UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

A Corpus and Semantic Parser for Multilingual Natural Language Querying of OpenStreetMap

Carolin Haas, Stefan Riezler

Heidelberg University
Institute for Computational Linguistics

June 14, 2016

Motivation: Natural Language Interface to OSM

OSM http://www.openstreetmap.org/

- community-built database of geographic data, containing over 3 billion data objects
- existing search tools are very simple:
  "*Where are 3 star hotels in Paris*" returns no answer, while objects with tags "*tourism=hotel*" and "*stars=3*" exist
- "*fuzzy*" language (e.g. "*closest hotel*") is not understood
- to find these objects, one needs to issue a complicated query:
  "area[name="Paris"]→.a;node(area.a)[tourism="hotel"][stars="3"];out;"

→ develop a semantic parser to be able to answer such questions

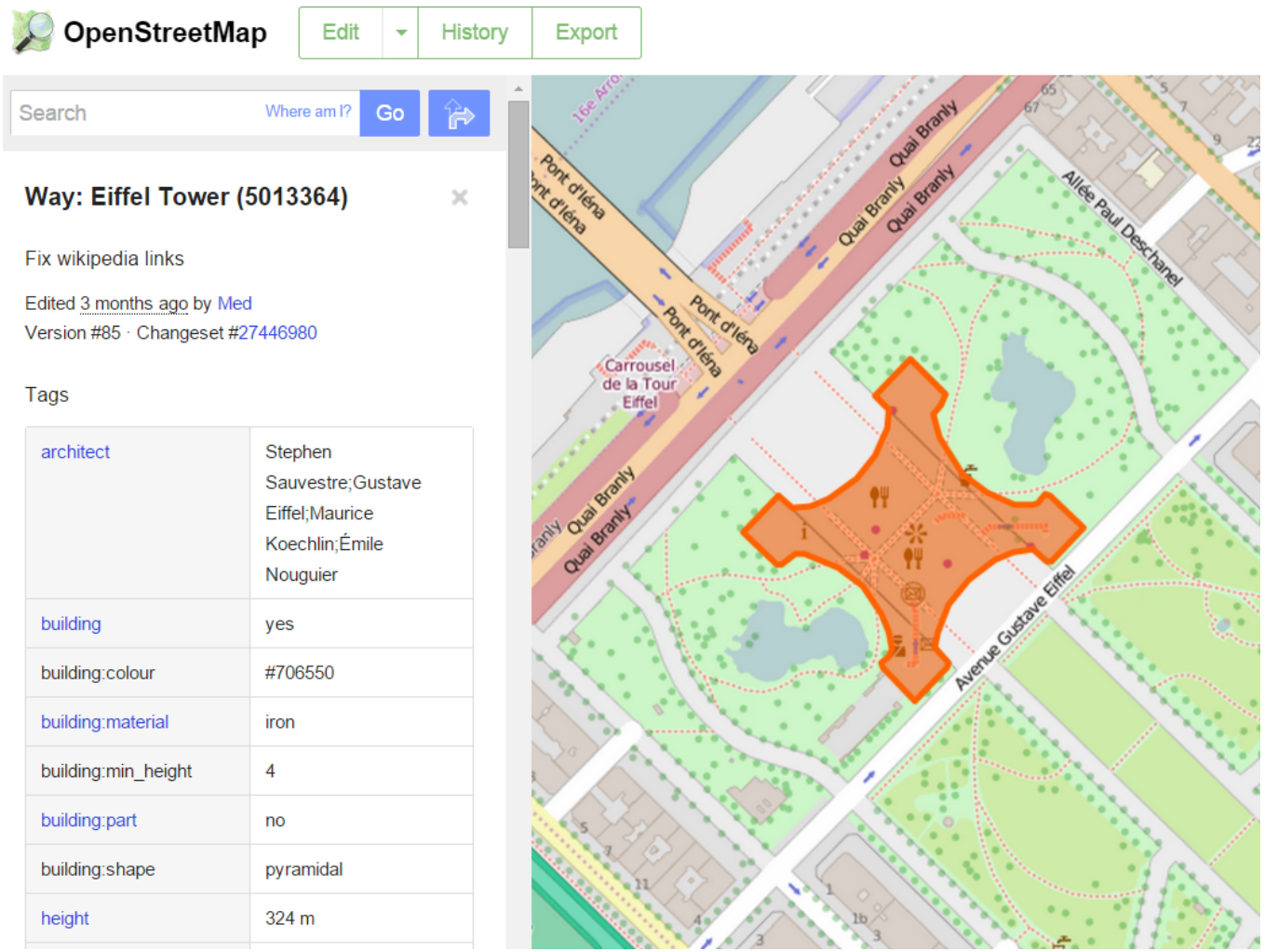
Contents

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

OpenStreetMap

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

contains nodes, ways and relations



Overpass

is a language to query the OSM database

```
area[name="Paris"]->.a;way(area.a)[name="Tour Eiffel"];out;
```

```
<?xml version="1.0" encoding="UTF-8"?>
<osm version="0.6" generator="Overpass API">
<note>The data included in this document is from www.openstreetm
<meta osm_base="2015-03-24T12:22:03Z" areas="2015-03-23T20:07:02

  <way id="5013364">
    <tag k="architect" v="Stephen Sauvestre;Gustave Eiffel;Mauri
    <tag k="building" v="yes"/>
    <tag k="building:colour" v="#706550"/>
    <tag k="building:material" v="iron"/>
    <tag k="building:min_height" v="4"/>
    <tag k="building:part" v="no"/>
    <tag k="building:shape" v="pyramidal"/>
    <tag k="height" v="324 m"/>
    <tag k="image" v="http://upload.wikimedia.org/wikipedia/comm
    <tag k="importance" v="international"/>
    <tag k="layer" v="2"/>
    <tag k="man_made" v="tower"/>
    <tag k="name" v="Tour Eiffel"/>
                ...
```

Pipeline

NL — *"Which hotels in Paris have wheelchair access?"*

via semantic parsing

MRL

```
query(area(keyval('name','Paris')),
nwr(keyval('tourism','hotel'),keyval('wheelchair','yes'))
findkey('name'))
```

extract Overpass + selection criteria

Overpass

```
area[name="Paris"]→.a;node(area.a)
["tourism"="hotel"]["wheelchair"="yes"];out;
```

execute query

XML

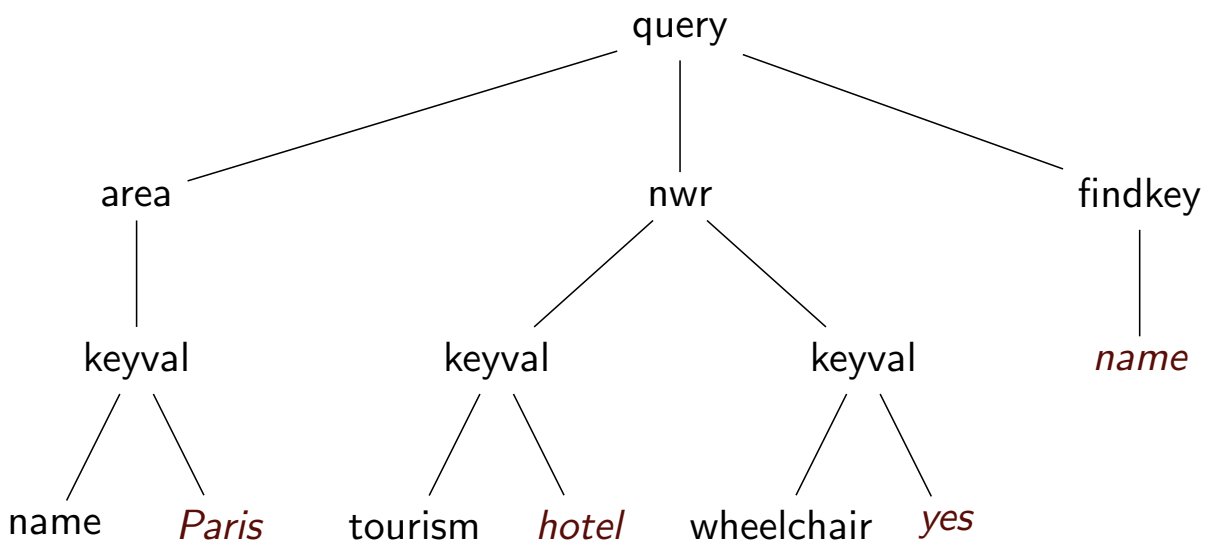
```
<node id="247435275" lat="48.8643625" lon="2.3363084">
<node id="252817440" lat="48.8301361" lon="2.3564885">
<node id="274261929" lat="48.8706664" lon="2.3304172">
  <tag k="name" v="Le Grand Hôtel Intercontinental"/>
  <tag k="tourism" v="hotel"/>
  <tag k="wheelchair" v="yes"/>
  <tag k="wikipedia" v="fr:InterContinental Paris Le Grand"/>
</node>
```

select relevant information

ANSWER — Hôtel Montpensier, Hôtel Ibis Paris Avenue d'Italie, Le Grand Hôtel Intercontinental, ...

NLMAPS' Machine Readable Language (MRL)

- NLMAPS' MRL is variable free & can be viewed as a tree
- Overpass with meta functions
- open vocabulary in *value* positions
- has a concise, unambiguous CFG



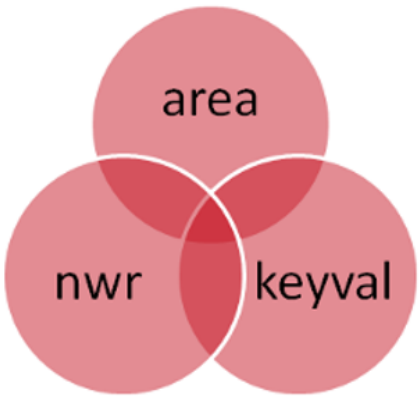
query(area(keyval(name,'Paris')),
nwr(keyval(tourism,'hotel'),keyval(wheelchair,'yes')), findkey(name))

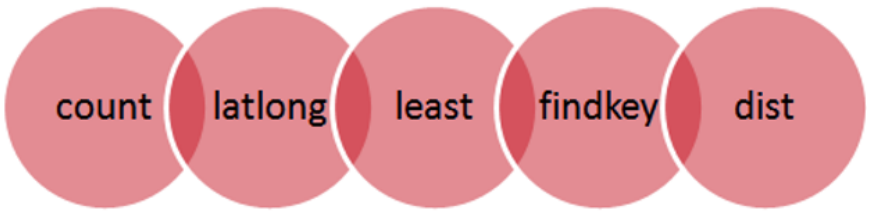
NLmaps Questions

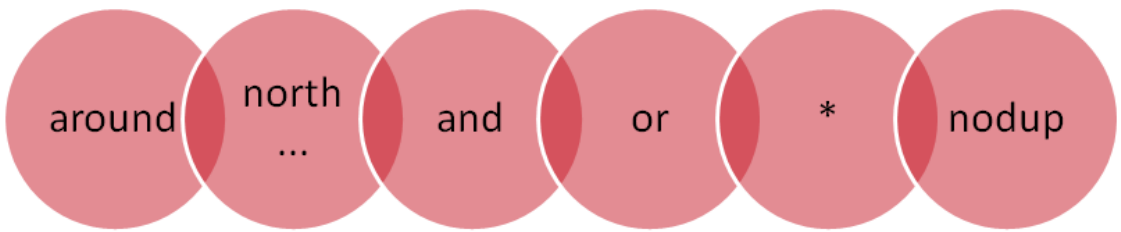
Example Questions

- What is the closest pharmacy from the Heidelberg castle?
- What are the websites and names of the museums or art centers in walking distance of the Eiffel Tower?
- What are the opening times of the Sainsbury's Local closest to the Edinburgh Waverley in Edinburgh?

MRL Components







Statistics

	NLmaps	Geoquery [1]	Free917 [2]
# sent.	2,380	880	917
tokens	25,906	6,660	6,785
types	1,002	296	2,038
avg. sent. length.	10.88	7.57	7.4
avg. types per sent.	0.42	0.34	2.22
avg. singleton per sent.	0.1	0.1	1.52
avg. NT per sent.	21	16	16
FRES	82.18	86.61	83.77

[1] (Zelle & Mooney 1996)
[2] (Berant et al., 2013)

Experiments Using SMT Semparse

(Andreas et al., 2014)

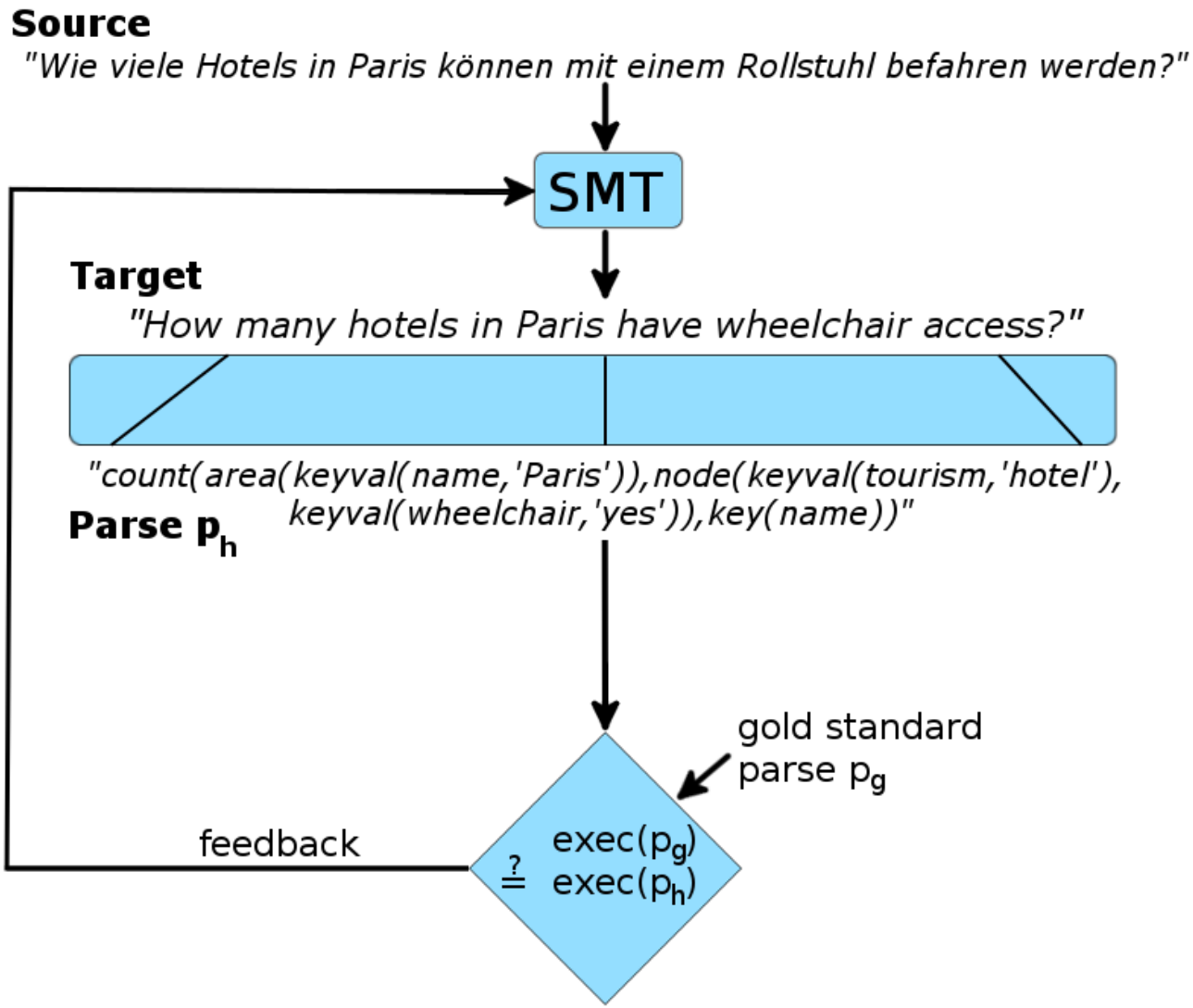
The tree-based MRL allows the formula to be broken into words
$\rightarrow$ semantic parsing can be treated as a monolingual MT task

Parsing Results

		Precision	Recall	F1
1	+intersect +stem +cdec	84.69	62.42	71.87
2	+intersect +stem +cdec +sparse	84.40	65.8	73.95[1]
3	+intersect +stem +cdec +pass +cfg	**89.45**	65.19	75.41[1]
4	+intersect +stem +cdec +sparse +pass +cfg	89.04	**68.3**	**77.3**[1,2,3]

Multilingual Database Access

(Riezler et al., 2014)



Multilingual Database Access: Idea

- prediction $\hat{y} = \arg\max_{y \in Y(x)} \langle w, \phi(x, y) \rangle$
- structured perceptron $w = w + \phi(x^{(i)}, y^{(i)}) - \phi(x^{(i)}, \hat{y})$
- $y^{(i)}$ might be unreachable

  Option 1 find translation that is as close as possible: $1 - BLEU(y^{(i)}, y)$

  Option 2 find alternatives by grounding them in semantic parsing, if answered correctly it is seen as another reference

- RAMPION (Gimpel & Smith 2012): Option 1
- EXEC: Option 2
- REBOL: Combines Option 1 & 2

Multilingual Database Access: Results

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

method	P	R	F1	BLEU
1CDEC	67.8	24.89	36.41	38.3
2EXEC	75.2	31.36	44.27[1]	40.85[1]
3RAMPION	78.21	38.75	51.82[1,2]	51.82[1,2]
4REBOL	80.76	41.02	**54.41**[1,2,3]	**51.88**[1,2]

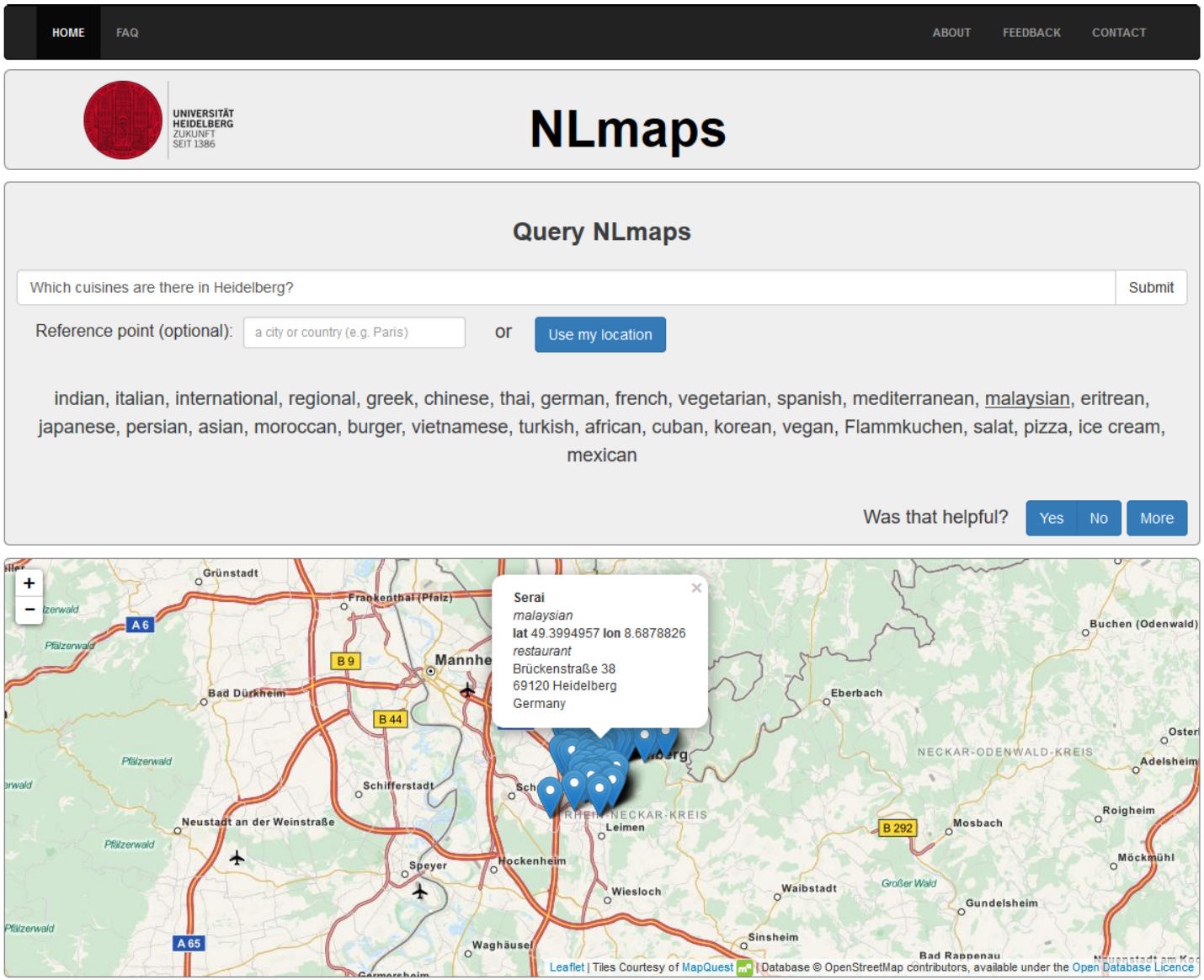
Table 1: Significant improvements are marked with superscript

Summary

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

- introduces the OSM database to QA
- a new complex corpus for QA
  where fuzzy language plays an important role
- MT based approach to semantic parsing works well for tree-based MRLs
- offers a new search tool for the OSM community
- can be extended to multilingual database access

http://nlmaps.cl.uni-heidelberg.de/



Stefan Riezler, Patrick Simianer, and Carolin Haas.
Response-based learning for grounded machine translation.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, 2014.

J. Berant, A. Chou, R. Frostig, and P. Liang.
Semantic parsing on Freebase from question-answer pairs.
In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

John M. Zelle and Raymond J. Mooney.
Learning to parse database queries using inductive logic programming.
In *Proceedings of AAAI*, Portland, OR, 1996.

Kevin Gimpel and Noah A. Smith.
Structured ramp loss minimization for machine translation.
In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies (HLT-NAACL), Stroudsburg, PA, 2012.