Text-based annotation of scientific images using Wikimedia categories

TIR 2018

Frieda Josi, Christian Wartena, Jean Charbonnier

University of Applied Sciences and Arts Hanover

September 4th, 2018

Scientific Image Search

NOA - Scientific Image Search

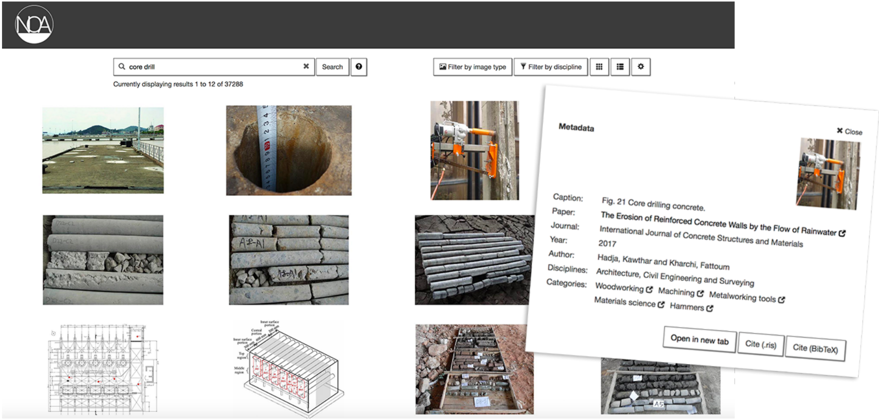
- Reuse of open access media
- Using Wikipedia categories for classification of images

Overview

Background	Data records and Wikipedia categories
Problem	Annotating scientific images
Solution 1/2	Term ranking
Solution 2/2	Category filtering and ranking
Evaluation	Wikimedia Commons categories
Application	Implementation in scientific image search

NOA - Scientific Image Search

- DFG-Project (2016-2019)



Currently displaying results 1 to 12 of 37288

Metadata

Caption: Fig. 21 Core drilling concrete.

Paper: The Erosion of Reinforced Concrete Walls by the Flow of Rainwater

Journal: International Journal of Concrete Structures and Materials

Year: 2017

Author: Hadja, Kawthar and Kharchi, Fattoum

Disciplines: Architecture, Civil Engineering and Surveying

Categories: Woodworking, Machining, Metalworking tools, Materials science, Hammers

NOA Search: `http://noa.wp.hs-hannover.de`

Data

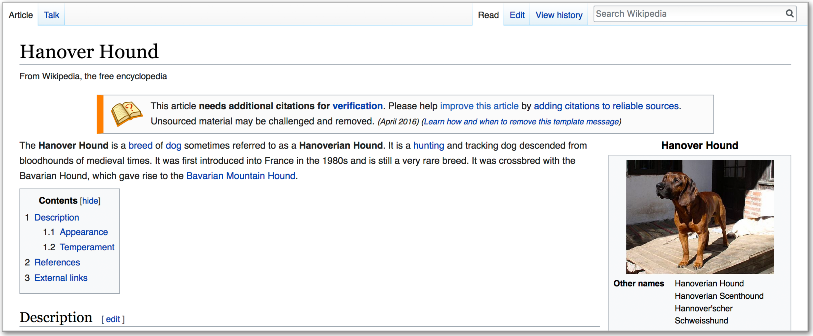
Data

- 3 million figures from 1 million open access papers
- Publisher: Springer, Hindawi, Copernicus, Frontiers etc.
- Subjects: Medicine, Science, Biology, Technology, Chemistry, ...

Data records for method development

- 397 data records: containing caption and sentences referring to the image
- Captions consist on average of 308 words

Wikipedia categories



Hanover Hound

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(April 2016) (Learn how and when to remove this template message)*

The **Hanover Hound** is a breed of dog sometimes referred to as a **Hanoverian Hound**. It is a hunting and tracking dog descended from bloodhounds of medieval times. It was first introduced into France in the 1980s and is still a very rare breed. It was crossbred with the Bavarian Hound, which gave rise to the Bavarian Mountain Hound.

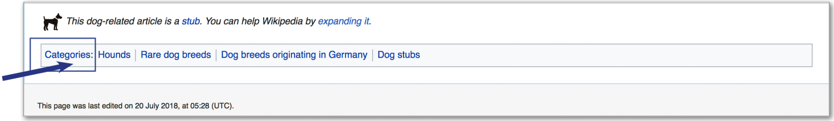
Contents [hide]

Hanover Hound

Other names	Hanoverian Hound Hanoverian Scenthound Hannover'scher Schweisshund

Description [edit]

• • •



This dog-related article is a stub. You can help Wikipedia by expanding it.

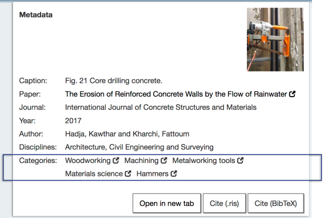
Categories: Hounds | Rare dog breeds | Dog breeds originating in Germany | Dog stubs

This page was last edited on 20 July 2018, at 05:28 (UTC).

Wikipedia categories

Wikipedia categories

- Using Wikipedia categories for classification
- Upload images to Wikimedia Commons
- Classification used as search field in NOA



Metadata

Caption:	Fig. 21 Core drilling concrete.
Paper:	The Erosion of Reinforced Concrete Walls by the Flow of Rainwater
Journal:	International Journal of Concrete Structures and Materials
Year:	2017
Author:	Hadja, Kawthar and Kharchi, Fattoum
Disciplines:	Architecture, Civil Engineering and Surveying
Categories:	Woodworking Machining Metalworking tools Materials science Hammers

Filter Categories

- Categories with meta information
- Hidden categories
- Container category
- List of regular expressions for filtering, e.g. all categories that contain the word *Wikipedia* or *stub* or *disambiguation*

Wikipedia categories

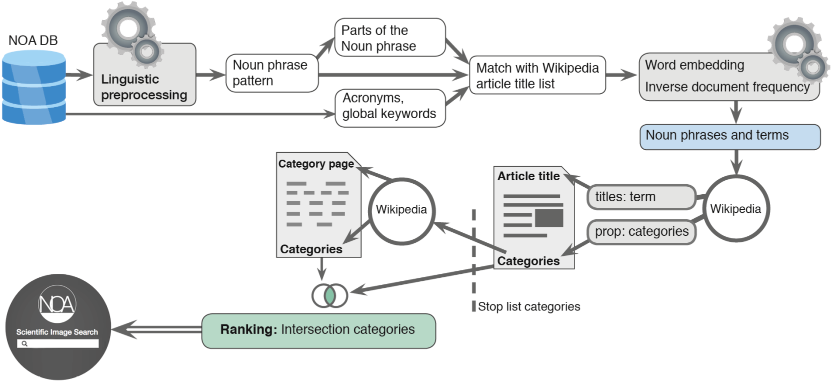
Size

After filtering: 5,1 Million categories!

Example

- Logan County, Colorado
- University of New England (United States)
- People from Apache Junction, Arizona
- Castles in the Hunsrück
- Cities in Sussex County, Delaware
- Theatres in Brighton and Hove
- Years in Bangladesh
- Armenian male stage actors
- Headlands of Greece
- Football competitions in Ivory Coast

Annotating scientific images



Noun phrase extraction

Linguistic preprocessing

- Tokenization, part of speech tagging (NLTK)
- Lemmatization (Wordnet lemmatizer)

Extracting terms for mapping with Wikipedia titles

- Noun phrases (Regular expression over POS tags)

$$\mathrm{NP} : (< \mathrm{CD} >)?(< \mathrm{JJ} >)* < \mathrm{N}(\mathrm{N}|\mathrm{P}).* > + \tag{1}$$

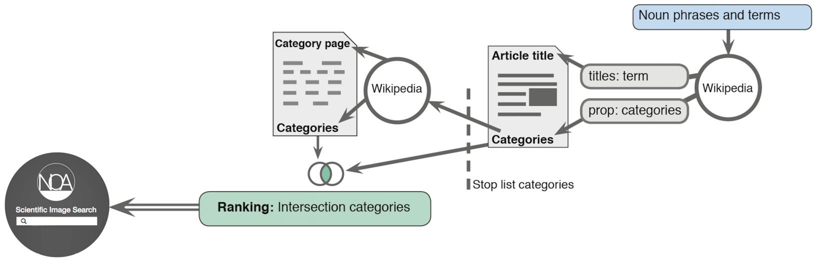
- POS tags: Penn Treebank Tagset
  (CD = cardinal number, JJ = adjective, NN = nouns, NP = proper nouns)

Examples

NP: deep fascia
POS tags: deep pos='JJ', fascia pos='NN'

Term mapping

Wikipedia API/ SQL-Dump



- Full noun phrase (NP) is a Wikipedia article title
- NP is further and further split into shorter phrases
- longer (and more specific) phrase will be used
- Smaller phrases used if the longer phrase is not found
- Pluralize words if the singular form was not found
- E.g. specific long phrase *Greenhouse gas* vs. *gas*

Term extraction

Example

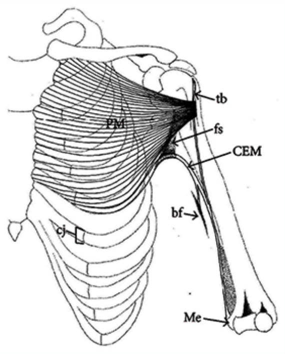


Figure 1: Schematic drawing of the left thorax and upper limb, demonstrating the chondroepitrochlearis muscle (CEM) inserting into the deep brachial fascia (bf) and the fibrous band (tuberoepicondylar band, tb) (PM: pectoralis major; fs: fascial sling; cj: costochondral junction; and Me: medial epicondyle).

Term/Noun phrases extraction

Example

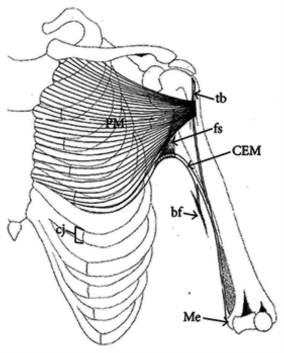


Figure 1: Schematic drawing of the left thorax and upper limb, demonstrating the chondroepitrochlearis muscle (CEM) inserting into the deep brachial fascia (bf) and the fibrous band (tuberoepicondylar band, tb) (PM: pectoralis major; fs: fascial sling; cj: costochondral junction; and Me: medial epicondyle).

Term/Noun phrases ranking

Inverse document frequency of noun phrases

$$idf = log \frac{\text{Number of data records in the corpus}}{\text{Number of data records containing NP}} \qquad (2)$$

Word embedding

Similarity of noun phrases with the captions

- For all words in the corpus that occur at least 5 times
- Word2vec model with window size 5 (CBOW model), embedding size of 300 and a minimum word occurrence threshold of 5.
- Cosine between key phrase vector and average vector of all words in the caption

Word embeddings

Goal

- Caption provides most precise image description
- Caption might not contain the required ’keywords’
- Context might provide more interesting candidates

Example term selection

Wikipedia terms found in the caption (c) and referring context (r) of an image.

Wikipedia Terms	src	idf	cos	Wikipedia Terms	src	idf	cos
axillary fascia	r	20.0	0.72	inch	r	10.9	0.33
griffith university	r	18.1	0.20	upper limb	c	10.8	0.65
brachial fascia	c	17.5	0.77	humerus	r	10.4	0.62
quartus	r	15.7	0.35	continuation	r	10.2	0.24
medical literature	r	14.4	0.26	fascia	c	10.0	0.75
common name	r	13.9	0.26	nomenclature	r	9.4	0.15
deep fascia	r	13.9	0.75	depiction	r	9.4	0.23
epicondyle	c	12.7	0.76	rib	r	9.3	0.59
joint capsule	r	12.4	0.58	informed consent	r	9.3	0.59
queensland	r	12.2	0.16	wood	r	9.2	0.24
cadaver	r	11.4	0.40	septum	r	9.1	0.56
axilla	r	11.3	0.56	thorax	c	9.1	0.58
biceps	r	11.1	0.69	...	...	...	...
tubercle	r	10.9	0.57	number	r	2.3	0.19

Ranking - Variants

3 Variants

1. 5 Terms with highest idf
2. 5 Terms with highest cosine similarity
3. 5 Terms with highest cosine similarity from 15 terms with highest idf

Category ranking

Definition of the weight for the category $w(c)$ as:

$$w(c) = \sum_{l=0}^{2} w_l \cdot r_l(c) \tag{3}$$

Association of term and category

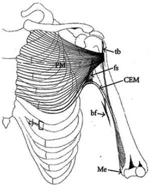
- **Level 0**: c is a category of the Wp. article with title k and $c = k$
- **Level 1**: c is a category of the Wp. article with title k and $c \neq k$
- **Level l**: c has a subcategory associated with k at level $l - 1$ and c != associated with k at level $l - 1$
- $\mathbf{r_l(c)}$: the number of keywords associated with c at level l

Weights

$w_0 = 1.2$, $w_1 = 1.0$ and $w_2 = 0.4$

Example category ranking

The article *Fascia* has the category *Fascia*, so the category *Fascia* is associated at level 0 with the keyword *fascia*.

Category	Value	Images
Fascia	3.0	
Muscular system	1.6	
Musculoskeletal system	1.6	
Soft tissue	1.2	
Connective tissue	1.2	
Tissues (biology)	1.2	
Elbow flexors	1.0	
Forearm supinators	1.0	
Muscles of the upper limb	1.0	
Shoulder flexors	1.0	
Skeletal system	1.0	
Medical Subject Headings	0.8	
...		

Evaluation - Data

Categories for evaluation

- 100 Images uploaded to Wikimedia Commons and manually annotated with categories.
- The images received 264 Wikimedia Commons categories
- Our annotation method uses Wikipedia categories

Uploaded images to Wikimedia Commons:
`https://commons.wikimedia.org/w/index.php?title=Special:ListFiles/Sohmen&ilshowall=1`

Evaluation - Method

Scope of literal consistency:

- Slight differences (Wikipedia vs. Wikimedia Commons)
- Singular and plural form of a category (soil vs. soils)

Reasons for semantic consistency:

- Useful categories, not completely wrong
- Suitable for annotation of images

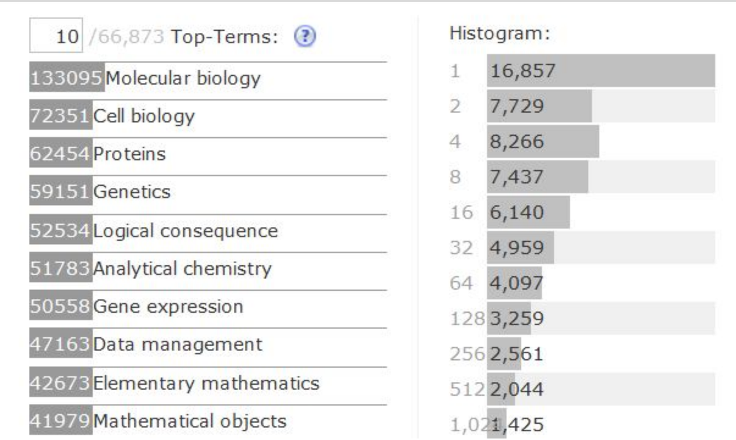
Commons Category	**Wikipedia Category**
Molecular biology	Molecular modeling
Temperature comparisons	Thermodynamics
Cochlear implants	Hearing
Robotics	Robots
Infectious disease control	Infectious diseases

Evaluation - Results

Precision and Recall

Method	Literal			Semantic		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Variant 1	0.036	0.015	0.021	0.42	0.36	0.39
Variant 2	0.054	0.059	0.057	0.40	0.40	0.40
Variant 3	0.053	0.058	0.055	0.42	0.40	0.41

Implementation in scientific image search



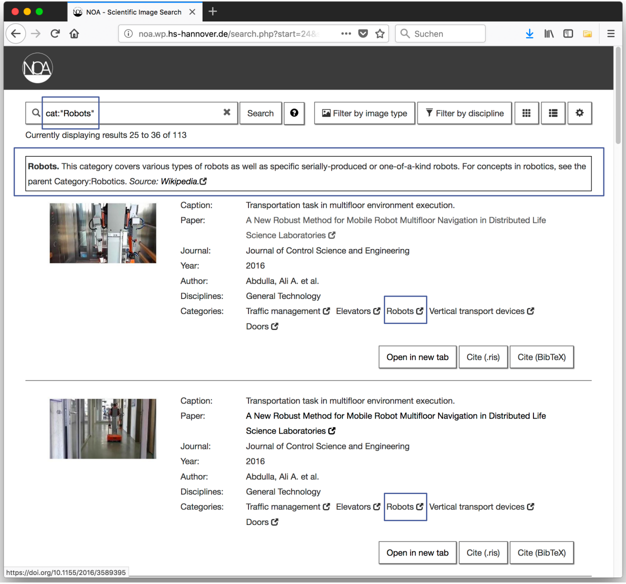
10 /66,873 Top-Terms: ?

133095	Molecular biology
72351	Cell biology
62454	Proteins
59151	Genetics
52534	Logical consequence
51783	Analytical chemistry
50558	Gene expression
47163	Data management
42673	Elementary mathematics
41979	Mathematical objects

Histogram:

1	16,857
2	7,729
4	8,266
8	7,437
16	6,140
32	4,959
64	4,097
128	3,259
256	2,561
512	2,044
1,02	1,425

Implementation in scientific image search



NOA - Scientific Image Search
noa.wp.hs-hannover.de/search.php?start=24&
Suchen
NOA

cat:"Robots" ✖ Search | Filter by image type | Filter by discipline

Currently displaying results 25 to 36 of 113

Robots. This category covers various types of robots as well as specific serially-produced or one-of-a-kind robots. For concepts in robotics, see the parent Category:Robotics. *Source: Wikipedia.*

Caption:	Transportation task in multifloor environment execution.
Paper:	A New Robust Method for Mobile Robot Multifloor Navigation in Distributed Life Science Laboratories
Journal:	Journal of Control Science and Engineering
Year:	2016
Author:	Abdulla, Ali A. et al.
Disciplines:	General Technology
Categories:	Traffic management Elevators Robots Vertical transport devices Doors

Caption:	Transportation task in multifloor environment execution.
Paper:	A New Robust Method for Mobile Robot Multifloor Navigation in Distributed Life Science Laboratories
Journal:	Journal of Control Science and Engineering
Year:	2016
Author:	Abdulla, Ali A. et al.
Disciplines:	General Technology
Categories:	Traffic management Elevators Robots Vertical transport devices Doors

References

Project-related links

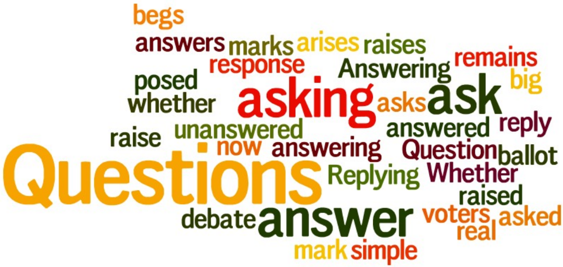
- NOA: `http://noa.wp.hs-hannover.de`
- Project information: `http://blogs.tib.eu/wp/noa/en`

Project-related publications

- NOA: A Search Engine for Reusable Scientific Images Beyond the Life Sciences (ECIR 2018)
- Discovery and efficient reuse of technology pictures using Wikimedia infrastructures (TPDL 2018)
- Using Word Embeddings for Unsupervised Acronym Disambiguation (Coling 2018)

Contact

Thanks for your attention!



- Frieda Josi: `frieda.josi@hs-hannover.de`
- Christian Wartena: `christian.wartena@hs-hannover.de`