Anaphora Resolution for Improving Spatial Relation Extraction from Text

Umar Manzoor[*] **Parisa Kordjamshidi**[*†]
[*]Tulane University, Computer Science Department, New Orleans, LA, USA
[†]Florida Institute for Human and Machine Cognition (IHMC), Pensacola, FL, USA
{umanzoor,pkordjam}@tulane.edu

Abstract

Spatial relation extraction from generic text is a challenging problem due to the ambiguity of the prepositions spatial meaning as well as the nesting structure of the spatial descriptions. In this work, we highlight the difficulties that the anaphora can make in the extraction of spatial relations. We use external multi-modal (here visual) resources to find the most probable candidates for resolving the anaphoras that refer to the landmarks of the spatial relations. We then use global inference to decide jointly on resolving the anaphora and extraction of the spatial relations. Our preliminary results show that resolving anaphora improves the state-of-the-art results on spatial relation extraction.

1 Introduction

Spatial relation extraction is the task of determining the relations that can exist among the spatial roles extracted from the text (D'Souza and Ng, 2015). In the recent years, significant progress has been made in spatial language understanding (i.e. mapping natural language text to a formal spatial meaning representation) (Kordjamshidi et al., 2017a; Kordjamshidi and Moens, 2015a). As a basic example consider the sentence, "A car is parked in front of a house." In this sentence *car* is a *trajector*, *house* is a *landmark* and *in front of* is a *spatial indicator*. Spatial indicators indicate the existence of spatial information in a sentence. Trajector is an entity whose location is described and landmark is a reference object for describing the location of a trajector.

Extraction of the spatial relations with a good accuracy is still challenging (Pustejovsky et al., 2015). Particularly, our investigation on the errors of the previous models shows that when in a sentence the landmark is expressed as a pronoun like *("it", "them", "him",..)*, the extraction of spatial relations becomes more difficult. For example, in the sentence, *"A narrow, rising street with colourful houses on both sides, among them a green house with balconies and a white car parked in front of it, and a blue-and-white church on the right"*, some of the spatial relations for this sentence will contain a landmark which is a pronoun such as $\langle R_1 \leftarrow$[*a green house*]$_{\text{tr}}$, [*among*]$_{\text{sp}}$, [*them*]$_{\text{lm}}\rangle$ and $\langle R_2 \leftarrow$[*a white car*]$_{\text{tr}}$, [*in front of*]$_{\text{sp}}$, [*it*]$_{\text{lm}}\rangle$. This issue is related to the well-known *anaphora resolution* problem which is also problematic for our goal of spatial relation extraction.

Anaphora Resolution which mostly appears as pronoun resolution, is the linguistic phenomenon by which the given pronoun is interpreted with the help of earlier or later items in the discourse (Mitkov, 2005). The pronoun word/phrase is referred as anaphor whereas the word/phrase to which it is referring is called antecedent, as both anaphor and antecedent are referring to the same object in the real world, they are termed co-referential (Mitkov et al., 2007). It might be possible that for some anaphor, the antecedent is not mentioned in the same sentence, for example, consider a sentence, *"there are a couple of trees in front of it"*, here *"it"* is referring to some object which is not mentioned in the sentence, however, the referring object might have been mentioned in another sentence of the document. Anaphora Resolution generally is recognized as a difficult problem in Natural Language Processing (Lee et al., 2017a; Marasovic et al., 2017).

The main research questions that we aim to address in this paper are, 1) whether the external knowledge from multimodal resources can help anaphora resolution in text. 2) whether the anaphora resolution can help in the spatial relation extraction from text (especially the relations in the form of triplet - Trajector, Spatial Indicator, Landmark). To answer these questions, we incorpo-

Proceedings of the First International Workshop on Spatial Language Understanding (SpLU-2018), pages 53–62
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

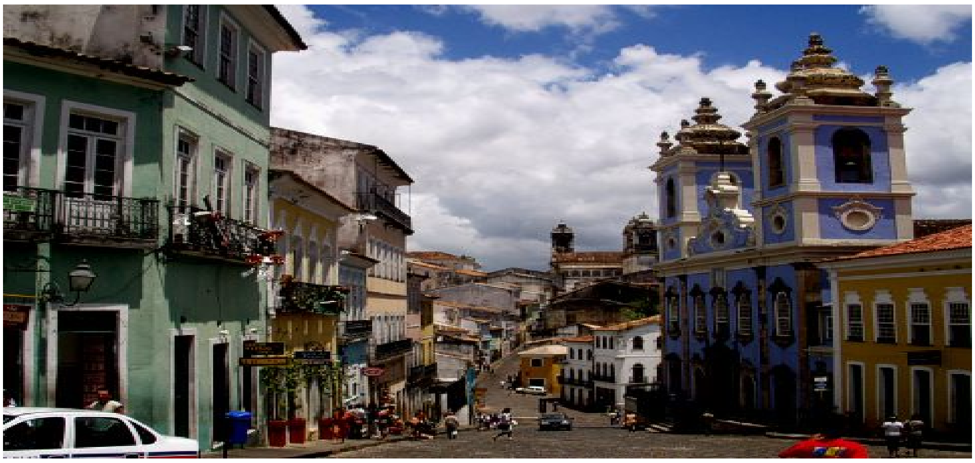

Figure 1: Image Textual Description: "A narrow, rising street with colourful houses on both sides, among them a green house with balconies and a white car parked in front of it, and a blue-and-white church on the right"

rated anaphora resolution for the pronouns in the sentence and proposed a global machine learning model to exploit the resolved pronouns. In the first step, we find the list of possible landmarks that can replace a pronoun in a relation (under consideration) with a specific candidate trajector and candidate spatial indicator. We used Visual Genome (Krishna et al., 2017) (an external) dataset for this purpose.

Visual genome dataset provides us a list of possible landmarks which can be used to resolve the anaphora by filtering them based on their similarity with the candidate landmarks that appear in the sentence. This information is used in the global inference model for joint prediction. We improve the spatial relation extraction from text by incorporating anaphora resolution to recognize landmarks in spatial relations which distinguishes our work from other works on anaphora resolution. The contribution of this paper includes a) exploiting external visual relation datasets to inject external knowledge into our models b) forming a joint model that imposes the consistency between the decisions made by separate relation classifiers that decide on a candidate spatial relation with pronoun landmark and candidate spatial relations with that pronoun replaced by candidate noun resolvants. c) obtaining state-of-the-art results on spatial information extraction by exploiting the anaphora resolution. This paper shows our preliminary efforts in the sense that we have not applied the existing work on anaphora resolution. We do not aim at improving the current techniques in that area but only show that such resolutions using visual resources can help spatial relation extraction.

The rest of this paper is organized as follows, first we describe the problem setting in Section 2; our proposed model for this problem is described in Section 3. The dataset used in tests, and evaluation results, are presented in Section 4. In Section 5, we briefly point to the related work in this area. Finally, Section 6 summarizes the conclusions and outlines directions for future work.

2 Problem Definition

The goal is to improve the extraction of spatial information from text by incorporating anaphora resolution for landmark candidates. We briefly define the spatial role labeling (SpRL) task which is based on a previous formalization of (Kordjamshidi et al., 2017b, 2011; Kordjamshidi and Moens, 2015b). Given a sentence S, segmented into phrases $P = [P_1, P_2, P_3, ...P_n]$ where P_i is the identifier of i^{th} phrase in the sentence, the goal of spatial role labeling is to find the phrases which carry spatial roles (i.e. trajector (Tr), spatial indicator (Sp), landmark (Lm)), as introduced in Section 1 and identify the links between them to form spatial realtion, $R = [Tr, Sp, Lm]$. Moreover, each Spatial relation is further classified into coarse-grained type - (*region*, *direction*, *distance*) and fine-grained types based on their coarse-grained types (e.g. *(region,EC), (region,DC), (direction,left), (direction,right)*).

Figure 2 shows an example of spatial roles, spatial relations and spatial relation type extracted from a given text. In this example, the location of *statue (trajector)* is described with respect to the *hill (landmark)* using the preposition *on (spatial indicator)*. In Figure 1, the caption shows the textual description of an image, featuring multiple spatial relations ($\langle R_1$ ←[*a green*

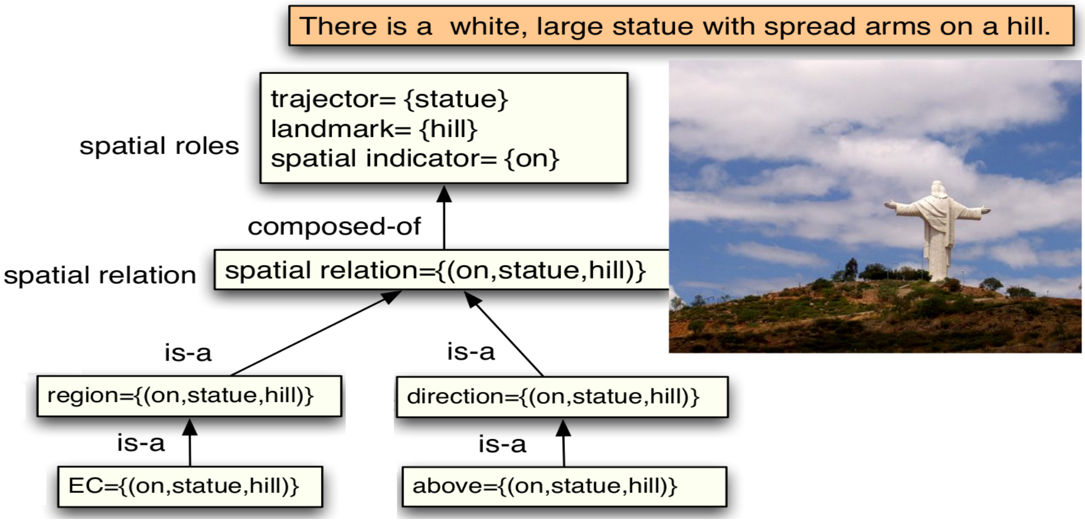

Figure 2: An Example of Spatial Roles and Relation Types.

house]$_{\text{tr}}$, [*among*]$_{\text{sp}}$, [*them*]$_{\text{lm}}\rangle$, $\langle R_2 \leftarrow$[*a white car*]$_{\text{tr}}$, [*in front of*]$_{\text{sp}}$, [*it*]$_{\text{lm}}\rangle$, $\langle R_3 \leftarrow$[*a blue-and-white church*]$_{\text{tr}}$, [*on the right*]$_{\text{sp}}$, [*None*]$_{\text{lm}}\rangle$ and $\langle R_4 \leftarrow$[*colorful houses*]$_{\text{tr}}$, [*on*]$_{\text{sp}}$, [*both sides*]$_{\text{lm}}\rangle$) where R_1, R_2 have pronoun landmark, and R_3, R_4 have implicit landmarks (i.e. not mentioned in the given sentence). $R_1 \rightarrow$landmark ([*it*]$_{\text{lm}}$), and $R_2 \rightarrow$landmark ([*them*]$_{\text{lm}}$) are referring to [*colorful houses*], and [*a green house*] respectively. R_1, R_2 belongs to a well known anaphora resolution problem where the given pronoun is interpreted with the help of earlier or later items in the discourse whereas R_3, R_4 belongs to coreference resolution problem (Lee et al., 2017b; Ng, 2010; Martschat and Strube, 2015) that aims at finding all expressions in the document that refer to the same entity.

The hypothesis of this paper is that how anaphora resolution for landmark candidates might help the inference for the extraction of roles as well as the relations from sentences. In this work, we proposed a model to address anaphora resolution for landmark candidates with the aim of improving the spatial relation extraction. In this paper, we assume that the antecedent (if any) of the anaphora (landmark here) is mentioned within the same sentence, therefore, cross-sentence anaphora resolution is not performed in this work.

3 Architecture

Depending on the description of the sentence, the spatial relations can contain pronoun landmarks (such as ”it”, ”them”, ”him”, ”her”). Consider the aforementioned spatial relations R_1 and R_2 extracted from sentence T, $R_1 \rightarrow$landmark ([*them*]$_{\text{lm}}$) and $R_2 \rightarrow$landmark ([*it*]$_{\text{lm}}$) are referring to [*colorful houses*] and [*a green house*] phrases of the sentence T respectively. The components of computing the anaphora resolution for pronoun landmark spatial relations is described in the following subsections.

3.1 Exploiting External Knowledge

Given a candidate spatial relation R with a pronoun landmark, we are interested in finding the possible landmark objects which can occur with the given trajector and spatial indicator. For this purpose, we used an external resource, that is Visual Genome relationship dataset (VG). This dataset contains the relation (preposition) between various subjects and objects – for details see section 4.1. Given R, similar relations are extracted from visual genome dataset V by matching preposition and subject with $R \rightarrow spatialIndicator$ and $R \rightarrow trajector\!-\!headword$ respectively, that is the candidate words for the sp and tr roles.

In this way, we obtain the list of possible landmark objects and their frequencies in the VG dataset. We compute the frequency ratio per object and this ratio is interpreted as the possibility score of a relation containing that landmark. In other words, the score R_S is computed as $R_S \leftarrow O_{R_i}/T_{V_R}$ where O_{R_i} is the frequency of having object i with the given trajector-spatial indicator pair, and T_{V_R} is the

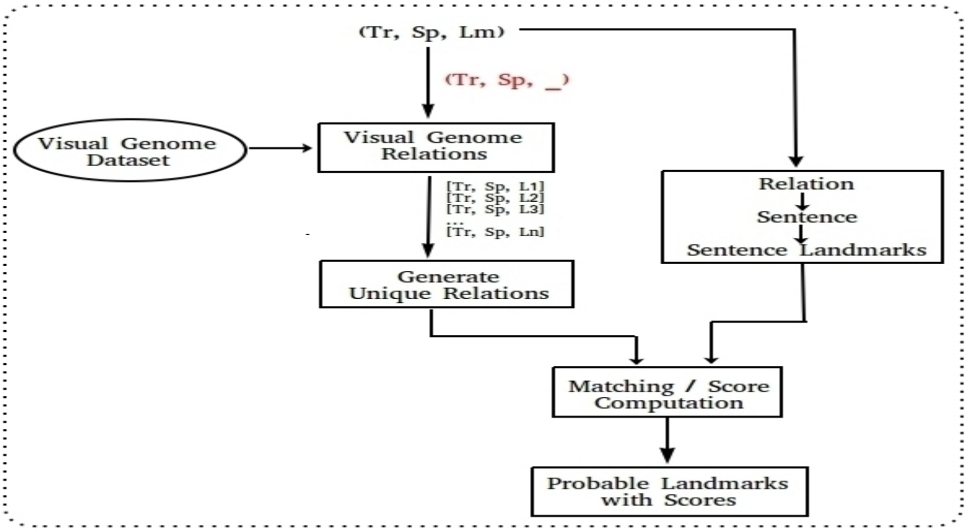

Figure 3: Probable landmark extraction model

total relations frequency for given trajector-spatial indicator pair in VG dataset. This will yield the set of possible triplets given the trajector-indicator pair with a score assigned to each triplet. We denote this set as, $U_R = [(U_{R_1}, S_{U_{R1}}), (U_{R_2}, S_{U_{R2}}), ..., (U_{R_n}, S_{U_{Rn}})]$ where U_{R_i} and $S_{U_{Ri}}$ is the i^{th} unique relation and its score respectively.

3.2 Scoring Landmark Candidate Resolvants

For each sentence we perform a pre-processing step based on the previous works and obtain a set of noun phrases that serve as the landmark candidates denoted by S_L. The aforementioned retrieved triplets from visual genome, U_R, can contain many landmarks which don't exist in our landmarks candidates set, therefore, in this step, we compute the similarity (using Google Word2Vec) score between each landmark in S_L with all U_R landmarks. The final score for each candidate landmark in the sentence will be the maximum score that is computed by averaging the similarity score and occurrence score of that landmark with respect to all U_R candidates. In this way we obtain a score for each candidate landmark in S_L.

3.3 Learning Model

We formulate this problem as a structured output prediction problem where given a set of input-output pairs as training examples, $E = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} : i = 1..N\}$, an objective function $g(x, y; W) = \langle W, f(x, y) \rangle$ is learned. This function is a linear discriminant function defined over combined feature representation of inputs and outputs denoted by $f(x, y)$. However, in this work, independent classifiers are trained per role and relations and only the predication is performed based on the global inference as in (Kordjamshidi et al., 2017a; Rahgooy et al., 2018) .
We construct a graph using the phrases $\{p_1, ..., p_n\}$ (i.e. each phrase is a node in the graph) and link these nodes to make composed concepts such as relations. A classifier is associated with each concept in the graph and the domain knowledge is encoded over these concepts by global constraints. Global reasoning is imposed over these classifiers to produce the final outputs by using these constraints. Furthermore, we used binary classifiers to classify the spatial roles and relations where trajector, landmark, spatial indicator are denoted by tr, sp, lm respectively and $sp.tr.lm$, $sp.tr.lm.\gamma$, $sp.tr.lm.\lambda$ denotes spatial relations, coarse-grained relations, and fine-grained relations. Additionally, we denote the *new-relation-classifier* described in section 3.5 by $sp.tr.lm_{NRC}$.

Each phrase in the sentence is described by a vector of linguistic features denoted by: $\psi_{phrase}(p_i)$ (e.g. word form, POS tag, headword POS tag, dependencyRelation, subCategorization, etc), these features are used by spatial role classifiers. The spatial relation is composed of three phrases (p_i, p_j, p_k), therefore, the combination of these phrases along with their descriptive vectors are used in the spatial relation feature set referred as: $\phi^{text}_{triplet}(p_i, p_j, p_k)$ (e.g. distance between trajector and spatial indicator, concatenation of trajector, spatial indicator, and landmark). These features are proposed by (Roberts and Harabagiu, 2012) and (Kordjamshidi et al., 2017a).

1	$\sum_i \sum_k sp_itr_jlm_k \geq tr_j$	Each tr candidate at least should appear in one relation
2	$\sum_i \sum_j sp_itr_jlm_k \geq lm_k$	Each lm candidate at least should appear in one relation
3	$\sum_j \sum_k sp_itr_jlm_k = sp_i$	Each sp candidate should appear in one relation
4	$\sum_j tr_j \geq sp_i$	For each sp we should have at-least one tr
5	$\sum_k lm_k \geq sp_i$	For each sp we should have at-least one lm
6	$sp_itr_jlm_k\gamma \leq sp_itr_jlm_k$	is-a constraints between relations and coarse-grained types
7	$sp_itr_jlm_k\lambda \underset{\lambda\in\Lambda_\gamma}{\leq} sp_itr_jlm_k\gamma$	is-a constraints between coarse-grained and corresponding fine-grained types where Λ_γ denotes the candidate fine-grained types related to coarse-grained type γ.
8	$sp_itr_jlm_{k_{NRC}} \leq sp_itr_jlm_k$	Spatial relation with pronoun candidate should be classified as true if anyone in top N of the anaphora-resolved triplets is classified as true.

Table 1: Model Constraints.

3.4 Constraints

The global constraints used in our proposed model is combination of previously proposed constraints (1-7) (Rahgooy et al., 2018) and new one (constraint 8) described in Table 3.3. In fact, the global inference is performed using integer linear programming techniques subject to these constraints.

3.5 Global Prediction Model

We obtain the output of each classifier in the model holistically by global reasoning that is by considering global correlations among classifiers, when calculating outputs. This goal is achieved by optimizing an objective function that is the summation of classifiers' discriminant functions. The global objective function for our model is on the basis of our previous work as follows,

$$\sum_{i\in C_{sp}} \langle W_{sp}, \phi_{sp_i}\rangle . sp_i + \sum_{i\in C_{tr}} \langle W_{tr}, \phi_{tr_i}\rangle . tr_i +$$

$$\sum_{i\in C_{lm}} \langle W_{lm}, \phi_{lm_i}\rangle . lm_i +$$

$$\sum_{i\in C_{sp}} \sum_{j\in C_{tr}} \sum_{k\in C_{lm}} \langle W_{sptrlm}, \phi_{sp_itr_jlm_k}\rangle . sp_itr_jlm_k +$$

$$\sum_{\gamma\in\Gamma} \sum_{i\in C_{sp}} \sum_{j\in C_{tr}} \sum_{k\in C_{lm}} \langle W_{sptrlm}, \phi_{sp_itr_jlm_k\gamma}\rangle . sp_itr_jlm_k\gamma +$$

$$\sum_{\lambda\in\Lambda} \sum_{i\in C_{sp}} \sum_{j\in C_{tr}} \sum_{k\in C_{lm}} \langle W_{sptrlm}, \phi_{sp_itr_jlm_k\lambda}\rangle . sp_itr_jlm_k\lambda +$$

$$\sum_{\tau\in\Upsilon} \sum_{i\in C_{sp}} \sum_{j\in C_{tr}} \sum_{k\in C_{lm}} \langle W_{sptrlm}, \phi_{sp_itr_jlm_{kNRC_\gamma}}\rangle . sp_itr_jlm_{kNRC} .$$

Each classifier is shown as a binary variable and Λ, Γ, Υ are the candidates for fine-grained relations, coarse-grained relations, and pronoun-landmark spatial relations respectively. The following model variations are designed to evaluate the performance of the proposed model. Furthermore, in all model variations, the CLEF 2017 mSprl dataset described in 4.1 is used for the training and evaluation of the classifiers.

- **Anaphora-Replacement** (A-Replacement): In this model, we replace the landmark phrase text of spatial relation where the landmark is a pronoun with the highest scored probable landmark (see 3.2), this approach is used for both training and testing. Furthermore, we train independent classifiers for spatial roles and relations classification. This is a learning only model where each classifier makes independent predictions. This model doesn't use any constraints, and is compared with similar (Rahgooy et al., 2018) baseline model in section 4.

- **Anaphora-Inference** (A-Inference): In this model, 1) we create an additional triplet classifier for classifying the relations that contain pronoun landmarks and we name it *new-relation-classifier (NRC)* and use it at the inference time, 2) joint prediction is performed using the constraints described in

	A-Replacement			M0		
	Precision	Recall	F1	Precision	Recall	F1
$Trajector$	53.24	67.66	59.59	54.22	62.05	57.87
$Landmark$	73.49	81.23	77.17	74.29	78.60	76.38
$SpatialIndicator$	94.60	96.98	95.78	94.60	96.98	95.78

Table 2: Spatial Roles - Comparison of A-Replacement with M0

	A-Inference			M0+C		
	Precision	Recall	F1	Precision	Recall	F1
$Trajector$	65.79	65.39	65.59	64.20	60.98	62.55
$Landmark$	84.69	78.60	81.53	79.09	82.28	80.65
$SpatialIndicator$	94.70	96.60	95.64	95.08	94.84	94.96

Table 3: Spatial Roles - Comparison of A-Inference with M0+C

3.4 to optimize the global objective function explained in section 3.5 which includes the *new-relation-classifier*. This implies that both relation classifier and the *new-relation-classifier* are assigned values jointly and should agree. For training the *new-relation-classifier*, we generate additional examples by replacing the pronoun landmarks in the ground-truth with the highest scored landmark from our candidate set, S_L. The original spatial relations with pronoun landmarks are also retained in the training. The training mechanism of remaining classifiers remains unchanged (i.e. trained on original spatial relations). In testing phase, we take the top N candidates from the scored landmarks generated in 3.2 for spatial relations with pronoun landmarks. In this way, we regenerate a set of candidate triplets by replacing the pronoun with the top probable landmarks. Our global inference decides jointly with using the original triplet classifier in a way that it satisfies the constraint that if anyone of these triplets is predicted as true, spatial relation classifier is forced at inference time to predict the spatial relation with the anaphora as true. See constraint number 8 in section 3.4. The experiments show that this simple idea can promote the relation extraction when anaphora occurs in the triplet candidates.

4 Experiments

4.1 Datasets

CLEF 2017 mSpRL dataset: Our model is evaluated on this dataset which is a subset of IAPR TC-12[1] Benchmark and annotated specifically for the SpRL task. The training set contains 761 and whereas test set contains 939 spatial relations respectively (Kordjamshidi et al., 2017b). The total number of spatial relations containing pronoun landmark in train and test is 44 and 129 respectively.

Visual Genome dataset (VG): Visual Genome dataset has seven main components (Krishna et al., 2017), one of it is ‘relationships‘ which contains the relationships between pairs of objects in the images. Each relation has two arguments, the first one is referred as subject whereas the latter one is referred as object. These relationships can be actions, spatial, prepositions, verbs, comparative or prepositional phrases. Visual genome dataset contains 108077 images whereas its relationships part contains 2316104 relation instances. This dataset is used to obtain the possible landmarks that can occur in a relationship with a given subject.

4.2 Experimental Results

In this section, we experimentally show the effectiveness of our proposed model in improving the spatial role/relation extraction. We use *Saul* (Kordjamshidi et al., 2015, 2016) to implement the models and solve the global inference of Section 3.5. The code is publicly available here[2].

We compare our approach with the state-of-the-art (Rahgooy et al., 2018). However, in the mentioned paper, the authors use visual data from the accompanying images to improve the models. In

[1] http://www.imageclef.org/SIAPRdata

[2] https://github.com/HetML/SpRL/tree/paper3

	Precision	Recall	F1
$M0$	65.64	60.23	62.82
$M0 + C$	70.04	66.55	68.25
$A\text{-}Replacement$	78.47	56.84	65.92
$A\text{-}Inference$	70.23	68.25	69.23

Table 4: Model Comparison - Spatial Relation Extraction

	A-Replacement			M0		
	Precision	Recall	F1	Precision	Recall	F1
$Region$	70.90	54.24	61.52	78.37	47.83	59.41
$Direction$	79.22	43.57	56.22	83.56	43.57	57.28

Table 5: Coarse-grained Spatial Relations - Comparison of A-Replacement with M0.

this paper, we use their best model (referred here as $M0$ -Baseline and $M0 + C$ -Baseline plus constraints) which is trained on text only and we ignore the visual information which is aligned with the text. The experimental results in Table 4 show that our baseline model (A-Replacement) is significantly better as compared to the state-of-the-art baseline model (M0). This shows that replacing the pronoun landmark candidates with our proposed model probable landmark has positive impact on extraction of spatial roles (as shown in Table 2) and relations. The improvement in the results is because the spatial roles predication is improved, which gives a more confidence to the model to classify the triplets as spatial relations which leads to more positive predictions and higher recall of the relations.

Furthermore, our second model (A-Inference) in which we train an additional *new-relation-classifier* by generating additional examples and perform joint inference further improves the results over the state-of-the-art model with constraints (M0+C). The experimental results in Table 3 show that adding constraints to our second model (A+Inference) significantly improves the classification of spatial roles (i.e. trajectors and landmarks), although the spatial indicators is slightly improved. Also these constraints help improving the coarse-grained spatial relations as shown in table 6, although it doesn't have any impact on distance category because the number of examples in test set is very small (i.e. three instances only).

Our results improve the state-of-the-art models for spatial relation extraction. Both proposed models significantly improves the extraction of spatial roles and relations (when compared with independent learning and with constrained models). However, the results of some of the categories in the fine-grained relations drops which are not reported here. These results are at the preliminary stage and we further analyze our models. Particularly, we will use existing anaphora resolution models to see how those could help and provide a more reasonable baseline. This baseline will help us to evaluate the advantage of the external visual knowledge more clearly. It will be interesting to investigate what caused this drop in fine-grained relation types. In addition to such further analysis, this work can be extended into two possible directions, 1) incorporate cross-sentence anaphora resolution for landmark candidates, and 2) incorporate co-reference resolution in general for all spatial relations.

5 Related Work

Our proposed model is a joint model for considering anaphora resolution to help spatial information extraction. Anaphora resolution is a fundamental problem in natural language processing and existing techniques can broadly be categorized into two types 1) Rule based models: apply rules to reduce candidate antecedents and resolve anaphora and 2) statistical models: use probabilistic models for the resolution of anaphora (Lee et al., 2017a). Early work (Hobbs, 1978; Asher and Wada, 1988; Lappin and Leass, 1994; Morton, 2000) focused on designing rule-based systems for anaphora resolution (the target was finding antecedents of pronouns only), however, these systems relied heavily on handcraft rules/weights. In early 2000, (Soon et al., 2001; Yang et al., 2003; Ng and Cardie, 2002) used statistical machine learning methods to resolve co-reference, these methods used a com-

	A-Inference			M0+C		
	Precision	Recall	F1	Precision	Recall	F1
Region	72.99	60.82	66.35	76.07	57.79	65.68
Direction	76.26	46.67	57.90	75.75	48.33	59.01

Table 6: Coarse-grained Spatial Relations - Comparison of A-Inference with M0+C.

mon strategy, that is, train a statistical model to measure the likeness of a pair as corefer. However, each candidate is resolved independently of the others which means how good a candidate antecedent is relative to others is not considered. To address this problem, (Denis and Baldridge, 2009) proposed a model by combining machine learning with global inference for performing the resolution jointly. Recently, (Park et al., 2016) proposed an mention pair model using deep learning and a system that combines both rule-based and deep learning-based systems using a guided MP model for co-reference resolution.

According to (Lee et al., 2017a), machine learning based models for anaphora resolution are relatively easy to build as compared to rule based models, however, a huge amount of handcrafted feature design is required in order to build a successful anaphora resolution model. Furthermore, the authors highlighted four key features of a ideal anaphora resolution system one of which is antecedent features should be learned automatically (i.e. minimum human design effort should be required). The proposed model doesn't require any handcrafting features or rules to implement the anaphora resolvers.

Join models have been proposed for resolving co-references with mention head detection using underlying integer linear programming as we do here (Peng et al., 2015). The main difference of our work compared to the above mentioned research works is that here we do not directly solve the anaphora resolution problem, but we use a kind of indirect supervision from an external multi-modal resource to help anaphora resolution and by means of that we solve our specific target problem. Our target problem of spatial information extraction has not been jointly performed with neither anaphora nor co-reference resolution tasks before. However, resolving co-references in the multi-modal setting has been investigated recently (Huang et al., 2017) in which text and video refer to the same scene and help each other in the resolution. As pointed above, this is different from using the vision modality as a source of distant supervision which is our aim in this work.

6 Conclusion

In this paper, we investigated the challenging issues of the extraction of spatial relations, that is, the triplets of (spatial indicator, trajector, landmark) from generic text. Particularly, We highlighted one important problem that is the issue of anaphoras accruing in the text that make recognizing landmarks and consequently recognizing the spatial relations difficult. In the presence of the anaphora recognizing the right link between the described objects in the text and extracting the relations correctly for any arbitrary pair of object becomes more challenging. Our proposed solution has been to use the external visual resources that can help to find out the most probable landmarks for a specific object and obtain the possible resolutions with a score. Using the scored resolutions we perform global inference to decide on both the anaphora resolution and spatial relation extraction jointly. Our best model improves the state-of-the-art results in all precision, recall and F1 metrics while having a more positive (about +2%) influence on the recall of the spatial relations extraction. While our preliminary experimental results show the advantage of anaphora resolution in spatial relation extraction, we will investigate more sophisticated baselines in the future to evaluate the advantage of external knowledge resources (that we used in this work) versus using the existing approaches for anaphora resolution in our models.

References

Nicholas Asher and Hajime Wada. 1988. A computtional account of syntactic, semantic and discourse principles for anaphora resolution. *Journal of Semantics*, 6(1):309–344.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based spatial relation extraction with expanding parse trees. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 758–768.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311 – 338.

De-An Huang, Joseph J. Lim, Fei-Fei Li, and Juan Carlos Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. *CoRR*, abs/1703.02521.

P. Kordjamshidi, D. Roth, and H. Wu. 2015. Saul: Towards declarative learning based programming. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Parisa Kordjamshidi, Daniel Khashabi, Christos Christodoulopoulos, Bhargav Mangipudi, Sameer Singh, and Dan Roth. 2016. Better call saul: Flexible programming for learning and inference in nlp. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3030–3040.

Parisa Kordjamshidi and Marie-Francine Moens. 2015a. Global machine learning for spatial ontology population. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30:3–21.

Parisa Kordjamshidi and Marie-Francine Moens. 2015b. Global machine learning for spatial ontology population. *Web Semant.*, 30(C):3–21.

Parisa Kordjamshidi, Taher Rahgooy, and Umar Manzoor. 2017a. Spatial language understanding with multimodal graphs using declarative learning based programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*, pages 33–43.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017b. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 367–376. Springer.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing*, 8:1–36.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, 20(4):535–561.

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

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017b. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045.

Ana Marasovic, Leo Born, Juri Opitz, and Anette Frank. 2017. A mention-ranking model for abstract anaphora resolution. *CoRR*, abs/1706.02256.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *TACL*, 3:405–418.

Ruslan Mitkov. 2005. *The Oxford Handbook of Computational Linguistics (Oxford Handbooks)*. Oxford University Press, Inc., New York, NY, USA.

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Le An Ha, and Viktor Pekar. 2007. Anaphora resolution: To what extent does it help nlp applications? In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Conference on Anaphora: Analysis, Algorithms and Applications*, DAARC'07, pages 179–190, Berlin, Heidelberg. Springer-Verlag.

Thomas S. Morton. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1396–1411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheoneum Park, Kyoung-Ho Choi, Changki Lee, and Soojong Lim. 2016. Korean coreference resolution with guided mention pair model using deep learning. *ETRI Journal*, 38(6):1207–1217.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *CoNLL*.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proc. of the Annual Meeting of the*

Association for Computational Linguistics (ACL), pages 884–894. ACL.

Taher Rahgooy, Umar Manzoor, and Parisa Kordjamshidi. 2018. Visually guided spatial relation extraction from text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics.

Kirk Roberts and Sanda M Harabagiu. 2012. Utd-sprl: A joint approach to spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 419–424. Association for Computational Linguistics.

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

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 176–183, Stroudsburg, PA, USA. Association for Computational Linguistics.