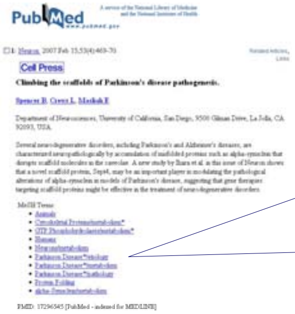
Automatic Indexing of the Biomedical Literature with a Controlled Vocabulary

Aurélie Névéol, Sonya E. Shooshan, Susanne M. Humphrey, James G. Mork and Alan R. Aronson
U.S. National Library of Medicine
Lister Hill National Center for Biomedical Communications

Indexing the Biomedical Literature in MEDLINE®

The MEDLINE database contains citations for 17 million articles in the biomedical domain. On average 2,500 citations are added weekly. The ~130 indexers at the National Library of Medicine (NLM) use Medical Subject Headings (MeSH® terms) to describe the content of articles.



PubMed www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

1: Neuron. 2007 Feb 15;53(4):469-70.

Related Articles, Links

Cell Press

Climbing the scaffolds of Parkinson's disease pathogenesis.

Spencer B, Crews L, Masliah E.

Department of Neurosciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

Several neurodegenerative disorders, including Parkinson's and Alzheimer's diseases, are characterized neuropathologically by accumulation of misfolded proteins such as alpha-synuclein that disrupts scaffold molecules in the synapse. A new study by Ihara et al. in this issue of Neuron shows that a novel scaffold protein, Sept4, may be an important player in modulating the pathological alterations of alpha-synuclein in models of Parkinson's disease, suggesting that gene therapies targeting scaffold proteins might be effective in the treatment of neurodegenerative disorders.

MeSH Terms:

- Animals
- Cytoskeletal Proteins/metabolism*
- GTP Phosphohydrolases/metabolism*
- Humans
- Neurons/metabolism
- Parkinson Disease/*etiology
- Parkinson Disease/*metabolism
- Parkinson Disease/*pathology
- Protein Folding
- alpha-Synuclein/metabolism

PMID: 17296545 [PubMed - indexed for MEDLINE]

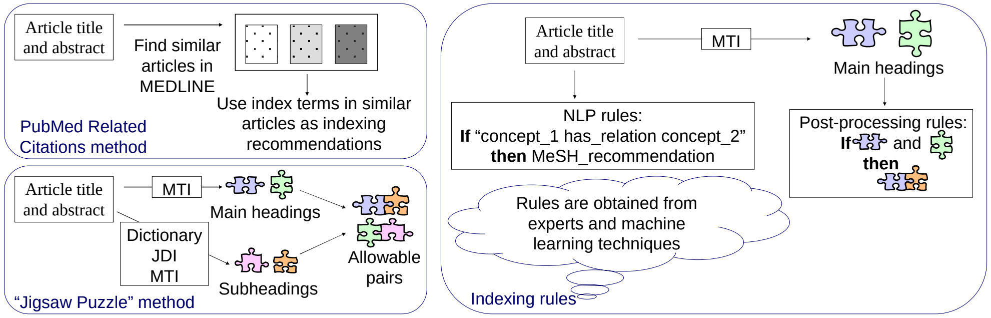
Challenges of MeSH indexing:

- **Scale**: MeSH 2008 includes 24,767 main headings (e.g. Humans, Parkinson Disease), 83 subheadings (e.g. etiology, metabolism), which amounts to **581,560 indexing terms** (e.g. Humans, Parkinson Disease/etiology)
- **Multiclass classification**: the number of indexing terms per article is not known in advance
- **Complex cognitive task**: There is no unique correct set of terms for a given article. Consistency among indexers is about 33.8% for main heading/subheading combinations (Funk et al., 1983)

Figure 1. A sample MEDLINE citation illustrating the use of MeSH indexing terms.

Creating an Automatic Indexing Tool for MEDLINE

The Medical Text Indexer, MTI, (Aronson et al., 2004) has been presenting NLM indexers with main heading recommendations since 2002. The objective of our work is to attach subheadings to the main headings in order to provide more accurate indexing recommendations. Statistical and Natural Language Processing techniques are used.



Article title and abstract

Find similar articles in MEDLINE

Use index terms in similar articles as indexing recommendations

PubMed Related Citations method

Article title and abstract

MTI

Main headings

Dictionary
JDI
MTI

Subheadings

Allowable pairs

"Jigsaw Puzzle" method

Article title and abstract

MTI

Main headings

NLP rules:
If "concept_1 has_relation concept_2"
then MeSH_recommendation

Post-processing rules:
If and
then

Rules are obtained from experts and machine learning techniques

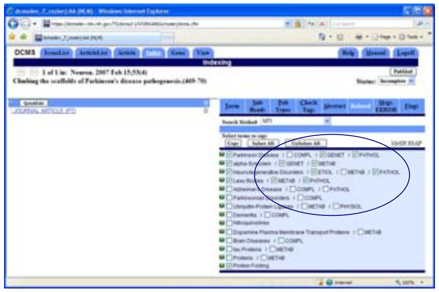
Indexing rules

Figure 2. Methods for subheading attachment in the Medical Text Indexer (MTI).

Results

The performance of the indexing recommendations produced by the subheading attachment methods varies depending on the method and the subheading considered. Overall, the best precision is achieved by a combination of the methods.

NLM indexers' feedback led to improving the recommendations, using an advanced combination process for recommendations coming from the different methods. The Index Section approved the final version of the indexing tool to be used daily at NLM.



DCMS

Indexing

1 of 1 in: Neuron. 2007 Feb 15;53(4)

Climbing the scaffolds of Parkinson's disease pathogenesis.(469-70)

Status: Incomplete

JOURNAL ARTICLE (PT)

Search Method: MTI

Select terms to copy:

☑ Parkinson Disease / □ COMPL / ☑ GENET / ☑ PATHOL
☑ alpha-Synuclein / ☑ GENET / ☑ METAB
☑ Neurodegenerative Disorders / ☑ ETIOL / □ METAB / ☑ PATHOL
☑ Lewy Bodies / ☑ METAB / ☑ PATHOL
□ Alzheimer Disease / □ COMPL / □ PATHOL
□ Parkinsonian Disorders / □ COMPL
□ Ubiquitin-Protein Ligases / □ METAB / □ PHYSIOL
□ Dementia / □ COMPL
□ [illegible]
□ Dopamine Plasma Membrane Transport Proteins / □ METAB
□ Brain Diseases / □ COMPL
□ tau Proteins / □ METAB
□ Proteins / □ METAB
☑ Protein Folding

Indexing method	Scope	P	R	F
Dictionary	83	26	35	30
MTI	82	25	14	18
JDI	83	25	27	26
Post-processing rules	19	39	8	14
NLP rules	20	17	3	5
PubMed Related Citations	83	35	**53**	**42**
Combination of Methods	83	**48**	30	36

Figure 3. Performance of the MeSH indexing recommendations shown to NLM indexers. Precision (P), Recall (R) and F-measure (F) were computed on a test corpus of 100,000 citations selected randomly from MEDLINE.

Conclusions

NLM's Medical Text Indexer now produces main heading/subheading recommendations in addition to isolated main heading recommendations. This new feature will be used to display automatic MeSH indexing recommendations in DCMS, the interface used by indexers to create MEDLINE citations.

Acknowledgments

This study was supported in part by the Intramural Research Programs of NIH/NLM. A. Névéol was supported by an appointment to the NLM Research Participation Program administered by ORISE through an inter-agency agreement between the U.S. Department of Energy and NLM.

References

Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ: The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004;11(Pt 1):268-72.

Funk ME, Reid CA, McGoogan LS: Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983 Apr;71(2):176-83.