WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Artificial Intelligence applied to IPC and Nice classifications

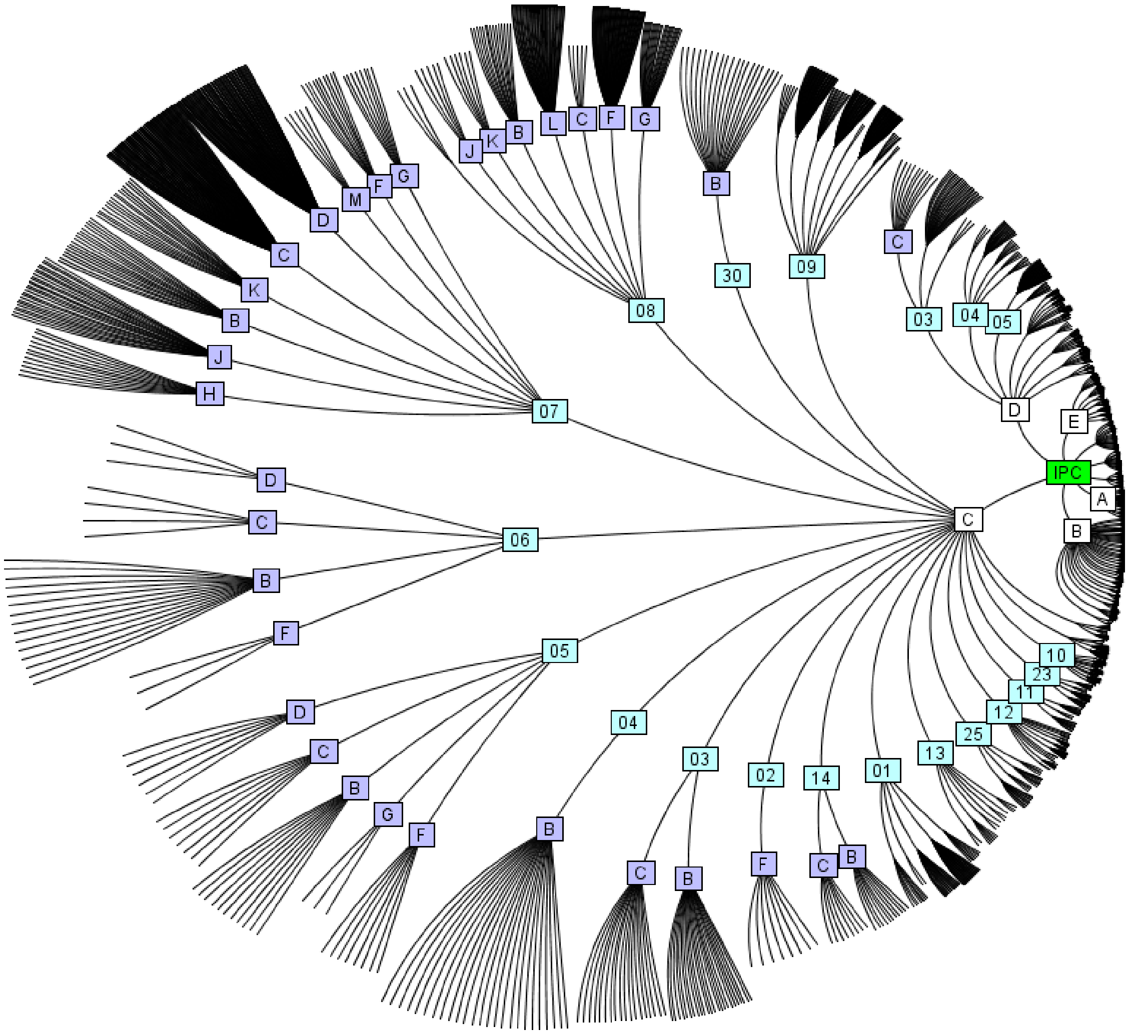
Geneva
May 25, 2018

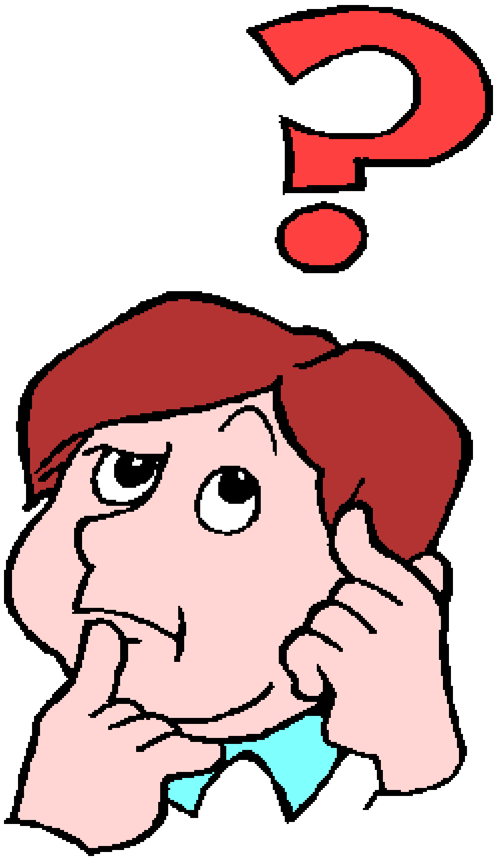
Patrick FIÉVET

IPCCAT-neural : automatic text categorization in the IPC

- **What is it about?**
  - Patent Classifications : **IPC** (and CPC)
  - Automatic text **CAT**egorization in the specific context of patent documents
  - Artificial Intelligence (AI) to mimic legacy patent classification practices

IPCCAT-neural : automatic text categorization in the IPC





WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

IPCCAT-neural : automatic text categorization in the IPC

- **Initial problems to be solved in 2002 (CLAIMS project):**
  - IPCs allotment in small Patent Offices
  - Languages: ES, **FR, EN**, DE, RU, ZH.
  - Automatic routing of patent/technical documents according to their technical domains based on a text input e.g. a patent abstract

IPCCAT-neural: construction phase

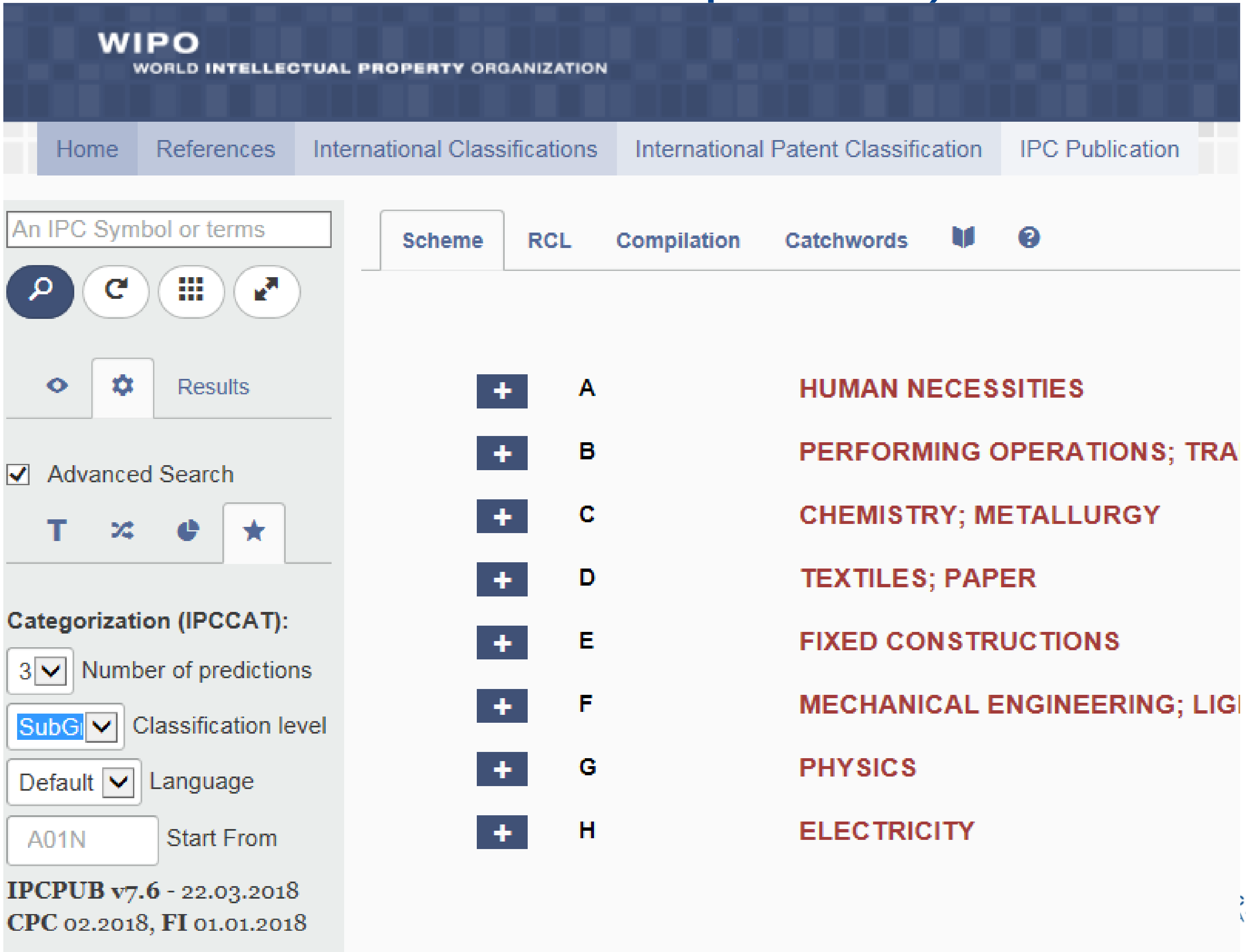
- Baseline of the solution (still valid in 2018):
  - **A Trained system** based on **neural networks (NN),**
  - Able to provide several predictions,
  - **that can be retrained** (new vocabulary, IPC revisions, patent reclassification).
  - **Data**: Training collection with **good IPC coverage** i.e. millions of already IPC classified patent documents (with at least Title and Abstract)
- Training /Testing phase : 80% / 20%
  - **Coverage and Precision assessment**: automated evaluation based on million of test cases

IPCCAT-neural : Production

- **Retraining with 100% of the collection**
  - **Web service**: returns 1 to 5 guessed IPCs with a numerical confidence level for each
  - User interface and API documentation through IPC publication platform (IPCPUB)

- **Potential cooperation agreement / FIT for the provision of IPCCAT to IPOs** (e.g. ES)

IPCCAT-neural : user interface (through IPC Publication platform)


WIPO
WORLD INTELLECTUAL PROPERTY ORGANIZATION
Home
References
International Classifications
International Patent Classification
IPC Publication
An IPC Symbol or terms
Scheme
RCL
Compilation
Catchwords
Results

☑ Advanced Search

Categorization (IPCCAT):

3 Number of predictions

SubGr Classification level

Default Language

A01N Start From

IPCPUB v7.6 - 22.03.2018
CPC 02.2018, **FI** 01.01.2018

+	A	HUMAN NECESSITIES
+	B	PERFORMING OPERATIONS; TRA
+	C	CHEMISTRY; METALLURGY
+	D	TEXTILES; PAPER
+	E	FIXED CONSTRUCTIONS
+	F	MECHANICAL ENGINEERING; LIGI
+	G	PHYSICS
+	H	ELECTRICITY

IPCCAT-neural : automatic text categorization in the IPC

- **Challenges/current solution:**
  - Availability of large and unified training collections with good IPC coverage: **WIPO DELTA XML** (currently computed from DOCDB XML)
  - Quality of IPCCAT-Neural (Absolute Vs. **Relative**): **imitates IPC actual usage in DOCDB**.

IPCCAT-neural : challenges

- **Precision versus Recall:**
  - One IPC is usually not enough for patent classification =>highest possible precision for the top prediction is not necessarily the best objective e.g. for prior art search
  - **IPCCAT precision based on three-guesses evaluation method**
  - Predictions of IPC symbols on the basis on a text input **with a confidence level for each**
  - Consideration for **additional feature (NN) to predict the number of IPCs** to be used

IPCCAT-neural : quality

- **IPCCAT quality is relative to IPC quality in its training collection:**
  - IPCCAT imitates human practices (good and bad ones)
  - Limited by patent documents fragments available for its training (currently Title & Abstract)
  - Potential added value from Full text data needs to be revisited (last assessed in 2003)

- **IPCCAT offers consistent and repeatable predictions**

IPCCAT-neural 2018

- Where are we today?

IPCCAT-neural 2018: text categorization in the IPC at subgroup level

- Automatic prediction in 99% of the IPC i.e. among **72,137 categories**

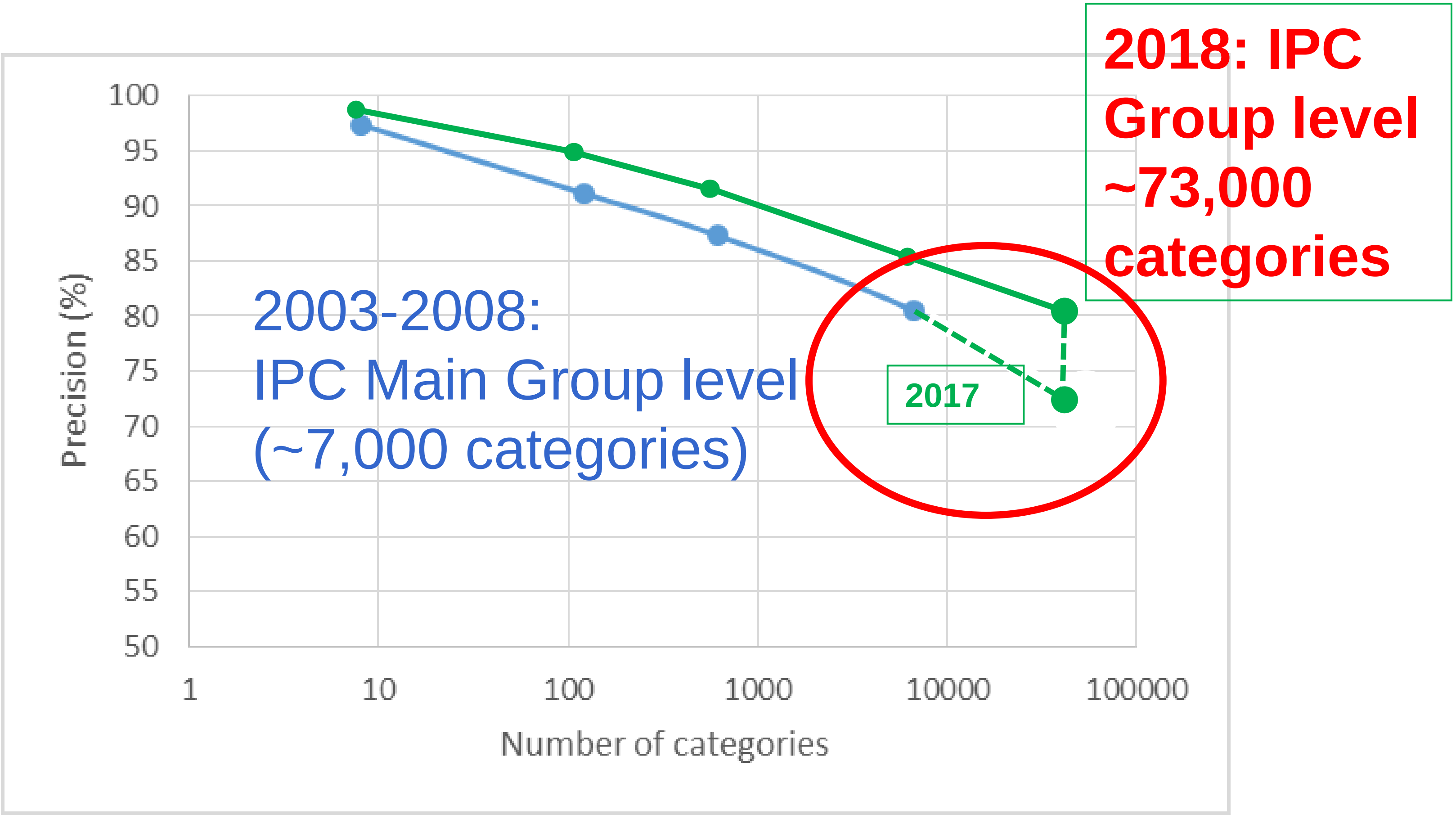
- Top-three guess **precision** > **80%**

IPCCAT-neural 2018: text categorization in the IPC at subgroup level

Training collection, IPC coverage and precision:

- Training collection: 27.7 million in EN and 4.4 in FR
- **Coverage of the IPC** (using IPC and CPC through concordance):
  - **99%** at subgroup level (**EN**)
  - 91% at subgroup level (FR)
- **Precision** (three guesses):
  - **82.5 % at subgroup level (EN)**
  - 72% at subgroup level (FR)

Evolution of IPCCAT R&D over years



2003-2008:
IPC Main Group level
(~7,000 categories)

2018: IPC Group level ~73,000 categories

IPCCAT-neural 2018

- Potential use of IPCCAT technology

IPCCAT-neural technology potential usage

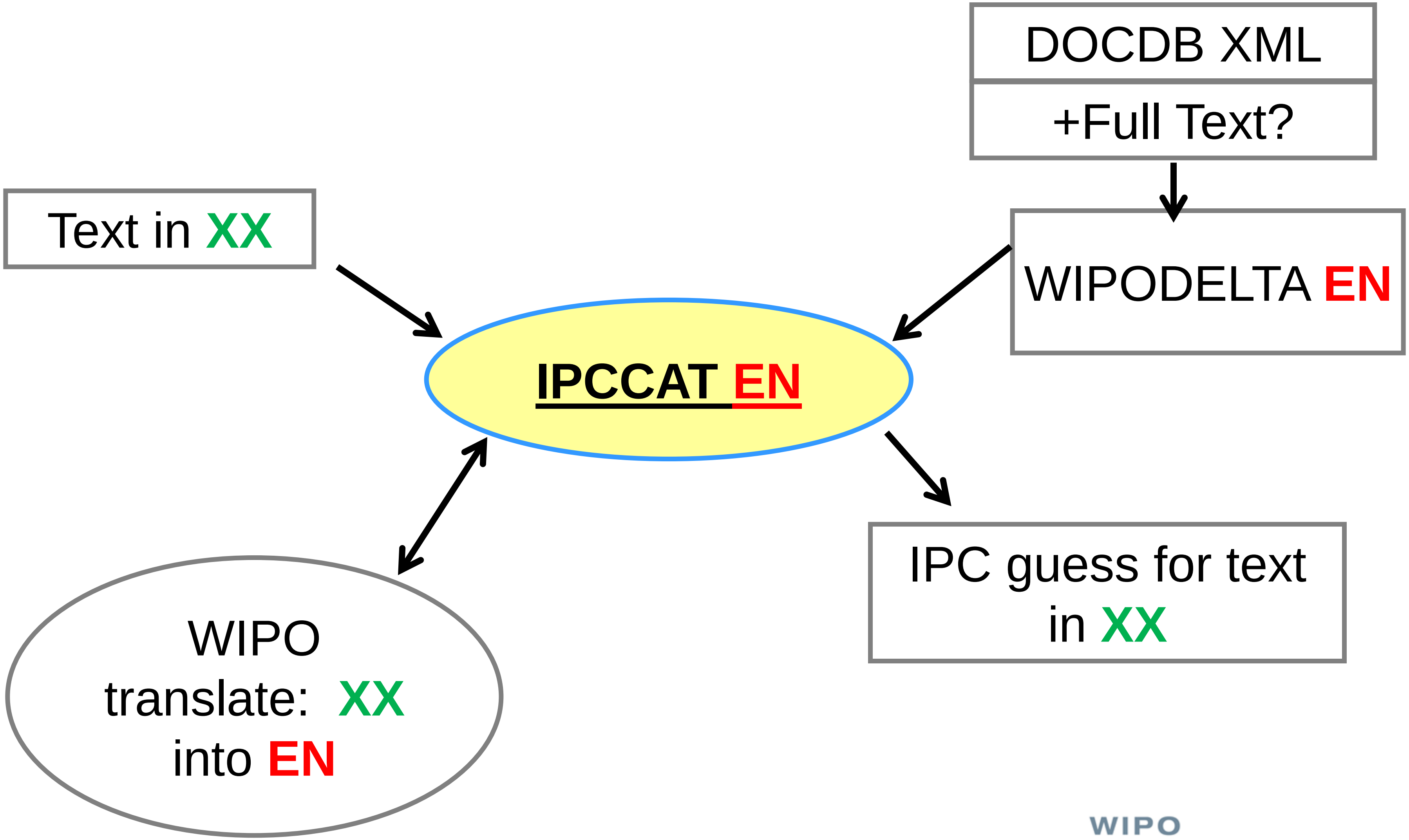
- **What it could be for?**
  - patent or NPL classification: improving consistency
  - Others: Massive extraction of documents according to training patterns (seeds) e.g. for EST…

- **Practical use of IPCCAT-neural**
  - Reduction of the backlog of IPC reclassification through automation of the residual IPC reclassification of patent documents after some years: **Potential alternative to IPC reclassification Default transfer**

IPCCAT-neural for IPC reclassification

- **Additional Challenges:**
  - **non-EN languages:**
    - Large training collections, with good IPC coverage
  - Consistency in IPC classification practices
  - Number of IPCs to be used for a given document

IPCCAT-neural cross lingual



DOCDB XML

+Full Text?

WIPODELTA **EN**

Text in **XX**

IPCCAT EN

IPC guess for text in **XX**

WIPO translate: **XX** into **EN**

Cross-lingual text categorization to assist IPC reclassification

- Chronology:

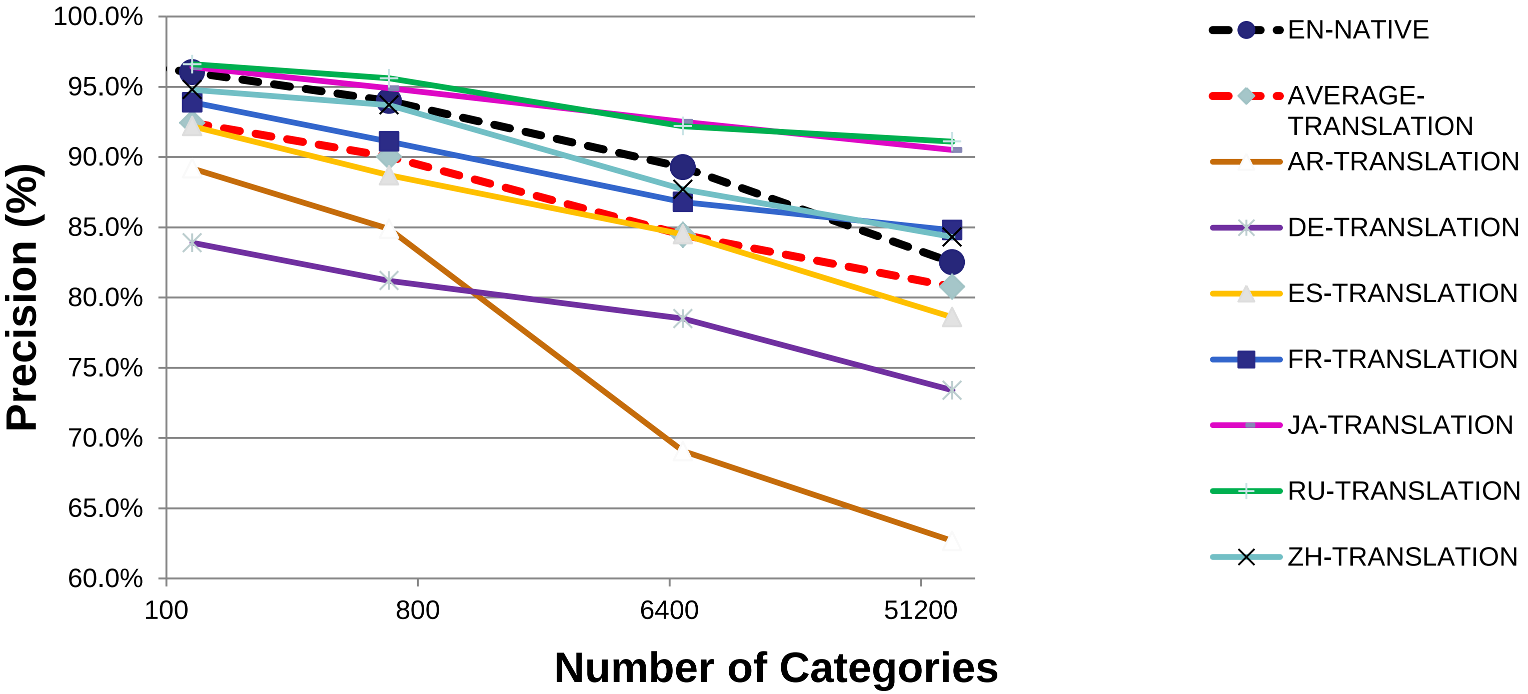
1. **Evidence** that text categorization works at IPC subgroup level with an **acceptable level of precision**: **Done**

2. Integration of IPCCAT neural at sub-group level into **IPCPUB v 7.6** **Done**

3. Confirmation that **Cross-lingual text categorization** can assist in other languages than EN, even in absence of large training collections: **Done**

IPCCAT-neural cross lingual prototype

- **Test with 1000 randomly selected patents in AR, DE, ES, FR, JA, RU, ZH**
- **Difficult to compare, not the same distribution of patents**



WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

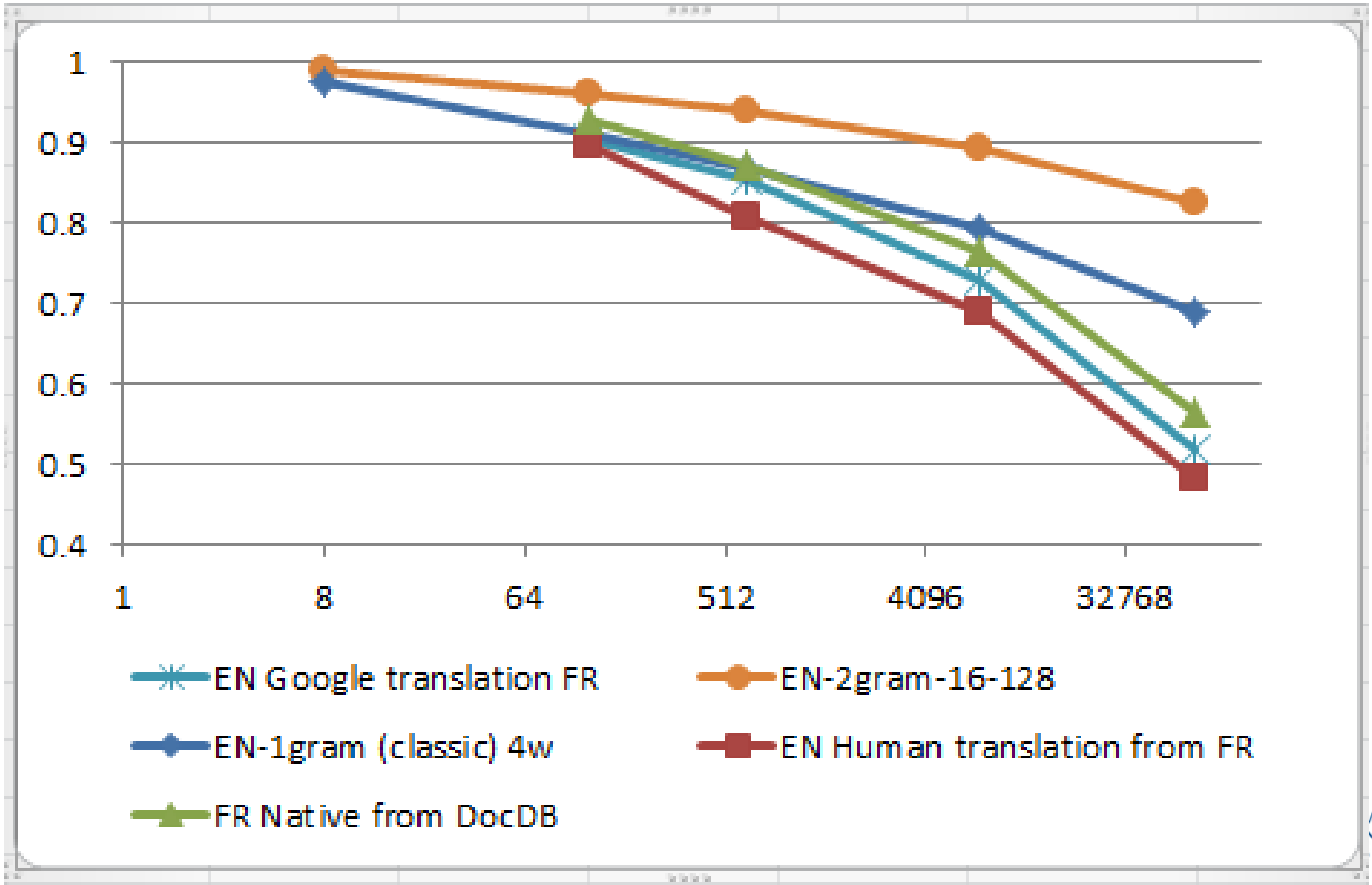
IPCCAT-neural cross lingual evaluation

- **IPCCAT trained in FR with a smaller corpus (5 million)**
  - **Vs.**
- **IPCCAT trained in EN with a bigger corpus (27 million) + automatic translation into FR**

- **Promising but …answer to come** (work in progress)

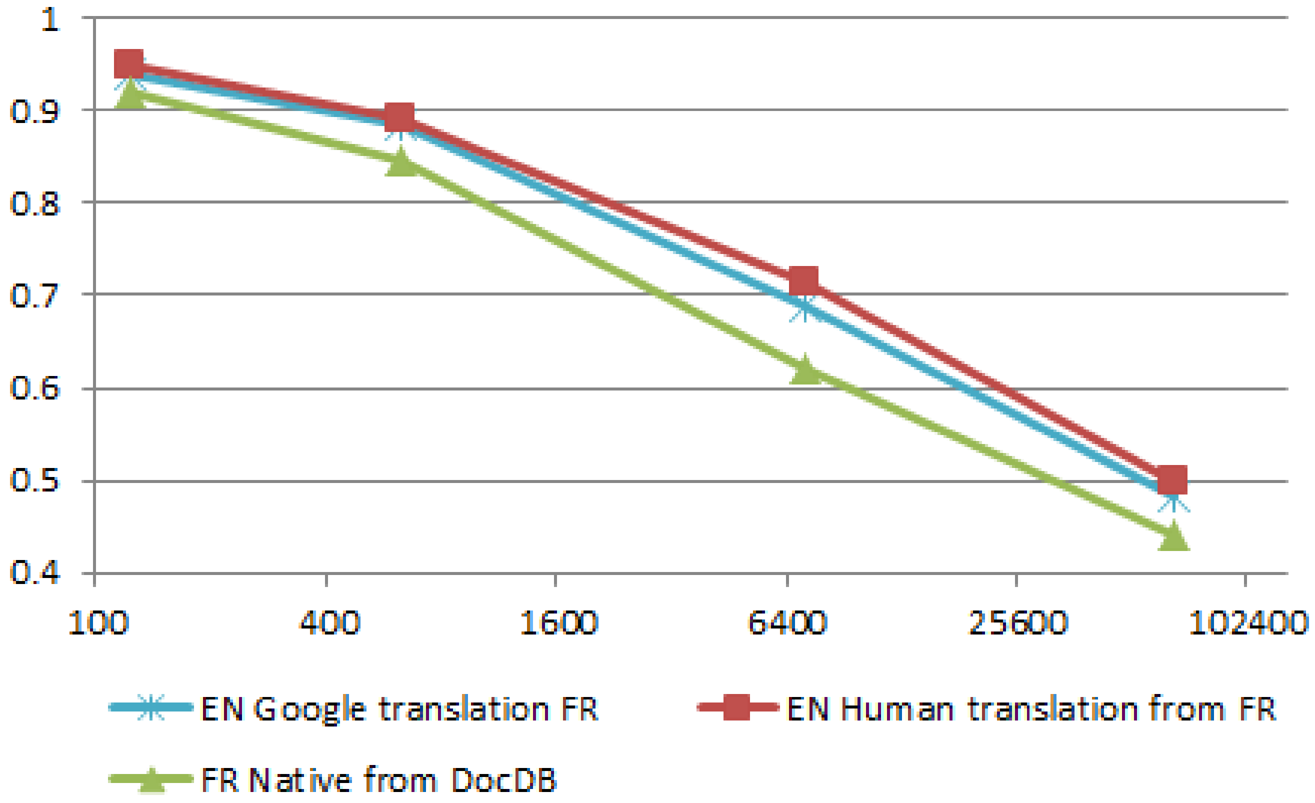
IPCCAT-neural cross lingual Vs. IPCCAT-neural FR

- **1000 randomly selected patent docs > sept 2017 in FR with (human?) translation in DocDB**



IPCCAT-neural cross lingual Test

- **1000 randomly selected patent docs > sept 2017 in FR with (human?) translation in DocDB within G06F subclass**



Cross-lingual text categorization to assist IPC reclassification

- **Chronology: (Still a long way to go)**

4. Incentives for R&D in automated text categorization: **WIPO DELTA** training collection: **Done**
5. Propose alternatives to Default Transfer e.g. guessed **number of symbols and IPC symbols** based on IPCCAT prediction and **related confidence levels, IPC-CE decisions**, resource planning, etc…: **2019-2020?**
6. **Development of** the production-scale solution integrating **cross-lingual text categorization and WIPO translate: 2019-2020?**
7. **Integration in IPC reclassification system (IPCWLMS) 2020?**

Incentive to R&D in text categorization: **WIPO-Delta collections & mycat**

- **Incentives for research and development institutes interested in automatic text categorization :**
  - WIPO DELTA 2018 **EN and FR** datasets available upon request
    - Fully specified XML format
    - **~50 million excerpts of patent documents classified in the IPC (and 4.7 million in FR)**
    - See **http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html**
  - **Open source: Mycat classifier available as on demand by the Olanto Foundation**

NCLCAT-neural 2017 Proof of Concept

- Potential use of AI for the Nice classification (NCL)

NCLCAT-neural 2017 Proof of Concept

- Potential use of AI for the Nice classification (NCL)
  - Cost-limited R&D to visit the **potential of AI in predicting the most appropriate NCL CLASS** on the basis on a text input (e.g. for TM applicants)
    - **Deep learning**
    - **Classic Neural Networks**
  - **Analysis and Prototype based on US and ES data**

NCLCAT-neural 2017 POC

- **Main outcomes** (details in the NCLCAT report)
  - **AI support to NCL is promising** and performs better than classic text search (Tests on US / ES collections)
  - **Prototype was done based on web service**
  - Automated testing on 40% of the collection indicates an **average accuracy ~98% for top 3 guesses**

NCLCAT-neural POC

- Other outcomes
- **Processing of the training collection is the real added-value**
  - Expensive data standardization and extraction should be improved
  - more investigation needed to address confusion between Classes (in particular Service Classes)

NCLCAT-neural POC

- **Some more outcomes**
  - A test on the ES collection using Mycat (classic Neural Network without recent improvements):
    - **Precision: 96.4%**, top 3 accuracy at **99.0%**
    - Deep Learning Vs. classic Neural Network: Not yet any evidence that convolutional Neural Network perform better
  - **Need for better and larger training sets** (e.g. Madrid collection)

Thank you for your attention!

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION