



Université 
de Montréal



Weather Warning Translation Prototype

Fabrizio Gotti

Philippe Langlais

Guy Lapalme

7 January 2010



Context of the collaboration

Université de Montréal - EC

2004 Corpus of weather forecasts ('02-'03) for a statistical machine translation system

Presentation at the CMC

2005 Corpus of weather warnings ('00-'05)

Master's thesis of Thomas Leplus

Article published in *Machine Translation*

2006 Demo system of warning translation

2008 Michel Jean: Contribution agreement for 3 years (\$25 000)

Submission for a MITACS seed project


2009 MITACS project accepted (\$40 000)

MITACS seed project

Multi-format Environmental Information Dissemination

- 3 major aspects
 - Text and graphical generation of environmental information
 - Data mining
 - Translation
- 3 core researchers
- Researchers and students at the RALI and LISA laboratories


The current translation prototype

- Translation of public weather warnings for Canada
- English ↔ French
- Available online rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html
- A demonstration 
- 2000 words/minute, the translation of a warning takes 10 s on average

Weather Translation Demo

http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html

RALI - Meteo Weather Translation Demo



rali Weather Warning Translation Prototype version 2

Source text

STRONG THUNDERSTORMS HAVE DEVELOPED IN THE PORTAGE LA PRAIRIE AREA. RADAR IMAGERY INDICATES POTENTIAL FOR LARGE HAIL AND DAMAGING WINDS.

A LINE OF THUNDERSHOWERS IS CURRENTLY MOVING ACROSS THE CHALEUR REGION. THESE THUNDERSHOWERS HAVE THE POTENTIAL TO DEVELOP INTO SEVERE THUNDERSTORMS. THE MAIN THREAT IS LOCALLY HEAVY DOWNPOURS.

SOUTHWESTERLY WINDS REACHING 90 KM/H WILL AFFECT THE LOCALITIES OF AUPALUK AND TASIUJAQ UNTIL EARLY THIS EVENING. THE OTHER ABOVE MENTIONED AREAS WILL SEE NORTHWESTERLY WINDS REACHING 80 TO 90 KM/H THIS EVENING, TONIGHT AND PART OF MONDAY.

Translation

Des orages forts se sont formés dans le secteur de Portage la Prairie. Le radar indique une possibilité de grosse grêle et de vents destructeurs.

Une ligne d'orages se déplace présentement sur la région chaleur. Ces orages pourraient devenir violents. Le principal risque associé à ces orages est de la pluie torrentielle par endroits.

Vents devenant du sud-ouest de 90 km/h souffleront sur les secteurs de Tasiujaq et Aupaluk jusqu'à tôt ce soir. Les autres régions mentionnées ci-dessus connaîtront des vents du nord-ouest atteignant 80 à 90 km/h ce soir, cette nuit et une partie de la journée lundi.

Les conditions s'amélioreront légèrement cet après-midi alors que les températures se modéreront, cependant un refroidissement éolien dépassant les moins 40 reprendra cette nuit et persistera jusqu'à

Translation direction:

Click to get sample source text in: [French](#) or [English](#)

Note This prototype has been developed and optimized to translate [Public Weather Warnings](#) published daily by Environment Canada.

Part of the [Multi-format Environmental Information Dissemination](#) project

Overview of translation steps

1. Segmentation and normalization

AT 4:50 PM MONDAY, THE WEATHER RADAR IS SHOWING
INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE, WITH
WINDS REACHING 90 KM/H.. =====

- Segmentation in sentences, then words
- Normalization of certain expressions

AT 4:50 PM MONDAY THE WEATHER RADAR IS SHOWING
INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE
WITH WINDS REACHING 90 KM / H .

Overview of translation steps

2. Serialization: identification of « tokens »

AT 4:50 PM MONDAY , THE WEATHER RADAR IS SHOWING INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE , WITH WINDS REACHING 90 KM / H .

- « Tokens » for names of months and days, for hours, numbers
- Increases the system's reliability

AT PM , THE WEATHER RADAR IS SHOWING INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE , WITH WINDS REACHING KM / H .

Overview of translation steps

3. Translation: the core of the prototype

AT __HOUR__ PM __DAY__ , THE WEATHER RADAR IS SHOWING INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE , WITH WINDS REACHING __NUM__ KM / H .

- The translation is performed by a *phrase-based statistical translator* (more later)
- Aided by a *translation memory*

A __HOUR__ __DAY__ , LE RADAR METEOROLOGIQUE MONTRE DES ORAGES INTENSES A PORTAGE LA PRAIRIE , AVEC DES VENTS ATTEIGNANT __NUM__ KM / H .

Overview of translation steps

4. Deserialization: translation of « tokens »

A ___ HOUR ___ DAY ___ , LE RADAR METEOROLOGIQUE
MONTRE DES ORAGES INTENSES A PORTAGE LA PRAIRIE ,
AVEC DES VENTS ATTEIGNANT ___ NUM ___ KM / H .

- Restoration of tokens identified earlier,
translated if need be (manually crafted rules)
- Faithful restoration of numbers, dates, ...

A **16H50 LUNDI** , LE RADAR METEOROLOGIQUE MONTRE
DES ORAGES INTENSES A PORTAGE LA PRAIRIE , AVEC
DES VENTS ATTEIGNANT **90** KM / H .

Overview of translation steps

5. Restoration of case and accents

A 16H50 LUNDI , LE RADAR METEOROLOGIQUE MONTRE
DES ORAGES INTENSES A PORTAGE LA PRAIRIE , AVEC
DES VENTS ATTEIGNANT 90 KM / H .

- Restoration of accents (for French)
- Restoration of case (lower/upper), using a list of places and a few simple rules

À 16h50 lundi , le radar météorologique montre
des orages intenses à Portage la Prairie , avec
des vents atteignant 90 km / h .

Overview of translation steps

6. Desegmentation and finishing touches

À 16h50 lundi , le radar météorologique montre des orages intenses à Portage la Prairie , avec des vents atteignant 90 km / h .

- Removal of unaesthetic spaces
- Insertion of apostrophes
(l Ontario → l'Ontario)

À 16h50 lundi_ le radar météorologique montre des orages intenses à Portage la Prairie_ avec des vents atteignant 90 km/h .

The translation engine

- Uses state-of-the-art *phrase-based statistical translation* (P. Koehn, F.-J. Och, D. Marcu)

AT ___ HOUR ___ PM ___ DAY ___ , THE WEATHER RADAR IS SHOWING
INTENSE THUNDERSTORMS IN PORTAGE LA PRAIRIE ,



A ___ HOUR ___ DAY ___ , LE RADAR METEOROLOGIQUE MONTRE DES
ORAGES INTENSES A PORTAGE LA PRAIRIE ,

- Relies on a (very large) phrase table

...

THE WEATHER RADAR AND A WATCHER ♦ LE RADAR METEO ET UN OBSERVATEUR

THE WEATHER RADAR AND A ♦ LE RADAR METEO ET UN

THE WEATHER RADAR IS SHOWING INTENSE THUNDERSTORMS ♦ LE RADAR METEOROLOGIQUE MONTRE
DES ORAGES INTENSES

THE WEATHER RADAR IS SHOWING THAT ♦ LE RADAR METEOROLOGIQUE MONTRE QU

THE WEATHER RADAR SHOWS A LINE ♦ LE RADAR METEOROLOGIQUE MONTRE QU UNE LIGNE

...

Training the translation engine

- The translation engine (*decoder*) must be trained on **numerous** and **high-quality** human translations



Blizzard conditions will be developing in Resolute this afternoon. This is a warning that blizzard conditions are imminent or occurring in these regions. Monitor weather conditions..Listen for



Blizzard à Resolute cet après-midi. Ceci est un avertissement indiquant qu'il y a ou qu'il y aura sous Peu du blizzard dans ces régions...Veuillez surveiller les bulletins météorologiques et ..

1. Alignment of documents which are translations of one another

2. Text extraction

Training the translation engine

3. Segmentation in sentences, then words

- Using task-specific rules

4. Sentence alignment

- RALI's aligner: Japa (<http://rali.iro.umontreal.ca/japa>)

BLIZZARD CONDITIONS WILL BE DEVELOPING
IN RESOLUTE THIS AFTERNOON .



BLIZZARD A RESOLUTE CET APRES
MIDI .

THIS IS A WARNING THAT BLIZZARD
CONDITIONS ARE IMMINENT OR OCCURRING IN
THESE REGIONS .



CECI EST UN AVERTISSEMENT
INDIQUANT QU IL Y A OU QU IL Y
AURA SOUS PEU DU BLIZZARD DANS CES
REGIONS .

MONITOR WEATHER CONDITIONS .



VEUILLEZ SURVEILLER LES BULLETINS
METEOROLOGIQUES ET LEURS MISES A
JOUR .

LISTEN FOR UPDATED STATEMENTS .



Bitext creation

- 2 types of text sources were used: weather forecasts (reports) and weather warnings
- Weather forecasts: 1 source
 - Years 2002-2003, made available by Rick Jones and Marc Besner
- Weather warnings: 4 sources
 - Years 2000-2004, provided by EC
 - Since June 2009, the RALI receives all weather warnings by e-mail
 - EC's Website « Public Weather Warnings for Canada »
 - ★ – **New:** Years 2005-2009, received in November 2009

All bitexts: statistics

Bitext	Pairs of sentences	Different sentence pairs	English		French	
			Words	Different words	Words	Different words
Forecasts RALI	4,345,626	349,236	31,608,117	3322	38,710,388	4395
Warnings RALI	104,971	69,670	1,736,637	5935	2,066,915	8144
Warnings email	45,959	7278	632,400	2967	779,376	3188
Warnings website	106,901	15,190	1,936,272	2941	2,281,993	3875
Total	4,603,457	441,155	35,913,426	8063	43,838,672	10,665
New warnings	~150,000	~100,000				

Text extraction: some difficulties

- Format corruptions, warnings broadcast in a single language, spelling and grammar

Alertes/avertissements

Côte centrale - secteurs côtiers

14h58 HNP le dimanche 29 novembre 2009

un avertissement de pluie pour

Côte centrale - secteurs côtiers est maintenu

Encore 50 à 70 mm de pluie jusqu'à lundi matin.

...

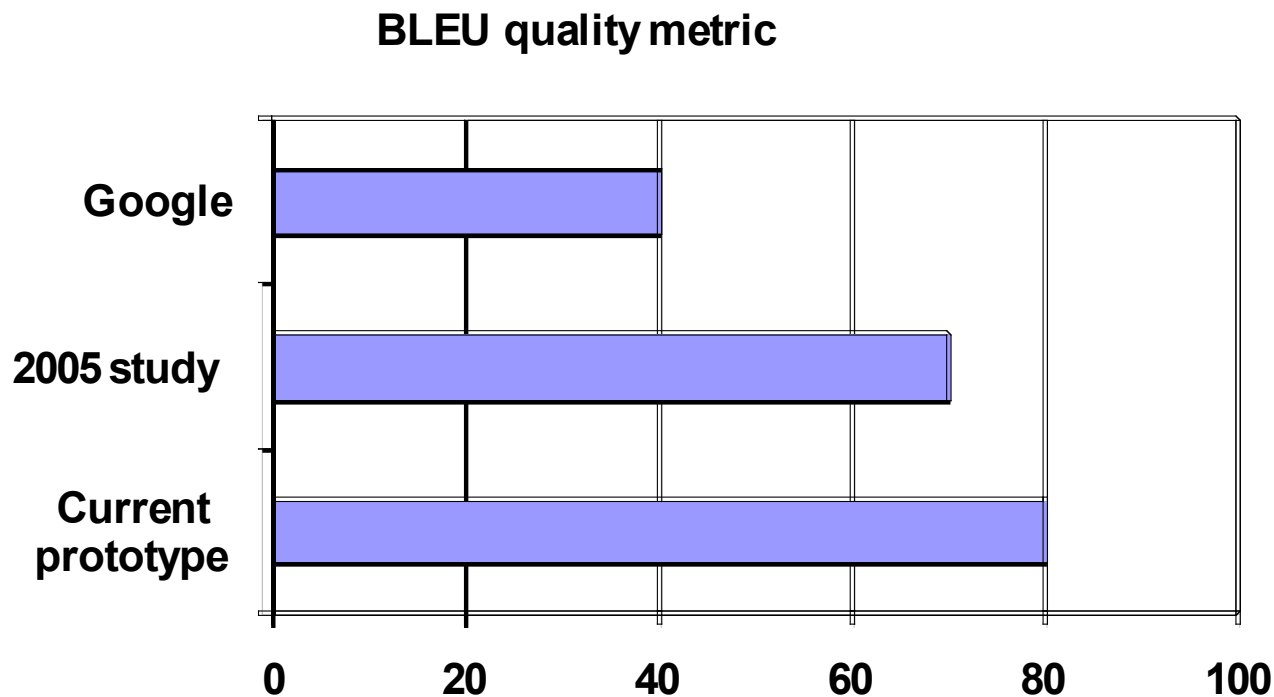
Meanwhile snow will continue across the inland areas and amounts up to 20 cm are forecast. The band of moisture will move out of the region Monday morning as the front moves to the south coast and precipitation will ease off. **Crossing.**

Le présent bulletin est un avertissement indiquant qu'il pleut ou qu'il pleuvra en abondance sur ces régions... Veuillez surveiller les bulletins météorologiques et leurs mises à jour.

Le front froid continue lentement de traverser la côte centrale et la pluie a cessé progressivement sur la côte. Toutefois dans l'intérieur les précipitations ont augmenté d'intensité et la **pluie à fait** place à la neige. La prochaine poussée venant du **pacifique** produira de la pluie le long de la côte à nouveau ce soir, y laissant **en 50 à 70 mm**. Entre-temps la neige persistera sur l'intérieur, **ou** l'on prévoit jusqu'à 20 cm. La bande d'humidité quittera la région lundi matin lorsque le front gagnera la côte sud et que les précipitations **Cesseront** progressivement..

Evaluation of the translation engine

- Methodology: comparison of the machine translation with the reference (human) translation



▲ *A score of 35 is considered good in our field*

The BLEU score is imperfect

- The BLEU score is an **automatic metric**, not a human judgment
 - Automatic measures are an **imperfect substitute** for human assessment of translation quality
- BLEU works by comparing the output of a system with a reference translation
 - **Problem:** If we were to use BLEU to compare the outputs of our prototype with those of the current system, the reference translation would be that of the current system : unfair evaluation...
- **A manual evaluation is more accurate, more meaningful**
 - Time-consuming, typically requires a large effort
 - The workload is distributed across a number of *human judges*

A manual evaluation campaign

1. Define the task the prototype should have to perform
2. Select performance criteria on which the prototype will be judged
 - Can be a comparison with the current system
3. Design an evaluation protocol, allowing the collection of data for those criteria
4. Proceed with manual evaluation, in conditions as « real-life » as possible, preferably performed by the eventual end-users of the system

A proposition for a campaign

1. Task definition: our prototype should act as a **gisting tool**, producing understandable warnings **very rapidly**

*Cette page a été traduite automatiquement.
Une version révisée sera bientôt disponible.*

Alertes/avertissements

Côte nord - secteurs côtiers
10h12 HNP le lundi 04 janvier 2010

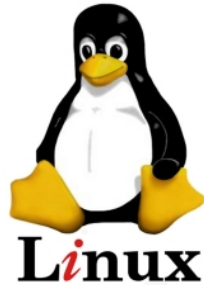
2. Performance criteria
 - Translation speed (average time to translate a warning, now 10s on average for the prototype)
 - Fidelity of the translation: is the meaning of the translation equivalent to that of the original?

A proposition for a campaign

3. Evaluation protocol

- Evaluation performed by human judges (preferably meteorologists) on a random sample of warnings
- The judges would be presented with the source text and our prototype's translation
- The judges measure how well the original meaning is transposed in the translation, using to-be-defined criteria
 - Examples: Is the duration of the warning present? Are the affected areas correctly translated?
 - In short, is the translation useful for the end user?

Current setup for the prototype



Web/Translation server

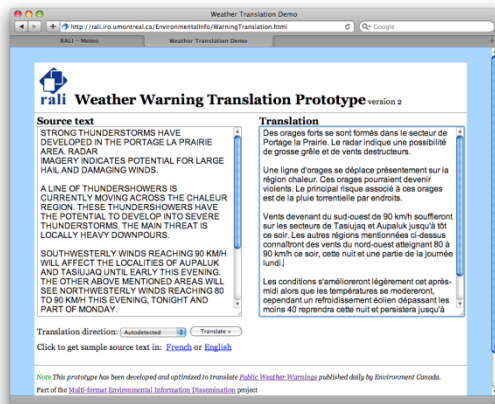
Linux 64 bits, 2 Gb RAM, CPU ~3 GHz

Software used

Moses decoder (LGPL License)

Language modeler Srm (Srm License)

Code from the RALI



Web client



Python client (batch processing of large quantities of text)

Research avenues

- Integration of new resources
 - New weather warnings (2005-2009) received in November 2009
 - Warnings for 2010
 - List of names of Canadian places in order to improve case restoration
 - Any set of translation rules that EC could disclose
 - Others

Research avenues

- Translation of other types of bulletins (marine, aviation, etc.)
 - Provided we can exploit numerous sample translations + other useful resources
- In December 2009, Gilles Ratté (National Coordinator - Aviation Program Development & Standards) sent us a few “forecaster notes”
 - These notes are short, technical texts accompanying TAFs (Terminal Aerodrome Forecasts)
 - Similar in some respects to “traditional” weather warnings
 - The prototype could be adapted to translate these texts
 - Difficulty: only the English version exists

Research avenues

- Translation of warnings to Inuktitut
 - Using the same translation engine
- Difficulty: no Inuktitut corpus specific to weather warnings available to train a prototype
- Possible approaches
 - Text from the Nunavut Parliamentary debates
 - Linguistic resources for Inuktitut:
www.inuktitutcomputing.ca
(National Research Council Canada)
 - Automatic morphological tools to process Inuktitut

Merci