University and Industry Partnership in NLP, is it worth the “trouble”?

Guy Lapalme
Université de Montréal
This talk

• is
  - a personal account of experience
  - limited to Natural Language Processing grants related industry in Université de Montréal
  - a team effort
  - vague on the amounts of money

• is not
  - big (or even small) science
  - applicable in every case
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
- Reac: automatic French accentuation
- SILC: language and coding detection
- Lexiquim: Québec text concordancer

**Information**
- Publications
- Textual Resources
- Contacts
- Members
- Collaborators

- RALI OLST-MITACS seminars (mostly in French)
- Courses by professors of the RALI
  - Fall 2009: IFT3335: Intelligence Artificielle
  - Fall 2009: IFT6810: Traitement Statistique des Langues Naturelles
- RALI in the press

A selection of international conferences in NLP
NLP at Université de Montréal

- Incognito (Informatique cognitive) (1984-1997)
  - Expert Systems
  - Information Retrieval
  - NLP projects
    - Spelling checkers
    - Dictionary editing of Meaning-Text Theory
    - Deterministic parsers
    - Text generation
Birth of the RALI

- Centre for Information Technologies Innovation (CITI) at Laval
  - Federal Government budget cuts in 1997
  - Almost privatized
  - Two teams were saved
    - Machine-Translation group (Pierre Isabelle et al.)
    - Computer tools for the handicapped
RALI - Today

- 3 professors
  - Philippe Langlais: machine translation
  - Jian-Yun Nie: information retrieval
  - Guy Lapalme: summarization, IE, etc...
- Adjunct professor: Atefeh Farzindar
- Students
  - 4 post-docs
  - 7 Ph.D.
  - 5 M.Sc.
- 1 Research associate
  - Fabrizio Gotti
Industrial collaboration
Expected benefits for industry

• Access to a rich talent pool
• Access to specific expertise
• Cost sharing using partnership governmental programs
Industrial collaboration

Expected benefits for universities

- Dealing with realistic problems
- Inspiration for new research
- Motivation for students
- Alternative funding opportunities
Types of industrial contributions

• Contracts (university overhead: 40%)
  – short term
• Licenses (uo: 40%)
  – existing technologies
• Grants (uo: 15%)
  – longer term, students involvement
• Partnership program (uo: 15% of indus. contrib.)
  – Governmental agencies (NSERC, Precarn, MITACS, Industry Canada) match an industrial contribution
• Networks of Centres of Excellence / CFI
Types of industrial contributions

• Contracts (university overhead: 40%)
  - short term
• Licenses (uo: 40%)
  - existing technologies
• Grants (uo: 15%)
  - longer term, students involvement
• Partnership program (uo: 15% of indus. contrib.)
  - Governmental agencies (NSERC, Precarn, MITACS, Industry Canada) match an industrial contribution
• Networks of Centres of Excellence / CFI

There are no known gold mine IT licenses

J Strother Moore (University of Texas, Austin)
SILC (Système d'Identification de la Langue et du Codage) automatically determines both the language in which a document is written, and the character encoding used. The current version recognizes close to thirty languages, and an average of three encodings per language.

It is now possible to integrate SILC in your application under the following systems: Windows, Linux, Solaris, MAC, HP-UX, AIX and SGI. SILC also exists in Java.

Liste of known languages and encodings

Type some text in the box below, in the language of your choice. The system needs at least a few words to be able to produce a reliable identification.

English cp1252
Chinese utf8
Japanese utf8
Spanish cp1252
German cp1252
Korean utf8
French cp1252
Italian cp1252
Portuguese cp1252
Dutch cp1252

Ein Unternehmen zu gründen, geht nicht von heute auf morgen, sondern ist ein Prozess. Eine geniale Idee alleine reicht dabei nicht aus, um erfolgreich zu sein.

Submit text
Ein Unternehmen zu gründen, geht nicht von heute auf morgen, sondern ist ein Prozess. Eine geniale Idee alleine reicht dabei nicht aus, um erfolgreich zu sein.

The language is German, the encoding is utf8

Ranking of candidates

<table>
<thead>
<tr>
<th>Language</th>
<th>Encoding</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>utf8</td>
<td>3.8506808</td>
</tr>
<tr>
<td>German</td>
<td>cp1252</td>
<td>4.559468</td>
</tr>
<tr>
<td>German</td>
<td>macintosh</td>
<td>4.559468</td>
</tr>
<tr>
<td>German</td>
<td>cp850</td>
<td>4.559468</td>
</tr>
<tr>
<td>Dutch</td>
<td>macintosh</td>
<td>9.825832</td>
</tr>
<tr>
<td>Dutch</td>
<td>cp850</td>
<td>9.825832</td>
</tr>
<tr>
<td>Dutch</td>
<td>cp1252</td>
<td>9.825832</td>
</tr>
<tr>
<td>Dutch</td>
<td>utf8</td>
<td>9.825832</td>
</tr>
<tr>
<td>English</td>
<td>cp1252</td>
<td>12.05258</td>
</tr>
<tr>
<td>English</td>
<td>utf8</td>
<td>12.05258</td>
</tr>
</tbody>
</table>
Creating a company does not happen overnight, but is a process. A brilliant idea alone is not enough to be successful.
SILC : commercialization

• *Que* - ALIS Technologies (1998 - 2001)
  - Binary and API in C (Java afterwards)
  - Integration into their web page translation service
  - Sold the whole system (source code and training corpora) to Oracle

• SILC - *Player*
  - COMScore, Mobikom, Excite

• Developed models for other languages
SILC : lessons learned

• Simple idea (in principle!)
• SILC apply vs SILC train-merge
• Licensing rights were long into making
  - no single year agreements, please...
  - payment schedule not always respected
• Simpler with a good licensing model
• Language recognition is now given in most text editors and MT web pages
C'est un tres beau systeme que celui-la.

Il reussit a tres bien deviner ou il aurait ete interessant d'ajouter d'ajouter les diacritiques appropries.
Réacc: French Accentuation

C'est un très beau système que celui-là.

Il réussit à très bien deviner où il aurait été intéressant d'ajouter d'ajouter les diacritiques appropriés.
Réacc

• Never managed to sell any license !!!
• Nice for demos, student projects and scientific articles but no market
  - Word plug-in
  - Mac application
• Also for Vietnamese
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, le papier est automatiquement configuré pour le format de papier et l'application.
Probabilistic model

\[ p(t|t',s) = p(t|t') \lambda(\Theta(t',s)) + M p(t|s) \left[ 1 - \lambda(\Theta(t',s)) \right] \]

source sentence: They are getting into seeding.

ongoing translation: Ils vont c \textit{ommencer}

\[ p(\text{commencer} | \text{they are getting into seeding}, 3) = 0.2 \]
\[ p(\text{commencent} | \text{they are getting into seeding}, 3) = 0.3 \]
\[ p(\text{cultivateurs} | \text{they are getting into seeding}, 3) = 0.1 \]

\[ p(\text{commencent} | \text{ils vont}) = 0.001 \]
\[ p(\text{commencer} | \text{ils vont}) = 0.1 \]
\[ p(\text{cultivateurs} | \text{ils vont}) = 0.0005 \]

translation probabilities

language probabilities

active vocabulary \( V \)
- actuellement
- agriculteurs
- amorcer
- commencent
- commencer
- cultivateurs
- elles
- ensemencement
- ils
- leurs
- semaines
- semences
- sont
TransType milestones

- TransTalk (CITI, 1995)
- Target-Text Mediated Interactive Machine Translation (Foster, 1997)
- Strategic grant (1997-2001) industrial partner: Machina Sapiens
- TransType2 (2002-2005) 5th Framework European Project IST-2001-32091

TransType2 Partners

- **Canadian (CRSNG-MDER)**
  - RALI - Université de Montréal
  - Société Gamma (Translation firm - *sister firm* of Terminotix)
  - CRTL

- **European (CEE)**
  - Atos Origin (Spain)
  - University of Aachen (Germany)
  - University of Valencia (Spain)
  - Xerox Research Centre Europe (France)
  - Celer Soluciones (Translation firm)
Experimentations in TransType2

• 3 language pairs
  - English↔{French, Spanish, German}

• 3 translation model (RALI,RWTH,ITI)

• 2 types of texts
  - technical
  - parliamentary (canadian et EC)

• Vocal interface

• Confidence estimation

• Adaptation
TransType2 - Evaluation

- reduction in the number of keys typed
- substitute production measure
- theoretical evaluation predicts 60% of saving
- practical evaluation on the order of 30%
- translators liked the idea
  - suggestion less intrusive than translation memory
  - no spelling error in suggestions
  - but it induces a literal translation
Lessons learned

• cognitive load for the translators
  - a single long suggestion is better than many shorter ones
• should limit the number of choice and frequency of appearance
• acceptance = typing 3 chars
• but no industrial application
  - ahead of its time
  - ATOS did not do its job
On leur donne des responsabilités moindres ? mieux adaptées à leurs capacités, j' imagine ? , mais ils continuent de faire partie du Cabinet, avec tous les avantages qui viennent avec.

Il examinera ces renseignements pour ensuite faire rapport à la Chambre ou au comité compétent, ce qui est tout à fait conforme aux pratiques courantes en ce qui a trait à ce genre de documents.

Cela détonne dans un pays qui veut, et dit vouloir, investir dans ses habitants.

Conformément à ces priorités, le budget mène à terme notre Plan d'action économique.

Dans le cadre de cette révision, on en profitera sûrement pour éliminer des programmes qui gêneront le gouvernement, qui ne sont pas conformes à son idéologie.

They get moved to lesser responsibilities, more in keeping with their capabilities, I suppose, but they stay in cabinet with all of the perks that go with being in cabinet.

That information will be reviewed by him. There will be a report back to the House or to a proper committee. This is very much in keeping with practices around documents such as these.

That is not in keeping with a country that is looking forward and saying that it wants to invest in its people.

In keeping with these priorities the budget completes our economic action plan.

When doing their review, they will certainly take advantage of it to eliminate programs that annoy the government, that are not in keeping with its ideology.
TransSearch - 1

• Initially (1992-1995) on Sun workstations
• Web version 1997
  - Hansard / Court Rulings 1986-1994
  - Free (as in free beer) for everybody
• Tried hard to convince industry to develop a
  Web service with this idea
• RALI launched its own commercial service in 2001
  - updated corpora (Hansard and Court Rulings)
  - revamped user and administration interface in Perl
  - kept the same textual data base (MG)
  - developed a business model (SAAS before its time)
• In two years
  - 1500 licenses (1000 at the Translation Bureau)
  - Enough to pay the development and maintenance
Deal with Terminotix in 2003
- administration and daily update of Hansard
- recruiting new customers
- service agreement
  - for adding new corpora
  - maintaining the server software
- added Spanish-English, Spanish-French corpora

Three-year deal renewed since

Win-win situation
• After years of evolution
  - hard to maintain and non incremental indexing
  - new research : translation spotting and its applications

• NSERC Collaborative Research and Development Grant
  - develop translation spotting
  - study new applications using *transspots*
  - embed these into a new system
165 translations of *in keeping with* in 670 occurrences

<table>
<thead>
<tr>
<th>conforme à</th>
<th>conforme à</th>
</tr>
</thead>
<tbody>
<tr>
<td>conforme à</td>
<td>conforme à</td>
</tr>
<tr>
<td>respect</td>
<td>respect</td>
</tr>
<tr>
<td>correspondant à</td>
<td>correspondant à</td>
</tr>
<tr>
<td>en conformité avec</td>
<td>en conformité avec</td>
</tr>
<tr>
<td>dans le sens de</td>
<td>dans le sens de</td>
</tr>
<tr>
<td>fidèle à</td>
<td>fidèle à</td>
</tr>
<tr>
<td>à fait</td>
<td>à fait</td>
</tr>
<tr>
<td>compte tenu</td>
<td>compte tenu</td>
</tr>
<tr>
<td>compatible avec</td>
<td>compatible avec</td>
</tr>
<tr>
<td>en fonction</td>
<td>en fonction</td>
</tr>
<tr>
<td>répondant</td>
<td>répondant</td>
</tr>
<tr>
<td>inscrit dans</td>
<td>inscrit dans</td>
</tr>
<tr>
<td>dans le cadre de</td>
<td>dans le cadre de</td>
</tr>
<tr>
<td>en vertu de</td>
<td>en vertu de</td>
</tr>
<tr>
<td>dans l'esprit de</td>
<td>dans l'esprit de</td>
</tr>
<tr>
<td>en logique avec</td>
<td>en logique avec</td>
</tr>
<tr>
<td>étant donné</td>
<td>étant donné</td>
</tr>
<tr>
<td>qui va dans le sens de</td>
<td>qui va dans le sens de</td>
</tr>
<tr>
<td>contraire</td>
<td>contraire</td>
</tr>
<tr>
<td>partie des</td>
<td>partie des</td>
</tr>
</tbody>
</table>

Fourth, the judge must be convinced that a conditional sentence is *in keeping with* the general principles of proportionality of the sentence.

It is a bill that is *in keeping with* the campaign and election commitments made by the Conservative Party of Canada to Canadians.

All this is *in keeping with* our Canadian values of compassion and generosity in times of need, a quality displayed across every community in Canada.

The reserve was *in keeping with* the established budgetary practice of setting aside policy reserves for specific contingent purposes.

I am certain that passing this bill, *in keeping with* Canada's policy of equal treatment of the parties, will contribute to reopening the roadmap to peace and will ensure a lasting peace between Israel and Palestine.

When the government introduced legislation, specifically Bill C-31 and Bill C-32, since, as explained the hon. Parliamentary Secretary to the Government House Leader, it was as a complement *in keeping with* Canadian practice...to confirm major changes in government organization through legislation.

I would simply ask that we preserve the current definition of marriage since it is wholesome for the common good, *in keeping with* the natural law and in conformity with God's design for the world.

This greater concentration in Africa is *in keeping with* Canada's
Automatic judgement summarization

• Area in which clients pay for summaries
• Usual approaches do not work well
• Extraction is preferable to abstraction
• A. Farzindar thesis (2005 - U de Montréal)
  - thematic segmentation
  - tabular presentation
NLP Technologies Inc. launched by A. Farzindar

- Web Services
  - Decision Express
  - Biblio Express
  - Statistic Express
  - Search Express

- High volume domains: immigration, tax, intellectual property

- Used by many lawyer offices and by judges from the Federal Court
ASLI
Automatic Summarization of Legal Information
June 2007-2008

• Industrial collaboration (Precarn)
  - NLP Technologies
  - RALI
  - Lawyers

• Automatic Summarization (French and English)
  - Finite state automata
  - Vocabulary separation from identification rules

• Machine Translation experimentation

• June 2007-2008
ISASLI

Intelligent system for Semantic processing, Automatic translation and Summarization of Legal Information (Jan-Dec 2009)

- Revision help - RevSum
- Statistical summarization
- Machine translation (another type of court CSST)
ASLI-ISASLI

Fruitful collaboration

- motivated industrial (entrepreneurial student !)
- willing to try new ideas
- willing to publish results
- MT proved to be a very successful *side effect*
Druide informatique
Statistical filtering of spelling corrections

• French spelling checker - Antidote
  - Excellent symbolic parser of French
  - tradeoff between noise and silence
  - clients are sometimes annoyed by too many suggestions

• 4 months contract

• Decision trees for 14 types of spelling errors:
  là vs la, er vs ez, verb modes, etc.

• Antidote HD (sept 2009) implements 8 classifiers
Other industrial partners

- Bell University Laboratories (1999-2003)
  - Automatic Investors E-mail processing
  - NSERC-RDC
    - 2 PhD + 2 MSc
- SOQUIJ & Lexum
  - anonymization of judgements (NOME)
- Environment Canada (2009-2011)
  - MITACS seed project for Multiformat Environmental Information Dissemination
What happened?

- SILC vs Reacc
- TransType
- TransSearch
- ASLI - ISASLI
- Druide
What happened?

- SILC vs Reacc
- TransType
- TransSearch
- ASLI - ISASLI
- Druide

Good research ideas are not enough!
Faculty, be aware that
Faculty, be aware that

- Need an internal champion within the industry
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a hurry
Faculty, be aware that

• Need an internal champion within the industry
• People move a lot within a company
• Discussions do not always result in concrete projects
• Takes time, especially with industrials who are in a hurry
• IP discussions between lawyers can be long and arduous
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a hurry
- IP discussions between lawyers can be long and arduous
- Big companies are not easier to work with
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a hurry
- IP discussions between lawyers can be long and arduous
- Big companies are not easier to work with
- Not a pot of gold
Faculty, be aware that

- Need an internal champion within the industry
- People move a lot within a company
- Discussions do not always result in concrete projects
- Takes time, especially with industrials who are in a hurry
- IP discussions between lawyers can be long and arduous
- Big companies are not easier to work with
- Not a pot of gold

Wait until you are tenured !!!
But

- Fun and stimulating
- Reality check
- Once industry decides to put *real* money, industrial grants are relatively easier to get than pure research grants
What I wish industry would understand

- scientific publications and conferences are good publicity
- the best benefits of a project is a privileged access to the students that worked on the project
- Each dollar spent in outside research should be matched by at least a dollar within the company
References

• J Strother Moore, Lawrence Snyder, Philip A. Bernstein, University-Industry Sponsored Research Agreements, CRA - Best Practice Memo, 2003.
Questions ?
Comments ?