Automated subject indexing with Annif and Finto AI
Putting DIY automated subject indexing into production

Oasma Suominen, Mona Lehtinen, Juho Inkinen
DCMI Virtual AI panel discussion
13 October 2021
Outline

1. Development of Annif
2. Quality of automated subject indexing
3. Community building
4. Annif deployments
5. Lessons learned
1. Development of Annif
Machine learning using existing metadata
Early prototype (2017) got people excited
Goals for Annif implementation (2018 → )

1. multilingual
2. independent of indexing vocabulary
3. support different subject indexing algorithms
4. CLI, Web user interface and REST API
5. community-oriented open source software
Lexical vs. associative algorithms for subject indexing

**Lexical approaches** (e.g.: Maui, MLLM, STWFSA)

match the **terms** in a document to **terms** in a controlled vocabulary

> "Renewable resources are a part of Earth’s natural environment and the largest components of its ecosphere."

**Associative approaches** (e.g.: fastText, Omikuji, SVC)

learn which **subjects** are correlated with which **words** in documents, based on training data

Lexical approaches need comparatively little training data.

Associative approaches need a lot more training data in order to cover each subject.
2. Quality of automated subject indexing
Comparison to “gold standard”

F1@5 scores for different test corpora and Annif API/model versions
Assessment by evaluators

At a workshop in 2019, **48 evaluators** evaluated subjects for **50 documents**. Subjects were given by either human indexers or four different algorithms.

The best ensemble algorithm (red bars) was not quite on the level of human indexers in quality scores (left), and significantly more of its suggestions were rejected (right).

3. Community building
Web site with form for testing at annif.org

Why AI ≠ Automated Indexing: What Is and Is Not Possible

Automated indexing is only as good as the training set, or rules that are available for the domain. It's important to learn what type of content a pre-trained algorithm has been trained on. Consider what type of content is readily available to train an algorithm—what's popular and what's available. Scholarly and historical content is not available in consumable formats at the large volume that is required for machine learning. There are exceptions such as science and medicine where large well documented collections are available. This panel will discuss the current state of automated categorization covering domains including research data, art history, and scientific publishing. The goal is to provide practical advice on how to take meaningful steps towards building the infrastructure needed for sustainable automated indexing.
Hands-on **Annif tutorial**
for those who want to use Annif on their own

Videos and exercises freely available on YouTube & GitHub!
4. Annif deployments
JYX repository, University of Jyväskylä
Students upload their Master’s and doctoral theses, Annif suggests subjects*

Keywords

<table>
<thead>
<tr>
<th>Keyword suggestions</th>
<th>[YSO]</th>
</tr>
</thead>
<tbody>
<tr>
<td>information management systems</td>
<td>[YSO]</td>
</tr>
<tr>
<td>metadata</td>
<td>[YSO]</td>
</tr>
<tr>
<td>connections (technical systems)</td>
<td>[YSO]</td>
</tr>
<tr>
<td>content management</td>
<td>[YSO]</td>
</tr>
<tr>
<td>multimedia (information technology)</td>
<td>[YSO]</td>
</tr>
<tr>
<td>digital libraries</td>
<td>[YSO]</td>
</tr>
<tr>
<td>XML</td>
<td>[YSO]</td>
</tr>
<tr>
<td>semantic web</td>
<td>[YSO]</td>
</tr>
<tr>
<td>open source code</td>
<td>[YSO]</td>
</tr>
<tr>
<td>open data</td>
<td>[YSO]</td>
</tr>
<tr>
<td>user-centeredness</td>
<td>[YSO]</td>
</tr>
<tr>
<td>archives (memory organisations)</td>
<td>[YSO]</td>
</tr>
<tr>
<td>seeking</td>
<td>[YSO]</td>
</tr>
<tr>
<td>Works</td>
<td>[YSO]</td>
</tr>
<tr>
<td>cloud services</td>
<td>[YSO]</td>
</tr>
<tr>
<td>electronic publications</td>
<td>[YSO]</td>
</tr>
</tbody>
</table>

*from YSO = General Finnish Ontology

Implemented using DSpace & GLAMpipe by Ari Häyrinen
Finto AI - automated subject indexing tool and API service

Finto AI suggests subjects for a given text. It’s based on Annif, a tool for automated subject indexing. Read more...

Launches in May 2020

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term "artificial intelligence" is often used to describe machines (or computer software) that can associate with the human mind, such that the machine can "think like a human.

As machines become increasingly intelligent, "the line between the "human" and "machine" are often removed from the definition. According to the Church-Turing Thesis, if a computer is intelligent, then a human is also intelligent. This paper shows how AI is whatever we choose it to be. There is no AI without human intervention. It is our responsibility to ensure that AI is used responsibly.

Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic games such as chess and Go, autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

ai.finto.fi

API service
Finto AI is also an API service that can be integrated to other systems. Lisätietoja | OpenAPI-kuvaus

Subject indexing

Vocabulary and text language

- YSO English

Maximum # of suggestions

- 10
- 15
- 20

Get subject suggestions

Suggestions

- artificial intelligence
- machine learning
- intelligence (mental properties)
- information technology
- computational science
- computer science
- computers
- computer-assisted teaching
- learning
- automation
Subject indexing for electronic deposits

In November 2020, the National Library of Finland started using Finto AI to suggest subjects when processing electronic deposits submitted through the individual submission form.

Implementation: Erik Lindgren, Mikko Merioksa, Satu Niininen
5. Lessons learned
Algorithms may be used *alone*, or in combinations, **ensembles**

Ensembles are nearly always better than individual algorithms
Start by experimentation, move slowly towards production

image credit: @kettudolls (IG)
With an API service such as Finto AI, implementing semi-automated indexing becomes easy; explaining it to users can be more challenging.

What is this? What should I do here?

Maybe it’s better to leave these alone...
Thank you!

Juho Inkinen

Mona Lehtinen

Osma Suominen

annif.org