SIMPLE SHIFT

OLANTO
Open Language Tools

Automatic Categorization : Future Perspectives

Jacques Guyot (jacques@simple-shift.com / jacques@olanto.org)

WIPO – Geneva – February 2017

Services & Researches

SIMPLE SHIFT

Open Language Tools

Simple-Shift

- A computer consulting company specializing in language engineering
  - Installation, maintenance, adaptation to the context of the organization
  - Have been installing CAT tools for more than 16 years, mainly for international organizations

Olanto

- Olanto is a non-profit foundation (Free Software - AGPL)
- Compete with nobody, but can be useful to every, is open to translators, terminologists , computer scientists, researchers, integrators , distributors, … for collaboration
- Software released or in development :

myCAT : concordancer and quote detector
myPREP : set of tools to prepare corpus (TMX, Bitext, Machine Translation training)
myPREP & myMT : set of tools to prepare corpus & statistical machine translation infrastructure
myTERM & How2Say: terminology manager based on TBX &terminological explorer for multilingual corpus
myCLASS : an automatic classifier for multilingual documents (https://www3.wipo.int/ipccat/)
mySEARCH : a multilingual search tool (using translation for requests).
Education: a translation environment for students.

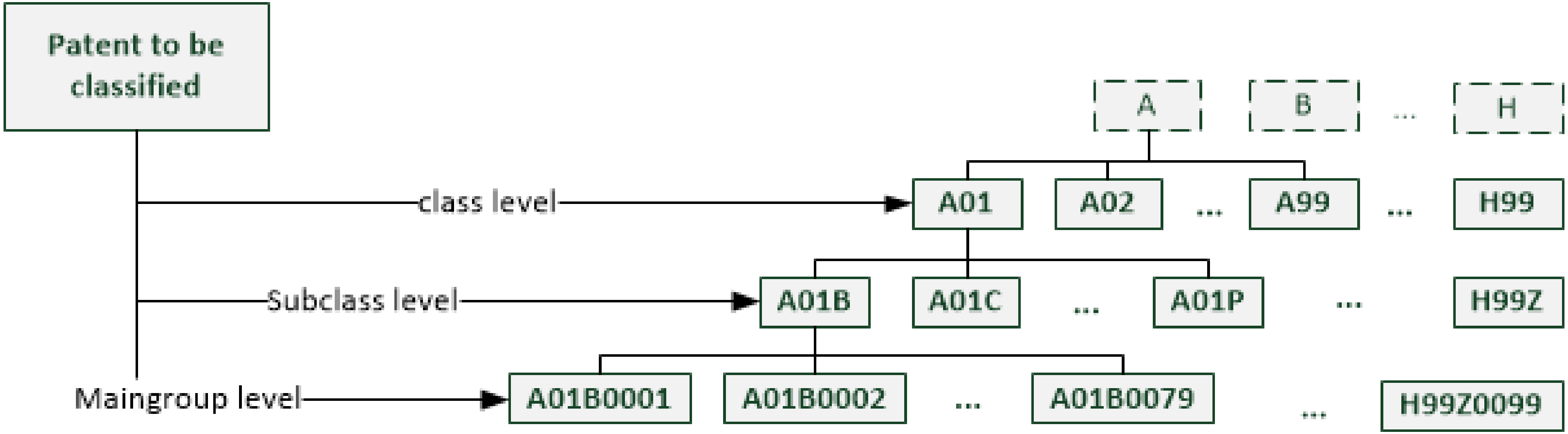
Presentation plan

- What was done at WIPO (since 2004)
- What can be done to improve IPCCAT
- Can IPCCAT be extended to other languages?

What is being done at WIPO

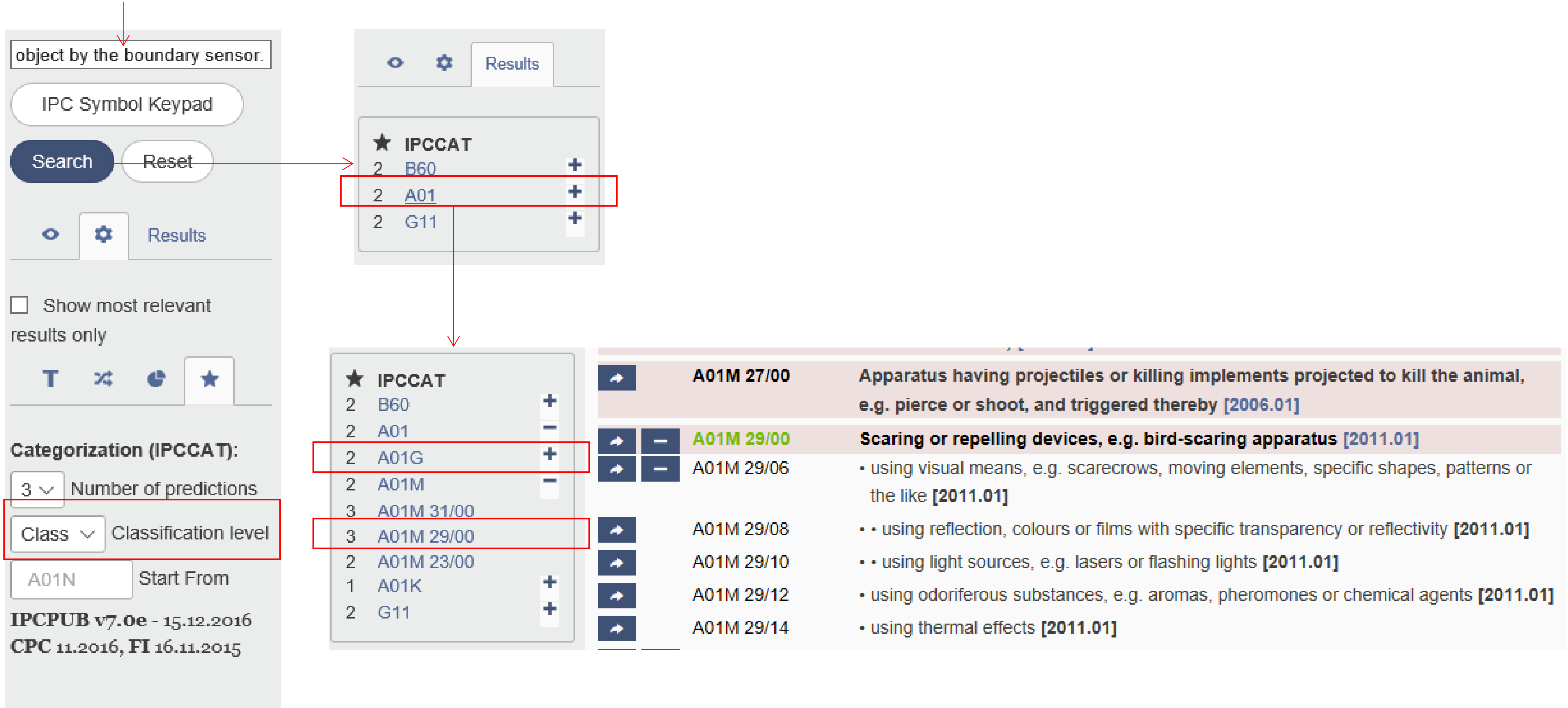
IPCCAT User interface available through IPC publication platform (IPCPUB):

- Copy the text to be classified
- Choose a classification level
- Have 3 guesses
- Select one
- Start again with a deeper level



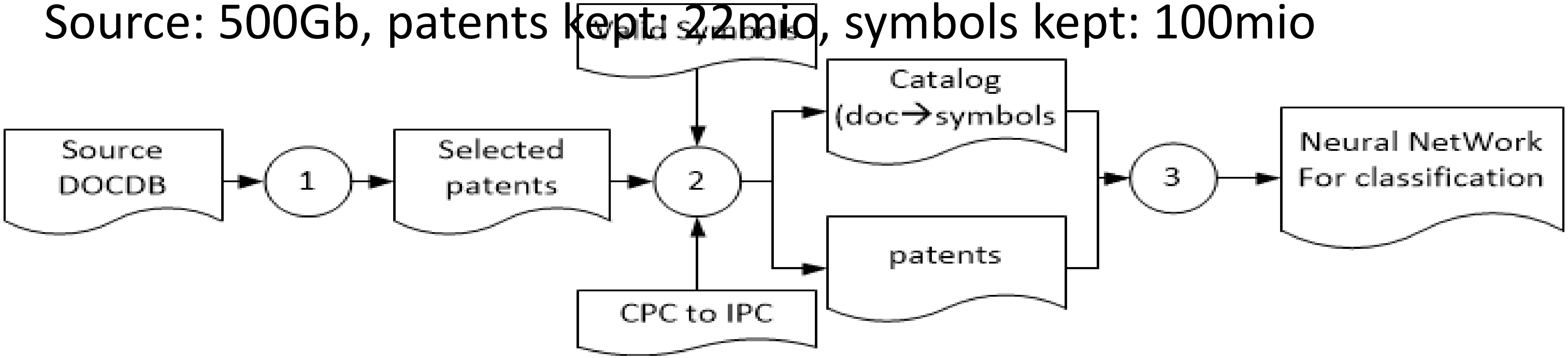
An example of use

A boundary control device, a boundary control system, and a method of conditioning the behavior of animals are provided. ….. upon sensing of the object by the boundary sensor.



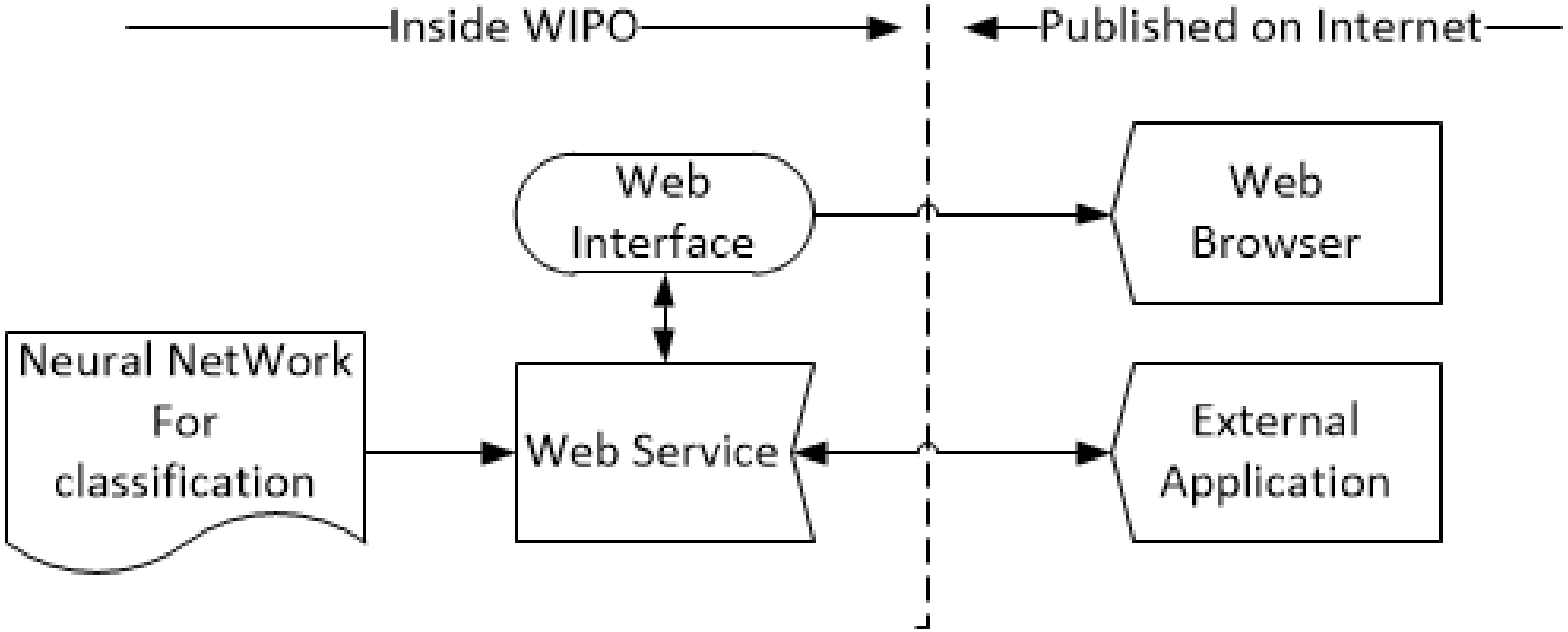
How it's done – Train a Neural Network

1. Select the English and French patents documents already classified. Keep only certain fields (title, abstract, symbols, ...)
2. Validate symbols to build the training corpus
3. Build a neural Network for each node of the classification hierarchy
4. Source: 500Gb, patents kept: 22mio, symbols kept: 100mio



How it's done – Published as a Web Service

1. Using the application through the WIPO interface with a browser
2. Using the Web Service through a specific application (developed externally)



What can be done to improve IPCCAT?

- **To Increase IPC coverage in the training corpus** (more symbols and at deeper level)
  Currently: 7,007 symbols among 72,981 in IPC 2017.01
- **To Increase IPCCAT accuracy**
  Currently: Top3 at main groups 80.5%
- **To Expand to other languages**
  Currently: English and French

Increase coverage (more symbols)

Add patents for uncovered symbols

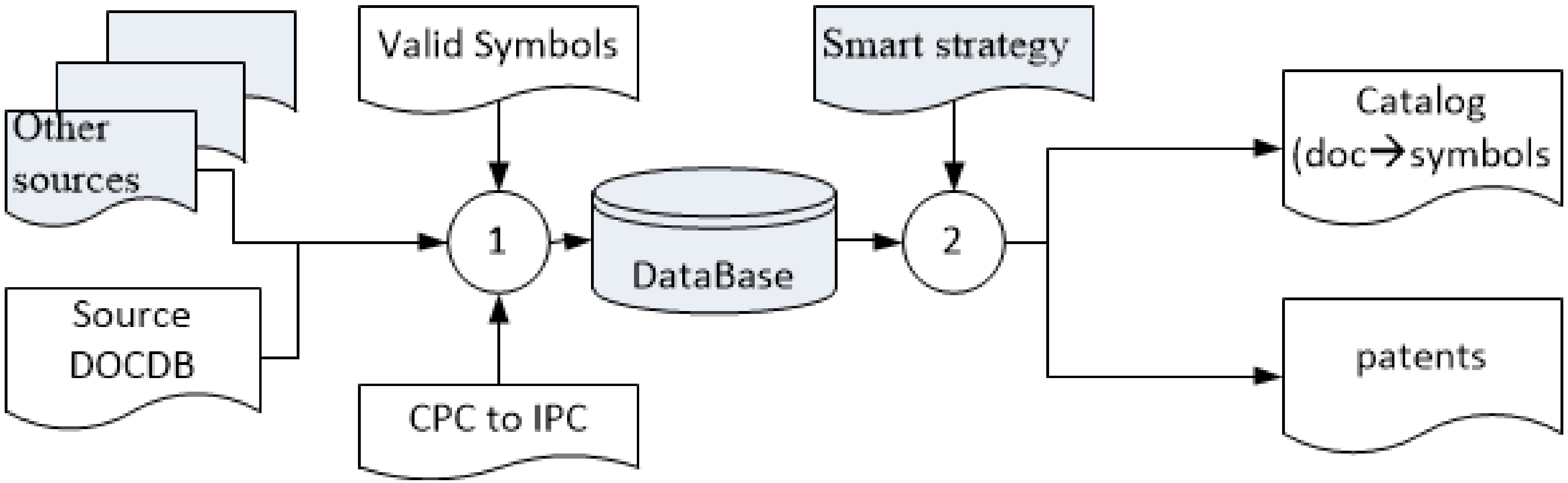
Improve the use of existing resources

→ Put all patents and symbols in a database

→ Extract the catalog with an intelligent strategy (CPC & IPC)

The experimented result at maingroup level (2016.01):

467 missing symbols and 310 in the improved version, ie 33% progress



Increase coverage (more symbols)

Add New sources for uncovered symbols

- Not easy to find reliable sources
- Not yet patent with this symbol, because too new

- Test with PatentScope

Examples of missing symbols	nb documents in Patent Scope	since
A23L0009	0	2016.01
A23L0015	0	2016.01
A23L0017	0	2016.01
A23L0025	0	2016.01
A23L0035	0	2016.01
A23P0020	0	2016.01
A42C0099	2	2006.01
A43D0057	2	2006.01
A43D0097	3	2006.01
A45D0097	16	2011.01

- Increase depth (group level)

In 2013, we conducted an experiment at the group level

Technically this is possible despite a network of 60 billion neurons

- Should improve coverage (see above)
- Must increase the accuracy by adding more examples for certain groups

Group	Stat 2013.01	Coverage
60 042	70 870	85%

Top 3 Average Precision (%)	
No Intermediate Step to Group	71%
Intermediate Step: From Class to Group	81%
Intermediate Step: From Main Group to Group	85%

Increase accuracy

For all techniques: **Add patents for under-populated symbols** (not enough examples for training)

Explore other approaches:

- **Support Vector Machine** (SVM)
  - Similar results - But very slow for training (100x)
- **Deep Learning**
  - Very good if the representation is hidden (sentiment analysis)
  - But no real improvement, for descriptive documents (without nuances) (https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf)
  - Need specialized machinery
  - To watch, see what emerges from this new technique

Increase accuracy

In 2010, we participated in a challenge organized by CLEF

(see http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-PiroiEt2010.pdf)

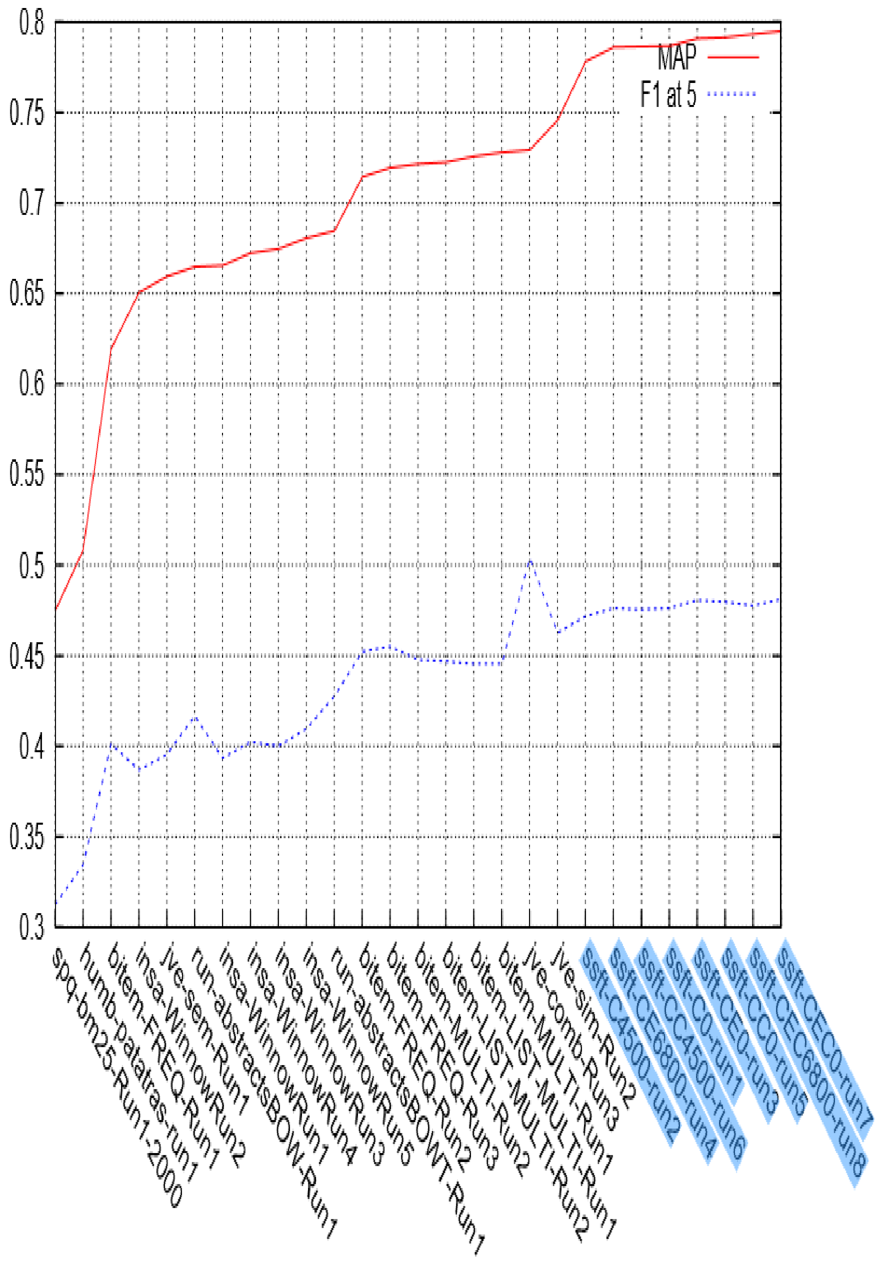
- 2 million patent corpus
- classification at main group level
- 12 participants

-> Our approach remains in front of all the others

Why?

- No language processing
- Keep all information

-> let the neural network do the job



Can IPCPUB be extended to other languages ?

The first version of IPCCAT had 4 languages EN, FR, DE, RU

- o But as we have seen above, It is difficult to maintain a training corpus with good coverage
- → Decide to maintain only English and French

What to do for other languages?

- o Automated translators have improved
- o The classification is not sensitive to syntax errors,
- o Only the correctness of the terminology is important

We decided to experiment the use of machine translation

Objectives of the experiment

- Compare several translation engines
- Choosing "difficult" languages
- Assess accuracy:
  - In the context of the interactive classification
  - In the context of reclassification
- Constraints: Have enough patents to do the tests

→ Translation engines google, yandex, WIPO-translate, Bing MS
→ Languages: **German, Russian, Chinese**
→ Maingroup for interactive classification A01B 1
→ For reclassification simulation A01B 1, A01B 3, A01B 49

Results for Interactive classification (A01B 1)

Source	nb patents	source	date	Mono class
RU	69	RUPAROM	2003	yes
DE	20	DEPAROM	2003	yes
ZH	20	PatentScope	recent	?

Precision Top 3 in % (The symbol is in the first three proposals)															
Task --> (EN %)	Class A01 (87%)			SubClass A01B (75%)			MainGroup A0B 1 (84%)			From class			From subclass		
	RU	**DE**	**ZH**	**RU**	**DE**	**ZH**	**RU**	**DE**	**ZH**	**RU**	**DE**	**ZH**	**RU**	**DE**	**ZH**
bing	94	100	95	88	100	85	58	100	75	74	85	90	88	100	95
google	94	100	100	94	90	75	62	85	70	84	80	80	100	100	100
yandex	94	85	95	90	75	85	68	75	75	84	70	90	94	80	95
wipo	94	85	95	91	90	80	61	95	65	75	80	80	96	95	100

Results for Interactive classification (A01B 1)

- → The automatic translation is sufficient to have honorable results (better than those of the trainings)
- → Between the translation machines there are differences.
- → But finally, as part of this test, they are not significant

Average of 5 tasks				Average
	RU	DE	ZH	RUDEZH
bing	81	97	88	89
google	87	91	85	88
yandex	86	77	88	84
wipo	83	89	84	85
Average	**84**	**89**	**86**	**86**

Results for reclassification

- We simulate the partition of a class into three parts
- T01B 0 / → T01B 1 /, T01B 3 /, T01B 49 /
- We train a neural network for this partition on english documents
- We use yandex for the translation from russian to english
- We use the first proposal for reclassification

	nb samples	Precision(first)
T01B 1	30	**87%**
T01B 3	30	**83%**
T01B 49	30	**70%**
average		80%

Translation can be an approach to reclassifying batches in foreign languages

Conclusion

- Neural networks are efficient and simple to implement.

  But we must remain vigilant on the new approaches
- Automatic translation is sufficiently efficient for classification tasks and allows access to automatic classification.

  But we have to test other languages (Arabic Spanish, Korean, ...)
- Emphasis should be placed on creating training corpuses
  - having sufficient examples for each symbol.
  - covering the maximum of the classification

  But we must remain relevant between effort and outcome
- Automatic classification at group level is possible

  But we must add this with caution

Thank you for your interest and attention