Patent Machine Translation (Handling large data with Moses)

Bruno Pouliquen (Bruno.Pouliquen@wipo.int) with the help of Marcin Junczys-Dowmunt
What this talk is about?

• Introduction
• Patent translation, what is specific about it?
• Our tool: Tapta
• Big model management with Moses
• Language specificities
• Our tool installed in various places
• User interfaces
• Quality / user acceptance
• Conclusion
Introduction: Basic facts about WIPO
World Intellectual Property Organization

Mission: promote the protection of intellectual property rights worldwide and extend the benefits of the international intellectual property system to all member states

Status: a specialized agency of the UN
Member states: 184
Observers: 250+
Staff: 950 from 101 countries

Translation of Patents:
PCT: translation of titles/abstracts and International Search reports (40 million words/year to EN and FR, 90% outsourced)
Patentscope: patent search engine
Patent translation, what is specific about it?

- A patent application is made up of a title, an abstract, a description and claims
  - Title: 8 words
  - Abstract: 114 words
  - Description: 6’428 words!
  - Claims: 726 words

- Meta information: office, classification (IPC), original language, filing date, publication date, inventor etc…
- Specific language, scientific terms, almost no repetitive text (Unknown phrase: “I am”, almost no proper names…)
- Usually only the title and the abstract are translated
WIPO SMT tool: TAPTA

- Fully automatic
  - preparation of data
  - training
  - evaluating/mert/binarization/etc.
- Domain aware
- Fast translations (on the fly)
- Free to use (open source + in-house development)
- Confidentiality
- Various User interfaces
- First goal: assimilation, online translation of patent applications on our search engine
- Additional goal: dissemination, integration in CAT tool, “translation accelerator”
TAPTA: First version - 2011

- Patent SMT: title+abstracts
- 180 M words (en-fr), 8M words
- Called TAPTA ("Translation Assistant for Patent Title and Abstract")
- Domain-aware: 32 domains encoded as factors in Moses
- Pouliquen et al. EAMT 2011
COPPA: Corpus Of Parallel Patent Applications

All English-French PCT application title and abstract (1990-2010)
Free for research
180 Million words in TMX format (8.7 Million translation units)

http://www.wipo.int/patentscope/en/data/products.html#coppa
Our system prepares the data for Moses, applies some post-processing (filter, pruning, binarization, optimization…) and offers various interfaces to translate.
TAPTA:

Moses' train model

IPC
A0.1
A01.2
A01.2.3
A02
B01
B01.1.2

IPC

en
fr

Filter
Sentence-align
Filter align

parallel segments

filter align

sentence-align

phrase table 0

language model

Moses decoder

tapta client

tapta server

en
fr

domain

antimicrobial coatings
enrobages antimicrobiens
CHEM

electronic transaction
transaction électronique
SPOR

ternary mixed ethers
éthers mixtes ternaires
CHEM

submarine
sous – marin
MARI

automatic translation
traduction automatique
DATA

automatic translation
translation automatique
BLDG

Moses’ train model

post-filter
prune

phrase table 0-0
phrase table 0,1-0

reordering model
language model
TAPTA: more than titles & abstracts

• Many offices, many languages (en, ja, fr, zh, de, ru, ko, es, pt…)
• Usually only the title and abstract are translated…
• But:
  • European Patent Office: translations of claims in English, French and German
  • One invention can be patented in more than one country (in various languages)…
  • Huge amount of data ~ Millions of patent applications
Zur Lösung dieses Problems sieht die Erfindung vor, dass bei dem eingangs genannten Tonträger wenigstens eine der Justierdurchbrechungen auf der Oberseite des Tonträgers von einem nach oben aus der Oberseite vorstehenden Wulstring, der z. B. ist, der Wulstring kann sich unmittelbar gelöst einen geringen Abstand von etwa 1,0 mm und einem flachen Absatz, was zur Pressung oder dem Durchbruch beitragt.

Die Wände der vergrößerten Justierdurchbrechungen sollten auf der den 6 mm haben. Die vergrößerten Durchbrechungen

Ein älterer bekannter Vorschlag (US Patentschrift 4,298,967) ist von der Praxis

Alternativ zu dieser Lösung kann auch ein Teil der Justierdurchbrechungen, z.

Selbstverständlich können auch alle Durchbrechungen eines Tonträgers mit

1,5 mm, insbesondere 1,2 mm, das heißt, der Wulstring sollte eine Höhe von 0,2 bis 0,3 mm, insbesondere von 0,3 mm, und eine Breite von ca. 1,1−1,5 mm, insbesondere 1,2 mm haben. Zwei Anordnungen wurden zum Durchbruch am unteren Rand des Wulstrings mit diesem Wulstring ausgestattet.

noch üblicherweise am Rand des größeren Wulstrings praktisch in der gleichen Länge und Form anliegen. Durchbrüche am unteren Rand des Wulstrings sind die erleichtern auch das Justieren der Füße des Tonträgers, nicht den größeren Wulstring praktisch in der gleichen Länge und Form anliegen. Durchbrüche am unteren Rand des Wulstrings sind die erleichtern auch das Justieren der Füße des Tonträgers, nicht den größeren Wulstring praktisch in der gleichen Länge und Form anliegen.

This is important as the sound reproducing device has to be actuated frequently also

To meet this object, the invention provides that in the sound carrier mentioned initially at

The walls of the enlarged adjustment perforations should have the same

If the two perforations in the lower edge of the sound carrier are used in this manner, which passes the sound reproducing device placed in position into the adjustment perforation/perforations in this position. A plane guiding funnel is formed in the exactly defined positions in which they would also be if all the perforations were, as the aligning feet to be inserted. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet.

as an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the outside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.

As an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the inside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.

Durchbrechungen eines Tonträgers mit hindernd auch das unbeabsichtigte Verhindern der Durchbrüche der Wiedergabe beim Studium von Lerntexten

Die Wände der vergrößerten Justierdurchbrechungen sollten auf der den 6 mm haben. Die vergrößerten Durchbrechungen

As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet.

This is important as the sound reproducing device has to be actuated frequently also

To meet this object, the invention provides that in the sound carrier mentioned initially at

The walls of the enlarged adjustment perforations should have the same

If the two perforations in the lower edge of the sound carrier are used in this manner, which passes the sound reproducing device placed in position into the adjustment perforation/perforations in this position. A plane guiding funnel is formed in the exactly defined positions in which they would also be if all the perforations were, as the aligning feet to be inserted. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet.

As an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the outside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.

As an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the outside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.

This is important as the sound reproducing device has to be actuated frequently also

To meet this object, the invention provides that in the sound carrier mentioned initially at

The walls of the enlarged adjustment perforations should have the same

If the two perforations in the lower edge of the sound carrier are used in this manner, which passes the sound reproducing device placed in position into the adjustment perforation/perforations in this position. A plane guiding funnel is formed in the exactly defined positions in which they would also be if all the perforations were, as the aligning feet to be inserted. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet. As a result, when centered, the feet are in each case nearly defined exactly by the aligned feet.

As an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the inside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.

As an alternative to this solution, the two lower adjustment perforations may be widened and elongated in the direction away from the other toward the outside or in the direction of the same toward the respective aligning tool to be inserted into them with large circular regions created in such a manner, so that said aligning tool has a narrow region and the other aligning foot is in a position to engage other adjustment perforations or perforations in this position.
2. The sound carrier according to claim 1, wherein each annular bead surrounds the adjustment perforation at a distance of about 1-1.5 mm leaving a shoulder between the perforation and the bead.

7. The sound carrier according to claim 6, wherein the coating is a color coating.

2. Tonträger nach Anspruch 1, dadurch gekennzeichnet, daß jeder Wulstring (10) die Justierdurchbrechung mit geringem Abstand von etwa 1-1,5 mm, insb. 1,2 mm, einen Absatz belassend, umgibt.

7. Tonträger nach Anspruch 6, dadurch gekennzeichnet, daß sich jede verbreiterte und verlängerte Justierdurchbrechung (8) vom breiten Bereich (13) zum engen Bereich (12) hin.
Big models: bitexts

Matching Chinese description/claims with US
~ 64 Million segments (en zh)
English size: 2’000 Million words, 10Gb

(more careful alignment/cleaning to be done in the future)
Big models: language models

English texts only ~ 1Tb (US+EPO+PCT descriptions/claims/titles/abstracts)
…all English Wikipedia is 44Gb

Currently we stick to 10Gb for the language model
Tapta and big models

Parallelization:
- Use mgiza
- For big models: split corpus in 4 parts, launch mgiza on each quarter
- For big models: stop at HMM iteration

Heavy compression
- Pruning
- Binarization (compact phrase table, kenlm)
- (without much loss in quality)
## Training and scalability

### Size reduction zh-en

<table>
<thead>
<tr>
<th></th>
<th>Phrase table 0-0</th>
<th>Phrase table 0,1-0</th>
<th>Reordering model</th>
<th>Language model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M rows</td>
<td>Gb</td>
<td>M rows</td>
<td>Gb</td>
</tr>
<tr>
<td>Basic</td>
<td>806</td>
<td>100</td>
<td>974</td>
<td>130</td>
</tr>
<tr>
<td>Pruned</td>
<td>551</td>
<td>69</td>
<td>623</td>
<td>83</td>
</tr>
<tr>
<td>Binarized</td>
<td>6.4</td>
<td>7.4</td>
<td>4.2</td>
<td>4.6</td>
</tr>
</tbody>
</table>

342 Gb

22.6 Gb (6.6%)
Language specificities

Tokenizer: Based on Lucene framework
• zh: adapted “SmartCn”
• ja: “kuromoji”
• de: decompounder (Junczys & Pouliquen, Eamt2014)
• ar: prefix splitting, removes shot vowels
• ko: decompounder
• Normalizes greek letters
• Groups references to figures (eg. “(1)” not “(_1_)_”)

Reordering
• de: pre-reordering (Junczys & Pouliquen, Eamt2014)
• ja: Simple naïve pre-reordering (more to be done)
Tapta: our tool installed in different places

And some projects going on to install Tapta prototype in other International institutions…
Our tool in different situations

• Adapted our code so that it can easily install and run
• Under version control, regression tests, Installation/administration documentation (100 pages)
• With installation instructions: ½ day to configure a new Linux server
• Runs on Linux:
  • Hardware: Amazon cloud, virtual machine, desktop PC, server
  • OS: Ubuntu/Suse/Centos/RedHat
Training and scalability: UN data

- All United Nation texts ~ 212 Million words, 10 M segments

<table>
<thead>
<tr>
<th></th>
<th>Phrase table</th>
<th>Reordering model</th>
<th>Language model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M rows</td>
<td>Gb</td>
<td>M rows</td>
</tr>
<tr>
<td>Basic</td>
<td>82</td>
<td>9.70</td>
<td>82</td>
</tr>
<tr>
<td>Pruned</td>
<td>19</td>
<td>2.20</td>
<td>19</td>
</tr>
<tr>
<td>Binarized</td>
<td>0.27</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

UN data

20Gb
1.12Gb (6%)
Our tool in production

- Install/publish/update/train/evaluate (robust) scripts
- Monitor tool
- Dashboard interface
- Anti robot policy (captcha)
- Statistics
- …
Hardware & OS

- Virtual, 4 Gb Ram, Suse SLES11, 4 cores, 250Gb
- Virtual, 16 Gb Ram, RedHat ent R6.2, 16 cores, 200Gb
- PC, 8Gb ram, Ubuntu 12.4, 8 cores, 350Gb
- Server, 16Gb, Centos R6.4, 16 cores, 400Gb
- Server, 11Gb Ram, RedHat ent. R6.5, 8 cores, 100Gb disk
- ... 
- Amazon cloud, 64 Gb Ram, Suse ent. 11, 8 cores, 400Gb
- ... 
- Server, 500Gb ram, RedHat ent. R6.5, 48 cores, 4T
Various user interfaces

- Java Swing
- Web interface
  - Gist translation
  - Interactive translation
- Hotkey
- Plugin (SDL studio, Worldserver, eLuna etc.)
- Tapta widget
User can specify the language pair (or let the system choose)
The system can “guess” the domain from the text, or the user can specify
This automatic translation is provided for information only, it may contain discrepancies or mistakes and does not have any juridical value.

- Please select segments in source text (with mouse or use "shift" and arrow keys)
- You can then select among the proposals
- Special keys: <escape> to undo (or press "undo" button) use CTRL to select non-contiguous segments

Segment: for organic electronic devices and photovoltaic cells.

Proposals: pour dispositifs électroniques organiques et des cellules photovoltaïques.
- pour dispositifs électroniques organiques et des cellules photovoltaïques.
- pour dispositifs électroniques organiques et des cellules photovoltaïques.

Source text:
Polymers which can be used in p-type materials for organic electronic devices and photovoltaic cells. Compounds, monomers, dimers, trimers, and polymers comprising formula (I) and/or formula (VIII) are prepared

Translated text:
Polymères qui peuvent être utilisés dans des matériaux de type p pour dispositifs électroniques organiques et des cellules photovoltaïques.
Tapta hotkey

- Access the translation server using the "F3" key: Select text, press F3, translation goes to clipboard
- Work only on PC (opensource AutoHotKey), but is a solution to integrate MT in any application
Plugins

SDL/Studio

Worldserver

Eluna
(United Nation internal CAT tool)

Multitrans?
Tapta widget

A script to be inserted at the beginning of any HTML page, translates inline text on the fly.
Tapta: translation quality

• Competitive!

• Better than Google and Microsoft translate
  • Working with our data
  • Based on good open source “Moses”

• Small team…
  but working with others…
Also the United Nations

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Tapta</th>
<th>Google</th>
<th>Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar-en</td>
<td>55.25</td>
<td>n/a</td>
<td>51.17</td>
</tr>
<tr>
<td>en-ar</td>
<td>44.10</td>
<td>33.74</td>
<td>28.94</td>
</tr>
<tr>
<td>en-es</td>
<td>61.81</td>
<td>53.39</td>
<td>46.86</td>
</tr>
<tr>
<td>en-fr</td>
<td>51.23</td>
<td>45.58</td>
<td>42.19</td>
</tr>
<tr>
<td>en-ru</td>
<td>50.85</td>
<td>39.67</td>
<td>38.96</td>
</tr>
<tr>
<td>en-zh</td>
<td>43.17</td>
<td>34.16</td>
<td>32.77</td>
</tr>
<tr>
<td>es-en</td>
<td>60.32</td>
<td>52.54</td>
<td>49.18</td>
</tr>
<tr>
<td>fr-en</td>
<td>53.36</td>
<td>46.46</td>
<td>43.39</td>
</tr>
<tr>
<td>ru-en</td>
<td>58.56</td>
<td>47.71</td>
<td>47.09</td>
</tr>
<tr>
<td>zh-en</td>
<td>42.31</td>
<td>36.55</td>
<td>30.60</td>
</tr>
</tbody>
</table>

The advisory services of the international committee of the red cross expressed its gratitude to all those who have contributed with organizations in drafting the manual, which was the product of intensive teamwork.

Bruno Prouilquin, Cecilia Elizalde, Marcin Junczys-Dowmunt, Christophe Mazenc, José García-Verdugo, **Large-scale multiple language translation accelerator at the United Nations**, MT Summit, Nice, France, September 2013
User acceptance

How Tapta is perceived among translators:
• When seen as a “translation accelerator”: very useful
• When seen as “replacement for translator”: useless
• When proposed as a copy-paste tool: not used
• When integrated in translator’s environment: used

Frustration: User has little impact on the MT output
Blacklist that we apply on the phrase table
Collect post-edition segments:
• quality estimation
• improving the MT
Conclusion/discussion

• MT contributes to information dissemination
• Moses easily supports huge models
• Tapta MT quality competitive
• Language dependent tools should be avoided in our context
• “User acceptance landscape is changing”
• Integration!
Future work on transliteration

- Continue improving quality/speed/costs
- Interactive translation (autosuggest? Casmacat?)
- Incremental
Future work

• Continue improving quality/speed/costs
  • OSM
  • Word cluster LM
  • Additional usage of meta data
• Interactive translation (autosuggest?)
• Incremental training
• Translation through pivot language
Future work

User feedbacks

• Take into account new translations
• Blacklist of phrases
• Collect post-editions
• …
Order differences complicate phrase extraction.

Reordering is helpful. E.g. English-to-Japanese

<table>
<thead>
<tr>
<th>Original</th>
<th>1</th>
<th>go</th>
<th>to</th>
<th>Tokyo</th>
<th>and</th>
<th>Kyoto</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reordered</td>
<td>1</td>
<td>go</td>
<td>to</td>
<td>Tokyo</td>
<td>and</td>
<td>Kyoto</td>
<td>.</td>
</tr>
</tbody>
</table>

I go to Tokyo and Kyoto.

私が東京と京都へ行く。
Thank you for your attention

شكرا لكم على اهتمامكم

Merci pour votre attention!

感谢您的关注

Grazie per la vostra attenzione!

Vielen Dank für Ihre Aufmerksamkeit!

Obrigado pela vossa atenção!

Dank je voor uw aandacht!

Dank u voor uw aandacht!

Dank je voor uw aandacht!

Thank you for your attention!


(2012) Marcin Junczys-Dowmunt : A Phrase Table without Phrases: Rank Encoding for Better Phrase Table Compression. *(EAMT 2012)*

(2011) Bruno Pouliquen & Christophe Mazenc: COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. *(MT Summit XIII)*.


**Tapta4UN:**

