Recent developments in the TransSearch bilingual concordancer

Guy Lapalme
RALI - Université de Montréal

joint work with
Fabrizio Gotti, Philippe Langlais, Julien Bourdaillet, Stéphane Huet, Jacques Steinlin

Industrial contributor: Terminotix
RALI's personnel includes computer scientists and linguists with considerable experience in natural language processing. It is the largest university NLP laboratory in Canada.

Research Projects

- Translation
- Information extraction
- Automatic summarization
- Information retrieval
- Judicial texts processing
- Environmental information

System demos

- Translation
  - TransSearch: bilingual concordancer
  - TransType: animation of an interactive translation session
- Reacc: automatic French accentuation
- SILC: language and coding detection
- Lexiquum: Québec text concordancer

Information

- Publications
- Textual Resources

Contacts

- Members
- Collaborators

RALI OLST-MITACS seminars (mostly in French)

- Courses by professors of the RALI
  - Fall 2011: IFT6810: Traitement Statistique des Langues Naturelles
  - Fall 2011: IFT3335: Intelligence Artificielle

RALI in the press

A selection of international conferences in NLP
RALI - Today

• 3 professors
  - Philippe Langlais : machine translation
  - Jian-Yun Nie : information retrieval
  - Guy Lapalme : summarization, generation, info extraction

• Adjunct professor : Atefeh Farzindar

• Students
  - 1 post-doc
  - 8 Ph.D.
  - 5 M.Sc.

• 1 Research associate
  - Fabrizio Gotti
Goals of RALI

- Applied NLP using symbolic and statistical methods
  - Tools for translators
  - Writing aids

- Collaborative work
  - Linguistics department
  - Industrial partners
  - Governmental partners
They can be taken for a ride.

Look what happens when they are taken for a ride.

It becomes too cumbersome and we think nobody is taking us for a ride.

They take it for a ride and leave it somewhere else.

They had been approached by people who they considered friends and trusted and they were taken for a ride.

That is why it is important to have full access to information about revenue if we do not want to be taken for a ride.
Timeline of TransSearch

1992  Michel Simard (CITI) on Sun machines
1995  First Web version (free)
1997  Web at RALI
1999  Paying subscribers - TSRali.com
2003  Transfer to Terminotix
2008  NSERC - CRD project for new version
2011  TransSearch with translation spotting in beta

Julien Bourdaillot, Stéphane Huet, Philippe Langlais and Guy Lapalme. 
<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emissions continue to rise and taxpayers are being taken along for the ride.</td>
<td>Les émissions continuent d'augmenter et c'est le contribuable qui est le dindon de la farce.</td>
</tr>
<tr>
<td>They are left with nothing. Now they are here illegally with no documentation. Canadians are being taken for a ride.</td>
<td>Ces personnes se trouvent ici illégalement, elles n'ont aucun document et nous, les Canadiens, sommes les dindons de la farce.</td>
</tr>
<tr>
<td>This would affect close to 400,000 Canadians, 80,000 of them Quebecers, who have been the ones taken for a ride.</td>
<td>Il s'agit d'une mesure qui toucherait près de 400 000 Canadiens, dont 80 000 Québécois, qui ont été les dindons de la farce.</td>
</tr>
<tr>
<td>I think that this is a prime example of a tainted system in which people who cannot afford to invest in sectors eligible for tax credits are urged to do so through all kinds of scams and end up being taken for a ride.</td>
<td>Je pense que c'est un exemple patent d'un système vicié, où des gens qui n'ont pas les moyens d'investir dans des domaines où on peut obtenir des crédits d'impôt se voient, par toutes sortes de subterfuges, invités à le faire et, en bout de ligne, ils se trouvent à être les dindons de la farce.</td>
</tr>
</tbody>
</table>
92 traductions de *take+ .. ride* dans 106 occurrences

<table>
<thead>
<tr>
<th>Expression</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>take+ .. ride</td>
<td></td>
</tr>
</tbody>
</table>

**monté un bateau**

Mr. Yvan Loubier (Saint-Hyacinthe-Bagot, BQ): Mr. Speaker, the Minister of Finance is *taking us for a ride*.

Ordinary Canadians are being *taken for a ride* by the same two old line parties which represent the corporate interest, the bottom line of this country, no matter what is going on.

It concerns the Meech Lake Accord. This week the Canadian Ethnocultural Council, the national umbrella organization representing multicultural associations across the country, had the following to say about the Accord. "Our worst suspicions about the process of the hearings were confirmed as the report merely rubberstamped the Accord...while we were *taken for a ride*...we cannot sit back and watch multiculturalism being set back for years".

M. Yvan Loubier (Saint-Hyacinthe-Bagot, BQ): Monsieur le Président, le ministre des Finances nous monte un bateau.

Les Canadiens ordinaires se font *monter un bateau* par les deux vieux partis qui représentent les grandes sociétés qui ont toujours gain de cause, quelles que soient les circonstances.

Le Conseil ethnoculturel du Canada, l'organisation-cadre nationale représentant des associations multiculturelles de tout le pays, a réagi cette semaine en ces termes au rapport du comité chargé d'étudier l'Entente: «Nos pires soupçons au sujet des audiences se sont trouvés confirmés étant donné que le rapport n'a fait qu'entériner l'Entente... Ou nous a *monté un bateau*... Nous ne pouvons voir sans protester le multiculturalisme reculer de plusieurs années.»
# TransBases (1986-2011)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>docs</th>
<th>K pairs</th>
<th>M words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansard</td>
<td>2 854</td>
<td>10 500,0</td>
<td>339,6</td>
</tr>
<tr>
<td>Senate</td>
<td>1 025</td>
<td>1 200,0</td>
<td>47,6</td>
</tr>
<tr>
<td>Canadian courts</td>
<td>13 068</td>
<td>3 200,0</td>
<td>142,6</td>
</tr>
<tr>
<td>Meteorological alerts</td>
<td>2 398</td>
<td>7,2</td>
<td>0,2</td>
</tr>
<tr>
<td></td>
<td>19 345</td>
<td>14 907,2</td>
<td>530,0</td>
</tr>
</tbody>
</table>
TransBase Creation

• Finding texts and their translation
• Cleaning and normalization
• Sentence alignment
• Indexing

• Web based querying
Query:A

Current
TransSearch

Translation Memory

Transpotting

Filtering

New
TransSearch

Merging

Display

Relevance feedback

x
y
x
z

x
y
x

x
y

x
y

x
y

x
y

x
y

x
y

x
y

x
y

x
y

Transpotting

Filtering

Merging

Display

Relevance feedback
Transpotting

- Word alignment combining HMM and IBM models
  - combine alignments from both directions
  - add a contiguity constraint
- Phrase-based transpotting
  - use the phrase table produced by MOSES

This is in keeping with that strategy.
La présente mesure est conforme à cette stratégie.
Postprocessing

on behalf of

<table>
<thead>
<tr>
<th>Transpot</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>au nom de</td>
<td>1424</td>
</tr>
<tr>
<td>au nom du</td>
<td>763</td>
</tr>
<tr>
<td>au nom des</td>
<td>683</td>
</tr>
<tr>
<td>de*</td>
<td>136</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>dans l’intérêt des</td>
<td>15</td>
</tr>
<tr>
<td>de la part de</td>
<td>13</td>
</tr>
<tr>
<td>dans*</td>
<td>13</td>
</tr>
<tr>
<td>part de</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>pour l’ensemble de</td>
<td>1</td>
</tr>
<tr>
<td>parler pour</td>
<td>1</td>
</tr>
<tr>
<td>loin que*</td>
<td>1</td>
</tr>
<tr>
<td>le bien*</td>
<td>1</td>
</tr>
</tbody>
</table>
Postprocessing

• Must deal with transpotting errors
  - between 20% and 40%
• User will focus on the first ten transpots
  - as many correct
  - as many diversified
Postprocessing

• Filtering bad transpots using supervised classification

• Merging variants
  - inflectional form of canonical words
  - ignore some grammatical words
  - start merging by considering the most frequent ones

• Pseudo-relevance feedback
  - search for more frequent transpots in sentences with less frequent ones
  - develop a new transfer table from this specialized corpus and find new alignments
Query:A

Translation Memory

Current TransSearch

Transpotting

Filtering

New TransSearch

Merging

Display

x 2
y 1
Evaluation corpora

• Hansard (8.3 million Fr-En sentence pairs)
• 5000 most frequent English queries to TransSearch between 2001 and 2007
• 5000 sentence pairs for each query
• Bilingual phrase lexicon for testing
  - 2,358 entries (284 DEV, 2,074 TEST)
  - average number of translations (3.6 DEV, 3.9 TEST)
Evaluation process

- **Transpotting**
  - only the ones with the source containing the query and the target containing a translation from the dictionary (180 000 DEV, 1.4M TEST)

- **Translation**
  - comparison with bilingual phrase lexicon

- **Classification**
  - hand annotation of 531 queries (12 144 examples)
Results

• Two best methods: PBM and C–HMM–bi
• Benefits from filtering and merging
• Limited impact of pseudo-relevance feedback

• But in the production version of TS3
  - at most 900 results for each query
  - transpotting
    • IBM-bi
  - filtering of transpots
    • 75% of grammatical words
    • only auxiliairies or numbers
Non web uses of TransSearch

- Web service
  - application embedding
  - call from a text processor (e.g. Word)
- Educational use
- Integration à la TransType
Other bilingual concordancers

• With transpotting
  - TotalRecall (Wu, 2003)
    • Chinese-English word alignment
  - LinearB (Callison-Burch, 2005)
    • basic translation spotting
  - Linguee.com
    • many language pairs of web text (free with ads)

• Without transpotting
  - Opus
    • multiple parallel corpora
  - WeBiText (CNRC/Terminotix)
    • meta crawler for bilingual texts
Why is TransSearch so popular?

- Complements a terminology bank
- Not a bilingual dictionary
  - 87% of queries have more than one word
- Ready-made solution to translation problems
  - idiomatic expressions
  - non compositional translations
  - non literal translations
Questions ?

7 traductions de *questions from the floor* dans 7 occurrences

<table>
<thead>
<tr>
<th>French Expression</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>questions à partir du parquet</td>
<td>1</td>
</tr>
<tr>
<td>répondu à leurs questions</td>
<td>1</td>
</tr>
<tr>
<td>les députés lui posaient des questions</td>
<td>1</td>
</tr>
<tr>
<td>questions</td>
<td>1</td>
</tr>
<tr>
<td>poser des questions</td>
<td>1</td>
</tr>
<tr>
<td>questions des spectateurs</td>
<td>1</td>
</tr>
<tr>
<td>l'auditoire à poser des questions</td>
<td>1</td>
</tr>
</tbody>
</table>

Each group gave presentations which lasted between five and ten minutes, and at the end of all the presentations, I invited *questions from the floor.*

Chaque groupe a fait une allocution variant entre cinq et dix minutes et, à la fin de tous les exposés, j'ai invité *l'auditoire à poser des questions.*
Retire later to live longer

Early retirement, as is commonly believed, does not help retirees to live longer and it may even shorten one's life. This is the conclusion of a study published on October 21 by the British Medical Journal. The research involved tracking more than 3,500 employees working for Shell Oil in Texas over a 26-year period. The workers retired at 55, 60 or 65 and were monitored to see what effect their age at retirement had on their lifespan. Researchers considered factors such as gender, socio-economic status in ascertaining whether retiring early is associated with better survival. It seems the findings have displaced the myth that spending our golden years at a leisurely pace away from the daily grind of the nine to five will increase our longevity. It appears that retiring late provides for a longer life.

The results were astonishing. The life expectancy of employees who retired at 55 was significantly reduced compared with those who retired at the age of 65. The researchers conclude that: “Retiring early at 55 or 60 was not associated with better survival than retiring at 65 a cohort of past employees of the petrochemical industry. Mortality was higher in employees who retired at 55 than in those who continued working.” Leader of the research team Shan Tsai said, “Although some workers retired at 55 because of failing health, these results clearly show that early retirement is not associated with increased survival. On the contrary, mortality improved with increasing age at retirement for people from both high and low socio-economic groups.”
Product overview

The machine is controlled from a liquid crystal color touch screen where you can view your operation and application settings.

Printer settings can be pre-programmed for specific production job types and when such a job type is selected, the printer is set up automatically for the paper type and application.

Aperçu de la machine:

La machine est contrôlée à partir d'un écran tactile à cristaux liquides vous permettant de visualiser vos paramètres d'application et d'opération.

Les paramètres de l'imprimante peuvent être préprogrammés pour des types de type de travail et lorsqu'il est sélectionné