Infomed

Indización Automática
IAt

Carlos Manuel
cmanuel@infomed.sld.cu

Alfonso Ali
ali@infomed.sld.cu

grupo de trabajo

Tecnologías de Información para el Acceso Equitativo a la Información

Presentación

Objetivo Geral

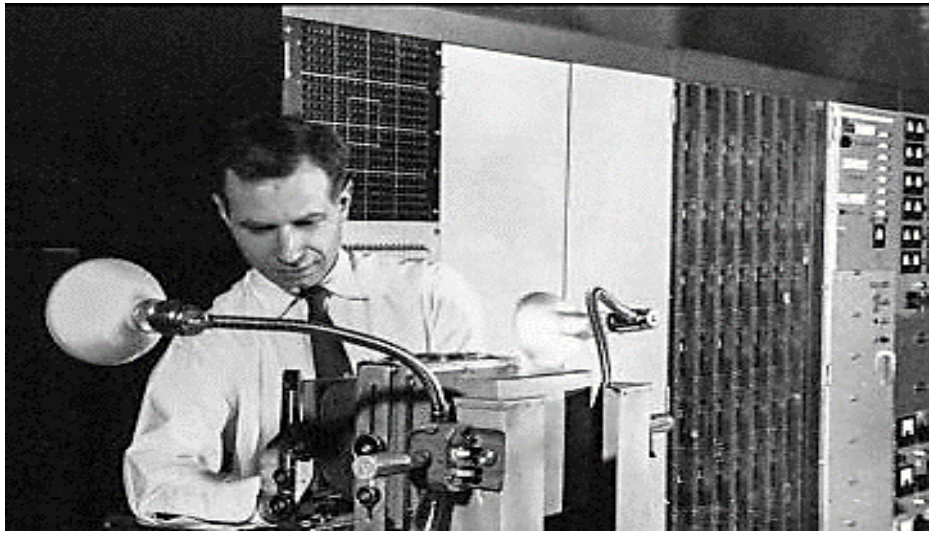
Subsidiar la definición de directrices y acciones estratégicas para la BVS, a traves de recomendaciones técnicas, con la visión de las tecnologías de información como medio para el acceso equitativo a la información.

Objetivos Específicos

1. Evaluar las variables tecnológicas internas y externas a la BVS;

2. Evaluar las tendencias mundiales en tecnologías de información enfocadas en gestión de información y conocimiento;

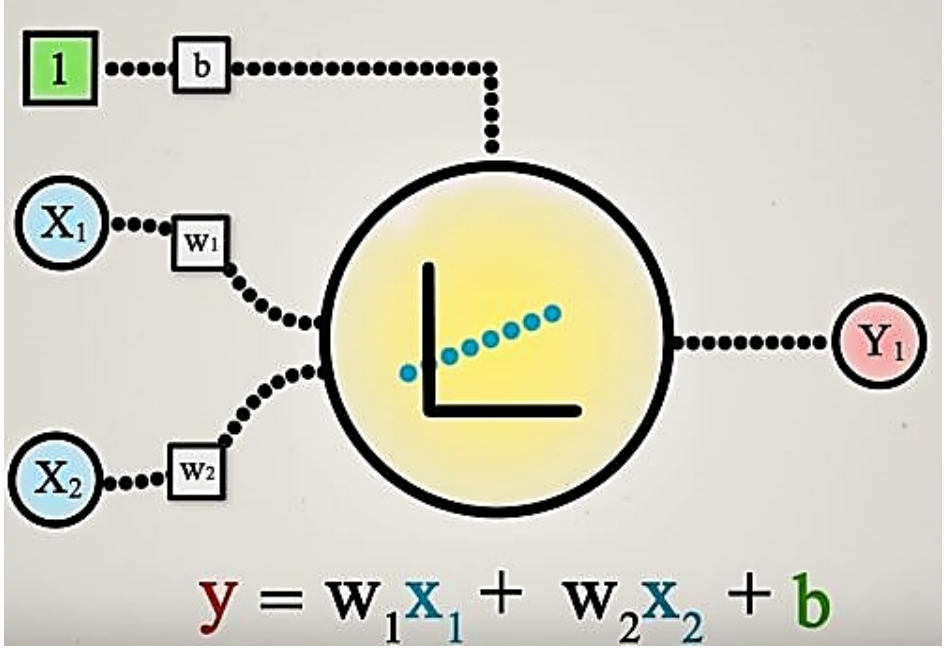
3. Discutir cuestiones relativas al acceso equitativo a las tecnologías digitales y a la información científica y técnica y como ellas influencian la estrategia de desarrollo y la utilización de la BVS.



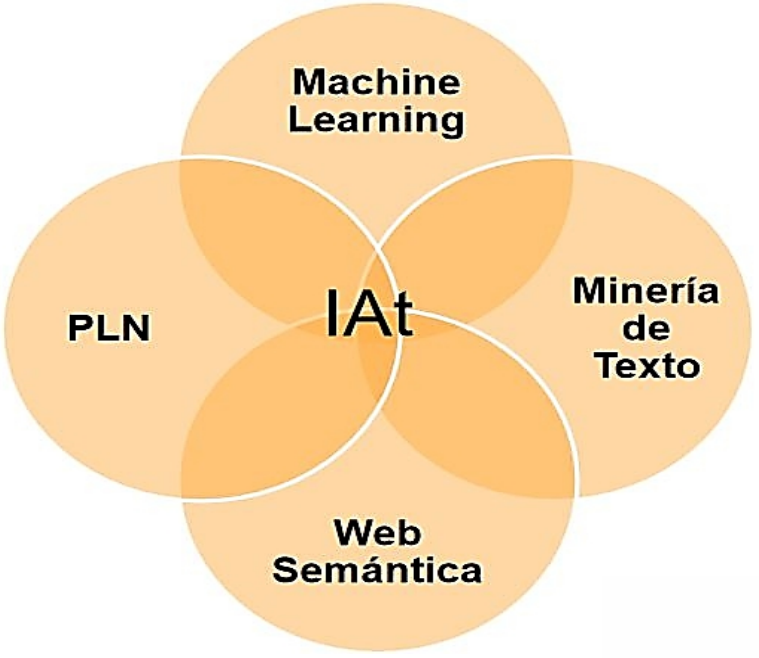
The New York Times 1958 “este es (Perceptrón) el embrión de un ordenador electrónico que espera poder caminar, hablar, ver, escribir, reproducirse y tener consciencia de su existencia”. Frank Rossenblatt

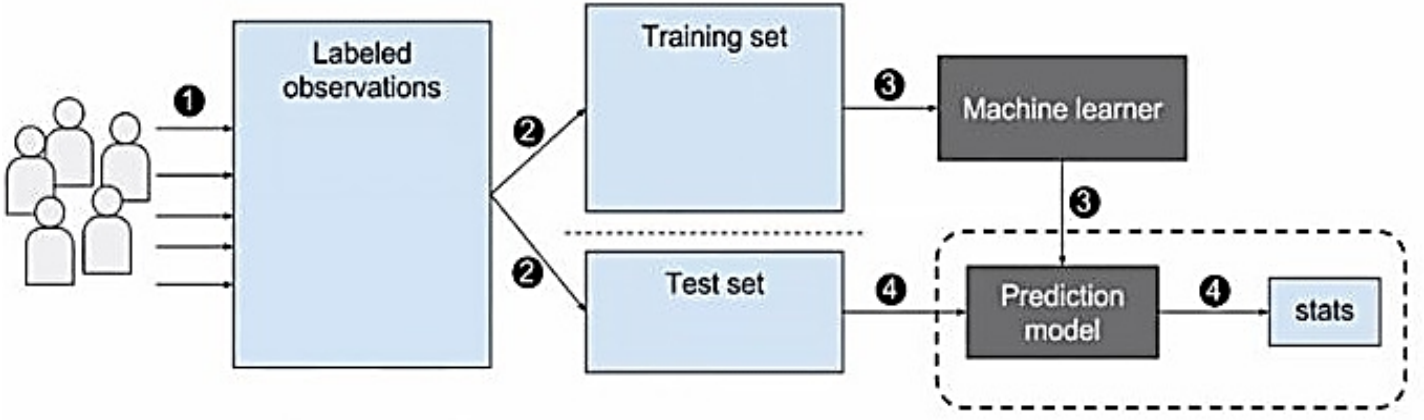
Inteligencia Artificial

1. Subdisciplina del campo de la Informática, que busca la creación de máquinas que puedan imitar comportamientos inteligentes.

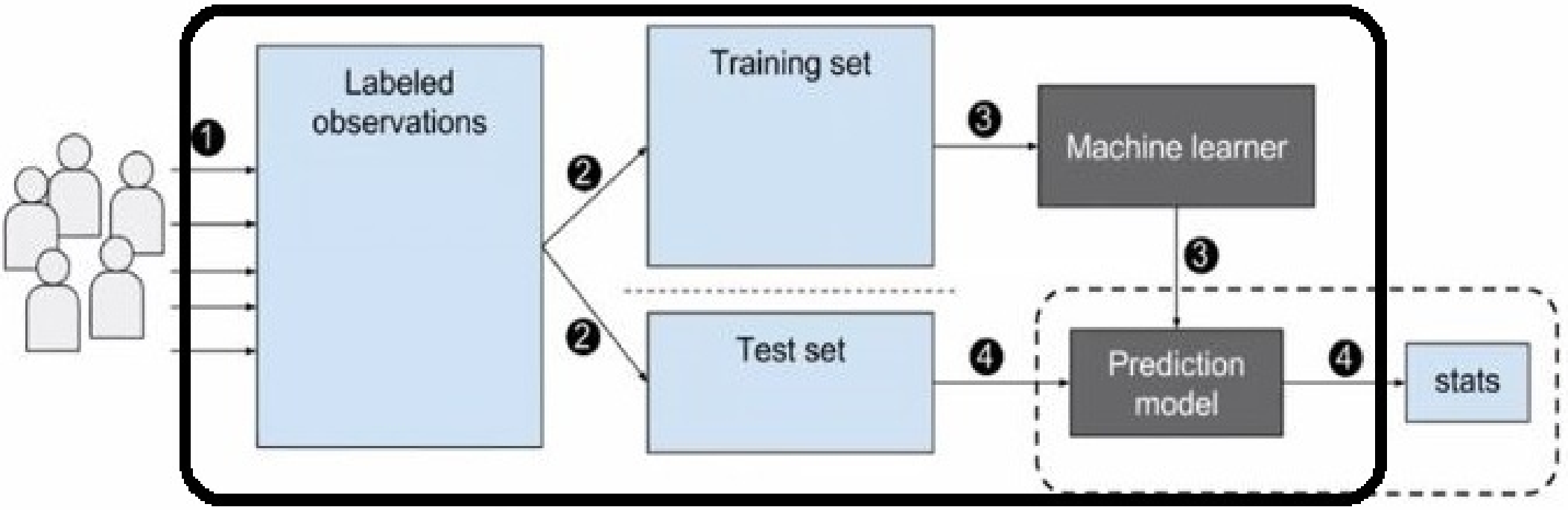


Relación de la Indización Automática con otras disciplinas





Algoritmos de Aprendizaje Profundo	Uso en PLN
Red neuronal NN	- Etiquetado de partes del discurso - Tokenización - Reconocimiento de entidades nombradas - Comprender el tipo de acción transmitida en las oraciones y todas sus partes.
Redes neuronales recurrentes RNN	- Sistemas de traducción automática - Sistemas que responden preguntas hechas en lenguaje natural - Descripción textual de imágenes - Análisis de sentimientos - Determinar si dos segmentos de texto tienen el mismo significado. - Determinar relaciones de pares de entidades etiquetadas - Aprender patrones a partir de ejemplos o datos de muestra
Red neuronal convolucional CNN	- Clasificación de oraciones y textos - Extracción y clasificación de relaciones. - Detección de spam - Categorización de consultas de búsqueda. - Extracción de relaciones semánticas


Labeled observations
Training set
Machine learner
Test set
Prediction model
stats
1
2
2
3
3
4
4

Recall (REC): ¿Qué porcentaje de casos positivos se identificó correctamente?

$$\frac{TP}{TP + FN}$$

Precisión (PRE): ¿Qué porcentaje de las predicciones positivas fueron correctas?

$$\frac{TP}{FP + TP}$$

En un centro de información se presenta una solicitud sobre un tema determinado, sobre el cual se han indizado 20 documentos relevantes. Al aplicar el índice para recuperar los documentos en respuesta a la solicitud se encuentran 14 de los documentos relevantes. Cuál es el recobrado?

a + c = 20 documentos relevantes en el sistema

a = 14 documentos relevantes recuperados

C_R = ?

$$C_R = \frac{a}{a + c} \cdot 100$$

$$C_R = \frac{14}{20} \cdot 100 = 70\%$$

A través de un índice se recuperan y entregan 30 documentos en respuesta a una solicitud. Después que el usuario revisa los documentos informa al centro que solamente 6 son relevantes a sus intereses.

Si los evaluadores por algún medio conocen que en la colección hay 10 documentos relevantes a la solicitud. Con qué coeficiente de precisión ha operado el índice?

a + b = 30 documentos recuperados (relevantes más no relevantes)

a = 6 documentos relevantes recuperados

C_P = ?

$$C_P = \frac{a}{a + b} 100$$

$$C_P = \frac{6}{30} 100 = 20\%$$

Precisión (PRE): $\frac{TP}{FP + TP}$

Recall (REC): $\frac{TP}{TP + FN}$

F1 – Score: $2 \times \frac{PRE \times REC}{PRE + REC}$

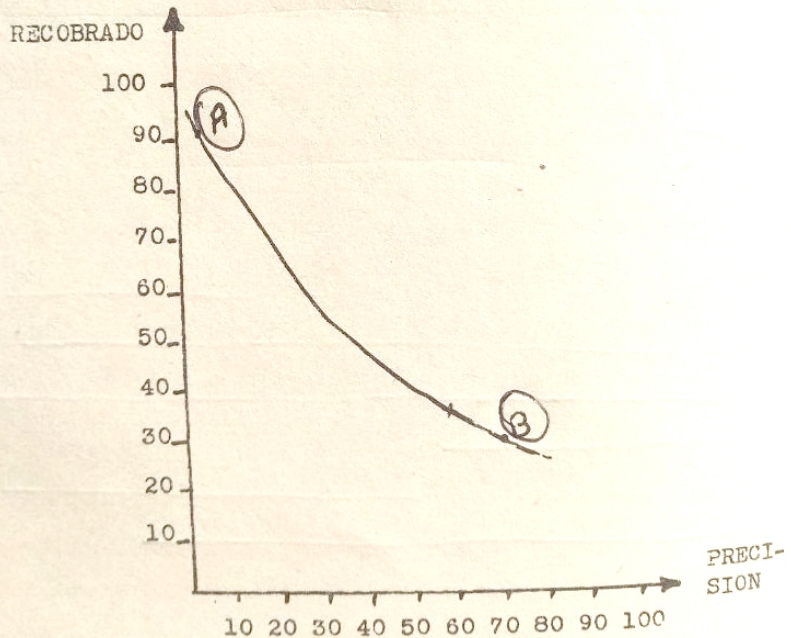
¿Qué tan bueno es mi modelo?

9.4 CURVAS DE RECOBRADO Y PRECISION

Como indicadores de la capacidad de recuperación de un índice se pueden utilizar las curvas de recobrado/precisión. En la Fig. 9.2 se representa una curva que se ha trazado con los resultados de los coeficientes de precisión y recobrado correspondientes a diferentes solicitudes de información, obtenidos aplicando un determinado índice. Por tanto, la curva expresa el comportamiento del índice ante distintas solicitudes.

Cuando la solicitud es muy general se tiene que utilizar para hacer la recuperación un término genérico del índice. Por ejemplo el punto A corresponde a una búsqueda genérica que tuvo un resultado con un 90% de recobrado y una muy baja precisión con un valor de 5%. A

En el otro extremo de la curva se tiene el punto B que representa los resultados de una recuperación en respuesta a una solicitud formulada de modo mucho más preciso, lo cual permite que se utilice un término específico del índice para realizar la recuperación de información. Los valores que se lograron en este caso fueron de 30% de recobrado y de 70% de precisión. B



1 The computation of MiF, MaF and EBF

Denote K as the size of all labels (MHs), and N as the number of instances (citations). Let y_i and $\hat{y}_i \in \{0,1\}^K$ be the true and predicted labels for instance i, respectively. For evaluating the performance of different models, we use the three groups of main metrics: Micro (recall, precision and F-Measure), Macro (recall, precision and F-Measure) and Example Based (recall, precision and F-Measure).

Micro F-measure: MiF

Micro F-measure (MiF) is the harmonic mean of micro-average precision (MiP) and micro-average recall (MiR):

$$\mathrm{MiF} = \frac{2 \cdot \mathrm{MiP} \cdot \mathrm{MiR}}{\mathrm{MiP} + \mathrm{MiR}}, \tag{1}$$

where

$$\mathrm{MiP} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} \hat{y}_i^k} \qquad \mathrm{MiR} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k}$$

Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials.

Luis Gasco[1], Anastasios Nentidis[2,3], Anastasia Krithara[2], Darryl Estrada-Zavala[1], Renato Toshiyuki Murasaki[4], Elena Primo-Peña[5], Cristina Bojo Canales[5], Georgios Paliouras[2] and Martin Krallinger[1]

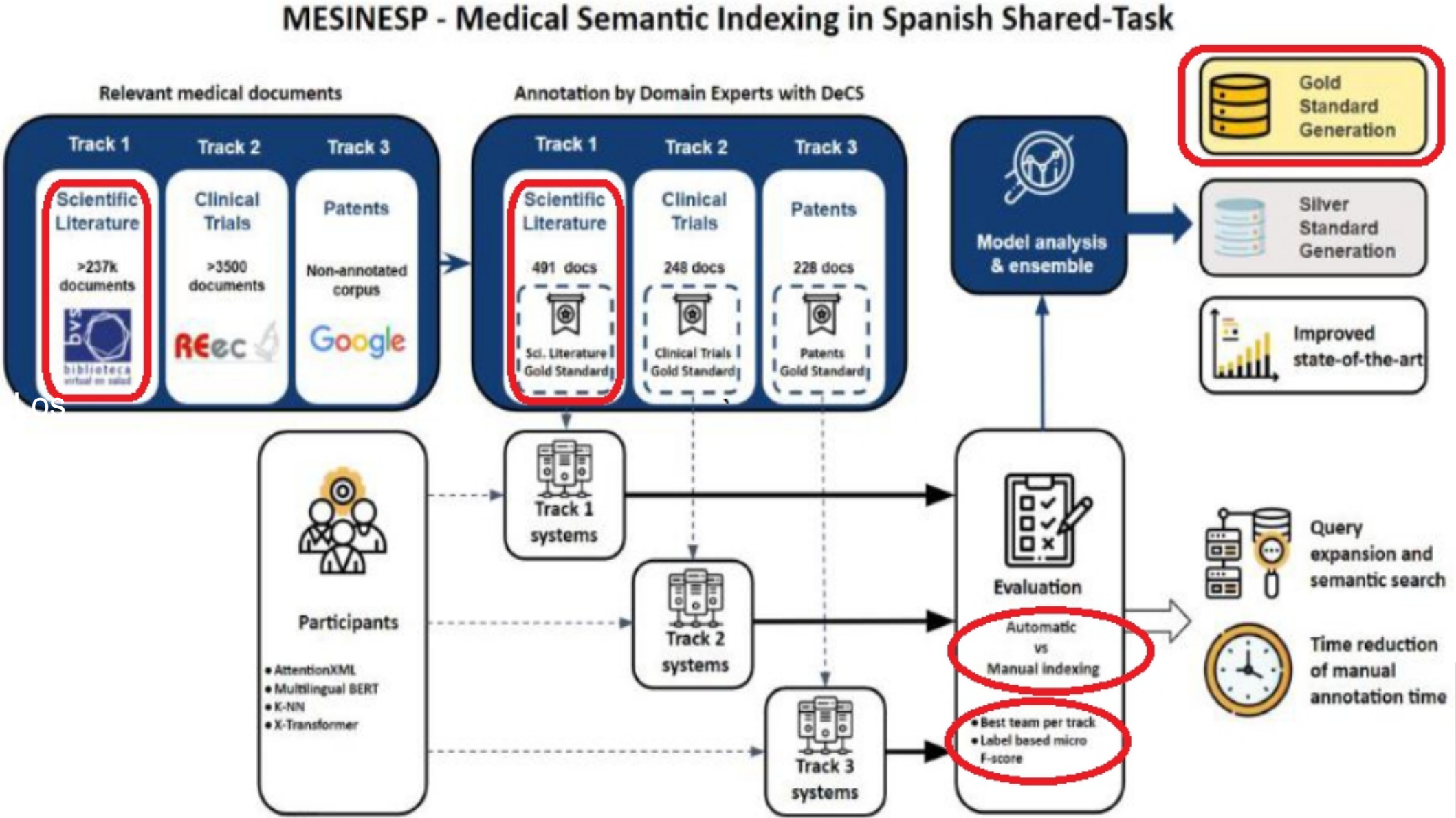
[1]*Barcelona Supercomputing Center. Barcelona, Spain*

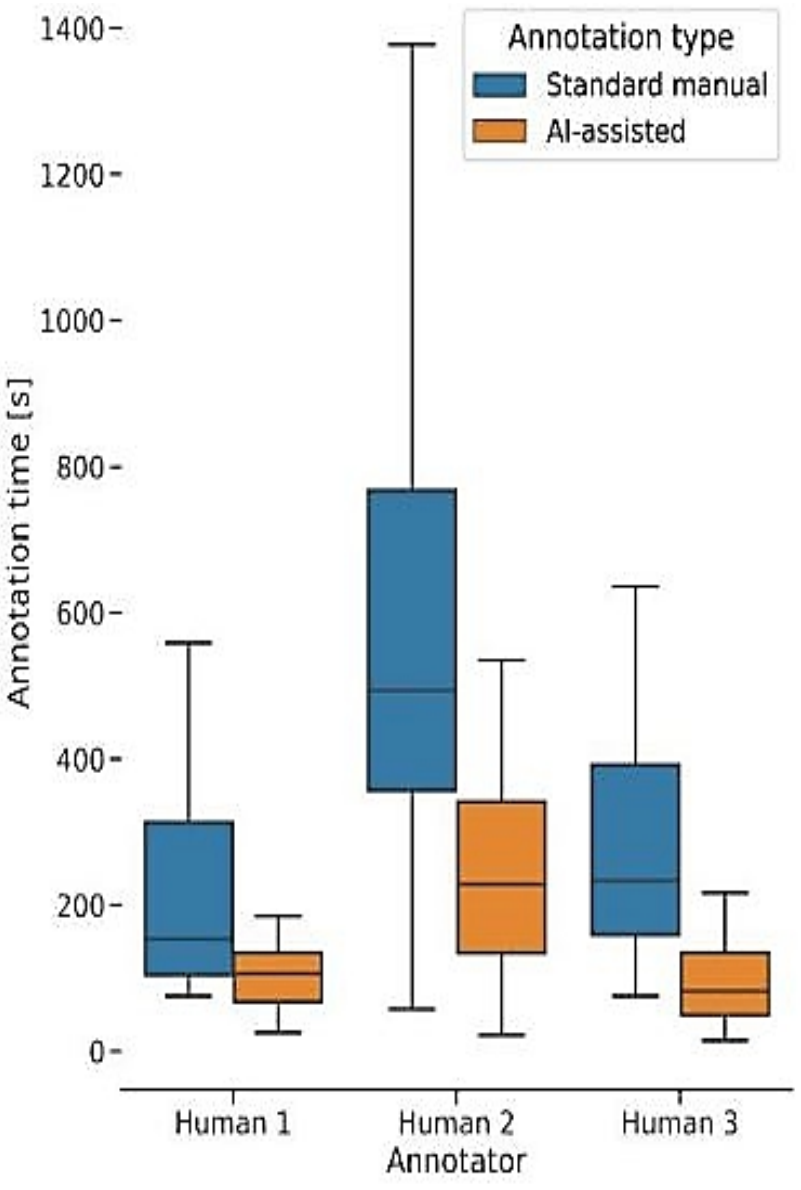
[2]*National Center for Scientific Research "Demokritos", Athens, Greece.*

[3]*Aristotle University of Thessaloniki. Thessaloniki, Greece*

[4]*Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud, Organización Panamericana de la Salud. São Paulo, Brasil.*

[5]*Instituto de Salud Carlos III. Biblioteca Nacional de Ciencias de la Salud. Madrid, Spain*


MESINESP - Medical Semantic Indexing in Spanish Shared-Task
Relevant medical documents
Annotation by Domain Experts with DeCS
Track 1
Track 2
Track 3
Scientific Literature
>237k documents
Clinical Trials
>3500 documents
Patents
Non-annotated corpus
491 docs
248 docs
228 docs
Sci. Literature Gold Standard
Clinical Trials Gold Standard
Patents Gold Standard
Model analysis & ensemble
Gold Standard Generation
Silver Standard Generation
Improved state-of-the-art
Participants
AttentionXML
Multilingual BERT
K-NN
X-Transformer
Track 1 systems
Track 2 systems
Track 3 systems
Evaluation
Automatic vs Manual indexing
Best team per track
Label based micro F-score
Query expansion and semantic search
Time reduction of manual annotation time



Como parte de la evaluación de los modelos utilizaron una herramienta llamada ASIT, que permite **registrar el tiempo que demora la indización,** midieron el tiempo promedio tomado por tres indizadores expertos para anotar 30 documentos de manera tradicional y comparar los tiempos con la indización realizada con la ayuda de descriptores predichos por el sistema BERTDeCS versión 4.

La preindexación de los documentos ha reducido los tiempos de anotación a más de la mitad.

ES UNA OPORTUNIDAD PARA INCORPORAR FUNCIONALIDADES A FI-ADMIN

Performance of each of the models generated by participating teams

				MESINESP-L		
Team	Country	Ref	System	MiF	MiP	MiR
Fudan University	China	-	BERTDeCS-CooMatInfer	0.4505	0.4791	0.4252
			BERTDeCS version 2	0.4798	0.5037	0.4581
			BERTDeCS version 3	0.4808	0.5047	0.4591
			BERTDeCS version 4	*0.4837*	0.5077	0.4618
			bertmesh-1	0.4808	0.5077	0.4591
Roche	Switzerland	[18]	bert_dna	0.3989	0.4662	0.3486
			pi_dna	*0.4225*	0.4667	0.3859
			pi_dna_2	0.3978	0.4520	0.3551
			pi_dna_3	0.4027	0.4348	0.3750
			bert_dna_2	0.3962	0.4820	0.3364
Lasige-TEAM (Universidade de Lisboa)	Portugal	[23]	LASIGE_BioTM_1	*0.2007*	0.1584	0.2738
			LASIGE_BioTM_2	0.1886	0.1489	0.2573
			clinical_trials_1.0	-	-	-
			clinical_trials_0.25	-	-	-
			patents_1.0	-	-	-
Vicomtech	Spain	[17]	Classifier	*0.3825*	0.4622	0.3262
			CSSClassifier025	0.3823	0.4509	0.3318
			CSSClassifier035	0.3801	0.4710	0.3186
			LabelGlosses01	0.3704	0.4526	0.3134
			LabelGlosses02	0.3746	0.4560	0.3179
			iria-1	0.3406	0.3641	0.3199
			iria-2	0.3389	0.3622	0.3185

En las Conclusiones

La principal métrica de evaluación utilizada para la tarea fue **MiF**

Los mejores resultados los obtuvo en las tres subpistas el equipo de la Universidad de Fudan de Shanghái. Todos sus modelos ocuparon el primer lugar en cada una de las subpistas.

Las posibles razones de la caída en el rendimiento con respecto a la tarea de inglés pueden estar asociadas a un **menor número de documentos de entrenamiento(237000)** y al hecho de que los documentos obtenidos de la BVS provienen de **dos bases de datos bibliográficas indexadas de manera ligeramente diferente.**

A pesar de la evolución positiva en el desempeño de la tarea MESINESP, se observa una caída en el desempeño del modelo con respecto a la tarea con documentos en ingles.

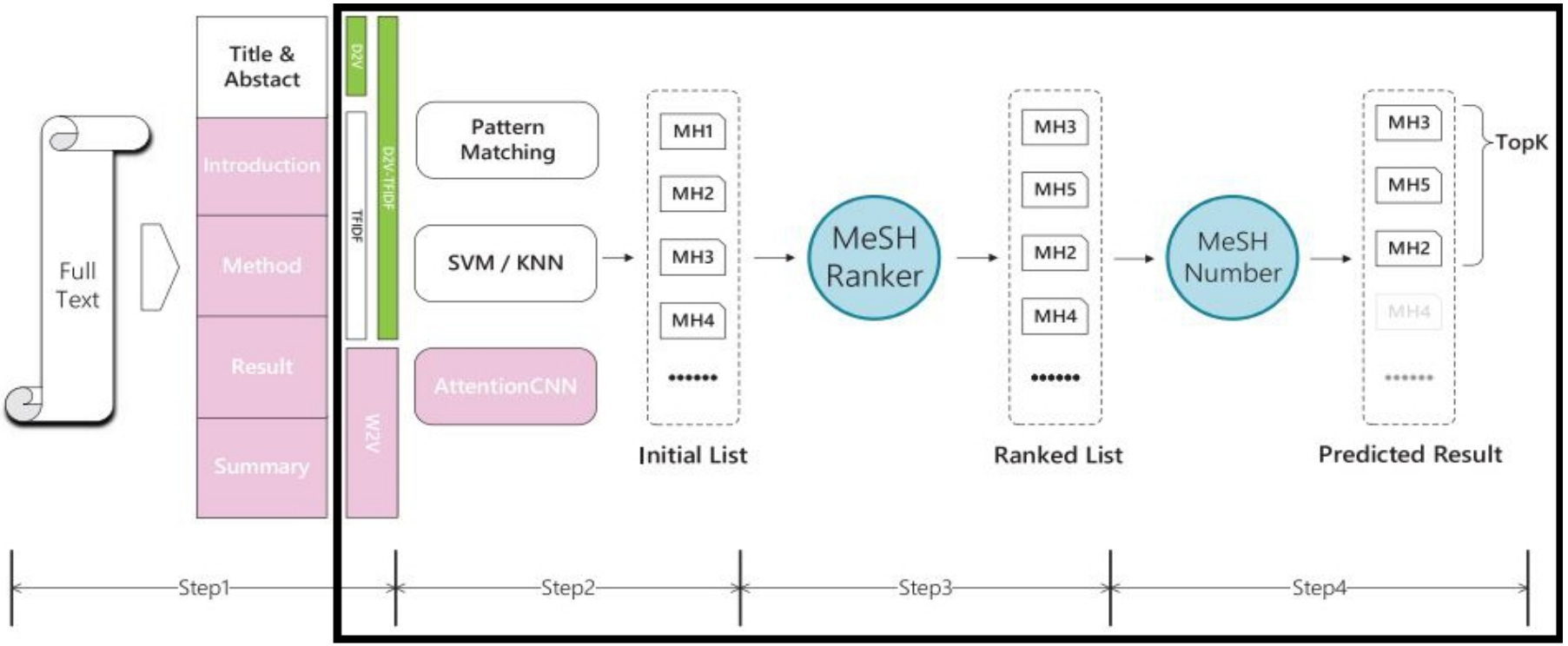
Data and text mining

FullMeSH: improving large-scale MeSH indexing with full text

Suyang Dai[1], Ronghui You[1], Zhiyong Lu [2], Xiaodi Huang[3], Hiroshi Mamitsuka[4,5] and Shanfeng Zhu[1,6,7,*]

[1]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China, [2]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA, [3]School of Computing and Mathematics, Charles Sturt University, Albury, NSW 2640, Australia, [4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto Prefecture, Japan, [5]Department of Computer Science, Aalto University, Espoo, Finland, [6]Shanghai Institute of Artificial Intelligence Algorithms and Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China and [7]Ministry of Education, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), China

FullMeSH se ha desarrollado y entrenado con un conjunto de 1,4 millones de artículos de acceso abierto a texto completo de PubMed Central.
Alcanzó una medida **Micro F del 66,76%** en un conjunto de prueba de 10 000 artículos.



Micro F-measure: MiF

Micro F-measure (MiF) is the harmonic mean of micro-average precision (MiP) and micro-average recall (MiR):

$$\mathrm{MiF} = \frac{2 \cdot \mathrm{MiP} \cdot \mathrm{MiR}}{\mathrm{MiP} + \mathrm{MiR}}, \tag{1}$$

where

$$\mathrm{MiP} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} \hat{y}_i^k} \qquad \mathrm{MiR} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k \cdot \hat{y}_i^k}{\sum_{k=1}^{K} \sum_{i=1}^{N} y_i^k}$$

Table 3. Overall results of FullMeSH

Method	MiP	MiR	MiF	EBP	EBR	EBF	MAP	MAR	MAF
MeSHLabeler	0.6636 (1.44e–70)	0.5946 (1.82e–72)	0.6273 (7.69e–76)	0.6624 (1.39e–66)	0.6096 (1.72e–71)	0.6214 (2.64e–73)	0.4935 (2.15e–75)	0.4420 (2.12e–83)	0.4663 (2.65e–82)
DeepMeSH	0.6808 (1.58e–66)	0.6142 (4.67e–75)	0.6458 (2.25e–74)	0.6793 (4.80e–65)	0.6302 (3.59e–73)	0.6407 (8.94e–72)	0.5484 (4.44e–59)	0.5171 (8.42e–66)	0.5323 (1.01e–64)
FullMeSH	0.7021	0.6363	0.6676	0.7017	0.6521	0.6629	0.5739	0.5456	0.5594

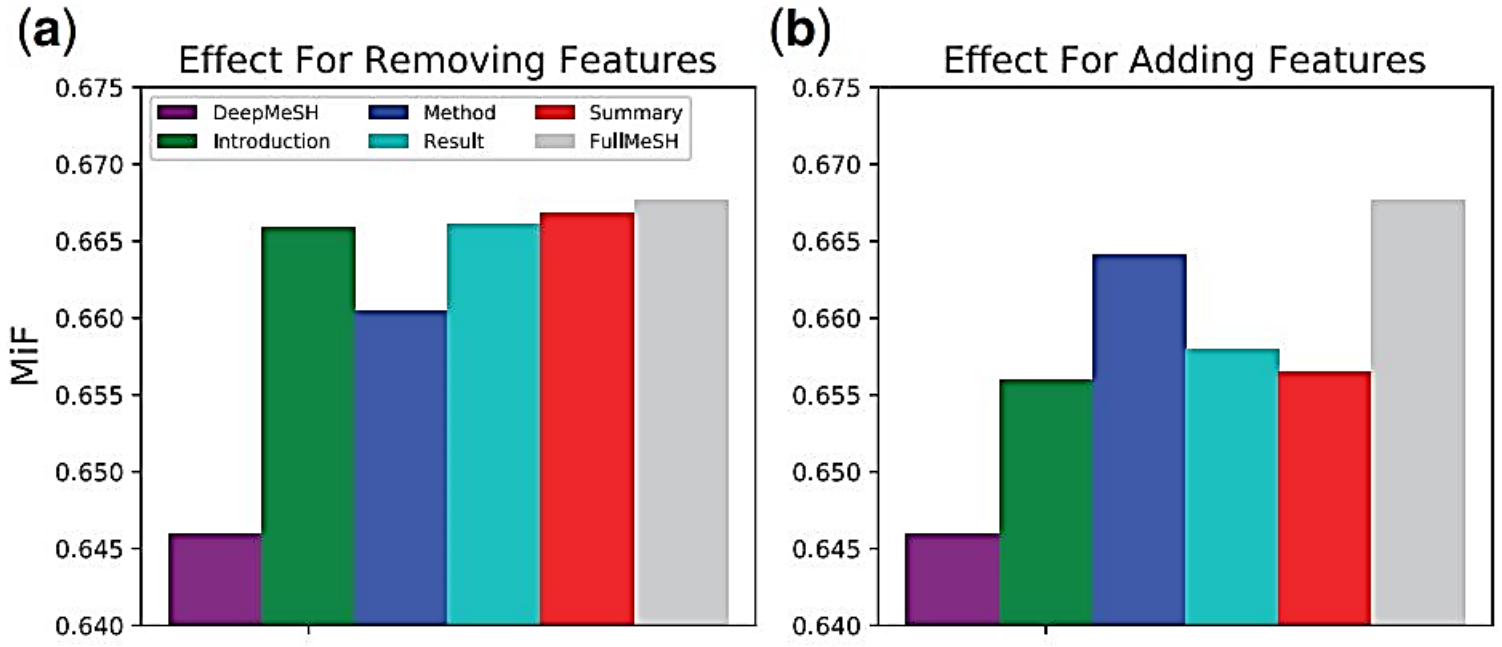


Fig. 3. Performance comparison among different sections (except Title and Abstract Sections)

CONCLUSIONES FULLMeSH

Los resultados experimentales demostraron que el uso de la información a texto completo es útil en cada paso:

(i) aumentando la cobertura de una lista de candidatos, lo que empuja el límite superior de rendimiento; (ii) mejorando en gran medida el rendimiento de clasificación utilizando información de cada sección y también una tecnología de aprendizaje profundo de vanguardia (iii) mejorando la predicción del número de Descriptores como salida.

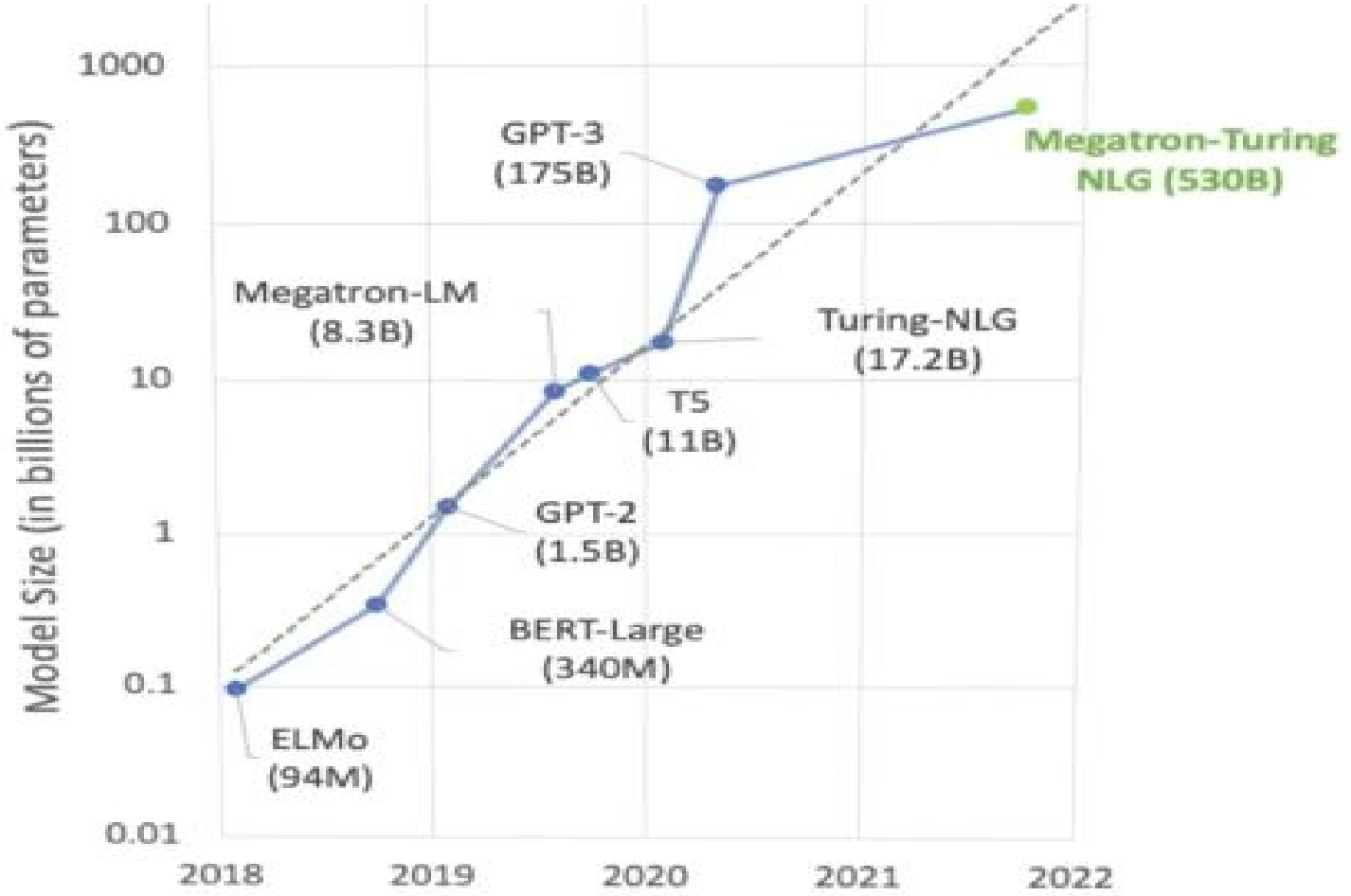
La superioridad de FullMeSH también se debe al marco de aprendizaje para clasificar , que puede integrar de forma eficiente y eficaz múltiples tipos de fuentes de datos (como varias secciones de los artículos), modelos (como KNN, SVM, concordancia de patrones y AttentionCNN) y representaciones (como TF-IDF, D2V y W2V), para un mejor rendimiento.

EFICIENCIA COMPUTACIONAL

Para entrenar todos los clasificadores y el modelo LTR (Learning to Rank), utilizaron un servidor de clúster con seis nodos. Cada nodo estaba equipado con dos CPU Intel XEON E5-4650 de 2,7 GHz y 128 GB de RAM. entrenar todos los clasificadores binarios para los descriptores y el modelo FullMeSH les tomó alrededor de 5 días.

Además, entrenaron el algoritmo AttentionCNN usando **3** GPU Nvidia Gerforce GTX 1080Ti, dedicando aproximadamente 10 h a cada sección. Por el contrario, tanto MeSHLabeler como DeepMeSH se entrenaron con un servidor con cuatro CPU Intel XEON E5-4650 de 2,7 GHz y 128 GB de RAM, lo que llevó alrededor de 5 y 7 días, respectivamente.

Observaron que FullMeSH necesita muchos más recursos computacionales para manejar el texto completo. Y que el coste computacional se puede reducir significativamente si se centran en secciones importantes de los artículos **como el Resumen.**



BERT
(Bidirectional Encoder Representations from Transformers)

Modelo de Lenguaje(LM) que se puede entrenar previamente con grandes cantidades de textos no estructurados y luego ajustarlos en una tarea posterior, método dominante en la comunidad de PNL.

El surgimiento de la serie Generative Pre-trained Transformer (GPT) y de grandes modelos de lenguaje (LLM), en particular GPT- 3, GPT- 4 y LLaMa2 pueden generar seguramente en el futuro mejores predicciones en la indización automática.

Annif and Finto AI: Developing and Implementing Automated Subject Indexing

Osma Suominen, Juho Inkinen, Mona Lehtinen

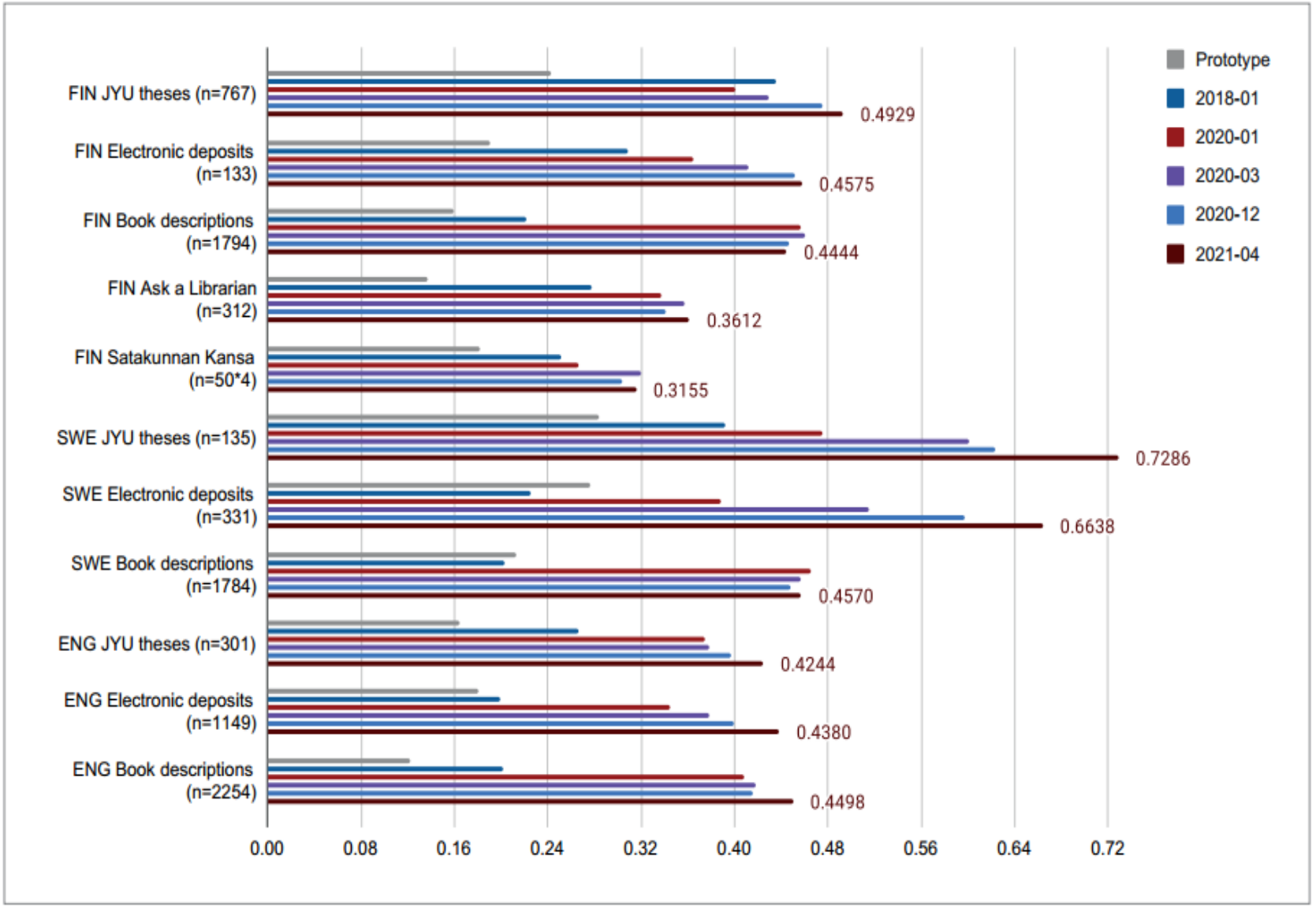
National Library of Finland, Finland

Preprint of article accepted on 23 June 2021 for publication in JLIS.it

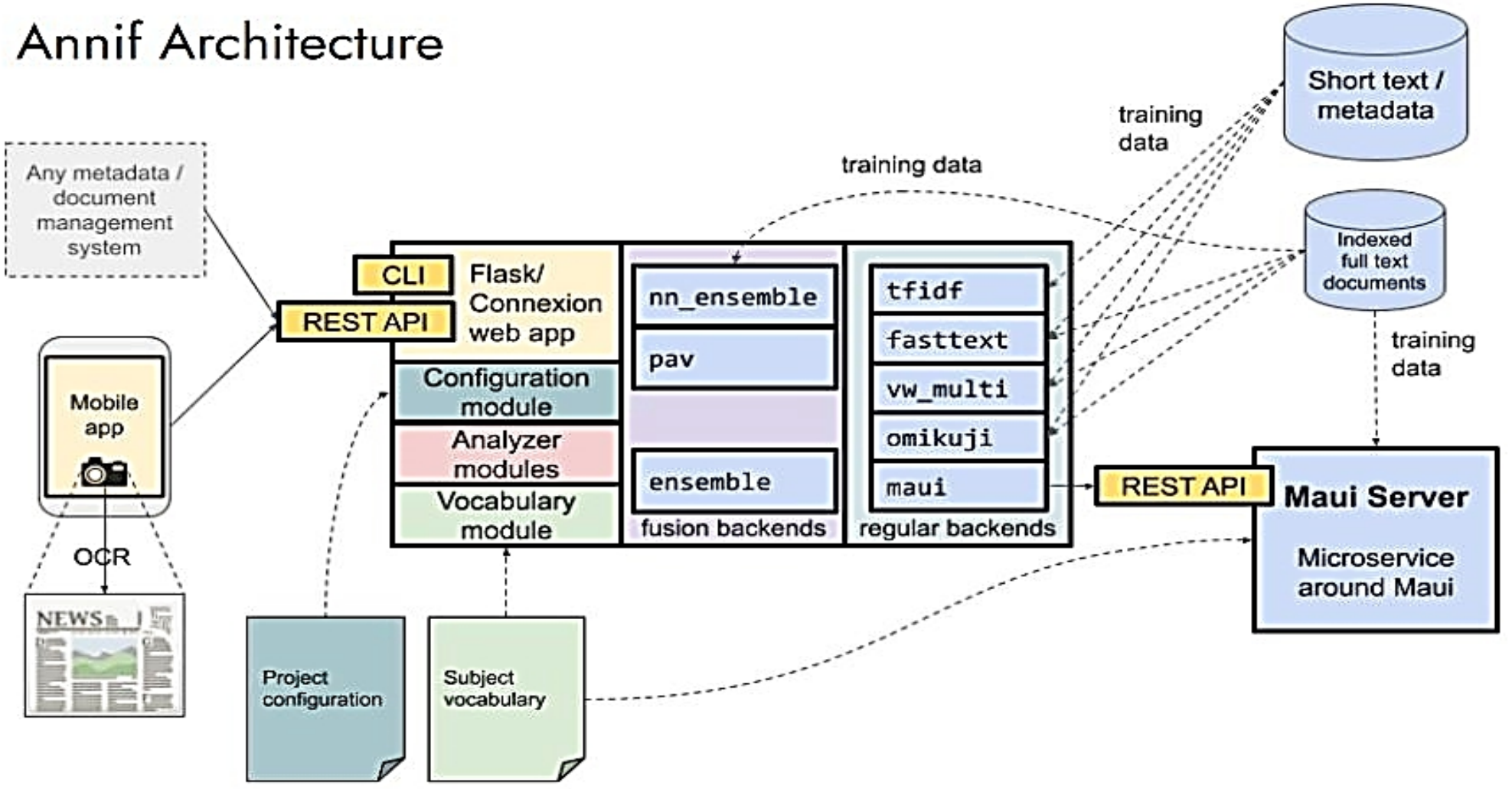
Algorithms used in Annif

lexical	Maui (using the Maui Server REST API) Maui is a lexical tool for automated indexing
associative	TF-IDF similarity (implemented with the Gensim Python library) baseline bag-of-words similarity measure and vector space model
	fastText (by Facebook Research) uses word embeddings and simulates a deep neural network architecture
	Parabel and Bonsai (implemented with the Omikuji Python library) tree-based algorithms for extreme multi-label classification (i.e., when the set of subjects is huge)

Métricas F1 que oscilan entre 0,3 y 0,5 como promedio.

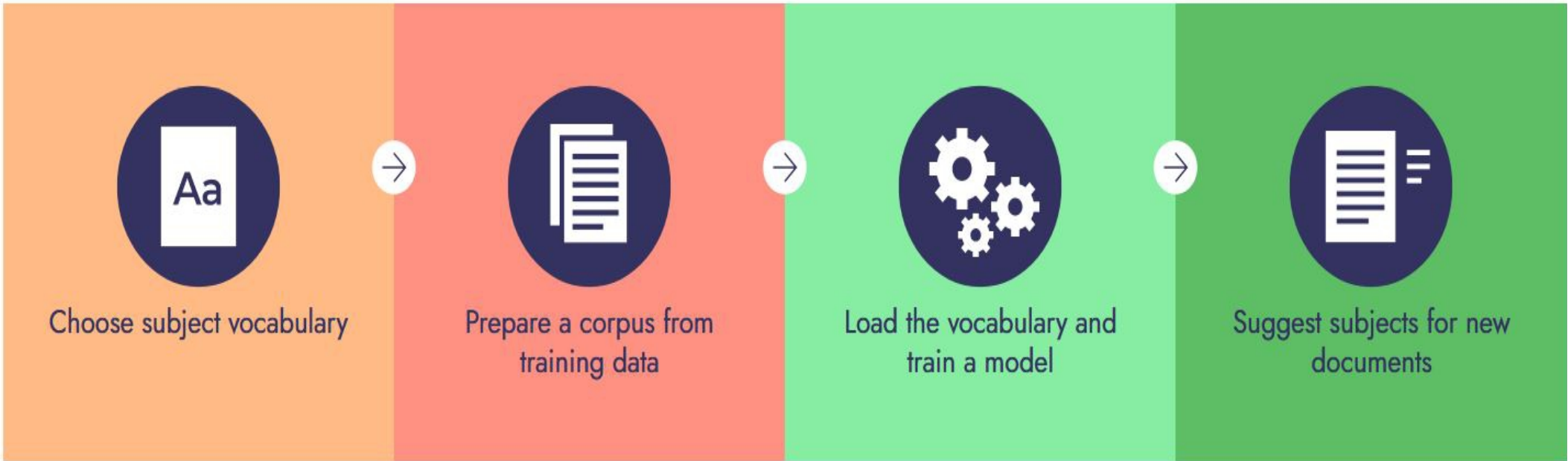


Annif Architecture



Choose a controlled subject vocabulary and train Annif on already indexed documents – it can then suggest subjects for new documents!

HOW TO USE ANNIF



finto ai

About Feedback suomeksi på svenska

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

API service
Finto AI is also an API service that can be integrated to other systems.
More information | OpenAPI description

Enter text to be indexed

Enter text | Upload file | Enter URL

Copy text here and press the button "Get subject suggestions"

You can also drop a file or a URL here

Subject indexing

Vocabulary and text language

YSO English (2023.6.Hypatia)

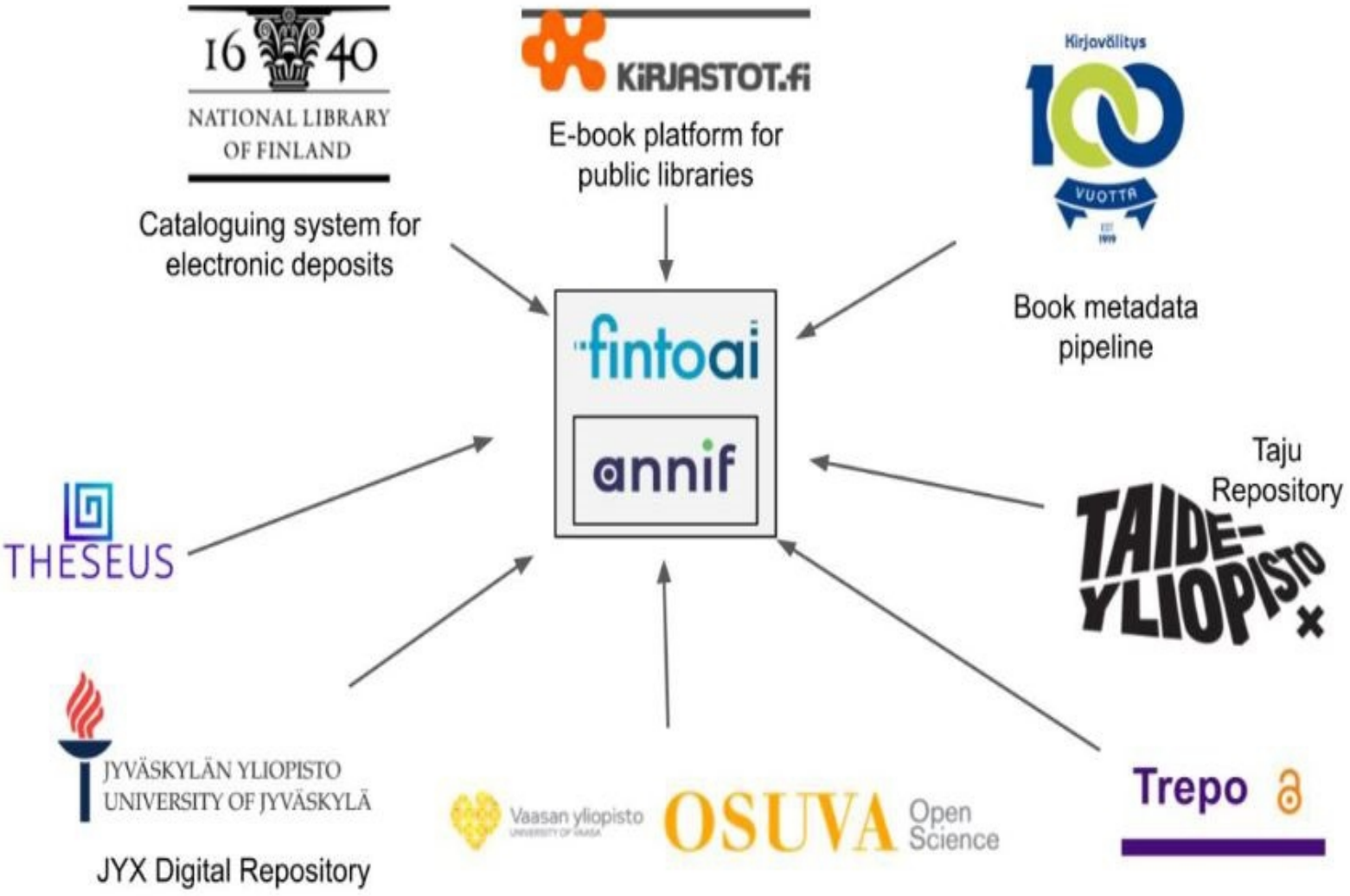
Maximum # of suggestions

10 15 20

Suggestions language

same as text language

Get subject suggestions



Usuarios institucionales de Finto AI.

Lineas de trabajo futuro

1- Terminar el proceso de poner en producción Annif para evaluar su rendimiento con texto en idioma español, para que el equipo adquiera capacidades de manejar aplicaciones de IA y para proporcionar lo mas pronto posible un mayor nivel de asistencia al proceso de indización con una herramienta para la indización semiautomática.

2- Instalar y asimilar aplicaciones de cluster de procesamiento paralelo para aprovechar la capacidad de procesamiento de que disponemos en horarios de poco uso y poder instalar modelos de IA de mayor rendimiento

3- Buscar alianzas y generar proyectos de indización automática con otras instituciones con capacidades de IA.

4- Identificar las aplicaciones donde con herramientas de IA se pueden brindar nuevas funcionalidades.

5- Identificar y preparar las colecciones de documentos a texto completo.