IC-SDV 2018

Automatic Text categorization in the International Patent Classification

IPCCAT-Neural

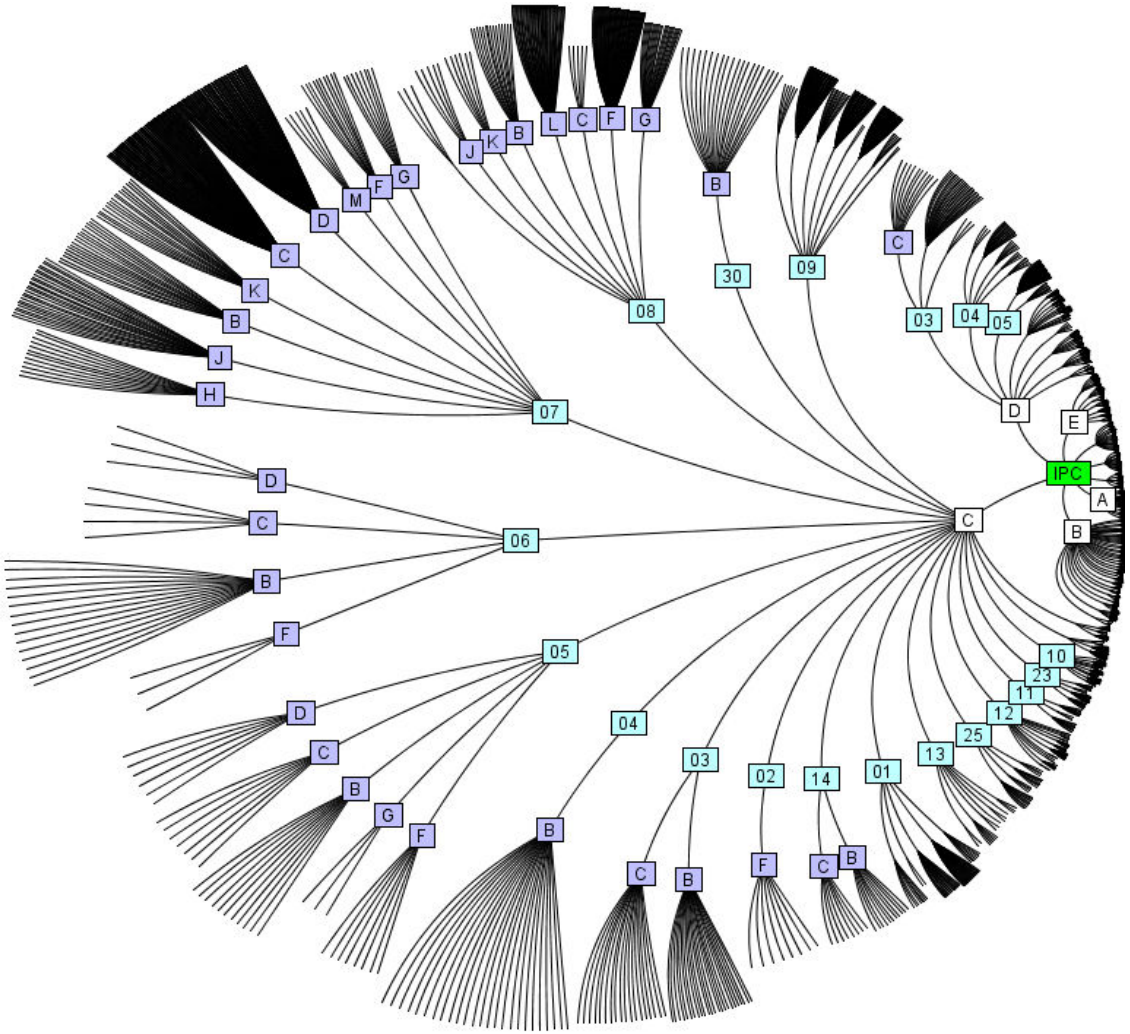
Nice
April 23, 2018

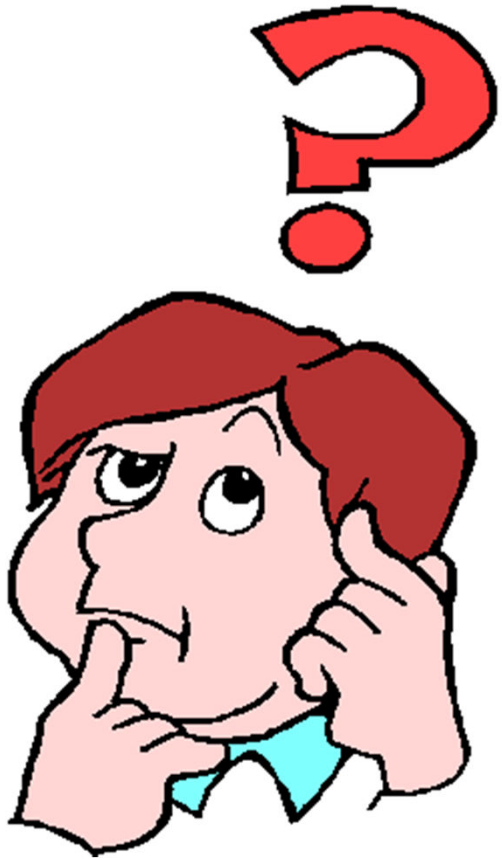
Patrick FIÉVET & Jacques GUYOT

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 patent classification practices by human beings

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:**
  - IPCs allotment in small Patent Offices
  - Automatic routing of patent/technical documents according to their technical domains

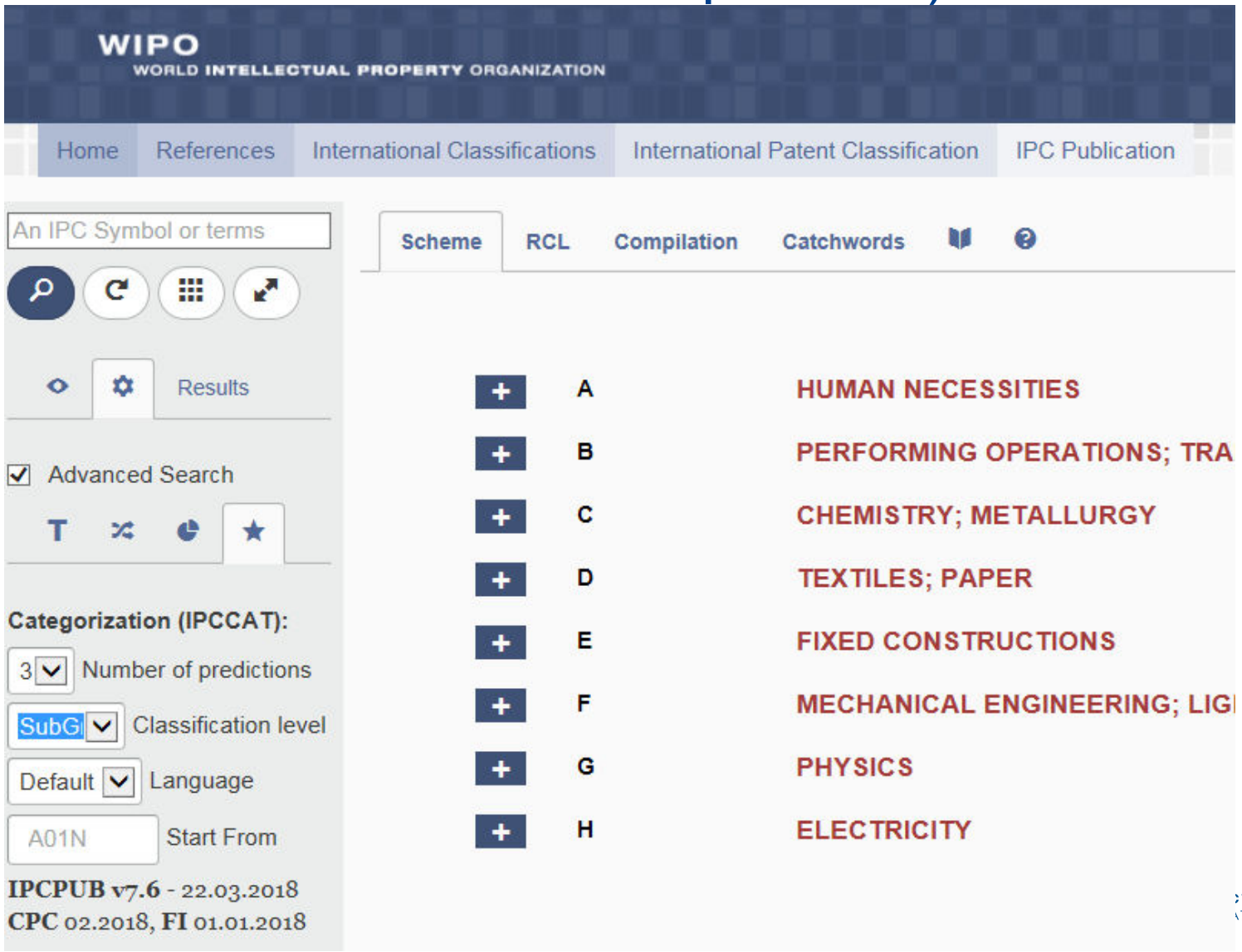
IPCCAT-neural : Principles

- Large collection (millions) of patent documents already classified, preferably with good-reputation practices
- Minimum number of documents for each IPC symbol
  - Use of the CPC (converted into IPC through concordance)
  - Complement with IPC symbols
- **Training /Testing phase : 80% / 20%**
  - Precision measure: Three-guesses evaluation on millions of test cases

IPCCAT-neural : Principles

- **Production use: 100% of the collection**
  - Web service: returns IPCs with a numerical confidence for each
  - User interface through IPC publication platform (IPCPUB) 5 confidence levels

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



IPCCAT-neural : automatic text categorization in the IPC

- **Challenges?**
  - IPC coverage Vs. availability of training collections
  - Precision Vs. Recall (e.g. for Prior art Search)
  - Absolute Vs. relative quality of IPCCAT-Neural

IPCCAT-neural : recall challenges

- **Recall: “First” IPC does not mean “Main” IPC**
  - One IPC is usually not enough for patent document classification but is enough for their routing
  - **Automatic Prediction** (or guess) of the **most appropriate IPC symbols** on the basis on a text input (e.g. patent abstract) **with a confidence level** associated to each of these predictions

IPCCAT-neural : precision challenges

- **Precision: “Mycat” based on classic neural networks**
  - **Trained system** based on (many) **neural networks**
  - No evidence that more recent technology e.g. Deep Learning would perform better than “classic” neural networks (because patent classification is not hidden information in the training collection)
  - Use of N-grams (terms with several words)

IPCCAT-neural : quality challenges

- **IPCCAT quality is relative to IPC quality in its training collection:**
  - IPCCAT Imitates human practices (good and bad ones)
  - Limited by patent documents fragments used during its training
- **IPCCAT offers consistent and repeatable predictions**
  - That human beings are usually not be able to achieve

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:

- **27.7 million in EN**

IPC coverage(using IPC and CPC through concordance):

- **99%** at subgroup level (**EN**)

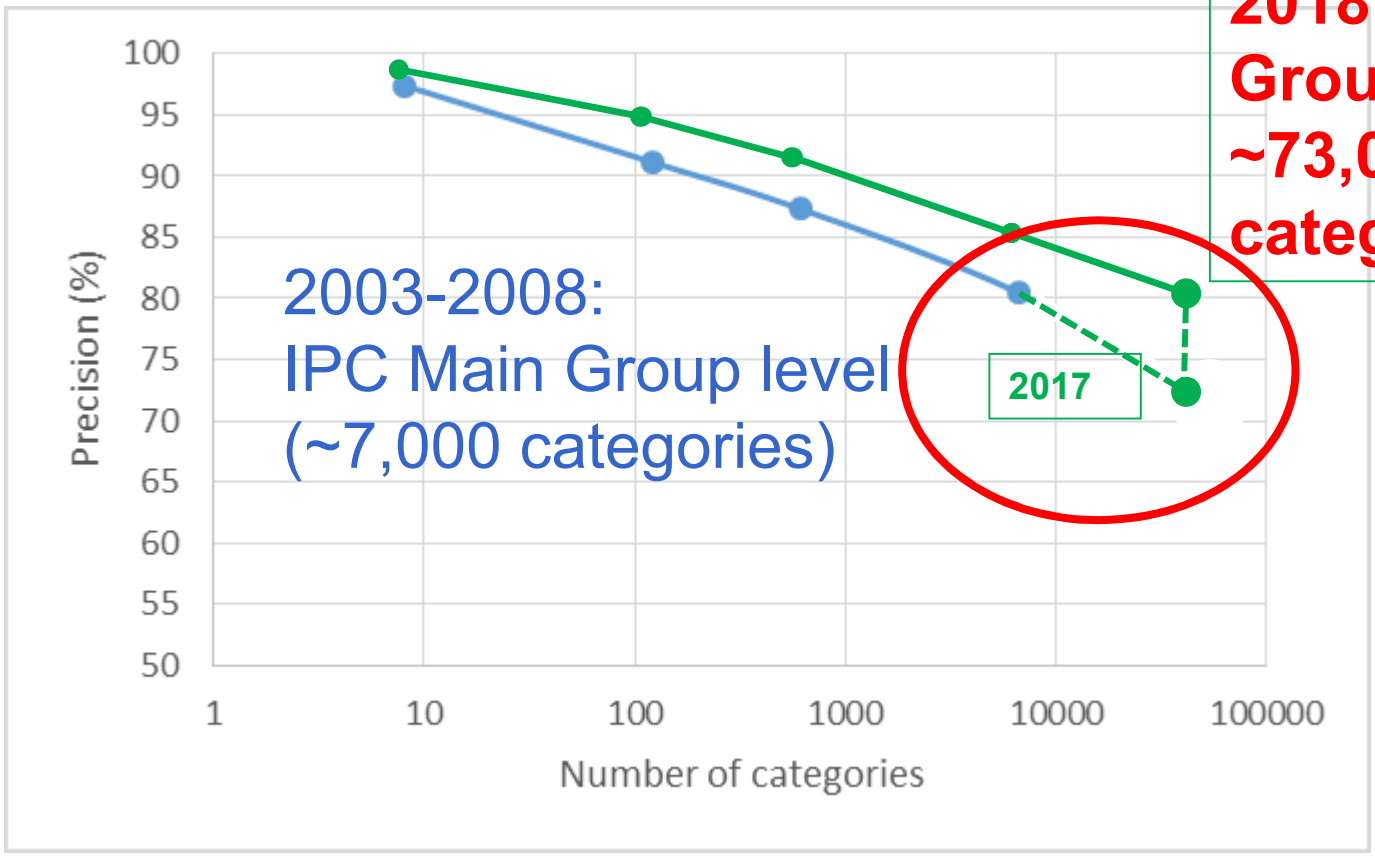
Top three guess evaluation of its Precision:

- **82.5%** based on **1.5 million of test cases (EN)**

IPCCAT-neural text categorization in the IPC at subgroup level

- **Why was It actually doable?**
  - Recent evolution of the IPCCAT classifier available on-demand as open source by the **Olanto foundation see** http://olanto.org/foundation
  - **Added value in data processing:**
    - Training based on patent documents computed from DOCDB XML excerpts (title + abstract)
    - Computation of both IPC and CPC classifications
    - Progress in computing power opens new R&D horizons e.g. GPU, text processing,…

Evolution of IPCCAT R&D over years



IPCCAT-neural 2018

- Potential use of IPCCAT technology

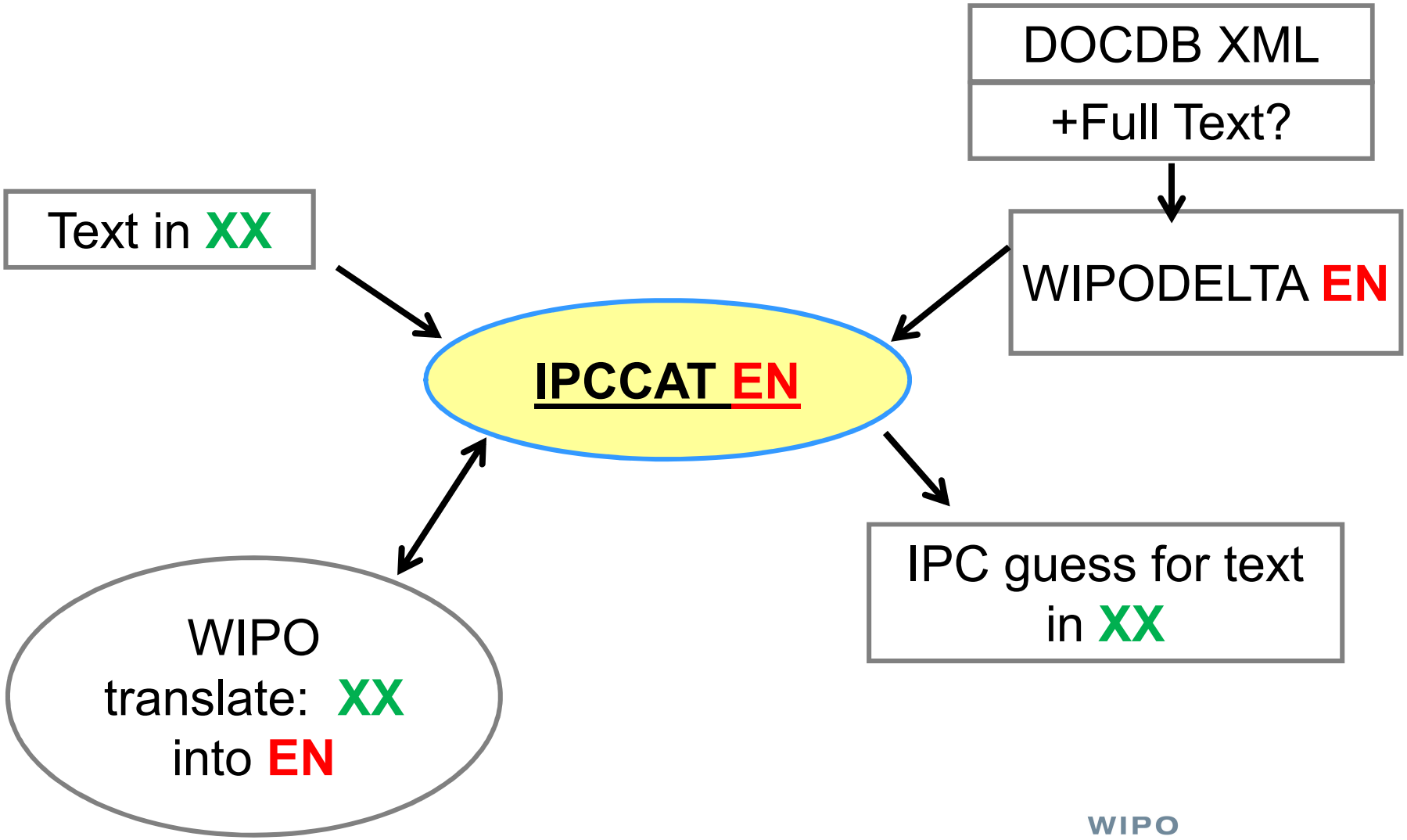
IPCCAT-neural practical use

- **What it could be for?**
  - Improvement of the consistency in patent classification
  - 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 language:**
    - Large training collection, with good IPC coverage
    - Consistency in legacy classification practices
    - Preserve possibility of language-dedicated solution
  - **Costs containment for WIPO**
    - Streamline production of training collection(s)
    - Streamline IPCCAT yearly retraining (new vocabulary)

IPCCAT-neural cross lingual

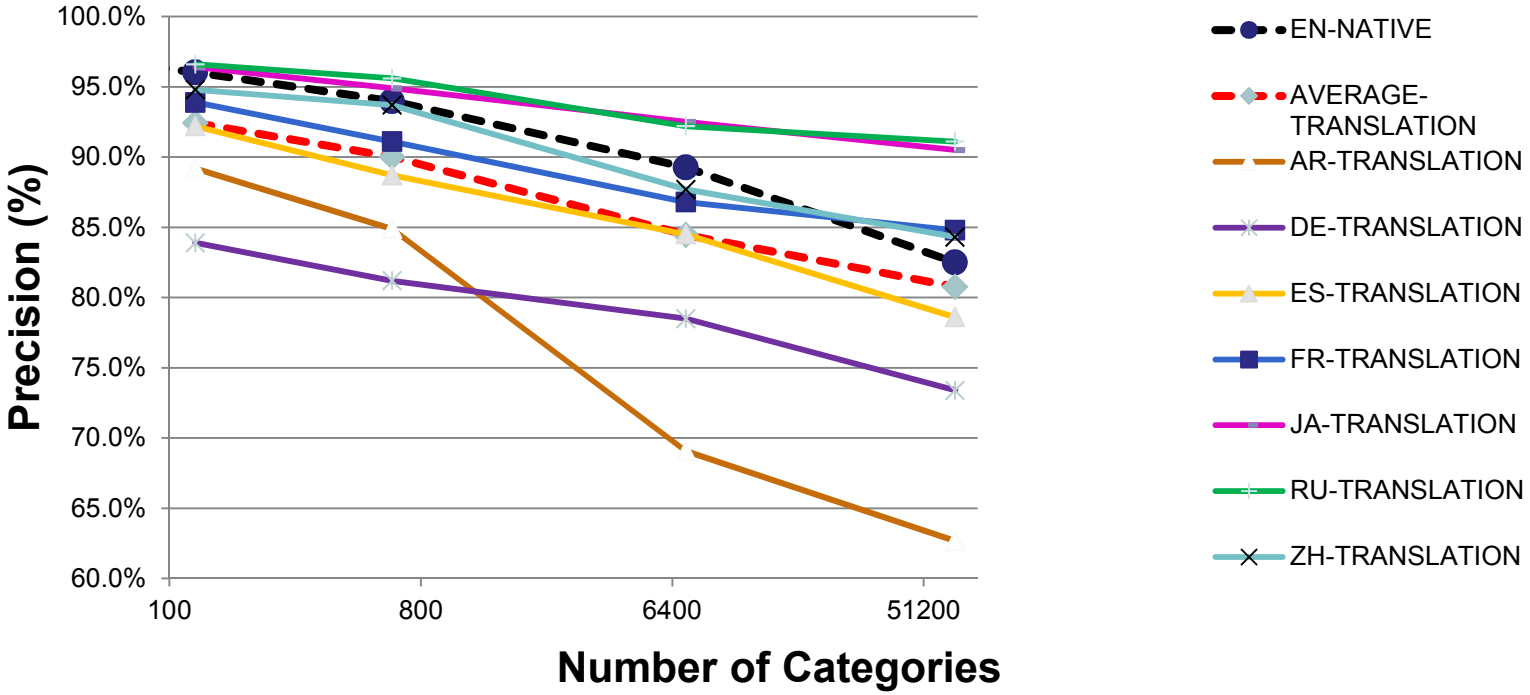


Cross-lingual text categorization to assist IPC reclassification

- **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

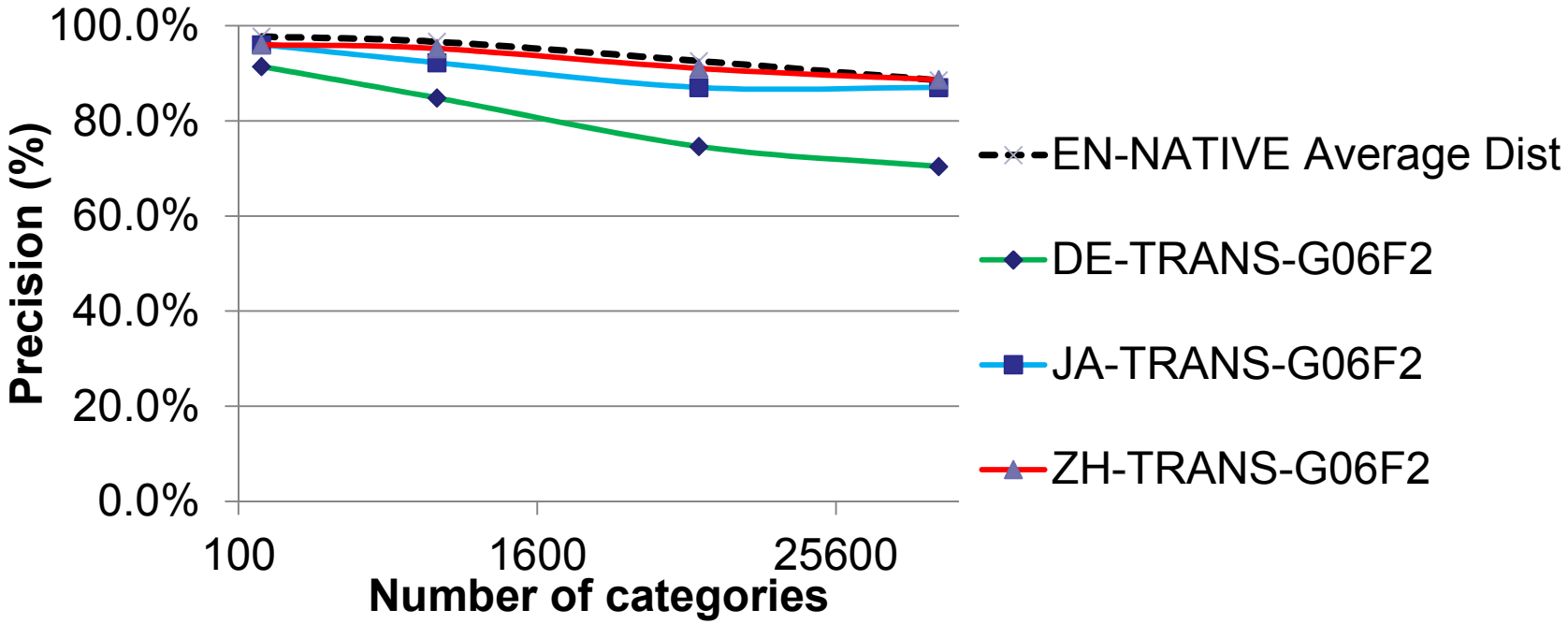
- **Test with 1000 randomly selected patents in AR, DE, ES, FR, JA, RU, ZH**
- **Difficult to compare, not the same distribution of patents**



WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

IPCCAT-neural cross lingual prototype

- **Test with 500 randomly selected patents from subclass G06F**
- **Losses due to translation are more visible for each language**
- **Needs to be evaluated for all languages**



WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

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. **more than one symbol** based on IPCCAT guesses **and confidence levels** for decisions, resource planning, etc.…: **2019**
  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 training collection**

- **Incentives for research and development institutes interested in automatic text categorization :**
  - WIPO DELTA 2018 EN dataset available upon request
  - Fully specified XML format
  - **~50 million excepts of patent documents classified in the IPC**
  - Complement the public WIPO-ALPHA & GAMMA datasets
  - See http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html

Thank you for your attention!

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION