The SpaceRef corpus
Documentation

Jana Götze & Johan Boye
KTH Royal Institute of Technology
Department for Speech, Music and Hearing
100 44 Stockholm, Sweden
{jagoetze,jboye}@kth.se

June 20, 2016

1 Introduction

This documentation describes the SpaceRef data that can be downloaded from `http://www.csc.kth.se/~jagoetze/data/spaceref`.

The data contains transcriptions, annotations, and GPS coordinates from a Wizard-of-Oz study in which 11 pedestrians walk along a route and describe their actions to the wizard via a speech-only interface. Pedestrians are wearing a headset and carry a smartphone in their pocket. The data was collected in February and March 2013 in Stockholm, Sweden.

2 Data collection

2.1 Setup and materials

The participants were equipped with an Android mobile phone (Motorola Razr) that ran an application which allowed us to record their GPS coordinates and speech signal. It also allowed to send messages from the experimenter to the participant via TTS. The experimenter sat in a laboratory and used an interface which allowed him to see the participant's position on a map and type messages (cf. Figure 1). Speech signal and GPS coordinates were automatically logged and time-stamped, thereby allowing to align speech transcriptions with a participant's GPS coordinates (see below).

The route that the participants were asked to walk was a round tour that started and ended on a university campus. The route was approximately two kilometers long and was given to the participants on an unlabeled map (Figure 2). The map had street and other names removed, as well as common symbols, e.g. for churches or bus stops, in order to force the participants to rely on information that they could perceive in the physical environment rather than on the map.

2.2 Task

The following main task was given to the participants (the full task description is part of the downloadable zipfile):

> Your task is to walk along the route that you are given on a map and describe this route to the system by giving instructions to it.
>
> The system is trying to follow your path description on its own map. If it doesn't understand where to go, it will ask you to clarify your instruction.
>
> The map that you are holding in your hands is not the same map that the system can see, your map serves only as a guide for you to go along a specified way. Do not describe the route on the map, but instruct the system according to what you see when you are walking.
>
> At the time of the experiment, the system has no access to the GPS data that is recorded on the phone. This data will later be used to compare the path that the system found according to your instructions to the route that you walked.

The wizard's task was to acknowledge what the pedestrian said (i.e. clicking on a button that send the message "okay" to the phone's TTS. If the pedestrian's speech was unintelligible to the wizard, he sent a "repeat" message. The wizard also encouraged the pedestrian to speak when there was a longer stretch of movement without explanation. The extension of "longer" was not further specified but was left for the wizard's interpretation. In that case, the wizard would typically issue a message that informed the pedestrian of the last successful action, the position ("I am at the intersection of . . . "), or a clarification ("Should I cross . . . "). If there was an obvious mistake, typically a left/right error, the wizard asked for repetition or clarification. All recordings were carried out with the same wizard.

3 Data

3.1 Files

The data consists of:

- Transcribed speech, segmented into utterances.
- Annotations: Utterances are tagged with a GPS position, each utterance contains information about which object(s) according to the city model was mentioned.
- Automatically logged GPS coordinates for each pedestrian.
- The city model of the area in which the data were collected, from the time at which the data were collected (Openstreetmap file).
- A file that contains information about which OSM entities were considered to belong to the major streets.
- Questionnaire answers from the participants.
- The task description as given to the participants.

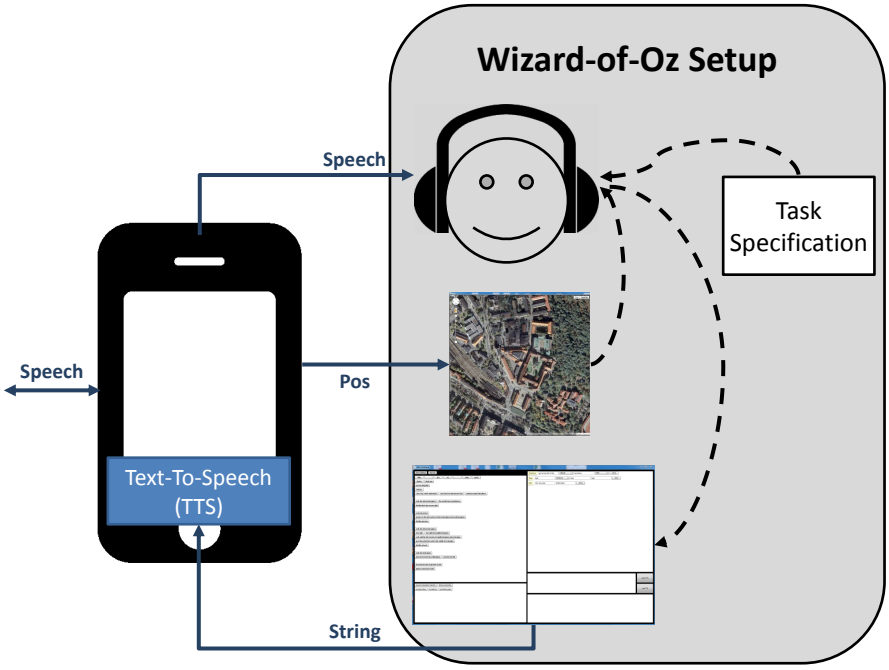

Figure 1: The Wizard-of-Oz architecture as used for the data collection

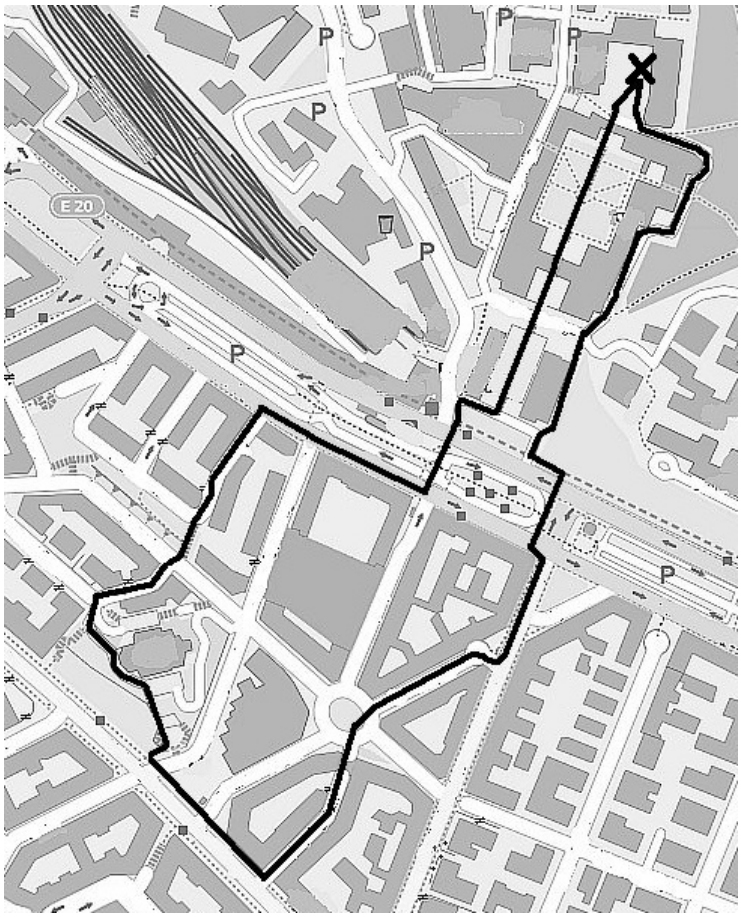

Figure 2: The unlabeled map that the pedestrians used to find their way

3.2 Data formats

3.2.1 Positional logs (geodata_log)

The GPS position is logged every second in the following format:

```
<2:1:10:4:54:161> <Geo-data    TimeStamp:2013-03-01 10:04:47.456 UTC
    time:1.362128689E12 Latitude:59.34863468333333 Longitude:18.07441215
    Speed:1.2346667 Bearing:279.5 Accuracy:16.0/>
```

3.2.2 Transcription

The **transcriptions** contain both pedestrian and wizard utterances. Wizard utterances are inserted from the log file and manually adjusted to the audio signal in which they were recorded together with the pedestrian's speech. Transcriptions were originally carried out with the Higgins Annotation Tool (HAT, `http://www.speech.kth.se/hat/`) and then transferred to the new xml format. All manual transcriptions (participant speech in transcription_X.xml) are in lower case and all numbers are spelled out. In the wizard speech, upper case and digits occur. No punctuation was used. Hesitations (*eh*, *ehm*, etc.) were not transcribed consistently (i.e. there might have been hesitations when none was transcribed).

The following are examples from the transcriptions:

- *down two steps*
- *so i'm going to walk about one hundred meters now southwest*
- *and to my right is the k t h administration building*
- *ahead of me is the bus station and a collection of trees*
- *and eh i soon eh walk behind under them*

When the recording was interrupted because the connection dropped, the segment is marked as CONNECTION_PROBLEM. When the user interrupted the recording (by calling the experimenter), this is marked as USER_INTERRUPTION.

3.2.3 Annotation

In the **annotations** (annotation_X.xml files), the wizard speech is removed, as well as some sub dialogs (see below). Special characters in names are replaced as follows:

ä	→	a
ö	→	o
å	→	a

The top-level element <States> consists of one <State> element per each utterance (wizard utterances are not contained in these files). Every <State> consists of three sub elements:

1. <Info> contains the complete transcribed utterance, the start time of the utterance, the GPS position of the pedestrian at the start time of the utterance, as well as three previous positions as extracted from the GPS logs (5, 10, and 15 seconds back).

```
<Info latitude="59.348" longitude="18.074" time="2:1:10:15:5:1156"
    utterance="there's a fountain in the middle of the park"><prev_pos_5sec
    latitude="59.348" longitude="18.074" /><prev_pos_10sec latitude="59.348"
    longitude="18.074" /><prev_pos_15sec latitude="59.348"
    longitude="18.074" /></Info>
```

2. <RESet> contains the set of referring expressions <RE> from the utterance. Each RE contains the string of the expression (NP, NB: this can be something else than a noun phrase, see below. The attribute *referential* ("y"/"n") tags whether this expression refers to something in the geographic environment (regardless of whether this object is contained in the city model). Then follow zero or more <referent> elements that specify the Openstreetmap ID of the referents according to the city model.

```
<RESet><RE NP="a fountain" NUM="singular" id="1" referential="y"><referent
    id="1607503578" /></RE><RE NP="the middle of the park" NUM="singular"
    id="2" referential="y"><referent id="20679943" /></RE></RESet>
```

3. <CandidateSet> specifies the list of candidates as it was computed using visibility information based on the GPS coordinates. This information is only present in the files ending "_cs.xml".

```
<CandidateSet><Candidate><OSM_ID id="20680082" /></Candidate>
<Candidate><OSM_ID id="20680069" /></Candidate>
<Candidate><OSM_ID id="8107989" /></Candidate>
<Candidate><OSM_ID id="20680216" /></Candidate>
<Candidate><OSM_ID id="116451953" /></Candidate>
<Candidate><OSM_ID id="20679991" /></Candidate>
<Candidate><OSM_ID id="20680049" /></Candidate>
<Candidate><OSM_ID id="1775793760" /></Candidate>
<Candidate><OSM_ID id="1775793761" /></Candidate>
<Candidate><OSM_ID id="1607503578" /></Candidate>
<Candidate><OSM_ID id="8107990" /></Candidate>
<Candidate><OSM_ID id="20680034" /></Candidate>
<Candidate><OSM_ID id="20680336" /></Candidate>
<Candidate><OSM_ID id="20680364" /></Candidate>
<Candidate><OSM_ID id="20680122" /></Candidate>
</CandidateSet>
```

For details on the choices of annotation, refer to [1] and [2].

3.2.4 City model: Openstreetmap

The city model file is an xml file downloaded from Openstreetmap at the time of the experiment. It specifies the *nodes* and *ways* in the area. More information on this data format can be found at `https://wiki.openstreetmap.org/wiki/Map_Features`. Copyright and license information can be found here: `http://www.openstreetmap.org/copyright`.

3.2.5 Paths

This file was used to facilitate annotation and include all way elements from OSM into a street. Many streets are represented by more than one OSM entity. This file assigns a shorter ID to these streets and then lists all OSM IDs that belong to that street. In the annotation_X.xml, the short IDs are used. Once the candidate set is inserted, this ID is expanded with all OSM entities contained in the candidate set.

```
1 uggelviksgatan 138623071 33057466 266080806
```

3.3 Synchronization

For synchronization of the speech files and the GPS logs, three files were used: the GPS log files, the system speech logs (everything that was sent to the TTS was automatically logged), and

the speech transcriptions. In the audio files containing the pedestrians' speech, the synthesized wizard utterances can be heard and the initial introductory lines are used as an offset to align the audio files with the two log files. The two log files are aligned in terms of time stamps and it is therefore straightforward to merge their information into one time series.

3.4 Interruptions

Some of the recordings were interrupted because the connection between the cellphone application and the wizard side broke down. The file *spaceref_meta.xlsx* contains details about the length of the interruptions. In the transcriptions, these are marked as CONNECTION_PROBLEM. During this time, the participant restarted the app as instructed by the experimenter.

Longer clarification dialogs, that were or were not successful, were excluded in the annotation.

4 Data summary

The following tables summarize different aspects of the data.

Table 1: Summary of the SPACEREF data

Number of participants	11
Male/female participants	9 / 2
Average age	27.4
Familiarity with the area (1^- to 6^+)	4.4
Voice application usage (1^- to 6^+)	1.8
Walking time	5h44m
Utterances	1,676
Referring expressions	1,323
Unique referents per participant	54.7
REs without referent on the map	58 (4%)

Table 2: Summary of referring expressions

Number of REs	1303	
Number of REs with 1 target	559	42.9%
Number of REs with 0 targets	526	40.4%
Number of REs with >1 target	218	16.7%
Av. number of targets if more than one	3.01	
Av. number of candidates per RE	33	

Table 3: Summary of the Openstreetmap data

Number of nodes	29,451
Number of ways	4,031
Number of unique tags	5,142
Number of unique tag keys	210
Average number of objects in candidate set	33

References

[1] Jana Götze and Johan Boye. 2016. *SpaceRef: A corpus of street-level geographic descriptions.* Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC).

[2] Jana Götze. 2016. *Talk the walk: Empirical studies and data-driven methods for geographical natural language applications.* Doctoral thesis. KTH Royal Institute of Technology. `http://kth.diva-portal.org/smash/record.jsf?pid=diva2:927120`