Exploring Spanish Corpora for Portuguese Coreference Resolution

André Ferreira Cruz
Faculdade de Engenharia
Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
Email: andre.ferreira.cruz@fe.up.pt

Gil Rocha
LIACC
Faculdade de Engenharia
Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
Email: gil.rocha@fe.up.pt

Henrique Lopes Cardoso
LIACC
DEI, Faculdade de Engenharia
Universidade do Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
Email: hlc@fe.up.pt

***Abstract*—The task of coreference resolution has attracted a great deal of attention in the literature due to its importance in deep language understanding, and its potential as a subtask in a variety of complex natural language processing problems. We experiment with different methods for generating training data and architectures for extracting meaningful mention representations. Coreference resolution in lesser-resourced languages is challenging, and transfer learning is a promising technique to overcome the comparatively smaller available corpora. We explore direct transfer learning from Spanish to Portuguese. We present state-of-the-art systems on both Spanish and Portuguese, and report promising results in a cross-language setting.**

I. Introduction

Coreference resolution is a natural language processing (NLP) task that comprises determining all linguistic expressions – or referring expressions – that refer to the same real-world entity. A referring expression (*i.e.* a *mention*) is either a noun phrase (NP), a named entity (NE), or a pronoun whose meaning is a reference to an entity or event in the real world – the *referent*. A grouping of referring expressions with the same referent is called a *coreference chain* or *cluster* [1].

The goal of a coreference resolution system is to output all the coreference chains of a given text. Addressing this problem typically requires addressing previous language processing tasks, such as parsing, named-entity recognition and part-of-speech tagging. Coreference resolution has a high-impact on several other NLP tasks, including textual entailment, summarization, information extraction, and question answering.

Figure 1 shows examples of sentences and their corresponding coreference chains. A classification algorithm could, for instance, use the hyponym/hypernym semantic relation between "bee" and "insect" to classify the two mentions as co-referent, and use world-knowledge to infer a strong relation between "Barack Obama" and "president". The handicaps machines exhibit when dealing with coreference resolution is evidenced by the use of this task in tests of machine intelligence, such as the "Winograd Schema Challenge" [2].

Despite some attempts to solve this problem with unsupervised methods [3], state of the art has consistently been driven by supervised machine learning [4], which presents a problem for low-resource languages (e.g. Portuguese). This predicament is sometimes tackled with transfer learning from models trained on large datasets of another language [5], and has been addressed in recent research tasks [6], [7].

The remainder of the paper is organized as follows. In Section II we explore the state of the art. In Section III we address the resources we use and the associated challenges. In Section IV we explore our experimental setup and methodology, as well as detail baselines, evaluation settings, and the performed experiments. In Section V we detail and discuss the results obtained by the systems described in this work. In Section VI we draw conclusions and discuss future work.

II. Related Work

With its roots in the 1960s, there have been numerous works on coreference resolution over time [4]. Additionally, it has been addressed in several tasks dating back to the sixth [8] and seventh [9] Message Understanding Conferences.

The typical architecture of a coreference resolution system includes a data preparation phase and a resolution phase (as seen in Fig. 2). Data preparation consists in the detection of mentions in the input text, followed by a feature extraction step that converts each data instance into an expressive feature-vector. The resolution phase consists in the binary classification of these instances as coreferent or not (or, in ranking systems, in the attribution of coreference scores), followed by the linking/clustering of mentions into the final coreference chains. These two steps of the resolution phase can be addressed either simultaneously or separately.

1. [**Bees**]$_0$ are critical to safeguarding [food supplies worldwide]$_1$. [These interesting **insects**]$_0$ have been hit hard by [climate change]$_2$.
2. [**Barack Obama**]$_0$, [the former **US president**]$_0$, has told [the country]$_1$ [he]$_0$'s ready for [a long vacation]$_2$.
3. [The city councilmen]$_0$ refused [the demonstrators]$_1$ [a permit]$_2$ because [they]$_1$ advocated violence.

Fig. 1. Coreference resolution examples. The third example was extracted from the "Winograd Schema Challenge" [2].

978-1-5386-9588-3/18/$31.00 ©2018 IEEE

Fig. 2. Typical architecture of a coreference resolution system. Figure based on one by Sapena *et al.* [1].

Performing classification and linking as two separate steps enables the use of global-optimization techniques in the linking phase, such as path-finding [10], clustering [11] or graph-partitioning algorithms [1]. In linking/clustering, it is also common to use heuristic-based approaches to link a mention with the best instance from a pool of positively identified antecedents (*e.g.* closest antecedent) [4]. Conversely, performing classification and linking simultaneously may lead to more informed decisions in the classification phase, as one can use features from a partially-formed coreference cluster to restrict future classifications (*e.g.* if the cluster has a well defined gender, only link new mentions with that same gender). Systems that use features related to the whole entity to make mention-wise linking decisions are called *entity-mention*.

In contrast, *mention-pair* models use only local information to classify mentions as coreferring or not. Similarly, mention-ranking models use mention-wise features to impose a ranking of candidate antecedents, determining a mention-pair score instead of classifying it as coreferring or not. Besides being simpler, these types of models can be followed by a clustering process that introduces global information into the problem.

Regarding the classification phase, traditional approaches were based on training linear models based on a set of hand-engineered features (*e.g.* string match among the two mentions, gender match, number match). More recently, Wiseman *et al.* [12] pioneered the use of a neural network to learn non-linear representations of raw data, improving the state-of-the art in this task. Following this trend towards *deep learning* models, Clark and Manning [13] and Wiseman *et al.* [14] further improved the state-of-the-art by incorporating global-level entity-based features into the non-linear model.

More recently, research has moved in the direction of end-to-end systems to solve the problem as a whole, with Lee *et al.* [15] jointly modeling mention-detection, coreference assessment, and a head-finding mechanism with impressive results, cutting reliance on external syntactic parsers. Although achieving state-of-the-art results, this model did still make locally informed mention-pair decisions. At the core of this work were vector embeddings representing spans of text in the document. These embeddings proved capable of representing the span's meaning, but still suffered from the fact that a single word's embedding was the same regardless of the context it was in [16]. In 2018, contextualized word embeddings were introduced by Peters *el al.* [17], improving state-of-the-art in several NLP tasks, including coreference resolution.

Following their own previous work, Lee *et al.* [18] tackled the lack of global information when assessing mention coreference by introducing an approximation to higher-order inference for coreference resolution, thus enabling the model to "softly consider multiple hops in the predicted clusters" [18] through an iterative process, and achieving the current highest score in the coreference CoNLL-2012 task [19].

Regarding coreference resolution in the Portuguese language, state-of-the-art has lacked behind more resourced languages, but direct comparison is complex as evaluation is obviously performed in different corpora. To the best of our knowledge, the coreference resolution systems reporting best results in an unrestricted Portuguese dataset are the works of Fonseca *et al.* [20] and of Rocha and Lopes Cardoso [21]. Both use hand engineered features and linear models for mention-pair classification with promising results.

Cross-lingual coreference resolution has been tackled by recent tasks, but these do not include the Portuguese language. Approaches in literature focus on projection-based coreference resolution [6], [7] and, most recently, direct transfer [5].

III. Resources

When using supervised machine learning techniques, as customary in the state-of-the-art for coreference resolution (see Section II), the availability of annotated corpora is an important requirement. Although large-scale corpora have been built for the English language, the most prominent being the OntoNotes 5.0 dataset [19], this type of corpora is not so mature for other languages. This scarcity poses a barrier to improving coreference resolution for lower-resourced languages, which we aim to overcome with transfer learning. A collection of available corpora for several languages is included in Table I, in chronological order.

The largest Portuguese dataset annotated with coreference information is Corref-PT [26], from late 2017. Corref-PT is approximately 8 times the size of the Summ-it++ corpus [25], a widely used resource in the Portuguese and Brazilian NLP communities. It is important to note that all Portuguese corpora

TABLE I
Coreference resolution corpora.

Corpus	*Language*	*#Tokens*	*#Docs*
MUC-6 [8]	English	25K	60
MUC-7 [9]	English	40K	67
ACE (2000-2004) [22]	English	960K	-
	Chinese	615K	-
	Arabic	500K	-
SemEval-2010 [23]	English	120K	353
	Catalan	345K	1138
	Dutch	104K	240
	German	455K	1235
	Italian	140K	143
	Spanish	380K	1183
OntoNotes v5.0 [19]	English	1.6M	2,384
	Chinese	950K	1,729
	Arabic	300K	447
Garcia *et al.* [24]	Portuguese	51K	97
	Galician	42K	57
	Spanish	46K	39
Summ-it++ [25]	Portuguese	20K	50
Corref-PT [26]	Portuguese	124K	182

discussed in this paper correspond to the Brazilian variant of this language, as European Portuguese corpora is even rarer.

For Spanish, another Latin language, the largest corpus available is the AnCora dataset [27], used as the Spanish section of the SemEval-2010 task [23], which resulted in the proposal of several coreference resolution systems [1], [28].

We use the AnCora corpus [27] for the Spanish language and the Corref-PT corpus [26] for the Portuguese language, which feature 380K and 124K tokens, respectively.

In addition to corpora resources, we use pre-trained word embeddings. This type of resources is widely used in the literature, serving as features to most recent state-of-the-art systems [13], [15], [17], [18]. In order to minimize traction in the model's context change (between datasets in different languages), we chose to use FastText multilingual word vectors [29]. These are 300-dimensional pre-trained word vectors whose vector spaces were aligned after training, meaning the Spanish and the Portuguese versions of a given word will have close vector representations in their respective embedding spaces. An additional advantage of FastText word vectors is their ability to predict representations for out-of-vocabulary words, which were not found frequently enough in the training phase but are found in the evaluation phase.

IV. Methodology

To tackle the task of coreference resolution we train deep neural networks with a variety of architectures. This section describes the training data generation, the linking algorithm used, the features used by the different models and their architectures, as well as the challenges faced in training. Furthermore, we discuss the direct transfer of learned model weights between Spanish and Portuguese.

Our experiments focus on the classification phase of the coreference resolution pipeline. We supply the model with gold-standard mention boundaries, and use a deterministic but proven linking algorithm: *closest antecedent* [30]. The *closest antecedent* algorithm consists in linking each mention with its closest positively identified antecedent, if one exists [30].

We follow the mention-pair model (described in Section II), as it is used in several recent state-of-the-art systems [15], and even more so in Portuguese [20], [21].

A. Training Set Creation

To transform the provided coreference annotations into a set of training instances suitable for the learning process, we create pairwise combinations of mentions by pairing each mention m_i with all its candidate antecedents m_j (mentions which appear before m_i). A learning instance is created for every pair $\langle m_i, m_j \rangle$: $\langle m_i, m_j, P \rangle$ if positively coreferent, or $\langle m_i, m_j, N \rangle$ if not coreferent.

This procedure generates a highly unbalanced dataset. On the Spanish AnCora dataset, 7,101,670 mention-pairs were generated from the 1183 documents, 190,834 of which were positive learning instances and 6,910,836 were negative learning instances, corresponding to a 2.7%/97.3% split. On the Portuguese Corref-PT dataset, 923,566 mention-pairs were generated, 45,659 of which were positive learning instances and 877,907 were negative learning instances, corresponding to a 4.9%/95.1% split. This class imbalance problem has been extensively studied in literature, and is usually tackled by using random undersampling of the majority class [20], [31]. We chose to perform our own study using one of the proposed architectures, aiming to identify which undersampling percentage is able to maximize performance on specific coreference metrics. Results are reported in Section V-A.

B. Evaluation

For comparison with SemEval-2010 systems, we report performance on SemEval-2010 metrics evaluated with the official SemEval scorer: MUC [32], B^3 [33], $CEAF_e$ [34], and $BLANC$ [35]. We also report performance on the official CoNLL metric [19]: the unweighted average of F1-scores of MUC, B^3 and $CEAF_e$. Additionally, due to using a recent corpus for Portuguese coreference resolution [26], we report the first coreference-specific results on the Corref-PT corpus. Direct comparison of our system with the works of Fonseca *et al.* [20] or Rocha and Lopes Cardoso [21] is not possible, as the performance of these systems on model-independent metrics is not reported, besides using a different corpus.

All results on the Spanish AnCora corpus are reported on the test portion of the dataset (as partitioned in SemEval-2010), using the development data for validation. Conversely, as the Corref-PT corpus is not split in training/test/development portions, we randomly select approximately 60% of documents for training, 20% for testing, and 20% for validation.

Since reporting single scores is insufficient to compare non-deterministic learning approaches [36], we report average scores of 5 runs with different random seeds. The performance on the mention-pairs test set corresponds to the sum of all experiments confusion matrices, reporting on precision, recall and F1 of the summed confusion matrix [37].

C. Feature Selection

All models receive as input a pair of 50-dimensional vectors, representing indexes in the embedding matrix of the words on both mentions $\langle m_i, m_j \rangle$, up to a maximum of 50 words per mention, as mentions can span multiple words. This way, in order to keep a constant-sized input, mentions which span more than 50 tokens/words (representing 0.10% of the total) are cropped, and mentions spanning less than 50 tokens are padded with a special embedding filled with zeros. Additionally, the distance in sentences and tokens between both mentions is also passed as input, binned into the buckets [1, 2, 3, 4, 5-7, 8-15, 16-31, 32-63, 64+], following Clark and Manning [13].

Despite having access to other useful features in the AnCora dataset (arguments and thematic roles, predicate semantic classes, and WordNet nominal senses), the Corref-PT dataset does not provide these features, and their use would mean employing error-inducing syntactic parsers on the latter corpus. As such, we chose to use only language-agnostic features that could be determined without errors at train and test-time: word-embeddings and distance features.

D. Architectures

In our experiments, we subdivide the models in two steps: the first concerns extracting representative features from mentions; the second focuses on assessing coreference affinity. The first is performed either using CNNs, LSTMs, or dense layers, following recent successes with these types of neural networks in NLP [38]; and the latter is performed using traditional dense layers. We experimented with 5 different model variations:

1) *Arch1*, composed by the following layers:
   a) an embedding layer whose vectors were obtained from a pre-parsed FastText file, containing the most common words found at training [29];
   b) the mentions embeddings are summed along the word axis, transforming (50, 300) tensors into (1, 300) tensors, 50 being the mention length, and 300 being the embedding dimensions;
   c) tensors from both mentions are stacked into a (2, 300) tensor and passed through a standard 1-D convolutional layer, with 64 output filters and window size of 2, outputting (1, 64) shaped tensors;
   d) this 64-dimensional representations are concatenated with the scalar distance features, then passed through a standard fully-connected layer (with 64 neurons), and a final *sigmoid*-activated fully-connected layer (with 1 neuron).
2) *Arch2*: this model's embedding layer was created by tokenizing the texts from the entire input dataset, loading the entire embeddings model, and leveraging FastText's ability to predict embeddings of out-of-vocabulary words [29] (not seen when learning the embeddings), thus ensuring that all words have some sort of representation, even though words seen during training will have a more representative embedding; remaining layers are the same as the previous architecture, *Arch1*.
3) *Arch2-dense*: embedding layer is the same as *Arch2*, and resulting embeddings are similarly summed into a pair of (1, 300) tensors, but these are then concatenated along with the distance features; these 602-dimensional tensors are then passed through two hidden layers (with 150 neurons, following Lee *et al.* [15]), and the final output layer (similar to the previous architectures).
4) *Arch-deep-CNN*: embedding layer is the same as *Arch2*, but instead of simply summing the word embeddings to form a mention embedding, the (50, 300) shaped tensors are passed through two 1-dimensional convolutional layers with 128 output filters each and windows size of 3; these tensors are then max-pooled along the whole time steps axis (1st axis), outputting 128-dimensional tensors, and followed by two hidden layers (with 150 neurons [15]), and the final output layer.
5) *Arch-biLSTM*: in order to better represent the time-dimension along the mention's tokens, this architecture feeds the (50, 300) tensors into a bidirectional LSTM layer [39]; the last state of this LSTM is then extracted and passed through two hidden layers (with 150 neurons), and the final output layer.

All hidden layers and convolutional layers are activated by a *relu* function [40]; all embedding layers, convolutional layers and LSTM units have a 40% dropout rate, and hidden layers have a 20% dropout rate [41]. These hyperparameters fit the problem well, and were based on Lee *et al.* [15], which uses an architecture with similar input, and then fine-tuned to our problem through extensive experimentation.

E. Baselines

To attest to our models' performance on popular coreference metrics, we developed 2 different random baseline approaches, and 1 deterministic baseline. Random baseline scores were averaged over 5 runs. Results are shown in Table II.

- *Rand1*: for every mention m_i, with 50% probability choose a random antecedent mention m_j, uniformly distributed between antecedents of m_i; otherwise select no antecedent for that mention.
- *Rand2*: for every mention m_i, with x probability, x being the percentage of coreferent mentions in the corpus, select a random antecedent mention m_j, uniformly distributed between antecedents of m_i; otherwise select no antecedent for that mention.
- *AlwaysNo*: set all mentions as having no antecedent (singleton mentions).

Additionally, we compare our models with the best reported systems on the Spanish AnCora dataset [1], [28]. Note that we only use the first column (token data) of the SemEval dataset, unlike reported systems which could use all columns.

F. Training

We use ADAM [42] for training, with learning rate starting at 0.001, and a batch size of 32 samples. The learning rate is reduced by a factor of 5 once learning stagnates for more than 3 epochs, based on the accuracy on validation data. Models are trained for up to 25 epochs, with early stopping after 6 epochs of non-improving performance on validation loss.

V. Results

This section outlines and explores the results from the experiments described in Section IV.

A. Undersampling

Aiming to improve model performance, we use the *Arch2* architecture, described in Section IV-D, to study which undersampling percentage is best suited for this task. Despite Fonseca *et al.* [20] using one-to-one undersampling (equal sampling of both classes), and measuring satisfactory results with this method [31], performance was reported on sampled mention-pairs, not model-independent coreference metrics. We train the model on the AnCora corpus, and report performance on macro-F1 of the mention-pairs test set (with original sampling) and the CoNLL metric.

Figure 3 presents the results of training the model on different undersampling percentages. We conclude that training the

TABLE II
RESULTS OF DIFFERENT ARCHITECTURES AND BASELINES ON ANCORA (ES) TEST SET, AND CORREF-PT TEST SET, WITH GOLD ANNOTATIONS.

		MUC			B^3			$CEAF_e$			*BLANC*			CoNLL Official
		Prec.	Rec.	$F1^1$	Prec.	Rec.	$F1^2$	Prec.	Rec.	$F1^3$	Prec.	Rec.	Blanc	$\frac{F1^1+F1^2+F1^3}{3}$
ES	*Rand1*	10.7	8.2	9.3	63.7	52.1	57.3	45.5	55.8	50.1	50.2	50.1	50.1	38.9
	Rand2	2.5	4.7	3.2	62.5	80.9	70.5	73.5	57.0	64.2	50.1	50.4	49.8	46.0
	AlwaysNo	0	0	0	62.2	100	76.7	89.2	55.5	68.4	50	48.8	49.4	48.4
	Relax [1]	14.8	73.8	24.7	65.3	97.5	78.2	66.6	66.6	66.6	53.4	81.8	55.6	56.5
	Sucre [28]	52.7	58.3	55.3	75.8	79.0	77.4	69.8	69.8	69.8	67.3	62.5	64.5	67.5
	Arch1	25.2	62.7	36.0	67.0	91.9	77.5	89.2	65.4	75.5	55.5	68.1	58.1	63.0
	Arch2	54.7	44.2	51.8	85.3	72.1	78.1	73.2	86.3	79.2	63.9	63.2	63.5	69.7
	Arch2-dense	37.4	69.9	48.7	71.5	90.2	79.8	88.9	69.3	77.8	63.4	64.0	63.6	68.8
	Arch-deep-CNN	37.7	66.4	48.1	70.9	88.1	78.5	87.9	69.6	77.7	62.0	65.9	63.6	68.1
	Arch-biLSTM	42.7	65.2	51.6	72.2	86.6	78.8	87.5	72.3	79.2	61.5	64.6	62.8	**69.9**
PT	*Rand1*	13.7	20.5	16.5	30.5	54.2	39.0	46.2	24.1	31.7	50.4	50.7	50.4	29.1
	Rand2	3.6	11.7	5.4	27.2	79.5	40.6	51.7	17.6	26.2	50.1	51.4	49.2	24.1
	AlwaysNo	0	0	0	26.4	100	41.7	52.2	13.8	21.8	50	47.3	48.6	21.2
	Arch1	43.8	55.4	48.9	46.0	57.6	51.2	49.5	31.2	38.3	57.6	55.7	56.4	46.1
	Arch2	46.8	59.7	52.5	46.97	62.6	53.7	55.1	34.5	42.4	58.3	60.9	59.3	**49.5**
	Arch2-dense	46.7	59.2	52.1	48.1	59.3	52.9	51.2	2.6	39.7	60.1	58.6	59.1	48.2
	Arch-deep-CNN	41.8	53.0	46.7	44.7	58.0	50.4	50.8	32.0	39.2	64.6	65.1	57.1	45.5
	Arch-biLSTM	46.8	58.2	51.8	48.4	58.9	53.1	51.4	33.4	40.5	59.1	58.5	58.7	48.4
Direct Transfer (ES to PT)	*Arch1*	0.6	46.7	1.2	26.7	99.5	42.1	52.5	14.0	22.1	50.1	66.4	48.7	21.8
	Arch2	56.9	60.9	58.7	58.6	39.7	45.8	33.0	28.0	29.7	52.3	50.6	41.2	**44.8**
	Arch2-dense	27.2	46.8	33.2	39.7	68.8	49.5	48.1	21.2	29.2	53.1	53.2	51.7	37.3
	Arch-deep-CNN	0.2	33.1	0.3	26.5	99.7	41.8	52.2	13.8	21.9	50.0	59.6	48.7	21.4
	Arch-biLSTM	5.4	40.0	9.6	28.1	92.0	43.1	51.9	15.2	23.5	50.8	56.2	50.5	25.4

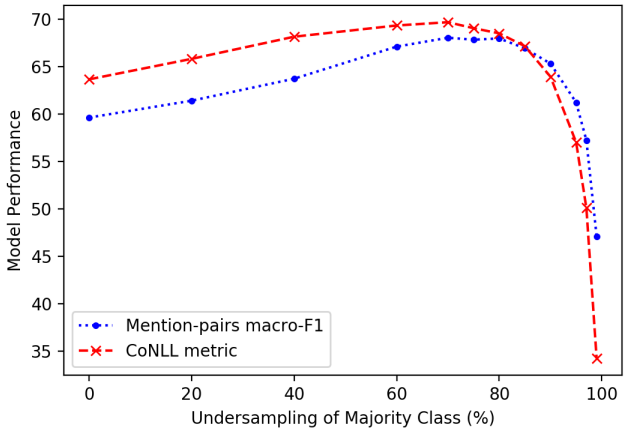

Fig. 3. Effects of undersampling of majority class on model performance.

model with 70% undersampling of the majority class improves the model's performance considerably, boosting it from 59.67 to 68.1 Macro-F1, and from 63.7 to 69.7 on the CoNLL metric, when compared to training with no undersampling.

B. Results

Using the best undersampling value from the previous experiment (70%), we train all proposed architectures on the Spanish AnCora corpus and report test-set results in Table II. The *Arch2* and *Arch-biLSTM* architectures have the best performance, clearly improving over the cited baselines.

Next, we select the most promising architectures and train them on the Portuguese Corref-PT corpus. It is important to note that random baselines perform substantially worse on the Corref-PT corpus than the AnCora corpus, due to the different class distributions. Despite this, and the considerably smaller dataset for Portuguese, we achieved promising results (see Table II). No cited baselines are shown because, to the best of our knowledge, none exists for this corpus.

Then, we experiment with the direct transfer of model weights from Spanish to Portuguese. Although results lack behind models trained directly on Portuguese data, knowledge transfer is clearly occurring, as the resulting models perform considerably better than random baselines, with special focus on good identification of positive coreference links, evidenced by the results on the *MUC* metric (58.7 F1 for *Arch2 versus 16.5* F1 for *Rand1*). Interestingly, the best performing models on this setting are the simplest architectures – *Arch2* and *Arch2-dense*, which have approximately an order of magnitude less parameters than the remaining two architectures. We associate these findings to the capability of more complex models to capture Spanish-specific characteristics (overfitting the training data), and the need for the simpler models to learn broader and more general characteristics.

VI. CONCLUSIONS AND FUTURE WORK

We have reported on a new state-of-the-art coreference resolution system on Spanish and the first, to the best of our knowledge, for Portuguese exploring the Corref-PT corpus. Additionally, we have studied the effect of undersampling on standard coreference metrics, providing a considerable boost to the system's performance.

We have also presented a first attempt at leveraging a Spanish corpus for coreference resolution in Portuguese. We show competitive results compared to an in-language model, which provides good indications towards further exploring transfer learning techniques to address less-resourced languages.

In future work, we expect to improve our results using a more context-aware architecture, more sophisticated clustering algorithms, and improved mention-wise representations. We will also deepen our research on the usage of transfer learning on coreference resolution.

ACKNOWLEDGMENTS

André Ferreira Cruz is supported by the Calouste Gulbenkian Foundation, under grant number 214508.

REFERENCES

[1] E. Sapena, L. Padró, and J. Turmo, "A constraint-based hypergraph partitioning approach to coreference resolution," *Computational Linguistics*, vol. 39, no. 4, pp. 847–884, 2013.

[2] H. J. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge." in *AAAI spring symposium: Logical formalizations of commonsense reasoning*, vol. 46, 2011, p. 47.

[3] V. Ng, "Unsupervised models for coreference resolution," in *Proc. of the Conf. on Empirical Methods in NLP*. ACL, 2008, pp. 640–649.

[4] ——, "Machine learning for entity coreference resolution: A retrospective look at two decades of research." in *AAAI*, 2017, pp. 4877–4884.

[5] G. Kundu, A. Sil, R. Florian, and W. Hamza, "Neural cross-lingual coreference resolution and its application to entity linking," in *Proc. of the 56th Annual Meeting of the ACL (Volume 2: Short Papers)*, vol. 2, 2018, pp. 395–400.

[6] H. Ji, J. Nothman, B. Hachey, and R. Florian, "Overview of tac-kbp2015 tri-lingual entity discovery and linking," in *Proc. of the Eighth Text Analysis Conf. (TAC2015)*, 2015.

[7] M. Ogrodniczuk and V. Ng, "Proc. of the 2nd workshop on coreference resolution beyond ontonotes (corbon 2017)," in *Proc. of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, 2017.

[8] B. M. Sundheim, "Overview of results of the muc-6 evaluation," in *Proc. of the 6th Conf. on Message Understanding*, ser. MUC6 '95. Stroudsburg, PA, USA: ACL, 1995, pp. 13–31.

[9] L. Hirschman and N. Chinchor, "Appendix f: Muc-7 coreference task definition (version 3.0)," in *Seventh Message Understanding Conf. (MUC-7): Proc. of a Conf. Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998.

[10] X. Luo, A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos, "A mention-synchronous coreference resolution algorithm based on the bell tree," in *Proc. of the 42nd Annual Meeting of the ACL*, 2004.

[11] M. Klenner and É. Ailloud, "Enhancing coreference clustering," in *Proc. of the Second Workshop on Anaphora Resolution*, 2008, pp. 31–40.

[12] S. Wiseman, A. M. Rush, S. Shieber, and J. Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution," in *Proc. of the 53rd Annual Meeting of the ACL and the 7th Int. Joint Conf. on NLP (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1416–1426.

[13] K. Clark and C. D. Manning, "Deep reinforcement learning for mention-ranking coreference models," in *Proc. of the 2016 Conf. on Empirical Methods in NLP*, 2016, pp. 2256–2262.

[14] S. Wiseman, A. M. Rush, and S. M. Shieber, "Learning global features for coreference resolution," in *Proc. of the 2016 Conf. of the NAACL: Human Language Technologies*, 2016, pp. 994–1004.

[15] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proc. of the 2017 Conf. on Empirical Methods in NLP*, 2017, pp. 188–197.

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 conf. on empirical methods in NLP (EMNLP)*, 2014, pp. 1532–1543.

[17] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of the 2018 Conf. of the NAACL: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 2227–2237.

[18] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," in *Proc. of the 2018 Conf. of the NAACL: Human Language Technologies*, vol. 2, 2018, pp. 687–692.

[19] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes," in *Joint Conf. on EMNLP and CoNLL-Shared Task*. ACL, 2012, pp. 1–40.

[20] E. Fonseca, R. Vieira, and A. Vanin, "Improving coreference resolution with semantic knowledge," in *Int. Conf. on Computational Processing of the Portuguese Language*. Springer, 2016, pp. 213–224.

[21] G. Rocha and H. Lopes Cardoso, "Towards a mention-pair model for coreference resolution in portuguese," in *Portuguese Conf. on Artificial Intelligence*. Springer, 2017, pp. 855–867.

[22] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation." in *LREC*, vol. 2, 2004, p. 1.

[23] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley, "Semeval-2010 task 1: Coreference resolution in multiple languages," in *Proc. of the 5th Int. Workshop on Semantic Evaluation*. ACL, 2010, pp. 1–8.

[24] M. Garcia and P. Gamallo, "Multilingual corpora with coreferential annotation of person entities." in *LREC*, 2014, pp. 3229–3233.

[25] A. Antonitsch, A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini, "Summ-it++: an enriched version of the summ-it corpus," in *of the Language Resources and Evaluation Conf. (LREC)*, 2016, pp. 2047–2051.

[26] E. Fonseca, V. Sesti, S. Collovini, R. Vieira, A. L. Leal, and P. Quaresma, "Collective elaboration of a coreference annotated corpus for portuguese texts," in *Proc. of II workshop on Evaluation of Human Language Technologies for Iberian Languages*, vol. 1881, Murcia, Spain, 2017, pp. 68–82.

[27] M. Recasens and M. A. Martí, "Ancora-co: Coreferentially annotated corpora for spanish and catalan," *Language Resources and Evaluation*, vol. 44, no. 4, pp. 315–345, Dec 2010.

[28] H. Kobdani and H. Schütze, "Sucre: A modular system for coreference resolution," in *Proc. of the 5th Int. Workshop on Semantic Evaluation*. ACL, 2010, pp. 92–95.

[29] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC 2018)*, 2018.

[30] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules," *Computational Linguistics*, vol. 39, no. 4, pp. 885–916, 2013.

[31] E. B. Fonseca, R. Vieira, and A. Vanin, "Dealing with imbalanced datasets for coreference resolution," in *The Twenty-Eighth International Flairs Conference*, 2015.

[32] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme," in *Proc. of the 6th conf. on Message understanding*. ACL, 1995, pp. 45–52.

[33] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *The 1st Int. Conf. on Language Resources and Evaluation Workshop on Linguistics Coreference*, vol. 1. Granada, 1998, pp. 563–566.

[34] X. Luo, "On coreference resolution performance metrics," in *Proc. of the conf. on human language technology and empirical methods in NLP*. ACL, 2005, pp. 25–32.

[35] M. Recasens and E. Hovy, "Blanc: Implementing the rand index for coreference evaluation," *Natural Language Engineering*, vol. 17, no. 4, pp. 485–510, 2011.

[36] N. Reimers and I. Gurevych, "Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging," in *Proc. of the 2017 Conference on Empirical Methods in NLP*, 2017, pp. 338–348.

[37] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, 2010.

[38] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.