WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Tapta:
A user-driven translation system for patent documents based on domain-aware Statistical Machine Translation

EAMT 2011

30/05/2011

Bruno Pouliquen, Christophe Mazenc, Aldo Iorio
WIPO (World Intellectual Property Organization)

Menu

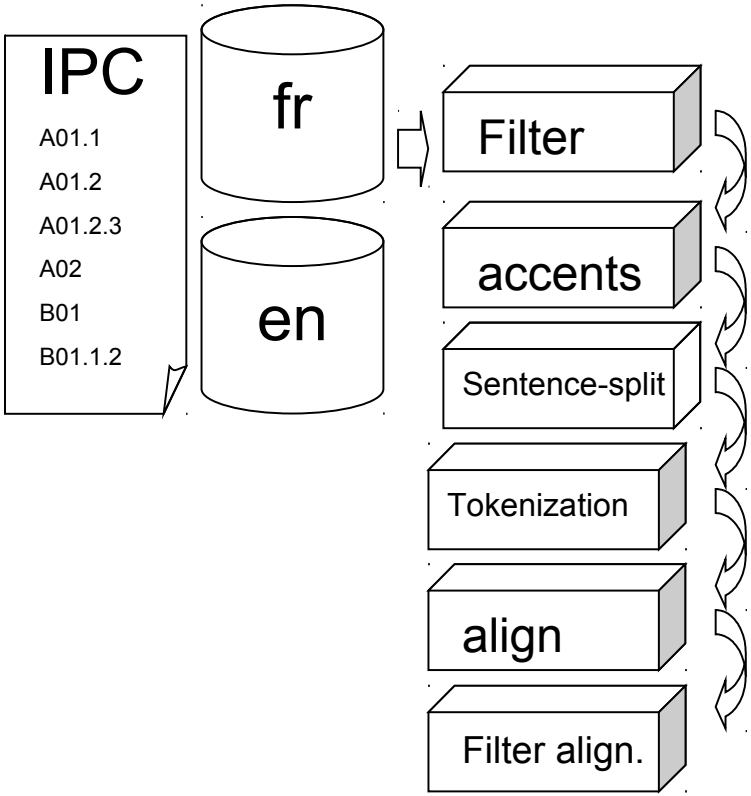
- Introduction
- Training our SMT
  - English French parallel corpus
  - SMT training
  - Optimization
- Translating GUI
  - Server
  - TAPTA GUI
- Results/Evaluation
  - Automatic evaluation
  - Human evaluation
- Conclusion

Introduction

- WIPO is a UN agency in charge of Intellectual Property
  - Patent Cooperation Treaty => facilitate patent protection in multiple countries
  - WIPO receives ~ 150,000 patent applications/year
  - Titles and abstracts must be available in English and French
- WIPO has a corpus of 1.8 M en-fr docs (> 8M segments)

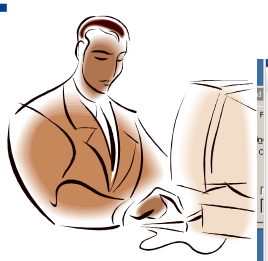
- Needs in CAT tools + huge parallel corpus => SMT
- **T**ranslation **A**ssistant for **P**atent **T**itles and **A**bstracts

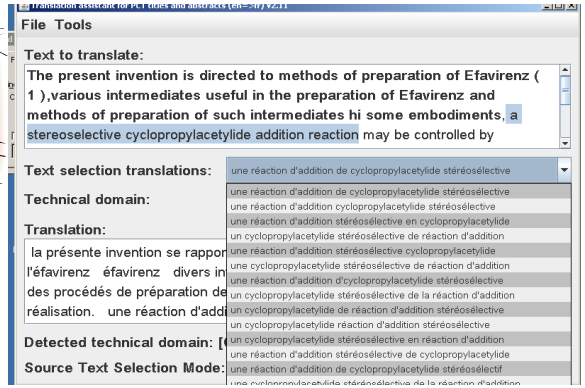
Our SMT



en	fr	domain
antimicrobial coatings	enrobages antimicrobiens	CHEM
ternary mixed ethers	éthers mixtes ternaires	CHEM
submarine	sous – marin	MARI
automatic translation	traduction automatique	DATA
automatic translation	translation automatique	BLDG

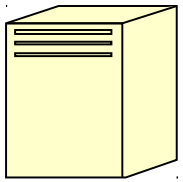
Parallel segments





Tapta client

Tapta server



Moses decoder

Moses decoder

Moses decoder

phrase table

language model

...

Moses' train model

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Training our SMT: corpus

- Previously translated patent applications (1,800,788 documents)
- Title+abstract
- Additional valuable information:
  - Classification (IPC)
  - Language of publication
  - Quality control passed
- **8’352’768** aligned pairs of segments (high quality)

Tokenization / sentence-splitting / Alignment

a·pesticide·and/or·a· fertilizer·and·the·like·scattered·or· sprayed·onto·a·golf·course·(1)·,· together·with·most·of·the·sprayed· water·or·rain· water·,·are· permated·into· a·turf· (15A)·,· a·thick·sand· layer·(15)·having· high·water·permeability·,·whereby· only·the·rain· water·filtered ·by·the· golf·course·is· discharged·to·the· outside

un·pesticide·et/ou·un·fertilisant·,· ou·un·produit·similaire·,· disperse· ou·pulverise·sur·un· parcours·de· golf·(1)·,·avec·la· majeure·partie· de·l'·eau· pulverisee·ou·de·l'·eau·de· pluie ,·qui·sont· impregnes·dans· un· gazon·(15A)·,·une·epaisse· couche·de·sable·(15)·ayant·une · forte·permeabilite·a·l'· eau·.·Selon· ce·procede·seule·l'·eau·de·pluie· filtree·par·le· sol·du·parcours·de· golf·se· deverse·a·l'·exterieur

SMT Training

- Using open-source Moses
- Training: > **8 Million segments**
- Small test set: 102 segments
- Big test set: newly published documents (9521 segments)

“domain-aware” SMT

- Text translated according to the technical domain
- Map IPC code to one of our 32 *domains*

[ADMN] Admin, Business, Management & Soc Sci
[AERO] Aeronautics & Aerospace Engineering
[AGRI] Agriculture, Fisheries & Forestry
[AUDV] Audio, Audiovisual, Image & Video Tech
[AUTO] Automotive & Road Vehicle Engineering
[BLDG] Civil Engineering & Building Construction
[CHEM] Chemical & Materials Technology
[DATA] Computer Sci, Telecom & Broadcasting
[ELEC] Electrical Engineering & Electronics
[ENGY] Energy, Fuels & Heat Transfer Eng
[ENVR] Environmental & Safety Engineering
[FOOD] Foods & Food Technology
[GENR] Generalities, Language, Media & Info Sci
[HOME] Home Contents & Household Maintenance
[HORO] Precision Mechanics, Jewelry & Horology
[MANU] Manufacturing & Materials Handling Tech

Domain information as *factor, eg:*

in|U the|U robotic|1 prosthesis|1 alignment|1 device|U provides|U automatic|1 translation|1 in|U two|U axes|1

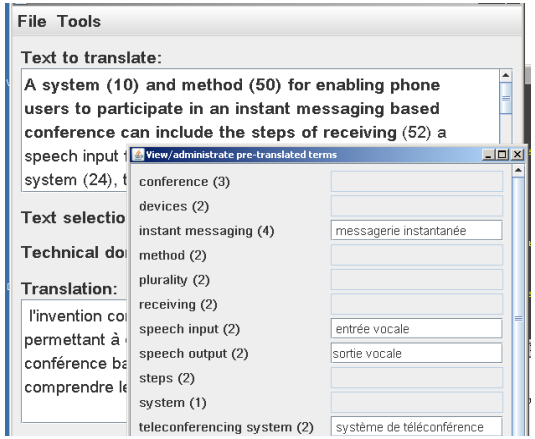
a|U chinese|2 to|U english|2 automatic|2 translation|2 method|U

[SCIE] Optical Engineering
[SPRT] Sports, Leisure, Tourism & Hospitality Ind
[TEXT] Textile & Clothing Industries
[TRAN] Transportation

SMT… more…

- Language model: irstlm (5 grams), kenlm, binarization
- Pruning of the phrase table, binarized (102M entries x 2)
- Improve the quality:
  - Get only segments out of QC-passed documents
  - Add segments that were originally in French
  - Create a sub-model (phrase tables)
  - Merge this sub-model with original (75-25%)

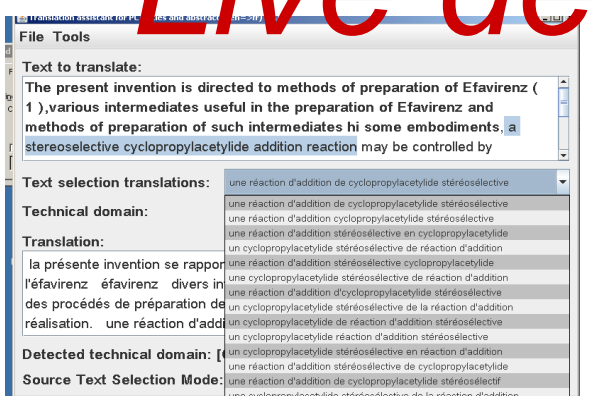
Graphical User Interface



traduction texte	5[qc][qc]
traduction textes	2
conversion texte	1
de traduction textes	1
transposition texte	1

traduction texte	WO2003071774	DATA	EN	...calls with optional voice to **text translation**	...de conférences téléphoniques avec une **traduction** facultative voix - **texte**
	WO2009036800	DATA	EN	...relates to a method of **text translation** comprising the steps of	...invention concerne un procédé de **traduction** de **texte** qui consiste à
	WO2009036800	DATA	EN	...electronic device capable of such **text translation** and a computer program comprising software instructions that , when executed , performs such a text ...	...électronique capable d' une telle **traduction** de **texte** et un programme informatique ...
	WO2003052624	DATA	EN	...voice to text conversion and/or **text translation** and/or the text to voice conversion ...	...en texte et/ou de la **traduction** de **texte** et/ou de la conversion ...
	WO2007045136	DATA	ZH	...multi - language speech and **text translation** server , the user terminal device consists of at least a first transmitting / receiving device and at least a second transmitting / receiving device ...	...contrôle et un serveur de **traduction** de paroles et de **texte** en plusieurs langues , ...
traduction textes	WO2002086737	DATA	EN	...user - directed web site **text translation**	...pour une fonction integree de **traduction** de **textes** de sites web orientée
	WO[illegible]4059712	DATA	[illegible]	...the invention is directed to **text translation** tools that are especially useful for translation of related electronic documents ...	...invention concerne des outils de **traduction** de **textes** qui sont spécialement utiles ...

Live demo???



WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Evaluation: automatic

Experiment	Speed Seconds (#docs)	BLEU score
Baseline	542 (51)	**40.11**
Pruned	164 (51)	41.11
Pruned+d=0.15+s=10	22 (51)	43.30
Pruned+BigLm	196 (51)	**51.71**
Pruned+BigLm w/oDomain	173 (51)	50.91
Only Qc	N/a (51)	28.29
Google translate	N/a (51)	34.09
Bing translator	N/a (51)	27.17
Pruned+BigLm	27140 (1390)	45.06
Pruned+BigLm w/oDomain	6934 (1390)	44.99
Pruned+BigLm_Qc_d=0.15 s=10	6378 (1390)	**45.07**

Reasonable speed: 5 sec/doc (without parallelization)

Real test: translate latest published documents

Evaluation: human

- 12 testers
  - Good French and English skills
  - Not translators
- Coached by 2 professional translator-revisers
- 516 translations produced during 13 days

	Translations produced	☺ Publishable		☹ Non Publishable	
With coaching (10 days)	403	243	**60.3%**	160	39.7%
Without coaching (3 days)	113	56	**49.6%**	57	50.4%

~ half of translation produced by non translators passed a high quality QC

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Second evaluation: compare with outsourcing agencies

Translations	translations QCed	Tapta		Agency		Same quality	Tapta better	Agency better
		P	**NP**	P	**NP**			
With coaching	388	350	38	344	44	111	193	84
Without coaching	112	103	9	98	14	23	60	29
Total	500	453	47	442	58	134	**253**	**113**
%	**100**	**90.6**	**9.4**	**88.4**	**11.6**	**26.8**	**50.6**	**22.6**

Conclusion

- SMT with reasonable results
- User-driven translation was successful
- Tapta was not adopted by WIPO professional translators
- But:
  - Adding Tapta in the workflow is investigated
  - Was judged as very valuable for non-translators
  - Accelerated training aid

Future work

- Other language pairs : Chinese, Korean, Japanese …
- Triangulation
- Incremental training

- Web Tapta: http://www.wipo.int/patentscope/translate/