

# HeidelPlace: An Extensible Framework for Geoparsing

Ludwig Richter, Johanna Geiß, Andreas Spitz and Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany



#### What is Geoparsing? Why HeidelPlace?

Motivation: Geoparsing is a key task in text processing and central to subsequent spatial analyses. It describes the process of identifying place mentions (topoynms) and linking them to unambiguous spatial references. For example, geoparsing "Heidelberg was founded in 1196 AD" may reveal a reference to the German town Heidelberg.

**Problem:** Several geoparsers are avail-





#### **HeidelPlace provides:**

• A generic gazetteer model supporting integration of place information from heterogeneous knowledge bases • A pipeline approach enabling implementation and combination of modules for specific geoparsing applications

### **Generic Gazetteer Data Model**



able, each with their own gazetteer and toponym recognition and resolution approaches. However, they often lack extensibility, implementations are not accessible, or they are fixed to a particular gazetteer. This makes adjustments to other application domains difficult and prevents easy experimental setup.

• GUIs for gazetteer browsing and testing developed modules

This makes HeidelPlace a unique and valuable tool for experimenting with new geoparsing approaches.

#### The Architecture of HeidelPlace

The architecture of HeidelPlace consists of three major components:

An Annotation Pipeline executes the geoparsing process on input documents. It utilizes the Stanford CoreNLP toolkit<sup>[1]</sup>. An annotation object contains metadata for a processed document. This enables passing information between annotators, which represent processing tasks that operate on a document. An annotation pipeline iteratively executes the tasks.



A Gazetteer serves as a knowledge base for the geoparsing modules. It employs a generic gazetteer data model, which provides a wide spectrum of place information. A flexible query system allows to efficiently search the gazetteer for places with certain characteristics.

## Outlook

#### Getting ready for practical use:

Include more modules for state-of-the-art methods, e.g., co-occurrence based toponym disambiguation<sup>[2]</sup>.



**Geoparsing Modules** unify the geoparsing process. For each step in the process, a module interface is defined that specifies expected in- and output and decides when the step can be executed.

**Gazetteer Browsing** 



This architecture forms a framework for geoparsing that can be easily extended and customized.

Performing & Analyzing Geoparsing

					- 🗆 X
Text document:					Recognition Modules:
My name is Ludwig. This has nothing to do with the town Ludwig.					GAZETTEER_LOOKUP
I study at Ruprechts-Karl-Universität in Heidelberg, Germany. New York Times is a newspaper. It is 5:30 nm	Linking Modules:				
U.S. is an abbreviation for the United States of America.	GAZETTEER_LOOKUP_EXACT				
The Simpsons live in Springfield.					
what about the place stellies nostel neidelberg?					HIGHEST_ADMIN_LEVEL_DISA 🔻
Running geoparsing step LINKING took 19ms					
Running geoparsing step LINKING took 19ms GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMII LINKING took 18ms	N_LEVEL_DISAMBIGUA	ATION		Linked Places	
Running geoparsing step LINKING took 19ms GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMII LINKING took 18ms My name is Ludwig . This has nothing to do with the town Ludwig L study at Ruprechts Karl-Universität in Heidelberg Germany	N_LEVEL_DISAMBIGUA	Type	^	Linked Places	
Running geoparsing step LINKING took 19ms GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMIN LINKING took 18ms My name is Ludwig . This has nothing to do with the town Ludwig I study at Ruprechts-Karl-Universität in Heidelberg , Germany . New York Times is a newspaper. It is 5:30 p.m.	N_LEVEL_DISAMBIGUA Named Entity Heidelberg Germany	Type LOCATION	^	Linked Places 28723184 1	
Running geoparsing step LINKING took 19ms         GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMIN         LINKING took 18ms         My name is       Ludwig         I study at Ruprechts-Karl-Universität in       Heidelberg         New York Times is a newspaper. It is 5:30 p.m.         U.S.       is an abbreviation for the         United States of America       .	N_LEVEL_DISAMBIGUA Named Entity Heidelberg Germany U.S.	Type LOCATION LOCATION	^	Linked Places 28723184 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
Running geoparsing step LINKING took 19ms GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMII LINKING took 18ms My name is Ludwig . This has nothing to do with the town Ludwig I study at Ruprechts-Karl-Universität in Heidelberg , Germany . New York Times is a newspaper. It is 5:30 p.m. U.S. is an abbreviation for the United States of America . The Simpsons live in Springfield .	N_LEVEL_DISAMBIGUA Named Entity Heidelberg Germany U.S. United States	Type LOCATION LOCATION LOCATION LOCATION	^	Linked Places 28723184 28723184 29729252 20524939	
Running geoparsing step LINKING took 19ms GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMIN LINKING took 18ms My name is Ludwig . This has nothing to do with the town Ludwig I study at Ruprechts-Karl-Universität in Heidelberg , Germany . New York Times is a newspaper. It is 5:30 p.m. U.S. is an abbreviation for the United States of America . The Simpsons live in Springfield . What about the place Steffies Hostel Heidelberg?	N_LEVEL_DISAMBIGUA Named Entity Heidelberg Germany U.S. United States Simpsons	Type LOCATION LOCATION LOCATION LOCATION LOCATION	^	Linked Places 28723184 2972318384 2972318384 29723183184 2972318384 2972318384 297238384 297238384 2972383864 297238384 29728384 29728384 29728384 29728384 29728384 297283848	

#### **Desktop App Geoparser Viewer:**

• Text input provided by user

#### **Further extensions of HeidelPlace:**

- Quantitative evaluation framework
- Gazetteer web service
- UIMA component
- Gazetteer data editor

### **Comparing Geoparsing Methods**

My name is Ludwig This has nothing to do with the town Ludwig	Named Entity	Туре	Linked Places	+
I study at Ruprechts-Karl-Universität in Heidelberg , Germany .	Heidelbergjes A	LOCATION	^ 3961938	
New York Times is a newspaper. It is 5:30 p.m.	Germany	LOCATION		
U.S. is an abbreviation for the United States of America .	New York Times	ORGANIZATION		
Simpsons     live in     Springfield       What about the place Steffies Hostel     Heidelberg	U.S.	LOCATION		
	United States	LOCATION		
	Simpsons	PERSON		Logist Man data @ Open StreetMan
My name is Ludwig . This has nothing to do with the town Ludwig	Named Entity		28723184	
STANFORD_NER, GAZETTEER_LOOKUP_EXACT, HIGHEST_ADMIN_LEVEL	DISAMBIGUATION	N		
I study at Ruprechts-Karl-Universität in Heidelberg , Germany . New York Times is a newspaper. It is 5:30 p.m.	Heidelberg	LOCATION	^ 28723184	- Star Start
	Germany	LOCATION		
U.S. is an abbreviation for the United States of America .	New York Times	ORGANIZATION		3
Sumpsons         Ive in         Springfield         .           What about the place Steffies Hostel         Heidelberg         .	U.S.	LOCATION		
	United States	LOCATION		4
	Simpsons	PERSON	~	Leaflet   Map data © OpenStreetMa
GAZETTEER_LOOKUP, GAZETTEER_LOOKUP_EXACT, HIGHEST_POPULATI DISAMBIGUATION took 91ms	ON_DISAMBIGUA	TION		
My name is Ludwig . This has nothing to do with the town Ludwig .	Named Entity	Туре	Linked Places	+ 2
I study at Ruprechts-Karl-Universität in Heidelberg, Germany. New York Times is a newspaper. It is 5:30 p.m. U.S. is an abbreviation for the United States of America. The Simpsons live in Springfield. What about the place Steffies Hostel Heidelberg?	Heidelberg	LOCATION	3961938	
	Germany	LOCATION		
	U.S.	LOCATION		
	United States	LOCATION		
	Simpsons	LOCATION		

#### **Desktop App Geoparser Viewer:**

#### **Desktop App Gazetteer Viewer:**

• Selection of search filters

- Detailed view of place information
- Intuitive navigation and visualization
- Selection of (pre-configured) modules
- Step-by-step geoparsing
- Output visualization for each step
- Interactive exploration of geoparsing
- Run multiple configurations at once
- Demonstrate handling of corner cases
- Understand differences of methods
- Experiment with module combinations

# References

[1] C. D. Manning, et al.: The Stanford CoreNLP Natural Language Processing Toolkit. ACL'14, 2014

[2] A. Spitz, J. Geiß, and M. Gertz: So Far Away and Yet so Close: Augmenting Toponym Disambiguation and Similarity With Text-Based Networks. GeoRich'16, 2016

# **Contact Information:**



Ludwig Richter

ludwig.richter@posteo.de

http://event.ifi.uni-heidelberg.de

This work was presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP '17), Sept. 7-11, 2017, Copenhagen, Denmark.