

NLP AND IR METHODS FOR HANDLING GEOSPATIAL INFORMATION IN TEXTUAL DOCUMENTS

BRUNO MARTINS

JULY 11TH, 2016

DOCUMENT GEOCODING

LINKING DOCUMENTS TO GEOSPATIAL COORDINATES



A screenshot of the Wikipedia page for Kraków. The page includes the Wikipedia logo, navigation links, and the main article text. The article text describes Kraków as the second largest and one of the oldest cities in Poland, situated on the Vistula River. It mentions its historical significance as the capital of the Kingdom of Poland, the Polish-Lithuanian Commonwealth, and the Free City of Kraków. The page also features a gallery of images showing various landmarks and buildings in Kraków. The coordinates are listed as 50°4'N 19°56'E.

Wikipedia page for “Kraków”

DOCUMENT GEOCODING

LINKING DOCUMENTS TO GEOSPATIAL COORDINATES



Latitude : 50°03'41"N

Longitude : 19°56'18"E

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search

Kraków

From Wikipedia, the free encyclopedia

Coordinates: 50°4′N 19°56′E﻿ / ﻿50.067°N 19.933°E﻿ / 50.067; 19.933

For other uses, see [Krakow](#) (disambiguation) and [Cracow](#) (disambiguation).

Kraków (Polish pronunciation: [ˈkrakuf] listen (help·info)), also **Cracow** or **Krakow** (US English /ˈkrɑːkaʊ/, UK English /ˈkrækaʊ/)^{[2][3]} is the second largest and one of the oldest cities in Poland. Situated on the Vistula River (Polish: *Wisła*) in the Lesser Poland region, the city dates back to the 7th century.^[4] Kraków has traditionally been one of the leading centres of Polish academic, cultural, and artistic life and is one of Poland's most important economic hubs. It was the capital of the **Crown of the Kingdom of Poland** from 1038 to 1569; the **Polish–Lithuanian Commonwealth** from 1569 to 1795;^[5] the **Free City of Kraków** from 1815 to 1846; the **Grand Duchy of Cracow** from 1846 to 1918; and **Kraków Voivodeship** from the 14th century to 1998. It has been the capital of **Lesser Poland Voivodeship** since 1999.

The city has grown from a **Stone Age** settlement to Poland's second most important city. It began as a hamlet on **Wawel Hill** and was already being reported as a busy trading centre of Slavonic Europe in 965.^[4] With the establishment of new universities and cultural venues at the emergence of the **Second Polish Republic** in 1918 and throughout the 20th century, Kraków reaffirmed its role as a major national academic and artistic centre. The city has a population of approximately 760,000, with approximately 8 million additional people living within a 100 km (62 mi) radius of its **main square**.^[6]

After the **invasion of Poland** at the start of **World War II**, Kraków became the capital

Kraków
Royal Capital City of Kraków
Stołeczne Królewskie Miasto Kraków

Wikipedia page for “Kraków”

TOPONYM RESOLUTION

LINKING INDIVIDUAL PLACE NAMES TO GEOSPATIAL COORDINATES



?

Kraków's historic centre, which includes the [Old Town](#), [Kazimierz](#) and the [Wawel Castle](#), was included as the first of its kind on the list of UNESCO [World Heritage Sites](#) in 1978.^[73]

Wikipedia page for “Kraków”

TOPONYM RESOLUTION

LINKING INDIVIDUAL PLACENAMES TO GEOSPATIAL COORDINATES



Latitude : 50°3'15.98"N

Longitude : 19°56'11.69"E

Kraków's historic centre, which includes the Old Town, Kazimierz and the Wawel Castle, was included as the first of its kind on the list of UNESCO World Heritage Sites in 1978.^[73]

Wikipedia page for "Kraków"

HANDLING GEOSPATIAL INFORMATION IN TEXT

- **Text and GIS Increasingly combined within DH research**
 - Cartographic visualization of information in document collections
 - Document retrieval according to geospatial constraints
 - Cross-links between resources
 - Spatial Humanities Project, Pelagios Project (*i.e.*, *Pleiades+Peripleo+Recogito*)
- **Most previous work leverages *gazetteer matching*, together with *heuristics* for resolving ambiguous toponyms**
 - Place prominence, relations towards other places in same document
- **Challenges**
 - Gazetteer coverage (*e.g.*, *vague regions, vernacular places, complete metadata*)
 - Toponym ambiguity (*i.e.*, *geo/geo or geo/non-geo*)
 - Toponyms change over time, different spellings, different borders, ...
- **State of the art methods from the NLP/IR communities still rarely considered in this practical application domain**

OVERVIEW

1. *Introduction and motivation*

2. Modern NLP/IR methods

- Named entity recognition
- Entity disambiguation

3. Language modeling methods

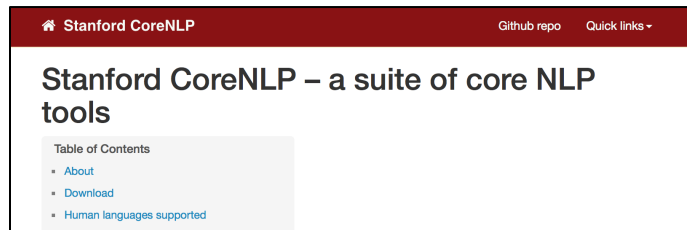
4. Conclusions

NAMED ENTITY RECOGNITION

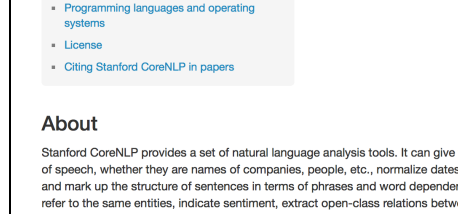
- **Delimiting spans of text that correspond to entities**
- **Within the NLP community the task is modeled as a sequence classification/tagging problem**
- **Models are learned from labeled sequences, and they can then assign probabilities to tagging decisions (and, consequently, also to sequences of tags)**
 - Hidden Markov Models
 - Conditional Random Fields
 - Deep Neural Networks (e.g., CNNs, RNNs, ...)
- **Current trends:** *avoid hand-engineered features, word embeddings, generalize across languages and domains*

NAMED ENTITY RECOGNITION RESOURCES

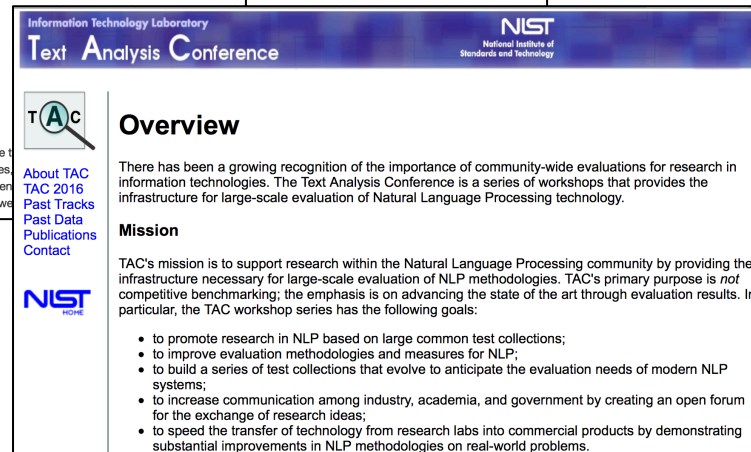
- **Stanford Core NLP and Stanford NER**
- **SENNA and systems inspired on SENNA**
- **Competition at the Text Analysis Conference**



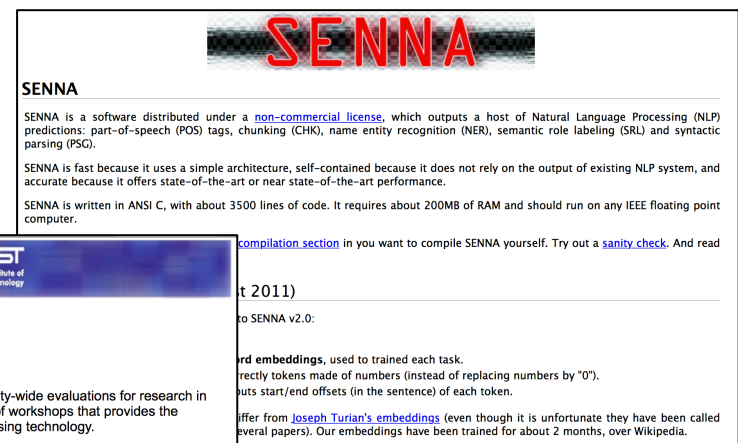
The screenshot shows the Stanford CoreNLP website. At the top, there is a navigation bar with "Stanford CoreNLP", "Github repo", and "Quick links". The main heading is "Stanford CoreNLP – a suite of core NLP tools". Below this, there is a "Table of Contents" section with links to "About", "Download", "Human languages supported", "Programming languages and operating systems", "License", and "Citing Stanford CoreNLP in papers".



The screenshot shows the "About" page of the Stanford CoreNLP website. It contains the following text: "Stanford CoreNLP provides a set of natural language analysis tools. It can give t of speech, whether they are names of companies, people, etc., normalize dates and mark up the structure of sentences in terms of phrases and word dependen refer to the same entities, indicate sentiment, extract open-class relations betwe".



The screenshot shows the "Overview" page of the Text Analysis Conference (TAC) website. The header includes "Information Technology Laboratory" and "NIST National Institute of Standards and Technology". The page title is "Text Analysis Conference". The "Overview" section states: "There has been a growing recognition of the importance of community-wide evaluations for research in information technologies. The Text Analysis Conference is a series of workshops that provides the infrastructure for large-scale evaluation of Natural Language Processing technology." The "Mission" section states: "TAC's mission is to support research within the Natural Language Processing community by providing the infrastructure necessary for large-scale evaluation of NLP methodologies. TAC's primary purpose is *not* competitive benchmarking; the emphasis is on advancing the state of the art through evaluation results. In particular, the TAC workshop series has the following goals:" followed by a list of goals: "to promote research in NLP based on large common test collections; to improve evaluation methodologies and measures for NLP; to build a series of test collections that evolve to anticipate the evaluation needs of modern NLP systems; to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in NLP methodologies on real-world problems."



The screenshot shows the SENNA project page. At the top, there is a large "SENNA" logo. Below the logo, the text reads: "SENNA is a software distributed under a [non-commercial license](#), which outputs a host of Natural Language Processing (NLP) predictions: part-of-speech (POS) tags, chunking (CHK), name entity recognition (NER), semantic role labeling (SRL) and syntactic parsing (PSG). SENNA is fast because it uses a simple architecture, self-contained because it does not rely on the output of existing NLP system, and accurate because it offers state-of-the-art or near state-of-the-art performance. SENNA is written in ANSI C, with about 3500 lines of code. It requires about 200MB of RAM and should run on any IEEE floating point computer." Below this, there is a "compilation section" with instructions: "in you want to compile SENNA yourself. Try out a [sanity check](#). And read t 2011) to SENNA v2.0: rd embeddings, used to trained each task. rectly tokens made of numbers (instead of replacing numbers by "0"). puts start/end offsets (in the sentence) of each token. offer from [Joseph Turian's embeddings](#) (even though it is unfortunate they have been called several papers). Our embeddings have been trained for about 2 months, over Wikipedia."

NAMED ENTITY DISAMBIGUATION

- **Link entities to a reference database (DB)**
- **Task is typically modeled as a candidate ranking problem, often also leveraging Wikipedia as the reference DB**
 - **Retrieve candidate disambiguations from a database**
 - Matching strings by similarity against Wikipedia concept names
 - **Rank according to likelihood of correct disambiguation**
 - ***Prior probability*** $P(\text{candidate}|\text{mention})$ from resources like Wikipedia
 - ***Context similarity*** between candidate and mention/document
 - ***Coherence*** between candidates (within same document)
 - Learn from examples to combine evidence and assign probability to candidates
- **Current trends: *global disambiguation, concept/entity embeddings***
- **Several studies proposed heuristics specific for toponyms**
 - Work my Mike Lieberman, Jochen Leidner, ...
 - Population, geospatial distance, ...



NAMED ENTITY DISAMBIGUATION RESOURCES

- **AIDA/YAGO**
- **Babelfy** (*entity linking and word sense disambiguation*)
- **Berkeley Entity Resolution** (*handles co-references*)
- **Competition at the Text Analysis Conference**

AIDA: Accurate Online Disambiguation of Named Entities in Text and Tables

News

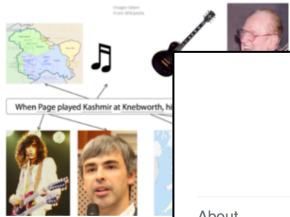
Stay up to date with AIDA news and releases, send a mail to: aida-news-subscribe@lists.mpi-inf.mpg.de

Overview

AIDA is a framework and online tool for entity detection and disambiguation. Given a natural language text or a Web table, it maps mentions of ambiguous names onto canonical entities (e.g., individual people or places) registered in the **YAGO2 knowledge base**.

You can try AIDA on any text you like in the [online demo](#). In case you are interested in using AIDA programmatically, the source code is available on github.com/yago-naga/aida.

To experimentally verify the quality of AIDA, we annotated nearly 1,400 newswire articles with the entities mentioned in each article. This collection is available for download (see [Downloads](#)).



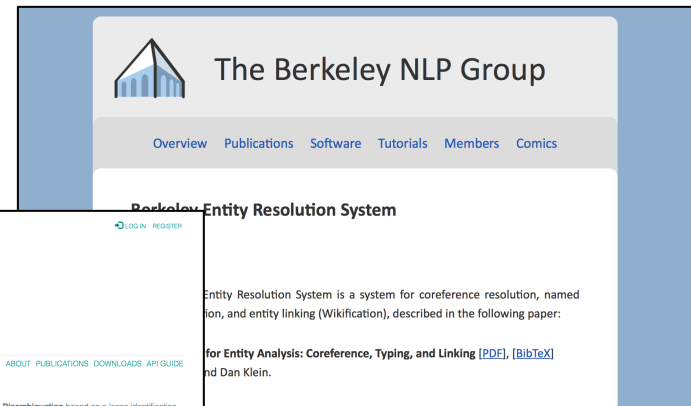
The Berkeley NLP Group

Overview Publications Software Tutorials Members Comics

Berkeley Entity Resolution System

Entity Resolution System is a system for coreference resolution, named entity resolution, and entity linking (Wikification), described in the following paper:

Coreference Resolution for Entity Analysis: Coreference, Typing, and Linking [PDF], [BibTeX] by Dan Klein.



fy B
Babelfy

ABOUT PUBLICATIONS DOWNLOADS API GUIDE

About

Babelfy is a unified, multilingual, graph-based approach to Entity Linking and Word Sense Disambiguation based on a loose identification of candidate meanings coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations. Babelfy is based on the **BabelNet 3.0** multilingual semantic network and jointly performs disambiguation and entity linking in three steps:

- It associates with each vertex of the **BabelNet** semantic network, i.e. either **concept** or **named entity**, a semantic signature, that is, a set of related vertices. This is a preliminary step which needs to be performed only once, independently of the input text.
- Given an input text, it extracts all the linkable fragments from this text and, for each of them, lists the possible **meanings** according to the **semantic network**.
- It creates a **graph-based** semantic interpretation of the whole text by linking the candidate meanings of the extracted fragments using the previously-computed semantic signatures. It then extracts a dense **subgraph** of this representation and selects the best candidate meaning for each fragment.

MODERN NLP/IR METHODS

- **Discussed methods handling named entities in general**
 - Provide very good performance on toponyms
 - Named entity recognition : accuracy around 90%
 - Entity linking : accuracy around 80%
 - Portable across tasks, languages, domains, ...
 - Methods actively developed in the NLP community, which now embraces open research and reproducibility of results
 - Robust software (although difficult to use by non experts)
- **Even if recognition leverages patterns in annotated data, disambiguation still depends on reference DB**
- **Some studies have specifically focused on handling toponyms and geospatial information...**

OVERVIEW

1. *Introduction and motivation*

2. *Modern NLP/IR methods*

- *Named entity recognition*
- *Entity disambiguation*

3. Language modeling methods

4. Conclusions

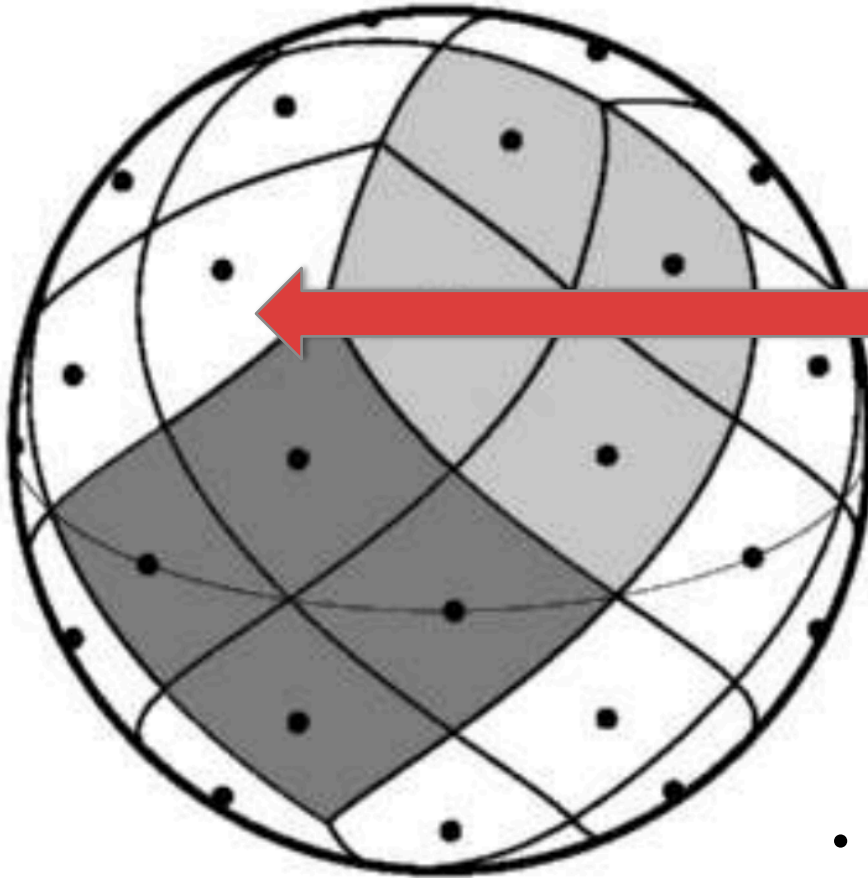
HANDLING GEOSPATIAL INFORMATION IN TEXTS



Kraków's historic centre, which includes the [Old Town](#), [Kazimierz](#) and the [Wawel Castle](#), was included as the first of its kind on the list of UNESCO [World Heritage Sites](#) in 1978.^[73]

Wikipedia page for “Kraków”

AN APPROACH BASED ON LANGUAGE MODELING



$\operatorname{argmax}_{\text{region}} P(\text{region}|\text{text})$

- Discretization of space
- Large datasets (e.g., Wikipedia)
- Standard language models

RELATED WORK

DOCUMENT GEOCODING

- **Several recent proposals based on language models (e.g., work by Baldrige et al.)**
 - **Discretize the surface of the Earth**
 - Regular grids versus hierarchical triangular meshes
 - **Train language models for each region of the discretization, with basis on available data (requires large datasets)**
 - Naïve Bayes models
 - Smoothed n-gram models
 - Discriminative classification methods
 - Neural language models (CNNs, RNNs, ...)
 - Assign region(s) most likely to generate test document
 - Many other variations (*e.g., smoothing, term selection, ...*)



RELATED WORK

TOPONYM RESOLUTION

- **Similar to document geocoding, considering text span around place reference**
 - *(often in combination with remaining text contained in the document, as back-off model)*
- **Avoid the use of gazetteers, instead relying on language models to better generalize**
 - Can handle vague geographic references (e.g., *downtown Kraków*)
 - Can handle relative references to places (e.g., *close to Kraków*)
 - Can assign text to multiple regions (e.g., raster representations)
 - ***Downside:*** Requires extensive amounts of training data

OVERVIEW

1. *Introduction and motivation*

2. *Modern NLP/IR methods*

- *Named entity recognition*
- *Entity disambiguation*

3. *Language modeling methods*

4. Conclusions

CONCLUSIONS

- Reviewed related work on the NLP/IR communities
- Described simple procedure, based on language models, for assigning text to geospatial locations
 - State-of-the-art results for document geocoding
 - Promising results in toponym resolution
- Can leverage existing resources (Wikipedia text)
- Language and domain independent
- Easy to implement (*out-of-the-box learning algorithms*)
- Efficient and easy to parallelize
- Also easy to extend...

MANY IDEAS FOR FUTURE WORK

- **Other statistical models and machine learning methods**
 - Novel neural network architectures
 - Structured sparsity (sentences, word clusters, ...)
- **Experiments with other reference datasets**
 - Many previous studies have leveraged Wikipedia
 - Other datasets: Perseus Civil War collection, SpatialML
 - **The DH community can help significantly here**
- **Explore cross-language/domain correlations**
 - Much more data is available for English newswire text
- **Extensions and applications in other related tasks**
 - Assignment to geospatial regions instead of coordinates
 - Resolving trajectories described within documents
 - Extracting place characteristics and relations between entities and places

QUESTIONS?

BRUNO MARTINS

JULY 11TH, 2016

THANKS TO MY STUDENTS...



(they actually did most of the work!)