Named Entities within Multiword Addresses

Struga poster proposal, Working Group 4, work in progress

Eduard Bejček and Pavel Straňák

Abstract

In this proposal we analyze inner structure of complex address data given in text, categorize it, annotate it and evaluate the annotation using statistic named entity recognizer.

1 Introduction

There are many approaches to annotation of named entities (NEs) described in ongoing or recent projects. Most of them distinguish several types of NEs or build an entire typology. Some NEs can be expressed using more than one word—thus fulfilling a requirement to be a multiword expression—and it is more common for some of these types than for others (however it is possible for almost all of them). In multiword named entities (MWNEs), there are relations of different kinds between the words. The relation can vary from rigid (word with spaces in fixed names) through relations that can be seen as almost traditional dependency ones (e.g. company names, events, books, …) to almost no relation at all in technical data inserted within the sentence (as bibliographic information or an address). This work deals with `address` and examines its structure in detail.

2 Motivation

If authors of a treebank push themselves into annotating relations between parts of an address in a traditional dependency way, it leads to paradoxes (see Figure 1).

The address in text seems to represent a set of attribute-value pairs, such as

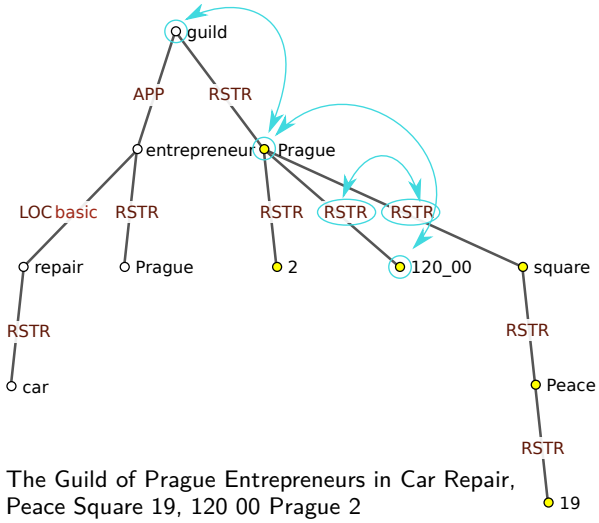

Figure 1: A complex address (translated into English) as it was annotated in PDT. For lack of a better option, the `addr:location` part was attached to the `addr:institution` part as a general modification of a noun. The same type of modification was assigned to the ZIP code or to the word "square". The later of both hardly specifies more precisely the city.

- `street="Baker street"` or
- `instit="H&W Detective Agency"` or
- `street_number="221"`

These values can again be classified as (multiword) named entities of several types (e.g. `location`, `institution` or `number` in the example above) and also grounded to entities they represent (e.g. through a Wikipedia link: `en.wikipedia.org/wiki/Baker_Street`).

Using this decomposition of NE, we can not only properly capture the information but also annotate nested NEs inside (see

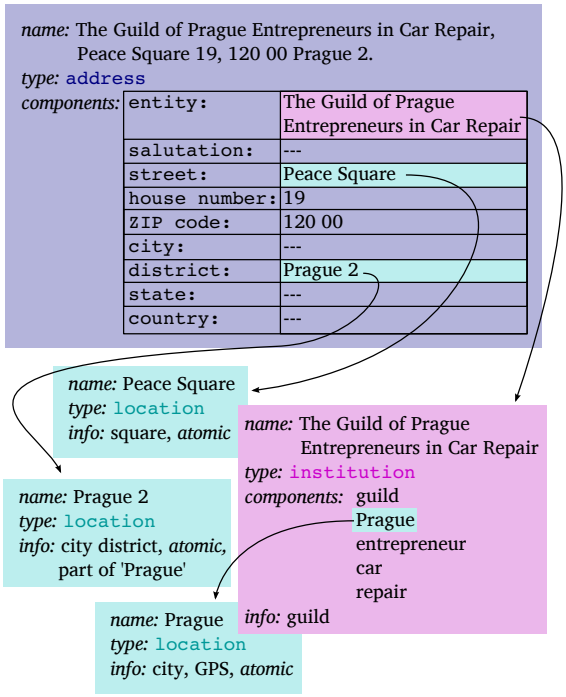

Figure 2: The scheme of the final annotation of the address from Figure 1 annotated as a set of attribute-value pairs. Some of them are marked as nested NEs with different colour and a link to some additional information in a lexicon.

Figure 2).

We try to categorize the attributes of an `address` MWNE, compile a list of them and annotate them as a subparts within `addresses` in PDT – Prague Dependency Treebank.

3 Annotation

There are several types of NEs already annotated in PDT, `address` being one of them. Our annotator deals with all of them trying to mark all their parts as one of these subtypes: • street, • street number, • ZIP code, • city, • district, • country, • phone number, • fax number, • name of person, • name of institution.

Now we show several problems or complex examples found during the annotation.

• Should we really annotate the name of an addressee (either a person or an institution) as part of an address? We consider it as another more precise specification (e.g. street → street number → flat number → tenant name) within an address.

• However, the addressee is not always part of the `address` when mentioned far from it, e.g. *"AllData Co.,* a data mining software startup, *John Smith St. 42, London"*. In this case, only the last part (*"John …"*) is an `address`; the company name is not annotated as a part of it.

• Nesting of NEs could be quite unexpected, e.g. `location` (different than in the rest of `address`) inside `addr:institute` part: *"Psychology Centre for* ***District 1****, Jerusalem St. 12, District 4, Zürich"*

• Is it better to annotate *"East End of London"* as an `addr:district` together, or split it into `addr:district`+`addr:city`?

• What should be annotated as an `addr:fax` in *"tel./FAX: (09) 795 1076"*? Only the number itself (even if it is also annotated as a `addr:phone`?) or together with *"FAX"* and with the colon?

4 Comparing with NE recognizer

The manual annotation described in previous section is compared with the output of named entity recognizer NameTag (Straková et al., 2013). The NE recognizer can find nested NEs as well as an entity of an `address` type. We show the most important cases of disagreement between our manual annotation and automatic NE recognizer. In general the NER is less granular in address parts, which we will show by comparing our annotation and the training corpus of the NER as well as the NER results on our data.

5 Conclusion

We describe our approach to annotation of addresses and their composition. We show application of our approach on a treebank by annotationg all the addresses in it and we present results in terms of both statistics and annalysis of problems. Finally we compare our approach with an approach previously adopted by authors of a named entity recogniser.