standard datasets. Further improved modified versions were created from this work [26], [27].

B. MACHINE-LEARNING APPROACHES

The increased availability of annotated corpora like MUC or ACE led to the development of machine learning approaches to coreference resolution.

The most fruitful family of machine-learning approaches dealt with coreference as a set of pairwise connections. It used a classifier to decide if two proper nouns (NP) were co-referent. First, these models created training instances to reduce the imbalance between co-referent and non-coreferent entities in the training samples. Types of classifiers proposed ranged from statistical learners [28] to random forests [29]. Finally, a NP partition was generated, in order to test the trained model on a test set to obtain the coreference chains. To handle this task, several clustering techniques were developed, i.e. best-first clustering [30], closest-first clustering [31], correlational clustering [32], Bell Tree beam search [33] and graph partitioning algorithms [34]. Other studies proposed to combine the classifier with effective partitioning using Integer Linear Programming [35] or to completely eliminate the classification phase [36].

Other types of proposed approaches were represented by entity-mention model, that focused on a single underlying entity of each referent in discourse [37], and by mention-pair model, which used a binary classifier to decide whether an antecedent was coreferent with the mention [38], [39].

C. DEEP LEARNING APPROACHES

As happened in all fields of NLP, the advent of Deep Learning rapidly subverted existing approaches. The possibility of representing words using vectors that embed their semantic relationships and the lowest number of required features drastically reduced the need to rely on manually-created features. These approaches captured dependencies between mentions using Recurrent Neural Networks (RNN) or Long short-term memory (LSTM). Limitations of deep learning approaches lie in their poor adaptability to the domain. They frequently need domain-specific adjustment before being used with satisfying results.

The first neural model proposed [1] focused on two critical aspects of coreference resolution: the identification of non-anaphoric references in texts and the ability to distinguish mentions from non-mentions. The model, trained on features extracted from BASIC [40], performed better than all existing approaches. Later developed models [2], [3] incorporated entity-level information produced by a RNN in order to exploit a global features about entity clusters. These models relied on including features defined on mention-pair clusters.

Another approach [41] used the neural mention ranking model [3] in order to replace the heuristic loss functions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

with reinforced-learning based policy gradient algorithm. Currently, the state-of-the-art approach [42], is based on end-to-end neural model with the construction of high-dimensional word embeddings to represent words of annotated documents. Although difficult to maintain because of its high-dimensionality, this system, based on LSTM, has as its strengths the ability to capture long term dependencies.

More recent approaches use BERT model to word representation [8], [11] or its modified version SpanBERT to create span representations and increase BERT's maximum segment limitations [12].

D. MULTILINGUAL APPROACHES

Even if the vast majority of coreference resolution approaches was focused on English language - as in many other NLP research areas - individual language-specific approaches was also developed.

Dedicated approaches for major European languages were proposed: German [43]–[45], Spanish [46], Portuguese [47] Czech [48], French [49] and many others. Concerning Italian, very few systems for coreference resolution were proposed so far [50], [51]

Evaluation campaigns like Semeval 2010 [52] and Conll 2012 [24] and the development of multilingual resources like Ontonotes [53] or ParCor [54] shifted the focus to systems that can be language-independent or able to be adapted to several languages simultaneously. However, current language-independent systems were based only on shallow approaches exploiting universal part-of-speech tagset [55] and universal dependencies [56]. Many issues are still unsolved and no system is yet able to handle language-specific features [57].

A recent approach that looks promising, in particular in case of low-resource languages, uses cross-lingual methodologies exploiting existing resources developed in other languages. In particular, for low-resource languages, projection-based techniques have been proposed [58], [59]. Projection is a technique consisting in automatically transferring annotations from a resource-rich language to a low-resource language across parallel corpora. In particular, these approaches have used English as source language, and they have been tested for Spanish and Italian [58], Portuguese and Spanish [60], German and Russian [61]. Other studies have tested a direct transfer learning between languages by using multilingual word embeddings, using a model trained on a language for other languages that share a common semantic space [62]: experiments have been carried out on Chinese, Spanish, Portuguese and English [47], [63] .

III. MATERIAL AND METHODS

Figure 1 shows the working process behind the proposed coreference resolution system. First, documents composing the OntoCorefIT dataset are given as input. Secondly, the end-to-end *c2f-coref* architecture proposed by [8], [42] and powered by the usage of ELECTRA to represent input tokens, is leveraged to calculate the coreference predictions. A detailed description of the coreference data set realized for the Italian language and the system architecture is provided below.

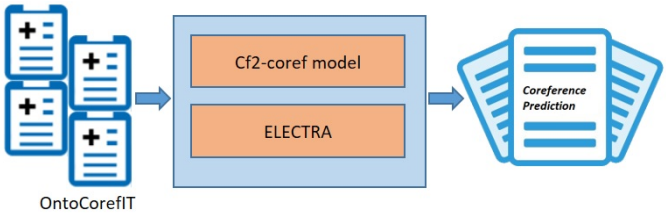


FIGURE 1. Working process of the proposed system.

A. DATASET

The dataset used here is OntoCorefIT, currently the largest dataset for coreference in the Italian language built from OntoNotes.

Even though other corpora exist for coreference in Italian, none of them is suitable for the purpose of this work. The first reason is given by the size. Corpora currently available for the Italian language are quite small (see Table 1), therefore they should not particularly suitable to be used as training set for a neural language model. On the contrary, OntoCorefIT consists of almost 55k sentences for a total of 704k tokens. A second reason relies in the domain-independence. OntoCorefIT corpus is based on the English version of OntoNotes, so it collects data from a set of heterogeneous domains. By contrast, other Italian existing resources are focused on specific domains: VENEX data come from financial newspapers [64] and spoken task-oriented dialogues [65]; I-CAB and LiveMemories are limited to a restricted material related to the Italian region of Trentino-Alto Adige/Südtirol.

Corpus	Size (words)	Annotation Scheme
OntoCorefIT	708k	CoNLL
Venex (2004)	40k	MATE/MMAX
i-Cab (2006)	250k	ACE
LiveMemories (2010)	250k	ARRAU

TABLE 1. Size comparison of Italian coreference datasets

OntoCorefIT has been automatically generated from OntoNotes dataset by using a cross-lingual approach, i.e. English utterances are automatically translated and refined in order to obtain utterances respecting Italian grammar and contextually, preserving original coreference and mentions. OntoCorefIT shares the same CoNLL-like annotation schema of original English Ontonotes, ensuring robustness in the evaluation being the *de facto* standard since CoNLL2012 to most recent state-of-the-art systems [42].

In more detail, starting from OntoNotes dataset, firstly, a multi-level translation has been performed to translate utterances from English to Italian, trying to preserve original mentions without losing in the translation the tokens compos-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

ing the mentions, their positions, and the verbal agreements involving them.

Secondly, the translated utterances are linguistically refined in order to improve their readability and, thus, produce an output text as close as possible to the Italian grammar. In particular, language-specific rules derived from the Theoretical linguistics have been introduced, covering most frequent phenomena in Italian sentences, i.e. gender and number agreement, subject pronoun deletion or inflection phenomena.

In particular, due the pro-drop nature of the Italian language, the explicit subject pronoun can always be removed. Thus, in case of a pronoun subject tagged as a mention, the tag has been shifted to the verb referred by the pronoun, otherwise it has been simply removed from the utterance.

In line with Ontonotes, all utterances have been successively annotated with Part-of-Speech (POS) tags in order to obtain a CoNLL-like format. Morphosyntactic analysis with POS-tagging has been automatically performed for the utterances in Italian. Thus, the lack of any supervision may have introduced some errors in tag assignment, in particular concerning cases presenting greater ambiguity in certain categories (e.g. pronouns and determiners). Italian, due to its specific features, is considered a more complex language than English [66]–[68] and, therefore, it exhibits greater criticalities in performing pos-tagging and parsing tasks [69], [70].

Table 2 reports an overview of the OntoCorefIT dataset, showing the total number of utterances (*utts*), coreferences (*coarefs*) and tokens for the three partitions into which the dataset is divided, namely *Train*, *Test* and *Dev*.

The subsets *Train*, *Test* and *Dev* are arranged into sets of documents each of which is composed of an ordered list of non-overlapping partitions of ordered utterances.

	Train	Test	Dev
utts	44073	5415	5363
corefs	40648	5039	4372
tokens	568641	71293	68796

TABLE 2. Overview of the OntoCorefIT dataset

B. SYSTEM ARCHITECTURE

The neural architecture of the proposed system is inspired by the end-to-end *c2f-coref* system proposed by [8], [42]. That end-to-end neural model has been very successful in the literature and it is the current state of the art for the English OntoNotes dataset [24].

In accordance with [8], [42], the task of coreference resolution is defined as a set of antecedent assignments y_i for each span i, with $1 \leq i \leq N$, belonging to a given document D that contains T tokens and $N = \frac{T(T+1)}{2}$ possible text spans.

In particular, all spans are considered as potential mentions and, for each span i, the set of antecedent assignments y_i, i.e. mentions preceding the span under examination and referring

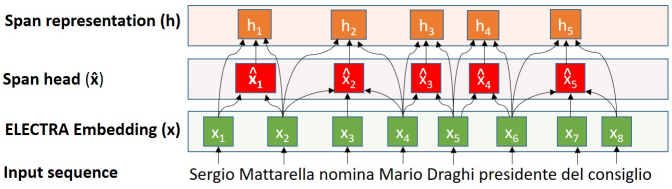


FIGURE 2. Span representation using ELECTRA and attention-based head words

to the same entity, is calculated. The set of of possible assignments for each y_i is $Y(i) = \{\epsilon, 1, ..., i-1\}$, which includes the dummy antecedent ϵ and all preceding spans. The dummy antecedent is used to cover the case when a span is not an entity mention, as well as the case when a span is an entity mention but is not coreferent with other spans. Grouping all spans connected by a set of antecedent predictions allows to define a final clustering.

To realize this task, the model proposed in [8], [42] learns a conditional probability distribution $P(y_1, ..., y_n|D)$ whose most likely configuration corresponds to the correct clustering. This distribution is calculated as the product of multinomials for each span:

$$\begin{aligned} P(y_1, ..., y_N|D) &= \prod_{i=1}^{N} P(y_1|D) \\ &= \prod_{i=1}^{N} \frac{exp(s(i, y_j))}{\sum_{y' \in Y(i)} exp(s(i, y'))} \end{aligned} \tag{1}$$

where $s(i, j)$ is a pairwise score for a coreference link between span i and span j in document D. This coreference score is computed as follows:

$$s(i,j) = \begin{cases} 0 & j = \varepsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \varepsilon \end{cases} \tag{2}$$

It is equal to 0 in case of dummy antecedent, otherwise it is the sum of three terms, namely $s_m(i)$ and $s_m(j)$ are the scores indicating that the spans i and j are mentions, and $s_a(i, j)$ is the score indicating that the span j is an antecedent for the span i.

The model predicts the best antecedent score if all non-dummy scores are positive, otherwise it vanishes.

Differently from [8], [42], each span i is given an embedding representation h_i by using ELECTRA as shown in Figure 2 and described in the next subsection.

Given these span representations, the scoring functions s_m and s_a are calculated, as shown in Figure 3, via feed-forward neural networks $FFNN_m$ and $FFNN_a$ as follows:

$$s_m(i) = w_m \cdot FFNN_m(h_i) \tag{3}$$

$$s_a(i,j) = w_a \cdot FFNN_a([h_i, h_j, h_i \circ h_j, \phi(i,j)]) \tag{4}$$

where $\cdot$ is the dot product; $\circ$ is element-wise multiplication; $FFNN$ is a feed forward neural network calculating

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

a non-linear mapping from input to output; $s_a(i,j)$ includes explicit element-wise similarity of each span e_i and a feature vector $\phi(i,j)$ containing information about speaker, genre and other syntactic metadata.

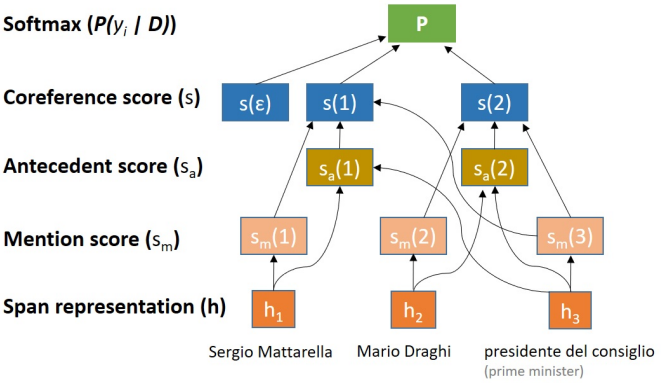


FIGURE 3. Calculation of mention and antecedent scoring functions

For the training, the marginal log-likelihood of all correct antecedents implied by the gold clustering is optimized:

$$log \prod_{i=1}^{N} \sum_{\hat{y} \in Y(i) \cap GOLD(i)} P(\hat{y}) \tag{5}$$

where GOLD(i) is the set of spans in the gold cluster containing the span i.

C. ELECTRA

Unlike previous work based on LSTM or BERT [4], [8], [11], the proposed system lies on a new NLM ispired by GAN networks [71], ELECTRA [15]. ELECTRA, has shown to be more compute-efficient than BERT, allowing to achieve better performance keeping the dimensions of the model unchanged.

In more detail, the pre-training step in ELECTRA is performed exploiting the masked language model (MLM) in a more efficient way than BERT. It uses two Transformer models, that share the same word embedding, namely a generator G and a discriminator D, and it is based on training D to to distinguish "fake" or replaced input tokens produced by G in the sequence. This approach, called replaced token detection (RTD), allows to use a minor number of examples without losing in performance.

In particular, for a given input sequence, where some tokens are randomly replaced with a [MASK] token, G is trained to predict the original tokens for all masked ones. On the other hand, G is given input sequences built by replacing [MASK] tokens with “fake” ones produced by G and it is trained to predict whether they are original or “fake”.

More formally, given an input sentence s of raw text χ, composed by a sequence of tokens $s = w_1, w_2, \ldots, w_n$ where w_t $(1 \leq t \leq n)$ represents the generic token (e.g. word, subword or character), both G and D firstly encode

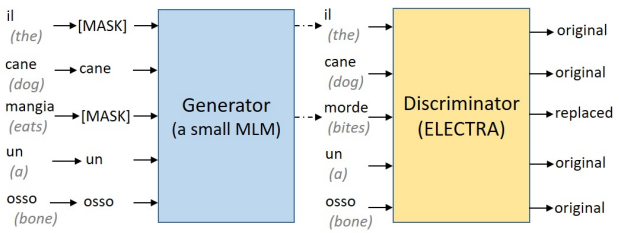


FIGURE 4. ELECTRA overview with replaced token detection. The generator G is usually a MLM trained with the discriminator D but it may virtually be any model producing an output distribution over tokens.

s into a sequence of contextualised vector representations $h(s) = h_1, h_2, \ldots, h_n$.

Then, for a given position t so that the corresponding $w_t = [MASK]$, the generator outputs the probability to generate a particular token w_t, with a softmax layer:

$$p_G(w_t|s) = \frac{e(w_t)^T h_G(s)_t}{\sum_{w'} exp(e(w')^T h_G(s)_t)} \tag{6}$$

where $e(\cdot)$ represents the embedding function.

On the other hand, the discriminator predicts whether w_t is the original or "fake", using a sigmoid layer:

$$D(s,t) = sigmoid(e(w_t)^T h_D(s)_t) \tag{7}$$

During the pre-training, G employs the following loss function:

$$\mathcal{L}_{Gen} = \mathcal{L}_{MLM} = \mathbb{E}(\sum_{i \in m} - \log p_G(w_i | s^{masked})) \tag{8}$$

where $m = m_1, m_2, \ldots, m_k$ are k random selected words and s^{masked} is the sentence with the masked words.

On the other hand, D uses the following loss function:

$$\begin{aligned} \mathcal{L}_{Dis} = \mathbb{E}(\sum_{t=1}^{n} & -\mathbb{I}(w_t^{corrupt} = x_t) \log D(s^{corrupt}, t) + \\ & -\mathbb{I}(w_t^{corrupt} \neq x_t) \log D(s^{corrupt}, t)) \end{aligned} \tag{9}$$

where $w_t^{corrupt}$ is the corrupted word within the corrupted sentence $s^{corrupt}$.

Finally, the following combined loss is minimised:

$$\min_{\theta_G, \theta_D} \sum_{s \in \chi} \mathcal{L}_{Gen}(s, \theta_G) + \lambda \mathcal{L}_{Dis}(s, \theta_D) \tag{10}$$

At the end of the pre-training, G is discarded and only the discriminator model D is used.

The main reason that improve ELECTRA efficiency respect to MLM BERT-like models is that predictions are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

calculated not only over masked tokens, but also for each token and the discriminator loss can be calculated over all input tokens.

In the proposed system, ELECTRA is used to represent each span by considering the embeddings for its boundary tokens as well as for a head token calculated over all its words. In more detail, for each span i, its representation h_i obtained by using ELECTRA is given by:

$$h_i = [x^*_{START(i)}, x^*_{END(i)}, \hat{x}_i, \phi(i)] \tag{11}$$

where $x^*_{START(i)}$ and $x^*_{END(i)}$ are the embedding representations of the boundary tokens, $\hat{x}_i$ represents a soft head word calculated over all the span tokens and $\phi(i)$ is a vector feature which encodes the span size.

The soft head word $\hat{x}_i$ is calculated by using an attention mechanism on the words of each span [72], as follows:

$$\alpha_t = w_\alpha \cdot FFNN_\alpha(x_t) \tag{12}$$

$$a_{\mathrm{i,t}} = \frac{exp(\alpha_t)}{\sum_{k=START(i)}^{END(i)} exp(\alpha_k)} \tag{13}$$

$$\hat{x}_i = \sum_{t=START(i)}^{END(i)} a_{\mathrm{i,t}} \cdot x_t \tag{14}$$

where $\hat{x}_i$ is is the weighted sum of the word vectors x_t belonging to the span i, and $a_{\mathrm{i,t}}$ are automatically learned weights.

IV. EXPERIMENTAL SETUP AND METRICS

Hereafter the experimental setup and the evaluation metrics are described in Section IV-A and IV-B respectively.

A. EXPERIMENTAL SETUP

To realize the proposed system, the implementation [2] of the coreference model proposed by [42] has been used, by exploiting ELECTRA embeddings to calculate span representations. ELECTRA model in its base (cased) version [3] has been tested, which is made available by Hugging Face Transformers [4] framework. This framework provides state-of-the-art Transformer-based architectures with thousands of pre-trained models in over a hundred languages for NLP tasks. In particular, this specific ELECTRA model has been pretrained on a corpus of 81GB, made of a recent Wikipedia dump, various texts from the OPUS [73] corpora collections and the Italian part of the OSCAR corpus [74].

In order to assess the effectiveness of the coreference resolution model integrating ELECTRA, further experiments have been arranged, where ELECTRA is replaced by BERT also considering its base (cased) version [5]. To the best of our knowledge, no other available implementations exist for the particular coreference resolution task in Italian. The choice of using BERT-base in the coreference resolution model as a valid option for comparison is justified also by the fact that it has proven to be effective in the coreference resolution task in English, as shown in [11]–[13].

Hyperparameter	Value
Epochs	24
Dropout	0.3
Learning rate	from 0.1 up to 0.00001
Loss	marginalized
Feature Embedding size	20
Max Span Width	30
Max training Sentences	6
Max segment Length	256
Dimensions Hidden State	256
Number of Attention Heads	12
Number of Hidden Layers	12
Hidden size	768
Number of Hidden Layers	12
Parameters	110M
Vocabulary Size	32102

TABLE 3. Hyper-parameters

In detail, the architectures of ELECTRA and BERT are characterised by 12 encoder layers, known as Transformers Blocks, and 12 attention heads (or Self-Attention as introduced in [75]), hence feed forward networks with a hidden size of 768. Each training session has been fixed of 24 epochs, with a learning rate of 0.1 . More architectural details and training hyper-parameters are reported in Table 3. All experiments have been performed performed on a deep learning workstation, with 40 Intel(R) Xeon(R) CPUs E5-2630 v4 @ 2.20GHz, 256 GB of RAM and 4 GPUs GeForce GTX 1080 Ti. The operating system is Ubuntu Linux 16.04.7 LTS.

Using the division of OntoCorefIT in the training, validation and testing datasets shown in the table 2, the results have been derived by averaging the performance of the coreference system integrating ELECTRA or BERT over five repetitions and finally reporting the arithmetic mean of the results, rounded to the second decimal place.

B. EVALUATION METRICS

Concerning evaluation metrics, there are still open issues in the literature and several metrics have been proposed, each of which tries to address biases of the earlier ones. For the purpose of this work, official metrics provided by the Conll 2012 shared task have been taken into account. In particular, the MELA metric has been adopted [76], which combines three different metrics addressing different dimensions: MUC [77], $B-CUBED$ [78] and $CEAF_e$ [79]. These metrics consider the true set of entities K (named key or key partition) obtained through manual annotation of the

[2] https://github.com/lxucs/coref-hoi

[3] https://huggingface.co/dbmdz/electra-base-italian-xxl-cased-discriminator

[4] https://github.com/huggingface/transformers

[5] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

Model	run	Identification of Mentions			Coreference									avg F1
					MUC			*B-CUBE*			$CEAF_e$			
		R	P	F1	R	P	F1	R	P	F1	R	P	F1	
ELECTRA cased-xxl	AVG	83,21	89,48	86,23	77,34	80,57	78,72	68,08	71,45	69,71	62,60	73,91	67,78	72,13
BERT cased-xxl	AVG	80,00	89,33	84,39	73,44	79,56	76,38	64,19	70,83	67,34	59,25	72,24	65,10	69,60

TABLE 4. Comparative results achieved with the coreference resolution system using ELECTRA and BERT models

entities, and the predicted (or response) set of entities R, i.e. answer partition produced by the system. In particular, they are defined as follows:

- MUC is a link-based metric. It compares the entities defined by the links in the key and the response. MUC considers a cluster of references as linked references, each reference is linked to one or more references. MUC metric primarily measures the number of link modifications required to make the result entity set R identical to the key entity set R. Precision is calculated as follows:

$$_{\text{MUC}}Precision = \frac{\sum_{r_j \in R} \frac{|r_j| - |P(r_j)|}{|r_j|}}{\sum_{r_j \in R}(|r_j| - 1)} \tag{15}$$

  while recall is equal to:

$$_{\text{MUC}}Recall = \frac{\sum_{k_i \in K} \frac{|k_i| - |P(k_i)|}{|k_i|}}{\sum_{k_i \in K}(|k_i| - 1)} \tag{16}$$

- $B-CUBED$ (B^3) is a mention-based metric. It first computes precision and recall for each mention, and then calculates the weighted average of these individual precision and recall scores to obtain global precision and recall. In particular, for each mention m of K, the Recall is computed by considering the fraction of the correct mentions included in the predicted entity that contains m. On the other hand, the precision is computed by exchanging the gold entities with the predicted ones. If K is the key entity containing mention m, and R is the response entity containing mention m, precision and recall for the mention m are calculated as:

$$_{\text{B}^3}Precision = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|}{|k_i|}}{\sum_{r_j \in R} |r_j|} \tag{17}$$

$$_{\text{B}^3}Recall = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|}{|k_i|}}{\sum_{k_i \in K} |k_i|} \tag{18}$$

- $CEAF_e$ is an entity-based metric. It is a particular instance of CEAF, a metric based on the assumption that each entity of K is mapped to a single entity of R, and vice versa. It uses a similarity measure to find the best one-to-one mapping between entities in K and entities in R. The best mapping is the one that maximizes the overall similarity of the entities, ϕ. In the case of $CEAF_e$, ϕ is given by the following equation:

$$\phi(k_i, r_i) = \frac{2|k_i \cap r_j|}{|k_i| + |r_i|} \tag{19}$$

  Recall is equal to the total similarity divided by the number of mentions in K:

$$_{\text{CEAF}}Recall = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{k_i \in K} \phi(k_i, k_i)} \tag{20}$$

  where g^* is a function that associates to every entity of K an entity of R, whereas K^* is the set of key entities included in the optimal mapping.
  Precision is the total similarity divided by the number of mentions in R:

$$_{\text{CEAF}}Precision = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{r_i \in R} \phi(r_i, r_i)} \tag{21}$$

According to [24], [80], the combination of these three metrics allows for a good compromise, balancing the limits of the various measures. It is worth noting that an evaluation metric must possess two properties, interpretability, which indicates the goodness of the detected entities and coreference, and discriminability, which allows to distinguish between good and not good decisions. $B-CUBED$ and $CEAF_e$ provide the property of discriminability, but not interpretability. On the other hand MUC benefits from interpretability, but it is the least discriminating among the three metrics. Since none of the three is reliable if taken individually, it is common practice to use the average of the three as the overall metric.

V. RESULTS AND DISCUSSION

This section presents and discusses the results achieved from both a quantitative and qualitative perspective.

A. QUANTITATIVE ANALYSIS

Table 4 reports results obtained using ELECTRA and BERT coupled with the coreference resolution system with respect to the three metrics described above. In particular, results are detailed with respect to two sub-tasks: *Identification of mentions* and *Coreference*, according to the criteria proposed in Conll 2012 shared task.

The first sub-task calculates the correctness of the mentions that are produced, without considering the coreference link structure, in other words, it is not verified that mentions refer to the right entity; the second one performs this check,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

evaluating the correctness of the expected coreference links between the key and answer mentions.

Concerning *Identification of mentions* task, the impact of ELECTRA can be seen in the average improvement of about 2 percentage points. Results of *Coreference* task according to three different metrics (columns MUC, B-CUBE and CEAF$_e$ of the table) show a slight drop in performance with respect to the first task in both models. This result is not surprising since the higher complexity of the coreference task justifies lower results.

With respect to the metrics used, MUC, being link-based, achieves better performance on precision and recall, respectively; while Ceaf$_e$ has the lowest scores, especially with regard to recall, which does not even reach 63% using the system with ELECTRA and and is even lower than 60% when BERT is used in place of ELECTRA. B-CUBE average scores are quite similar to those obtained with Ceaf$_e$, even if there is a change in behaviour: the recall is higher than Ceaf$_e$, to the detriment of precision which achieves the values of 71,45% and 70,83% when the system uses ELECTRA and BERT respectively.

Even there is still no agreement in the literature on the reliability of coreference resolution metrics, the best result obtained with MUC is consistent with other work on the English language [11]. One possible explanation for this result is that MUC has been proven to be robust despite being the least discriminating metric [81]. A further reason is that MUC is particularly suitable on data with many links per mention, as in the case of OntoCorefIT.

B. ERROR ANALYSIS

A further error analysis is reported, which assesses the performance of the coreference resolution system integrating ELECTRA model with respect to the different Parts-Of-Speech (POS) each mention belongs to.

Table 5 reports the frequency of occurrence of single-token coreferences occurring in the OntoCorefIT dataset, grouped in relation to their POS ordered in a decreasing manner.

Numbers in the table confirm findings in other studies [82], [83] that indicate pronouns as the primary POS used to co-refer to an entity.

Part-of-Speech	Tag	Percentage
Pronouns	PRON	33, 6%
Proper Nouns	PROPN	10, 6%
Verbs	VERB	8, 6%
Determiners	DET	6, 8%
Nouns	NOUN	1, 08%
Adverbs	ADV	1, 04%

TABLE 5. Distribution of most frequent single-token coreferences grouped with respect to their POS in the OntoCorefIT dataset

Table 6 outlines the numbers of single-token mentions that are predicted (p) or not predicted (np) correctly by the system using ELECTRA or BERT, grouped with reference to the three most frequent POSs associated, namely pronouns (PRON), proper nouns (PROPN) and verbs (VERB).

		ELECTRA		BERT	
POS	*tot*	*p*	*np*	*p*	*np*
PRON	1053	991 (94,1%)	62 (5,8%)	983 (93,3%)	70 (6,6%)
personal	*553*	*531 (96%)*	*22 (3,9%)*	*523 (94,5%)*	*30 (5,4%)*
possessive	*362*	*354 (97,7%)*	*8 (2,2%)*	*352 (97,2%)*	*10 (2,7%)*
demonstrative	*138*	*106 (76,8%)*	*32 (23,1%)*	*108 (78,2%)*	*30 (21,7%)*
PROPN	639	605 (94,6%)	34 (5,3%)	597 (93,4%)	42 (6,5%)
VERB	525	392 (74,6%)	133 (25,3%)	372 (70,8%)	153 (29,1%)

TABLE 6. Predicted (*p*) and not-predicted (*np*) mentions for each POS.

Table 6 shows discrepancies in the effectiveness of predictions in relation to different part-of-speech categories for the coreference resolution system using ELECTRA or BERT models. Concerning most frequent categories, namely pronouns and proper nouns, the system integrating ELECTRA model reaches a percentage of wrong predictions that is lower than 5,8% and 5,3% respectively. On the other hand, the system integrating BERT model obtains scores that are slightly lower, with 6,6% of not predicted pronouns and 6,5% of not predicted proper nouns. Concerning verbs, the system has an error rate of more than 25% when using both the models, but the impact on the overall results is low due to the fact that verbs account for only 8,6% of the total POS categories occurring in the dataset (see Table 5). This result can be explained also considering the criteria used to construct the OntoCorefIt dataset. Indeed, mentions have been frequently shifted to verbs - following the approach already proposed in [84], [85] - when the subject could be omitted, as in the case of personal pronouns with subject role (*pronoun-dropping*), adding a higher ambiguity with respect to final identification of mentions belonging to the verb POS.

A more detailed analysis of pronouns shows a different trend between correctly predicted mentions and not. Indeed, the pronouns category is the most representative (33,6% of single-token coreferences) and it includes various types of pronouns with different functions and syntactic structures, ranging from subject and object pronouns to possessives and demonstratives.

Substantial differences can be noted between the predicted and erroneous pronoun types (as shown in Table 6). More than 80% of correctly predicted pronouns belongs to possessives and personal pronouns: respectively 535 and 362 out of a total of 1053 for the system using ELECTRA while 983 and 523 for the system adopting BERT. In particular, when ELECTRA is used instead of BERT, the system predicts both types of pronouns with a greater accuracy, with an increase of two percentage points in the case of personal pronouns.

Mention predictions are much more inaccurate in case of demonstratives; the system using both the models achieves even 80% of correct predictions. This is also the only case where the system using BERT performs better than when ELECTRA is adopted, achieving 108 of the 138 correctly predicted mentions (78,2%), while stopping with ELECTRA at 106 (76,8%). This is probably explained by the complexity

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

PoS	ELECTRA	BERT
PRON	Uno dei suoi compiti come **presidente**... Pensi che lei lo sta salvando per il libro? (One of **her** tasks as president... You think she is saving him for the book?) Non ha previsto il risultato di **lui**. Ma ha rifiutato ciò (He did not anticipate **his** findings. But he refused it)	Questi hanno scritto oggi... Lei non ha condiviso le note con **loro** (They wrote today... She did not share the notes with **them**) Ho pensato a lungo a questo... Molti criticano *ciò* (I thought about this for a long time... Many people criticise *this*)
PROPN	Credo che la Cina sia un concorrente... La **nazione** controlla un sacco del debito (I believe that China it is a competitor. The country controls a lot of the debt) Elia è venuto da noi. *Ha dato* un sacchetto... (Elijah came to us. [He] *gave* a bag...)	L'ex avvocato di Clinton... Sono mosse accuse contro di **lui** (Clinton's former lawyer... Allegations are made against **him**) Iraq ha trionfato sul male dell'Occidente ...il *Paese* crede questo (Iraq triumphed over the evil of the West ... the *Country* believes this
VERB	Abbiamo visto Leo... **Pensava** che potesse dividere gli Stati Uniti. (We saw Leo... He **thought** it could divide the United States) [Lei] Non ha visto alcuna luce... [Lei] *Diventerà* la tutrice legale (She saw no light... She *will become* the legal guardian)	Non aveva Ken Starr a confutare... molti hanno seguito il **processo** (He didn't have Ken Starr to refute...many followed the **trial**) Ci riferivamo a esso... è sempre difficile *farlo* (We were referring to it... it is always difficult *to do that*)

TABLE 7. Examples of correct (bold) and incorrect (italic) predictions with respect to single-token mentions (underline) for most frequent Part-of-Speech categories.

of this pronominal category in the Italian language. Demonstratives present in fact several difficulties in Italian, both at morphological and a syntactic level. They are characterized - like possessives - by several inflected form, depending on gender and number and on the initial letter of the word they refer to. In addition, they are usually involved in long-distance dependencies and subject-object clause construction, i.e. "Ed è **questo** che ho deciso di chiedere" (*And **that** is what I decided to ask* in English). Finally, a particular case within the group of demonstratives is the pronoun "ciò" ("that" in English) that occupies 15% of the wrong prediction. "Ciò" is invariant neuter pronoun and its lack of inflection produces ambiguities and leads to a number of challenges in some situation, i.e. subject-verb agreement.

C. QUALITATIVE ANALYSIS

To deepen the analysis on the typology of errors, a qualitative analysis of the performance of the coreference system comparing the results achieved by using ELECTRA and BERT language models has been carried out with reference to OntoCorefIT dataset.

Table 7 shows a snippet containing an example related to most frequent POS in the dataset for each language model used. **Bold** text indicates correct prediction with respect to underlined mentions, while incorrect assignments are shown using *italic* text. Square brackets indicate dropped subject pronoun in Italian.

Concerning pronouns, the system using ELECTRA shows a correct prediction in a interrogative sentence with a relative clause. The singular feminine third-person personal pronoun "lei" (*her*) acting as a mention, has the role of subject into the relative clause "pensi che lei lo sta salvando" (*You think she is saving him*) which refers to the noun "presidente" (*president*) in the main clause. But the second sentence contains an error "Non ha previsto il risultato di lui... Ma ha rifiutato ciò" (*He did not anticipate his findings, But he refused it*), even if it has a simple syntax with no subordinates. The sentence has a negative construction and the mention is the neuter pronoun "ciò" (*it*) in a object position. A similar behaviour can be observed for the same POS category for the system using BERT model (third column of the table 7). The correctly predicted mention occurs as subject "Questi hanno scritto oggi" (*They wrote today*), while the system with BERT fails the correct assignment when the mention occurs as indirect object introduced by a preposition "a questo" (*about this*).

A similar behaviour can be observed for the proper nouns POS. The system using ELECTRA is able to correctly predict the proper noun with object function inside the relative clause introduced by the preposition "Credo che la Cina" (*I believe that China*), while proper noun "Elia" (*Elijah*) is not correctly referred by verb "[Egli] Ha dato" (*He gave*). Maybe the error in an elementary sentence such as this one composed of the simple Subject-Verb-Object (SVO) order is given by the dropping of the subject, which remains implicit in Italian, by making the correct prediction more difficult. The system's behaviour with BERT for this POS category is more ambiguous, as shown with two example utterances presenting a similar syntax. In the first example, the proper noun acting as subject is into a prepositional phrase "L'ex avvocato di Clinton" (*Clinton's former lawyer*) and it is correctly predicted. The second utterance has a construction even simpler with the mention as subject in preverbal position at the beginning of the sentence "Iraq". Despite this, it is not correctly predicted. However, it is important to note that the error rate for proper nouns is the lowest of all the POSs (5,3% and 6,6% respectively), therefore the typology of not predicted sentences is not very representative.

The scenario is different concerning the verb POS, the grammatical category for which there is a smaller gap between correct predictions and incorrectly predicted relationships (see table 6). Even more than in the previous PoS, two factors seem to influence whether or not the prediction is correct: position of the mention (subject/object) and subject expressed or omitted. For instance, in the example utterance "Abbiamo visto Leo... Pensava che potesse dividere gli Stati Uniti" (*We saw Leo... He thought it could divide the United States*), the system with ELECTRA correctly associates the verb mention holding the relative clause "**Pensava** che..." (*He **thought***), to the noun "Leo". This sentence may appear more articulated, since it does not have a linear syntax and it presents a subordinate clause. However, it should be noted that the coreference "Leo" is in the object position, therefore the prediction must not refer to an implicit element that has been shifted to the adjacent verb, as in the next example. For

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

the sentence "Diventerà la tutrice legale" (*She will become the legal guardian*) the system with ELECTRA generates a wrong prediction. The verb "diventerà" (*will become*) is tagged as mention, since the subject pronoun [lei] (*She*) is not expressed in the Italian sentence. In addition, this sentence presents two omitted subject pronoun, both for mention and for coreference, thus increasing the difficulty in making a correct prediction. By contrast, the system's behaviour with BERT is slightly better in this category. An example of correctly predicted mention for the verbs category is represented by a simple SVO utterance with a negative construction "Non aveva Ken Starr..." (*He didn't have Ken Starr...*). The last utterance "Ci riferivamo a esso... è sempre difficile farlo" (*We were referring to it... it is always difficult to do that*) shows an error of the system using BERT since a particular case of verb PoS is occurred. Indeed, the utterance contains a dative construction with a clitic pronoun "Ci" (literally *us*) preceding the mention "riferivamo" (*were referring*) and an enclitic form merged with the verb in the form of suffix -lo for the coreference "farlo" (*to do that*).

VI. CONCLUSION AND FUTURE WORK

In this paper a coreference resolution system for the Italian language has been presented. The system is based on an end-to-end architecture and it uses ELECTRA as a language model to represent input tokens. For the training step, OntoCorefIT dataset has been used. It is the biggest coreference resolution dataset for the Italian language, built using a cross-lingual methodology starting from OntoNotes. As far as is known, there are currently no existing works for Italian based on neural networks for coreference, nor applications that use ELECTRA as language model.

Therefore, the performances of the system integrating ELECTRA have not been assessed with a direct comparison with other existing solutions for the Italian Language, but they have been evaluated by considering BERT in place of ELECTRA in its base (cased) version. Experiments with the system using ELECTRA have achieved better results than using BERT. Scores have been calculated according to *de facto* standard metrics proposed in Conll2012 task, with respect to the sub-tasks *Identification of mentions* and *Coreference*, showing an increase of at least two percentage points with the system with ELECTRA with respect to BERT. In detail, the system with ELECTRA or with BERT achieves a F_1 score of 72,13% and 69,60% for *Identification of mentions* task and of 86,23% 84,39% for *Coreference* task, respectively.

According to the results, an error analysis has been carried out aiming at evaluating which grammatical category (PoS) is mostly error-prone for single-token coreferences. Worse performances have been shown for the VERB category, since in OntoCorefIT dataset, mentions have been frequently shifted to verbs, in case the subject could be omitted, to accommodate the syntactic constructions of Italian. Hence, this has added a higher ambiguity in the final recognition. A deeper analysis has been conducted with respect to mentions belonging to the pronouns category, showing better performances with possessives and personal pronouns with respect to demonstratives, due to a high variability related to morphological and a syntactic aspects. Finally, a qualitative analysis has been conducted to compare the behaviour of the system when ELECTRA or BERT is used as language model, also reporting examples of correct and incorrect predicted mentions.

Further studies could be conducted on the possibility of using other language models specifically for the coreference resolution task. Furthermore, from a strictly linguistic point of view, a deeper analysis not only limited to grammatical categories but also to other linguistic features affecting performances could be carried out as future work.

REFERENCES

[1] Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. 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). Association for Computational Linguistics, 2015.

[2] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[3] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. arXiv preprint arXiv:1609.08667, 2016.

[4] Changki Lee, Sangkeun Jung, and Cheon-Eum Park. Anaphora resolution with pointer networks. Pattern Recognition Letters, 95:1–7, 2017.

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

[6] Sarah E Blackwell. Testing the neo-gricean pragmatic theory of anaphora: The influence of consistency constraints on interpretations of coreference in spanish. Journal of Pragmatics, 33(6):901–941, 2001.

[7] Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 1396–1411, 2010.

[8] Kenton Lee, Luheng He, and Luke Zettlemoyer. 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, 2018.

[9] Hongliang Fei, Xu Li, Dingcheng Li, and Ping Li. End-to-end deep reinforcement learning based coreference resolution. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 660–665, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Ben Kantor and Amir Globerson. Coreference resolution with entity equalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 673–677, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 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, November 2019. Association for Computational Linguistics.

[12] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77, 2020.

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

[14] Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In Proceedings of the human

language technology conference of the NAACL, Companion Volume: Short Papers, pages 57–60, 2006.

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

[16] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8:842–866, 2020.

[17] Jerry R. Hobbs. Resolving pronoun references. Lingua, 44(4):311–338, Apr 1978.

[18] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse,. Jan 1995.

[19] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. Computational linguistics, 20(4):535–561, 1994.

[20] Ruslan Mitkov. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. Applied Artificial Intelligence, 15(3):253–276, Mar 2001.

[21] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

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

[23] Sameer Pradhan, Lance Ramshaw, Mitch Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–27, 2011.

[24] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 1–40, 2012.

[25] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In Proceedings of the 2009 conference on empirical methods in natural language processing, pages 1152–1161, 2009.

[26] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. A multi-pass sieve for coreference resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 492–501, 2010.

[27] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational linguistics, 39(4):885–916, 2013.

[28] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In Sixth Workshop on Very Large Corpora, 1998.

[29] HEEYOUNG LEE, Mihai Surdeanu, and DAN JURAFSKY. A scaffolding approach to coreference resolution integrating statistical and rule-based models. Natural Language Engineering, pages 1–30, 2017.

[30] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 104–111, 2002.

[31] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4):521–544, Dec 2001.

[32] Andrew McCallum and Ben Wellner. Conditional models of identity uncertainty with application to noun coreference. Advances in neural information processing systems, 17:905–912, 2004.

[33] Xiaoqiang Luo. On coreference resolution performance metrics. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05, 2005.

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

[35] Jenny Rose Finkel and Christopher D. Manning. Enforcing transitivity in coreference resolution. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08, 2008.

[36] Eraldo Fernandes, Cicero dos Santos, and Ruy Luiz Milidiú. Latent structure perceptron with feature induction for unrestricted coreference resolution. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 41–48, 2012.

[37] Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. Improving noun phrase coreference resolution by matching strings. Lecture Notes in Computer Science, page 22–31, 2005.

[38] Xiaofeng Yang, Jian Su, and Chew Lim Tan. A twin-candidate model for learning-based anaphora resolution. Computational Linguistics, 34(3):327–356, Sep 2008.

[39] Pascal Denis and Jason Baldridge. Specialized models and ranking for coreference resolution. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08, 2008.

[40] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1971–1982, 2013.

[41] Kevin Clark and Christopher D Manning. Improving coreference resolution by learning entity-level distributed representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 643–653, 2016.

[42] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, 2017.

[43] Yannick Versley. Using the Web to resolve coreferent bridging in German newspaper text. Citeseer, 2007.

[44] Michael Strube, Stefan Rapp, and Christoph Müller. The influence of minimum edit distance on reference resolution. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 312–319, 2002.

[45] Don Tuggener. Incremental coreference resolution for German. PhD thesis, University of Zurich, 2016.

[46] Fernando Acerenza, Macarena Rabosto, Magdalena Zubizarreta, Aiala Rosá, and Dina Wonsever. Coreference resolution between sources of opinions in spanish texts. In 2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI), pages 1–8. IEEE, 2012.

[47] André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. Exploring spanish corpora for portuguese coreference resolution. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), pages 290–295. IEEE, 2018.

[48] Anna Nedoluzhko and Jiří Mírovskỳ. How dependency trees and tectogrammatics help annotating coreference and bridging relations in prague dependency treebank. In Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 244–251, 2013.

[49] Emmanuel Lassalle and Pascal Denis. Leveraging different meronym discovery methods for bridging resolution in french. In Discourse Anaphora and Anaphor Resolution Colloquium, pages 35–46. Springer, 2011.

[50] Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. Bart: A multilingual anaphora resolution system. In Proceedings of the 5th international workshop on semantic evaluation, pages 104–107, 2010.

[51] Massimo Poesio, Olga Uryupina, and Yannick Versley. Creating a coreference resolution system for italian. In LREC, 2010.

[52] Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 1–8, 2010.

[53] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA, 23, 2013.

[54] Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. Parcor 1.0: A parallel pronoun-coreference corpus to support statistical mt. In 9th International Conference on Language Resources and Evaluation (LREC), MAY 26-31, 2014, Reykjavik, ICELAND, pages 3191–3198. European Language Resources Association, 2014.

[55] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2089–2096, 2012.

[56] Joakim Nivre. Towards a universal grammar for natural language processing. In International conference on intelligent text processing and computational linguistics, pages 3–16. Springer, 2015.

[57] Sandra Kübler and Desislava Zhekova. Multilingual coreference resolution. Language and Linguistics Compass, 10(11):614–631, 2016.

[58] Altaf Rahman and Vincent Ng. Translation-based projection for multilingual coreference resolution. In Proceedings of the 2012 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 720–730, 2012.

[59] Yulia Grishina and Manfred Stede. Knowledge-lean projection of coreference chains across languages. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 14–22, 2015.

[60] André FT Martins. Transferring coreference resolvers with posterior regularization. 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 1427–1437, 2015.

[61] Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtskỳ. Projection-based coreference resolution using deep syntax. In Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), pages 56–64, 2017.

[62] Gourab Kundu, Avirup Sil, Radu Florian, and Wael Hamza. Neural cross-lingual coreference resolution and its application to entity linking. arXiv preprint arXiv:1806.10201, 2018.

[63] Hannaneh Hajishirzi, Leila Zilles, Daniel S Weld, and Luke Zettlemoyer. Joint coreference resolution and named-entity linking with multi-pass sieves. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 289–299, 2013.

[64] Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, et al. The italian syntactic-semantic treebank: Architecture, annotation, tools and evaluation. 2003.

[65] A Bristot, L Chiran, and R Delmonte. Verso un'annotazione xml di dialoghi spontanei per l'analisi sintattico-semantica. XI Giornate di Studio GFS, Multimodalita'e Multimedialit nella comunicazione, pages 42–50, 2000.

[66] John H McWhorter. The worlds simplest grammars are creole grammars. Linguistic typology, 5(2-3):125–166, 2001.

[67] Charles A Ferguson. Simplified registers and linguistic theory. Exceptional language and linguistics, 49:66, 1982.

[68] Dominique Brunato and Felice Dell'Orletta. On the order of words in italian: a study on genre vs complexity. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 25–31, 2017.

[69] Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 146–182, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[70] Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. Cross-framework evaluation for statistical parsing. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 44–54, 2012.

[71] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.

[72] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[73] Jörg Tiedemann. Opus–parallel corpora for everyone. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, 2016.

[74] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache.

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017.

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

[77] Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995, 1995.

[78] A BAGGA. Algorithms for scoring coreference chains. In Proc. Linguistic Coreference Workshop at the first Conf. on Language Resources and Evaluation (LREC), Granada, Spain, May 1998, 1998.

[79] Xiaoqiang Luo. On coreference resolution performance metrics. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 25–32, 2005.

[80] Lluís Màrquez, Marta Recasens, and Emili Sapena. Coreference resolution: an empirical study based on semeval-2010 shared task 1. Language resources and evaluation, 47(3):661–694, 2013.

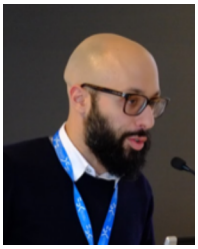
[81] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 632–642, 2016.

[82] Mark-Christoph Müller. Fully automatic resolution of'it','this', and'that'in unrestricted multi-party dialog. 2008.

[83] Mijail Alexandrov Kabadjov. A comprehensive evaluation of anaphora resolution and discourse-new classification. PhD thesis, Citeseer, 2007.

[84] David McKelvie, Amy Isard, Andreas Mengel, Morten Baun Møller, Michael Grosse, and Marion Klein. The mate workbench – an annotation tool for xml coded speech corpora. Speech Communication, 33(1):97 – 112, 2001. Speech Annotation and Corpus Tools.

[85] Kepa Joseba Rodrıguez, Francesca Delogu, Yannick Versley, Egon W Stemle, and Massimo Poesio. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In Proceedings of LREC, pages 157–163, 2010.



RAFFAELE GUARASCI is currently a researcher at the Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He received his M.Sc. in Computational Linguistics from University of Pisa, and a Ph.D. degree in Computational Linguistics from the University of Salerno. He is a member of the "Cognitive Systems" laboratory at CNR-ICAR. He has been adjunct professor in the PhD program in Computational Linguistics & Social Media at the University of Salerno and he has been on the program committee of national and international conferences. His current research interests range from the interaction between neural language models and linguistics theoretical and cognitive issues to the application of Artificial Intelligence methods based on Machine/Deep Learning to Natural Language Processing tasks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3105278, IEEE Access

ANIELLO MINUTOLO is currently a researcher at the Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He received his M.Sc. in Computer Science Engineering from University of Naples "Federico II", and a Ph.D. degree in Information Technology Engineering from the University of Naples "Parthenope". Since 2018, he has been a contract professor of Informatics at the University of Naples "Federico II", Faculty of Engineering. His current research interests include Artificial Intelligence, Decision Support Systems, Dialog Systems, and Knowledge management, modelling and reasoning. He has been involved in different national and European projects, he has been on the program committee of some international conferences and workshops and, moreover, is currently a member of the editorial board of some international journals.

EMANUELE DAMIANO is currently a graduate research fellow at the Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He received his MSc degrees in Computer Science from the University of Naples "Federico II". He is a member of the "Cognitive Systems" laboratory at CNR-ICAR and been involved in different national projects. His research focuses on Artificial Intelligence approaches based on Machine/Deep Learning applied to Natural Language Processing and Question Answering.

GIUSEPPE DE PIETRO is currently the Director of the Institute for High Performance Computing and Networking, CNR, and an Adjunct Professor with the College of Science and Technology, Temple University, Philadelphia. He has been actively involved in many European and national projects, with industrial co-operations. He is the author of over 200 scientific articles published in international journals and conferences. His current research interests include cognitive computing, clinical decision support systems, and software architectures for e-health. He is also a KES International Member. He is also involved in many program committees and journal editorial boards.

HAMIDO FUJITA (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Manchester, Manchester, U.K., in 1979, and the master's and Ph.D. degrees in information engineering from Tohoku University, Sendai, Japan, in 1985 and 1988, respectively. He is currently a Professor of Artificial Intelligence with Iwate Prefectural University, Takizawa, Japan, as a Director of Intelligent Software Systems. He received Doctor Honoris Causa from Óbuda University, Budapest, Hungary, in 2013 and also from Timisoara Technical University, Timisoara, Romania, in 2018, and a title of Honorary Professor from Óbuda University, in 2011. He is an Adjunct Professor of Computer Science and Artificial Intelligence with Stockholm University, Stockholm, Sweden; University of Technology Sydney, Ultimo, NSW, Australia; National Taiwan Ocean University, Keelung, Taiwan, and others. He has supervised Ph.D. students jointly with the University of Laval, Quebec City, QC, Canada; University of Technology Sydney; Oregon State University, Corvallis, OR, USA; University of Paris 1 Pantheon-Sorbonne, Paris, France; and University of Genoa, Genoa, Italy. Dr. Fujita is the recipient of the Honorary Scholar Award from the University of Technology Sydney, in 2012. He has four international patents in software system and several research projects with Japanese industry and partners. He is the Editor-in-Chief for Knowledge-Based Systems. He is the Vice President of International Society of Applied Intelligence, and currently Editor-in-Chief of Applied Intelligence (Springer). He is also Highly Cited Researcher in Cross-field for the year 2019 and Computer Science for the year 2020 respectively by Clarivate Analytics. He has given many keynotes in many prestigious international conferences on intelligent system and subjective intelligence. He headed a number of projects including intelligent HCI, a project related to mental cloning for healthcare system as an intelligent user interface between human users and computers, and SCOPE project on virtual doctor systems for medical applications.

MASSIMO ESPOSITO is currently a researcher at the Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He received his M.Sc. in Computer Science Engineering (Cum Laude) from University of Naples Federico II' in March 2004. He received a University 1st level Master degree, named European Master on Critical Networked Systems in December 2007, and a Ph.D. degree in Information Technology Engineering in April 2011 from the University of Naples Parthenope. Since 2012, he has been a contract professor of Informatics at the University of Naples "Federico II", Faculty of Engineering. Since 2016, he has been responsible of the laboratory "Cognitive Systems" at ICAR-CNR. His current research interests are in the field of Artificial Intelligence (AI) and are focused on AI algorithms and techniques, mixing deep learning and knowledge-based technologies, for building intelligent systems able to converse, understand natural language and answer to questions, with emphasis on the distributional neural representation of words and sentences,and on specific natural language tasks such as part of speech tagging, sentence classification and open information extraction. He has been involved in different national and European projects, he has been on the program committee of many international conferences and workshops and, moreover, is currently a member of the editorial board of some international journals. He is author of over 100 peer-reviewed papers on international journals and conference proceedings.