Automatic Categorization: Future Perspectives

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

WIPO – Geneva – February 2017
Services & Researches

Simple-Shift

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

Olanto

o Olanto is a non-profit foundation (Free Software - AGPL)
  o 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

<table>
<thead>
<tr>
<th>Examples of missing symbols</th>
<th>nb documents in Patent Scope</th>
<th>since</th>
</tr>
</thead>
<tbody>
<tr>
<td>A23L0009</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A23L0015</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A23L0017</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A23L0025</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A23L0035</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A23P0020</td>
<td>0</td>
<td>2016.01</td>
</tr>
<tr>
<td>A42C0099</td>
<td>2</td>
<td>2006.01</td>
</tr>
<tr>
<td>A43D0057</td>
<td>2</td>
<td>2006.01</td>
</tr>
<tr>
<td>A43D0097</td>
<td>3</td>
<td>2006.01</td>
</tr>
<tr>
<td>A45D0097</td>
<td>16</td>
<td>2011.01</td>
</tr>
</tbody>
</table>
• 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

<table>
<thead>
<tr>
<th>Group Stat 2013.01 Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 042</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top 3 Average Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Intermediate Step to Group</td>
</tr>
<tr>
<td>Intermediate Step: From Class to Group</td>
</tr>
<tr>
<td>Intermediate Step: From Main Group to Group</td>
</tr>
</tbody>
</table>
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)
  - 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

- 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?**

- Automated translators have improved
- The classification is not sensitive to syntax errors,
- 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)

<table>
<thead>
<tr>
<th>Source</th>
<th>nb patents</th>
<th>source</th>
<th>date</th>
<th>Mono class</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU</td>
<td>69</td>
<td>RUPAROM</td>
<td>2003</td>
<td>yes</td>
</tr>
<tr>
<td>DE</td>
<td>20</td>
<td>DEPAROM</td>
<td>2003</td>
<td>yes</td>
</tr>
<tr>
<td>ZH</td>
<td>20</td>
<td>PatentScope</td>
<td>recent</td>
<td>?</td>
</tr>
</tbody>
</table>

## Precision Top 3 in %
(The symbol is in the first three proposals)

<table>
<thead>
<tr>
<th>Task --&gt; (EN %)</th>
<th>Class A01 (87%)</th>
<th>SubClass A01B (75%)</th>
<th>MainGroup A0B 1 (84%)</th>
<th>From class</th>
<th>From subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>bing</td>
<td>RU  94 DE 100 ZH 95</td>
<td>RU  88 DE 100 ZH 85</td>
<td>RU  58 DE 100 ZH 75</td>
<td>RU  74 DE 85 ZH 90</td>
<td>RU  88 DE 100 ZH 95</td>
</tr>
<tr>
<td>google</td>
<td>RU  94 DE 100 ZH 100</td>
<td>RU  94 DE 90 ZH 75</td>
<td>RU  62 DE 85 ZH 70</td>
<td>RU  84 DE 80 ZH 80</td>
<td>RU  100 DE 100 ZH 100</td>
</tr>
<tr>
<td>yandex</td>
<td>RU  94 DE 85 ZH 95</td>
<td>RU  90 DE 75 ZH 85</td>
<td>RU  68 DE 75 ZH 75</td>
<td>RU  84 DE 70 ZH 90</td>
<td>RU  94 DE 80 ZH 95</td>
</tr>
<tr>
<td>wipo</td>
<td>RU  94 DE 85 ZH 95</td>
<td>RU  91 DE 90 ZH 80</td>
<td>RU  61 DE 95 ZH 65</td>
<td>RU  75 DE 80 ZH 80</td>
<td>RU  96 DE 95 ZH 100</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Average of 5 tasks</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RU</td>
<td>DE</td>
</tr>
<tr>
<td>bing</td>
<td>81</td>
<td>97</td>
</tr>
<tr>
<td>google</td>
<td>87</td>
<td>91</td>
</tr>
<tr>
<td>yandex</td>
<td>86</td>
<td>77</td>
</tr>
<tr>
<td>wipo</td>
<td>83</td>
<td>89</td>
</tr>
<tr>
<td>Average</td>
<td><strong>84</strong></td>
<td><strong>89</strong></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>nb samples</th>
<th>Precision(first)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01B 1</td>
<td>30</td>
<td>87%</td>
</tr>
<tr>
<td>T01B 3</td>
<td>30</td>
<td>83%</td>
</tr>
<tr>
<td>T01B 49</td>
<td>30</td>
<td>70%</td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>80%</td>
</tr>
</tbody>
</table>

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