Morphoogle - A Multilingual Interface to a Web Search Engine

Morphosaurus

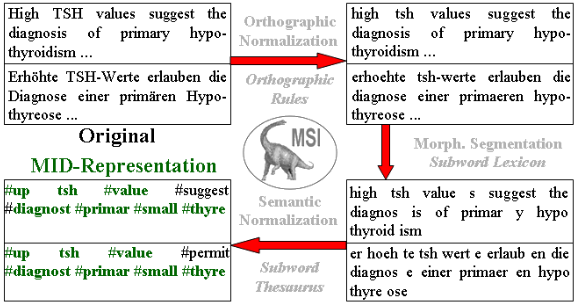
Freiburg University Hospital [1] Jena University [2] Pontificial University of Paraná [3]

MediLOG JULIE HTMP

Philipp Daumke[1], Stefan Schulz[1,3], Kornél Markó[1,2]

Morphoogle is a query translation and expansion tool which allows users to search for biomedical content in different languages in the web. This approach is based on **Morphosaurus**, a system which transforms medical text into a language independent interlingua.

This underlying procedure is named **Morpho-Semantic Indexing (MSI),** a term normalization methodology developed by the authors, which deals with various morphological processes in different languages. MSI uses a special type of dictionary, whose entries consist of **subwords**, i.e. semantically minimal units. Subwords are grouped into **language independent equivalence classes**, represented by Morpheme identifiers (MIDs). A morphosyntactic parser extracts subwords from texts and assigns MIDs in a three step procedure (cf. Figure 1).


High TSH values suggest the diagnosis of primary hypothyroidism ...

Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...

Original

high tsh values suggest the diagnosis of primary hypothyroidism ...

erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose ...

high tsh value s suggest the diagnos is of primar y hypo thyroid ism

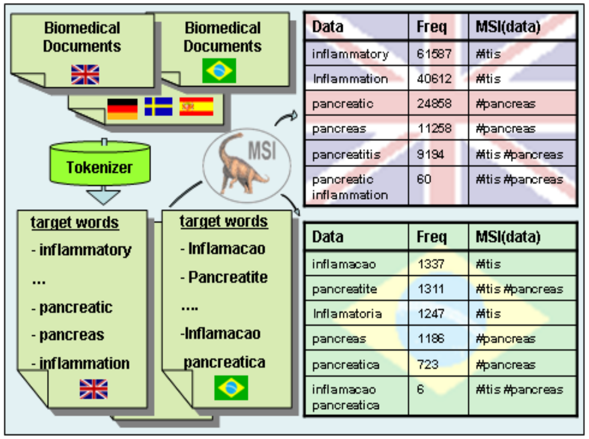
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose

MID-Representation

#up tsh #value #suggest #diagnost #primar #small #thyre

#up tsh #value #permit #diagnost #primar #small #thyre

Figure 1: Morpho-Semantic Indexing (MSI)



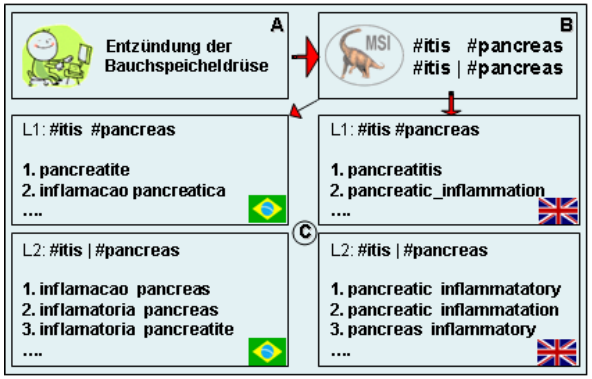
target words
- inflammatory
...
- pancreatic
- pancreas
- inflammation

target words
- Inflamacao
- Pancreatite
....
-Inflamacao pancreatica

Data	Freq	MSI(data)
inflammatory	61587	#itis
Inflammation	40612	#itis
pancreatic	24858	#pancreas
pancreas	11258	#pancreas
pancreatitis	9194	#itis #pancreas
pancreatic inflammation	60	#itis #pancreas

Data	Freq	MSI(data)
inflamacao	1337	#itis
pancreatite	1311	#itis #pancreas
Inflamatoria	1247	#itis
pancreas	1186	#pancreas
pancreatica	723	#pancreas
inflamacao pancreatica	6	#itis #pancreas

Figure 2: Generation of target word databases



Entzündung der Bauchspeicheldrüse

#itis #pancreas
#itis | #pancreas

L1: #itis #pancreas

1. pancreatite
2. inflamacao pancreatica
....

L1: #itis #pancreas

1. pancreatitis
2. pancreatic_inflammation
....

L2: #itis | #pancreas

1. inflamacao pancreas
2. inflamatoria pancreas
3. inflamatoria pancreatite
....

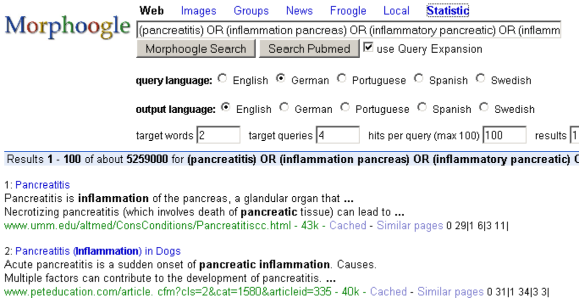
L2: #itis | #pancreas

1. pancreatic inflammatatory
2. pancreatic inflammatation
3. pancreas inflammatory
....

Figure 3: Output of the dictionary
(here in Portuguese and English)

Morphoogle

We acquired **domain and language specific corpora** from various medical sources in the WWW. Using a tokenizer we then created large **lists** of surface words, bigrams and trigrams **of adjacent words** containing their frequencies within these corpora (target words). All target words are translated to a set of MIDs and stored in language specific databases. These databases consist of about 3 M entries each (cf. Figure 2).



(pancreatitis) OR (inflammation pancreas) OR (inflammatory pancreatic) OR (inflamm

☑ use Query Expansion

query language: ○ English ● German ○ Portuguese ○ Spanish ○ Swedish

output language: ● English ○ German ○ Portuguese ○ Spanish ○ Swedish

target words 2 target queries 4 hits per query (max 100) 100 results 1

Results **1** - **100** of about **5259000** for **(pancreatitis) OR (inflammation pancreas) OR (inflammatory pancreatic) (**

1: Pancreatitis
Pancreatitis is **inflammation** of the pancreas, a glandular organ that **...**
Necrotizing pancreatitis (which involves death of **pancreatic** tissue) can lead to **...**
www.umm.edu/altmed/ConsConditions/Pancreatitiscc.html - 43k - Cached - Similar pages 0 29|1 6|3 11|

2: Pancreatitis (**Inflammation**) in Dogs
Acute pancreatitis is a sudden onset of **pancreatic inflammation**. Causes.
Multiple factors can contribute to the development of pancreatitis. **...**
www.peteducation.com/article. cfm?cls=2&cat=1580&articleid=335 - 40k - Cached - Similar pages 0 31|1 34|3 3|

A user can send a **query** via a web interface. Again, this query is firstly altered to a **set of corresponding MIDs**. This MID set is used to create a list of possible reading variants (partitions) (cf. Figure 3). Each partition consists of one or more subwords which are now **compared to the relevant databases**. All matching records are finally sorted using several heuristics and sent to a search engine (e.g. Google) (cf. Figure 4).

www.morphosaurus.net -> Web Tools -> Morphoogle

Contact: Philipp Daumke, Freiburg Medical University, Dept. Medical Informatics, Freiburg, Germany, http://www.morphosaurus.net, email: daumke@coling.uni-freiburg.de