SpaceRefNet: a neural approach to spatial reference resolution in a real city environment

Dmytro Kalpakchi
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
dmytroka@kth.se

Johan Boye
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
jboye@kth.se

Abstract

Adding interactive capabilities to pedestrian wayfinding systems in the form of spoken dialogue will make them more natural to humans. Such an interactive wayfinding system needs to continuously understand and interpret pedestrian's utterances referring to the spatial context. Achieving this requires the system to identify exophoric referring expressions in the utterances, and link these expressions to the geographic entities in the vicinity. This *exophoric spatial reference resolution* problem is difficult, as there are often several dozens of candidate referents. We present a neural network-based approach for identifying pedestrian's references (using a network called *RefNet*) and resolving them to appropriate geographic objects (using a network called *SpaceRefNet*). Both methods show promising results beating the respective baselines and earlier reported results in the literature.

1 Introduction

Remember yourself being lost in a completely unfamiliar city without knowing the local language or acquaintances that can help? Being close to desperate, you ask a passerby for a help and get an answer similar to the following:

> Just go forward until you see a McDonald's on the corner. There you turn right and keep straight until the old Gothic style church. A tall glass building near it is exactly what you need.

Such wayfinding instruction is a typical example of how humans guide each other in a city, relying mostly on *landmarks* in the vicinity (Cornell and Greidanus, 2006; Goodman et al., 2004; May et al., 2003; Denis, 1997; Lynch, 1960).

On the contrary, a current generation of navigation systems aiding pedestrian wayfinding generally makes use of quantitative information based on GPS signals, e.g. distances, cardinal directions and street names. The same instruction rephrased by such system would sound as follows:

> Head north on West Avenue. Turn right at the corner. Continue 150 meters straight until East Avenue 29. You've reached your destination.

Such instructions are presented to a pedestrian as a sequence on a screen (possibly voiced as well) supplemented by a map with a moving marker indicating pedestrian's position.

The approach presented above, referred to as *turn-by-turn navigation*, does not resemble a human wayfinding process and thus can be perceived as unnatural and more complicated than it should. In our opinion, making pedestrian's experience more natural should be based on the following two observations.

First, a wayfinding is an inherently interactive process, e.g. we need to know if a person is lost, if the instruction is not clear enough, etc. Human guide guarantees such interactivity, since wayfinding happens in a dialogue, hence a wayfinding system should interact with a pedestrian by means of a spoken dialogue.

Second, humans have difficulties understanding instructions based on quantitative characteristics of a spatial environment (such as distance or angles) (Ross et al., 2004), (Moar and Bower, 1983). Such instructions make humans less confident in their ability to reach the goal correctly. Hence, they tend to rely more on qualitative ones, such as salient geographical objects (*landmarks*), by simply referring to them (Denis, 1997). Such approach can be called *landmark-by-landmark navigation*. Furthermore, landmarks can be used not only when giving route descriptions, i.e. serving as a guide, but also when being guided. For instance, when giving a reassuring confirmation to

Proceedings of the SIGDial 2019 Conference, pages 422–431
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

the guide, such as “Yes, I can see a tall glass building that you’ve mentioned before”, or describing the proximal surroundings when got lost (“I believe I’m lost, but I see a pizzeria to my right”).

A prerequisite for providing such interaction capabilities is being able to identify the landmarks referred to by phrases as “a tall glass building” or “a pizzeria to my right”. Such kind of phrases is called *referring expressions* (RE) and the landmarks these phrases refer to are called *referents*. A task of matching a referring expression with its referent(s) is called *reference resolution* (RR). Guiding humans in a real city environment requires resolving exophoric spatial references, i.e. those referring to spatial objects outside of the discourse. The focus of this paper is on designing the method for solving this task.

The main contribution of this paper is a new method for resolving exophoric spatial REs, consisting of two substeps:

- a method for identifying exophoric spatial REs in spoken utterances;
- a method for resolving exophoric spatial REs to the appropriate referents, represented as 0, 1, or more geographic entities.

2 Background

Pedestrian wayfinding is an interactive, problem-solving process by which people use environmental information to locate themselves and navigate from place to place (Vandenberg et al., 2016). Despite the ubiquity of wayfinding for pedestrians, the navigation systems aiding the process, usually mobile applications, generally use methods offering *a turn-by-turn navigation*, described in the previous section. Such approach limits possibilities for interaction with the system along the route and forces the user to pay constant attention to the map on the screen. Such design can also lead to an increasing spatial anxiety (an anxious feeling when navigating in unfamiliar environments), which was shown by several studies (Hund and Minarik, 2006; Lawton and Kallai, 2002) to negatively influence pedestrian’s wayfinding performance.

In this paper we suggest to remove pedestrian’s dependency on the digital maps by interacting with a pedestrian by means of a spoken dialogue offering *a landmark-by-landmark navigation*. In fact, a number of studies (Cornell and Greidanus, 2006; Goodman et al., 2004; May et al., 2003; Denis, 1997; Lynch, 1960) have confirmed that humans reason about a spatial environment in qualitative terms, mostly relying on landmarks. As stated in (May et al., 2003), pedestrians were observed to use distances and street names much less frequently than landmarks when describing a city environment. Such approach have been observed to be more efficient for older people, who tend to find a way quicker when using a landmark-based navigation aid (Goodman et al., 2004). Pedestrians with cognitive impairment have been observed to rely on landmarks during navigation as well (Sheehan et al., 2006).

As previously stated, a landmark-based navigation requires being able to resolve exophoric spatial references. Exophoric reference resolution is not a new task in itself, but it has primarily been explored in unrealistic environments containing distinct objects that can be described by a relatively small number of visual features, e.g. recognizing one of 36 Pentomino puzzle pieces in (Kennington and Schlangen, 2015), one of 7 Tangram puzzles in (Funakoshi et al., 2012) or an object in a 3D treasure-hunt game in (Engonopoulos et al., 2013). Only recently the research started to focus on resolving references to objects in real environments. (Schlangen et al., 2015) try to identify objects in the images taken from different locations around the world. (Götze and Boye, 2017) deal with reference resolution in a complex city environment. (Chen et al., 2019) present a TOUCHDOWN dataset, where the agents navigate in a real-life visual urban environment trying to find a hidden object based on a number of cues formulated in a natural language. The presented task is then called *spatial description resolution*, i.e. given a set on instructions find the referred place, whereas *reference resolution* aims at resolving *all* references mentioned in the given instructions as well.

A number of research papers on exophoric reference resolution (eRR) decompose the problem into three subtasks: identifying referring expressions (RE), constructing a search space of candidate referents and resolving the found references. Hence, the descriptions of the existing eRR methods are decomposed in the same way.

As mentioned above, most of the studies on eRR have been conducted in an unrealistically small toy domain, hence REs can be identified manually, as in (Engonopoulos et al., 2013), (Funakoshi et al., 2012) or (Kennington and Schlangen, 2015). (Schlangen et al., 2015) and (Götze and Boye, 2017) addressed RR in a realistic domain, but all REs were manually annotated as well. (Schutte et al., 2010) worked on resolving REs in simple manipulation instructions, e.g. "hit that red button", and identified REs using a set of simple regular expressions. (Prasov and Chai, 2010) used syntactic parsing on a word confusion network, constructed out of n-best list of alternative speech recognition hypotheses. All non-pronominal NPs were then detected and said to be a set of exophoric REs.

In most research studies, the search space of candidate referents is the same for all utterances and consists of a limited number of objects, e.g. (Kennington and Schlangen, 2015), (Funakoshi et al., 2012), (Engonopoulos et al., 2013), (Matuszek et al., 2014). In these studies all candidate referents (candidate set) have a limited number of distinct properties (color, shape, size, etc) and hence each object in the search space is either represented as a combination of such properties or simply as a numeric identifier (as the search spaces are very small). (Schlangen et al., 2015) worked with resolving references to a much more diverse real-life objects in images containing object segmentations. The referred objects come from over 80 different categories and only around 2% of the objects comprise geographical entities, e.g. benches, traffic lights, fire hydrants, etc., that are of interest in the present article. The candidate set for every referring expression was set to contain all object segmentations of the given image and every candidate is encoded using a deep convolutional neural network augmented with a number of extra features. Similarly, (Götze and Boye, 2017) have dealt with a constantly changing candidate set of diverse *geographical* objects in a pedestrian's vicinity. Each geographical object was then represented by a pedestrian's position and a number of properties inferred from OpenStreetMap (OSM) (Haklay and Weber, 2008).

In most of the studies, eRR itself is solved by taking the stochastic approach by training a generative probabilistic model to estimate the distribution over a set of candidate objects and then find the most probable intended referent as:

$$O^* = \arg\max_{O} P(O|U, S), \qquad (1)$$

where U is a representation of an utterance constituting RE, S is the search space of possible referents, O is an object in the search space, O^* is the predicted referent. Such a stochastic approach is pursued, for instance, in (Kennington and Schlangen, 2015), (Engonopoulos et al., 2013), (Matuszek et al., 2014), (Funakoshi et al., 2012), (Schlangen et al., 2015), (Götze and Boye, 2017).

3 Approach

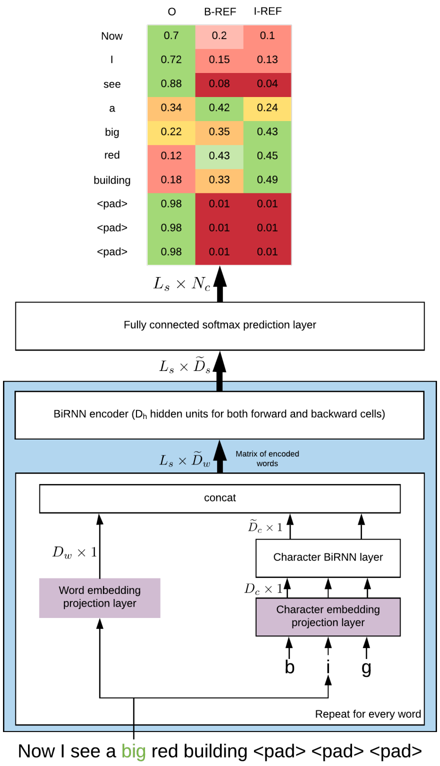
Also in this paper, spatial reference resolution is seen as a three-stage problem. First, referring expressions should be identified in the utterances and encoded into a numerical representation. We refer to this stage as *spatial referring expressions identification (sREI)*. This is achieved by a neural network, referred to as *RefNet*. Then the candidate set of referents should be constructed (described further in Sect. 4). Finally, the found referring expressions should be resolved to the appropriate referents, which we call a *spatial reference resolution (sRR) stage*. This adds a spatial dimension to the first task, hence a method name *SpaceRefNet* (also a neural network).

sREI (Sect. 3.1) is seen as a classification problem, where each word is to be assigned one of the three labels, **B-REF** (beginning of RE), **I-REF** (inside of RE), **O** (outside of RE), inspired by the BIO labeling strategy for named entity recognition (NER).

sRR (Sect. 3.2) is seen as a set of binary classification problems, each assigning a pair of an RE and a candidate object to either the positive class, if the candidate is predicted as a referent for the RE, or to the negative class, otherwise. Both stages use the same dataset, described further in Sect. 4, pre-processed in different ways.

3.1 Referring expression identification

Let us now describe the way *RefNet* operates (see Fig. 1). We start by padding (with a special word `<pad>`) or trimming every utterance to some fixed sentence length L_s. Each utterance is fed into RefNet word by word, as a part of a training batch. Each word is encoded using pre-trained D_w-dimensional distributional word embeddings (we are using GloVe (Pennington et al., 2014)). Additionally, each word is split into characters, mapped to the pre-trained $\widetilde{D}_c$-dimensional character embeddings, trained on the SpaceRef corpus using the Random Indexing technique (Kanerva et al., 2000) for a character level. These character embeddings are then fed into a bidirectional recurrent neural network (BiRNN) with gated recurrent units (GRUs), having rectifier activation functions (ReLUs) and H_c-dimensional hidden states. This BiRNN produces $(D_c = 2 \times H_c)$-dimensional *character-level word embeddings* by concatenating the hidden states of forward and backward GRUs.



	O	B-REF	I-REF
Now	0.7	0.2	0.1
I	0.72	0.15	0.13
see	0.88	0.08	0.04
a	0.34	0.42	0.24
big	0.22	0.35	0.43
red	0.12	0.43	0.45
building	0.18	0.33	0.49
<pad>	0.98	0.01	0.01
<pad>	0.98	0.01	0.01
<pad>	0.98	0.01	0.01

$L_s \times N_c$
Fully connected softmax prediction layer
$L_s \times \widetilde{D}_s$
BiRNN encoder (D_h hidden units for both forward and backward cells)
$L_s \times \widetilde{D}_w$
Matrix of encoded words
concat
$\widetilde{D}_c \times 1$
$D_w \times 1$
Character BiRNN layer
Word embedding projection layer
$D_c \times 1$
Character embedding projection layer
b
i
g
Repeat for every word
Now I see a big red building <pad> <pad> <pad>

Figure 1: RefNet architecture diagram. The purple blocks specify the pre-trained layers; thick arrows emphasize that 2D tensors of dimensionality specified to the left of arrows are passed; the blue block denotes RefNet encoder. (*Best viewed in color*)

The motivation behind taking character-level embeddings into account is that some words in an RE will inevitably lack word vectors. In such cases, the corresponding word embeddings are assigned to be zero vectors, leaving character-level embeddings as the only source of information. This amendment should be particularly helpful in at least the following two cases:

- if an RE is a proper name of a geographical object, pronounced in the language, different from the dominant language of the utterance, e.g. “Bahnstraße”, “Östvägen”;
- if an RE is a composite name with one of the constituents being recognized as a valid RE, e.g. “supermegamarket”.

The final *word encoding* is then a concatenation of the word embedding and the character-level word embedding, resulting in a $(D_w + D_c)$-dimensional vector. These word encodings are then collected into a sentence representation, which is a $L_s \times D_w$ matrix. This sentence representation is fed row-wise as a sequence into another BiRNN (with forward and backward GRUs with ReLUs having H_s hidden units).

In order to incorporate the contextual information, we want to represent a sentence as a matrix, the i^{th} row of which contains the information about the sub-sentence up until, and including, the i^{th} word. To clarify, let us say the sentence “I see a building” is being processed (the padding step is omitted for the sake of brevity), then we are interested in vectorizing all its sub-sentences in the forward direction, i.e. “I”, “I see”, “I see a”, “I see a building”, and in a backward direction, i.e. “building”, “building a”, “building a see”, “building a see I”. To achieve that, we concatenate forward and backward memory cells (which are equivalent to hidden states in case of GRUs) at each time step i. This results in $(D_s = 2 \times H_s)$-dimensional sub-sentence representations, which are concatenated into $L_s \times D_s$ matrix, referred to as *sub-sentence encoding*. The sub-network used for obtaining such encoding will be referred to as *RefNet encoder* (blue block in Fig. 1).

The rationale behind using sub-sentence encoding is that the same word can be either a part of RE or not, depending on the preceding and succeeding words. Consider the following two passages:

1. “You can see a train station to the right, it is for commuter trains and is called City’s Eastern.”
2. “You can see a train departing from the second track. It one of the city’s eastern parts.”

Finally, the sub-sentence encoding is fed into the softmax layer, which produces a $L_s \times 3$ matrix with i^{th} row representing a probability distribution over the possible labels, i.e. **O**, **B-REF**, **I-REF**, for the i^{th} word. RefNet is trained by minimizing the cross-entropy loss using Adam optimization method, presented in (Kingma and Ba, 2014).

3.2 Reference resolution

The resolution step implies matching a textual referring expression with a candidate geographical object. For performing such reference resolution (sRR) we employ another neural network architecture, dubbed as *SpaceRefNet* (see Fig. 2).

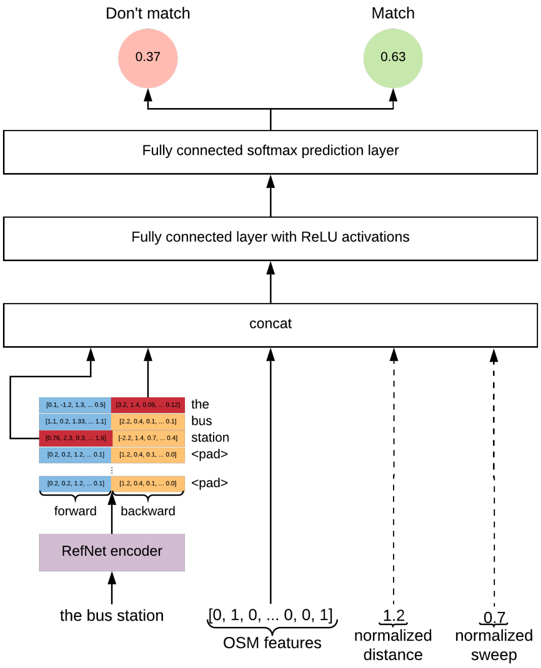


Figure 2: SpaceRefNet architecture diagram. The purple block is the pre-trained layer; the dashed arrows denote optional connections (*Best viewed in color*)

SpaceRefNet takes as an input a referring expression (RE) and a candidate geographical object, denoted as *the candidate*. The RE is encoded using the pre-trained RefNet encoder, resulting in a $L_s \times D_s$ matrix, containing forward and backward $\frac{D_s}{2}$-dimensional encodings for every sub-sentence. The final RE encoding is then the concatenation of the vectors containing forward and backward sub-sentence encodings for the whole sentence excluding paddings (the selected vectors are shown in a dark red in Fig. 2), resulting in D_s-dimensional vector. The input candidate is fed as only an OSM representation, or together with the distance and/or sweep features (see details in Sect. 4). The vectors obtained after encoding both RE and candidate are then concatenated and passed to the fully connected layer with N_h hidden units having rectifier activation functions. The final fully connected softmax prediction layer produces the probability of a match between the RE and the candidate.

SpaceRefNet is trained by optimizing the *weighted cross-entropy loss* using the Adam optimization method. Weights for the loss function are introduced, because the SpaceRef dataset has a high class imbalance – it has much more negatives (when a candidate and an RE mismatch) than positives (when a candidate and an RE match). To counteract this, a contribution of each data point (a candidate and an RE) to the global loss is adjusted using class-dependent multiplication factors (negatives receive lower weights than positives), allowing us to penalize the network more for the mistakes made on positive data points.

Such an architecture allows handling the cases when an RE has any number of referents (0, 1, or more) in the candidate set, which is an advantage compared to previously developed methods that required more ad-hoc solutions, e.g. setting an experimentally selected probability threshold in (Götze and Boye, 2017).

4 Data and processing

The utilized data consists of three datasets:

- a slightly corrected version of a publicly available *SpaceRef* dataset (Götze and Boye, 2016) (used for RefNet and SpaceRefNet training);
- a number of walks, containing the subjects' descriptions of their vicinity, which is referred to as *WalksRef* dataset[1] (used for RefNet training);
- a number of dialogues with manually annotated REs, referred to as *DialogsRef*, taken from the publicly available *Cornell Movie-Dialogs Corpus* (Danescu-Niculescu-Mizil and Lee, 2011) and *DailyDialog* corpus (Li et al., 2017) (used for RefNet training only).

The SpaceRef dataset contains descriptions of immediate geographical environment given by pedestrians following predefined routes. REs in the spoken utterances were manually annotated. GPS information representing a physical context is also available.

Referring expressions (REs) in SpaceRef are mostly noun phrases (NPs). Some example utterances with the referring expressions (underlined) include:

[1] is publicly available at https://traktor.csc.kth.se

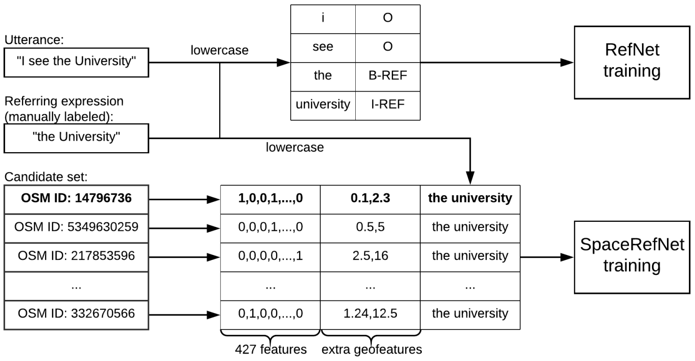


Figure 3: Data processing for training. The rows in **bold** denote a positive data point for SpaceRefNet training, i.e. the one where RE describes the given OSM entity.

- **indefinite and definite NPs**, e.g. "... walking down some stairs", "there is a fountain to my left";
- **NPs with interjections**, which should be excluded from an RE, e.g. "I am near eh the red eh brick building";
- **demonstratives**, e.g. "... standing to the right of this building";
- **proper names**, e.g. "... I am now passing 7-Eleven store".

However, not all NPs in these categories are REs, for instance,

- in the utterance "Do you know if there is a subway station nearby?", "a subway station" is not an RE, since it has no intention of referring to a specific geographic object;
- in the utterance "This architectural style I like the most", a demonstrative "this architectural style" is not an RE;
- in the utterance "The statue in front of the library portrays Carl Linnaeus", a proper name "Carl Linnaeus" is not an RE either.

SpaceRef and WalksRef contain mostly the utterances with at least one RE in them. Hence, the number of negative examples (NPs that are not REs), was not sufficient for training the neural network. With this in mind, the DialogsRef corpus was annotated providing more negative examples to improve the robustness of the trained models.

The candidate sets were *regenerated* for each referring expression by first computing lines-of-sight around the pedestrian location in 1 degree steps using a "visibility engine", inspired by (Boye et al., 2014). The lines-of-sight were computed in every direction between -100 and 100 degrees with respect to the pedestrian's walking direction. The closest OSM nodes and ways, intersecting with these lines-of-sight, were included into the candidate set as OSM identifiers.

Each candidate referent is then encoded using the following features:

- 427 binary *OSM type features*, as described in (Götze, 2016, Subsection 4.3.2).;
- *the distance feature*: the logarithm of a distance between pedestrian's and object's locations;
- *the sweep feature*: a number of lines-of-sight intersecting with an object divided by 360.

The last two features are referred to as *extra geofeatures* and a numeric vector consisting of these 429 features – as *geoencoding*.

The available data were transformed differently for training RefNet and SpaceRefNet (see Fig. 3). RefNet training requires the data to be labeled using BIO-REF labeling strategy (as mentioned before), i.e. each word in an utterance is either at the beginning of an RE and gets a label *B-REF*, or inside an RE and gets a label *I-REF*, or is not a part of an RE and gets a label *O*. SpaceRefNet training requires labeling of tuples (RE, OSM entity) with a binary label (1 if RE describes this OSM entity, 0 otherwise).

Finally, note that the training data for SpaceRefNet are heavily (and necessarily) skewed: for every referring utterance from the user, there will be about 30 candidate referents to consider, and in most cases all of them but one are not the referent the user intended. Thus, there will always be many more negative examples than positive examples in any dataset.

5 Models for comparison

5.1 Referring expression identification baseline

The REs are mostly represented by the noun phrases (NP), so the natural baseline is just returning every found NP as a candidate RE. The baseline was implemented as follows:

- a part-of-speech (POS) tag was defined for each word in an utterance using the Stanford POS tagger for English (Toutanova et al., 2003), to be more specific, the *wsj-0-18-bidirectional-distsim* version was used;
- the POS-tagged utterance is then parsed using NLTK RegexpParser (Bird et al., 2009), supplied with the following grammar:

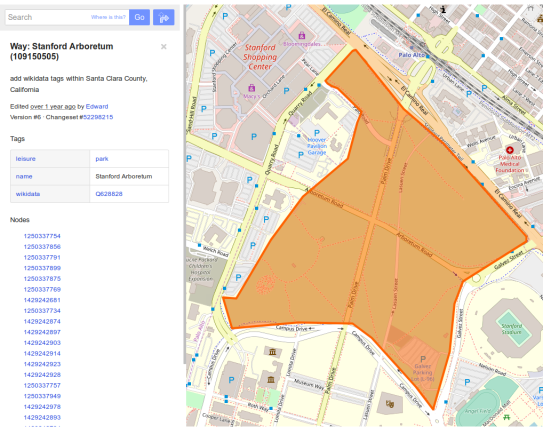
```
NP: {
  (<DT>?(<RB.*>*<JJ>*)*<NN.*>+<IN>*)+
}
```

- all found NPs are returned as REs.

5.2 Reference resolution baseline

The natural RR baseline is just querying the OSM database and checking for geographical objects with an OSM property containing at least one word from the utterance (except stop words) either in a property key or value. For example, consider two utterances, (1) "a very nice big park" and (2) "a huge green area", are being matched with the geographical object "Stanford Arboretum" (see Fig. 4). The utterances are first split by space and then all the stop words are removed. The result would be as follows: (1) {very, nice, big, park} and (2) {huge, green, area}.

Each word is then checked against all properties of the OSM object (both keys and values are checked). The first utterance will then be matched with "Stanford Arboretum", because the "leisure" tag has value "park", which is part of the utterance. The second utterance will not be matched, since none of the words matches any of the property keys or values.



Way: Stanford Arboretum (109150505)

add wikidata tags within Santa Clara County, California

Edited over 1 year ago by Edward
Version #6 · Changeset #52298215

Tags

leisure	park
name	Stanford Arboretum
wikidata	Q628828

Nodes

1250337754
1250337856
1250337791
1250337899
1250337875
1250337769
1429242681
1250337734
1429242874
1429242897
1429242903
1429242914
1429242923
1429242928
1250337757
1250337949
1429242978
1429242893

Figure 4: OpenStreetMap (OSM) representation of "Stanford Arboretum"

6 Experimental results

In all experiments the networks were trained for a maximum of 100 epochs with the early stopping (patience of 5 epochs).

6.1 Spatial RE identification

A RefNet was trained on the SpaceRef, WalksRef and DialogRef corpora. The data was split into training set (around 90% of the data containing around 90% of REs) and a test set (the remaining data). RefNet has a large number of hyper-parameters making a grid search computationally infeasible. Instead, some of the hyper-parameters were fixed to the values found through manual experiments and the others were found using a random search (Bergstra and Bengio, 2012). The hyper-parameter space was searched during 60 random trials evaluating RefNet's performance for each of hyper-parameter's combination using 5-fold cross-validation.

The model's performance was assessed by computing precision and recall, which can be done in several ways. The most straightforward way is to consider a word and its BIO-REF label as one datapoint, and compute precision and recall based on this. However, our aim is to measure how well the network identifies full referring expressions. Therefore, we've considered one data point being a tuple of all words in the referring expression together with their respective labels. The datapoint is then considered as a true positive only if *all* the words in the RE are correctly labeled (*exact match*). In order to assess the magnitude of method's errors, we say that a *partial match* occurs if the starting word is correctly labeled and there are no more than 2 errors in the rest of the expression.

During the experiments the batch size was fixed to 128, the maximum sentence length set to 100 and maximum word length to 30. The best-performing RefNet model found after the performed hyper-parameter search had 24 hidden units on the character-level BiRNN layer, 51 hidden unit on the sentence-level BiRNN layer, a learning rate of 0.002. A regularization in the form of dropout was applied with the probabilities of keeping the input, the state and the output being 0.7, 0.75, 0.95 for the character-level BiRNN and 0.8, 0.95, 0.95 for the sentence-level BiRNN respectively. The found RefNet model achieved the following performance (averaged over 5 folds):

- a precision of 0.7846 (partial precision of 0.8083);
- a recall of 0.6608 (partial recall of 0.784).

Evaluating the same model on the test set resulted in a better performance compared to the baseline (see Table 1).

Metric	Baseline	RefNet
Correct sentences (%)	21.02	77.08
Precision	0.1204	0.5457
Recall	0.2997	0.5531
Partial precision	0.1663	0.7204
Partial recall	0.4142	0.7302

Table 1: Performance of different methods for solving spatial referring expression identification (sREI) task on the test set

6.2 Spatial reference resolution

SpaceRefNet was trained exclusively on the SpaceRef corpus. The data were split into training set (around 80% of the data), validation set (around 10% of the data) and a test set (the remaining data). SpaceRefNet has a smaller number of hyper-parameters than RefNet, but a higher-dimensional input data (429 dimensions + RE encoding size). Hence, the combination of random search with cross-validation becomes computationally infeasible. Given the nature of SpaceRef data, i.e. the subjects walking along the routes in the same vicinity, the datapoints are more homogeneous compared to DialogRef and WalksRef used for RefNet training. Keeping in mind everything mentioned above, hyper-parameter space was searched using a combination of random and manual search relying on the performance on the held-out validation set.

During random search, the batch size was fixed to 256. The best found SpaceRefNet model had 32 hidden units, negatives weighted with 0.25 and positives – with 1 in the loss function, a learning rate of 0.001 and used both distance and sweep features. The model's performance was evaluated by computing precision, recall and F1-score for positives (matches between an RE and a candidate) and negatives (mismatches). The performance on the validation set was:

- for positives, precision of 0.5854, recall of 0.4444, F1-score of 0.5053;
- for negatives, precision of 0.9860, recall of 0.9748, F1-score of 0.9804;

Evaluating the same model on the test set resulted in a better performance compared to the baseline and previously reported results in the literature (see Table 2). Additionally a percentage of completely correctly labeled sentences is provided.

Metric	Baseline	WAC	SpaceRefNet
Prec. (p)	0.5588	0.40	0.6105
Rec. (p)	0.2043	0.45	0.6237
F1 (p)	0.2992	0.42	0.6170
Prec. (n)	0.9757	0.98	0.9883
Rec. (n)	0.995	0.98	0.9876
F1 (n)	0.9853	0.98	0.9879

Table 2: Performance of different methods for solving spatial reference resolution (sRR) task on the test set ("(p)" stands for positives, "(n)" stands for negatives), "WAC" stands for words-as-classifiers method (results reported in (Götze and Boye, 2017)).

7 Discussion

The designed methods have shown promising results in solving exophoric spatial reference resolution (sRR) beating the respective baselines and earlier reported results in the literature. It should be noted that sRR is a complicated task with non-trivial subproblems. Identifying REs in spoken utterances gets complicated because of multiple challenges:

- *unclear sentence segmentation* in spoken utterances results in the utterances like "I am passing the shop **on my left on my right** there is a bank", the phrase "on my left" describes the RE "the shop", whereas the phrase "on my right" describes "the bank";

- *ASR errors* can lead to the utterances like "I'm crossing the street on my **rights**";
- *interjections and self-corrections* result in utterances like "there is another shop eh called ehm jer- jersey shop".

A problem arises because of the possible differences in the interpretation. Consider the utterance "on my right is the embassy of Poland in an old fantastic villa". Depending on the interpretation, one might find either two REs "the embassy of Poland" and "an old fantastic villa" referring to the same geographic object or only one RE "the embassy of Poland in an old fantastic villa" referring to the same object. Such interpretation differences have not been considered while evaluating *RefNet*.

Resolving spatial references is even more tricky, since each found RE has mostly only one correct referent out of 30 candidates on average, making data very unbalanced. Furthermore, one RE can have multiple referents, e.g. the streets often consist of many different parts in OSM, or have no referents, e.g. some specifics about the geographical objects ("a big clock on the wall of the university"), or outdated information.

Ongoing work includes incorporating this reference resolution model into our wayfinding spoken dialogue system and collecting more data to improve the model.

Acknowledgements

This work was supported by Vinnova (Sweden's Innovation Agency) within the project 2018-01672, and by PTS (The Swedish Post and Telecom Authority) within the project 17-10324.

References

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Johan Boye, Morgan Fredriksson, Jana Götze, Joakim Gustafson, and Jürgen Königsmann. 2014. Walk this way: Spatial grounding for city exploration. In *Natural interaction with robots, knowbots and smartphones*, pages 59–67. Springer.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Edward H Cornell and Elaine Greidanus. 2006. Path integration during a neighborhood walk. *Spatial Cognition and Computation*, 6(3):203–234.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Michel Denis. 1997. The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de Psychologie*, 16:409–458.

Nikos Engonopoulos, Martin Villalba, Ivan Titov, and Alexander Koller. 2013. Predicting the resolution of referring expressions from user behavior. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1359.

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. 2012. A unified probabilistic approach to referring expressions. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*, pages 237–246. Association for Computational Linguistics.

Joy Goodman, Phil Gray, Kartik Khammampad, and Stephen Brewster. 2004. Using landmarks to support older people in navigation. In *International Conference on Mobile Human-Computer Interaction*, pages 38–48. Springer.

Jana Götze. 2016. *Talk the walk: Empirical studies and data-driven methods for geographical natural language applications.* Ph.D. thesis, KTH Royal Institute of Technology.

Jana Götze and Johan Boye. 2016. Spaceref: A corpus of street-level geographic descriptions. In *LREC*.

Jana Götze and Johan Boye. 2017. Reference resolution for pedestrian wayfinding systems. In *International Conference on Geographic Information Science*, pages 59–75. Springer.

Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Alycia M Hund and Jennifer L Minarik. 2006. Getting from here to there: Spatial anxiety, wayfinding strategies, direction type, and wayfinding efficiency. *Spatial cognition and computation*, 6(3):179–201.

Pentii Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference 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)*, volume 1, pages 292–301.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Carol A Lawton and Janos Kallai. 2002. Gender differences in wayfinding strategies and anxiety about wayfinding: A cross-cultural comparison. *Sex roles*, 47(9-10):389–401.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Kevin Lynch. 1960. *The image of the city*, volume 11. MIT press.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pages 2556–2563.

Andrew J May, Tracy Ross, Steven H Bayer, and Mikko J Tarkiainen. 2003. Pedestrian navigation aids: information requirements and design implications. *Personal and Ubiquitous Computing*, 7(6):331–338.

Ian Moar and Gordon H Bower. 1983. Inconsistency in spatial knowledge. *Memory & Cognition*, 11(2):107–113.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Zahar Prasov and Joyce Y Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481. Association for Computational Linguistics.

Tracy Ross, Andrew May, and Simon Thompson. 2004. The use of landmarks in pedestrian navigation instructions and the effects of context. In *International Conference on Mobile Human-Computer Interaction*, pages 300–304. Springer.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2015. Resolving references to objects in photographs using the words-as-classifiers model. *arXiv preprint arXiv:1510.02125*.

Niels Schutte, John Kelleher, and Brian Mac Namee. 2010. Visual salience and reference resolution in situated dialogues: A corpus-based evaluation.

Bart Sheehan, Elizabeth Burton, and Lynne Mitchell. 2006. Outdoor wayfinding in dementia. *Dementia*, 5(2):271–281.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Ann E Vandenberg, Rebecca H Hunter, Lynda A Anderson, Lucinda L Bryant, Steven P Hooker, and William A Satariano. 2016. Walking and walkability: Is wayfinding a missing link? implications for public health practice. *Journal of physical activity and health*, 13(2):189–197.