Prototype de traduction des avertissements météo

Fabrizio Gotti

Philippe Langlais

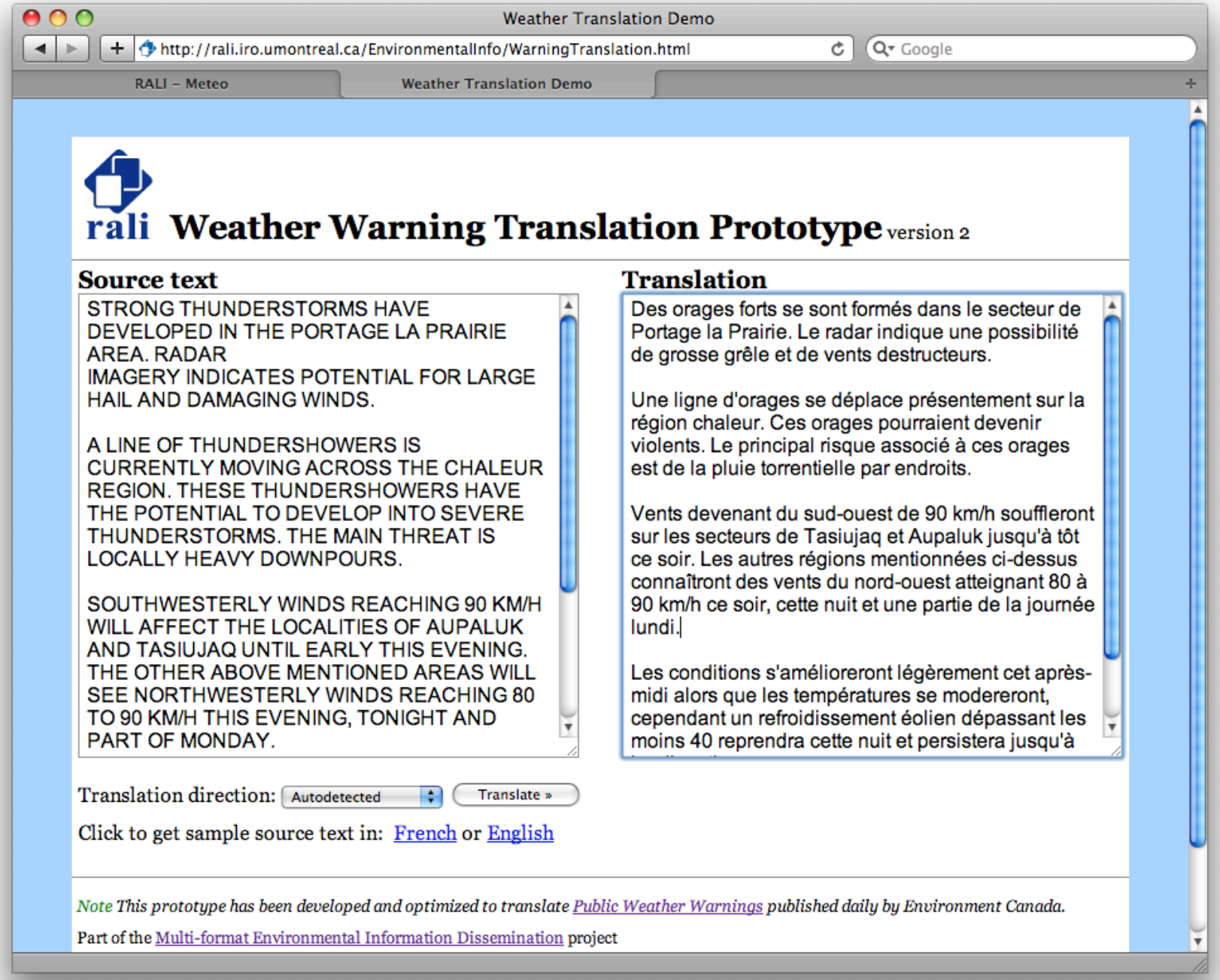
Guy Lapalme

Plan de la présentation

1. Démonstration du prototype actuel
2. Explication de son fonctionnement
3. Textes utilisés pour entraîner le prototype
4. Quelques difficultés rencontrées
5. Déploiement chez Environnement Canada
6. Pistes intéressantes

Le prototype actuel

- Traduction des avertissements météo publics (*alertes météo*) pour le Canada
- Anglais ↔ français
- Disponible en ligne
  rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html
- Démonstration 


Weather Translation Demo
http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html
Google
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 modereront, cependant un refroidissement éolien dépassant les moins 40 reprendra cette nuit et persistera jusqu'à
Translation direction: Autodetected
Translate »
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

Survol des étapes de traduction

1. Segmentation et normalisation

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

- Segmentation en phrases, puis en mots
- Normalisation de certaines expressions

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

Survol des étapes de traduction

2. Sérialisation: identification des « tokens »

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

- « Tokens » pour les noms de mois, les jours, les heures, les nombres
- Augmente la fiabilité du système

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

Survol des étapes de traduction

3. Traduction: le cœur du système

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

- La traduction est faite à l'aide d'un *traducteur statistique à base de segments*
- Aidé d'une *mémoire de traduction*

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

Survol des étapes de traduction

4. Désérialisation: traduction des « tokens »

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

- Restauration des tokens isolés plus tôt, avec traduction le cas échéant (règles manuelles)
- Restauration fidèle des nombres

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

Survol des étapes de traduction

5. Réaccentuation + restauration de la casse

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

- Restauration des accents (pour le français)
- Restauration des majuscules, à partir d'une liste de noms de lieux et de règles simples

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

Survol des étapes de traduction

6. Dé-segmentation et peaufinage

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

- Suppression des espaces inesthétiques
- Ajout des 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**.

Le moteur de traduction

- Principe de la traduction statistique à base de segments (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 ,
```

- À partir d'un (très grande) table de segments

```
...
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
...
```

Entraînement du moteur de traduction

- Le décodeur (traducteur) doit être *entraîné* sur des traductions humaines **nombreuses** et **de qualité**

1. Alignement des documents en relation de traduction

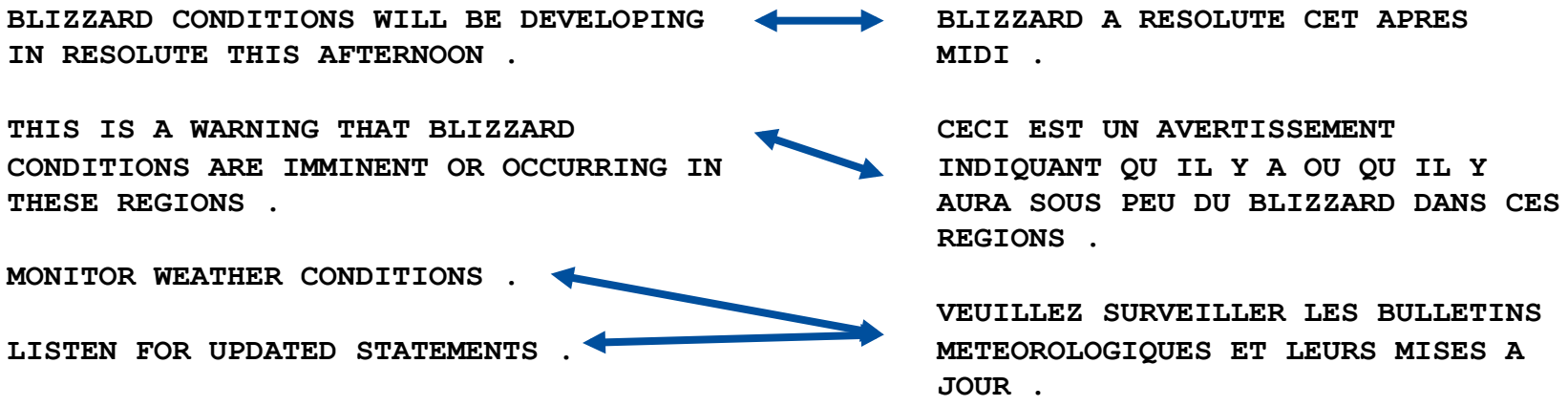
```
Blizzard conditions will
be developing in
Resolute this
afternoon.This is a
warning that blizzard
conditions are imminent
or occurring in these
regions. Monitor weather
conditions..Listen for
```

```
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 ..
```

2. Extraction du texte pur

Entraînement du moteur de traduction

3. Segmentation en mots et phrases
   - Ensemble de règles conçues pour l'occasion
4. Alignement des phrases
   - Aligneur du RALI: Japa (http://rali.iro.umontreal.ca/japa)



Création des bitextes

- 2 types de données utilisées: les bulletins météorologiques et les alertes
- Bulletins météorologiques
  - Les années 2002-2003, rendues disponibles par Rick Jones et Marc Besner
- Alertes météorologiques: 4 sources
  - Les années 2000-2004, fournies par le Bureau d'EC
  - « Dérivation » du système actuel afin de recevoir les alertes par courriel au RALI, depuis juin 2009
  - Site Web d'EC « Avertissements météo publics pour le Canada »
  - **Nouveau:** les années 2005-2009, reçues en nov.

Tous les bitextes: statistiques

Textes	Nb. paires phrases	Nb. paires différentes	Anglais		Français	
			Nb. mots	Nb. mots différents	Nb. mots	Nb. mots différents
Bulletins RALI	4 345 626	349 236	31 608 117	3322	38 710 388	4395
Alertes RALI	104 971	69 670	1 736 637	5935	2 066 915	8144
Alertes courriel	45 959	7278	632 400	2967	779 376	3188
Alertes Web	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	10665
Nouvelles alertes	~150 000	~100 000				

Extraction du texte: difficultés

- Corruptions de format, alertes unilingues, orthographe et grammaire

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.

…

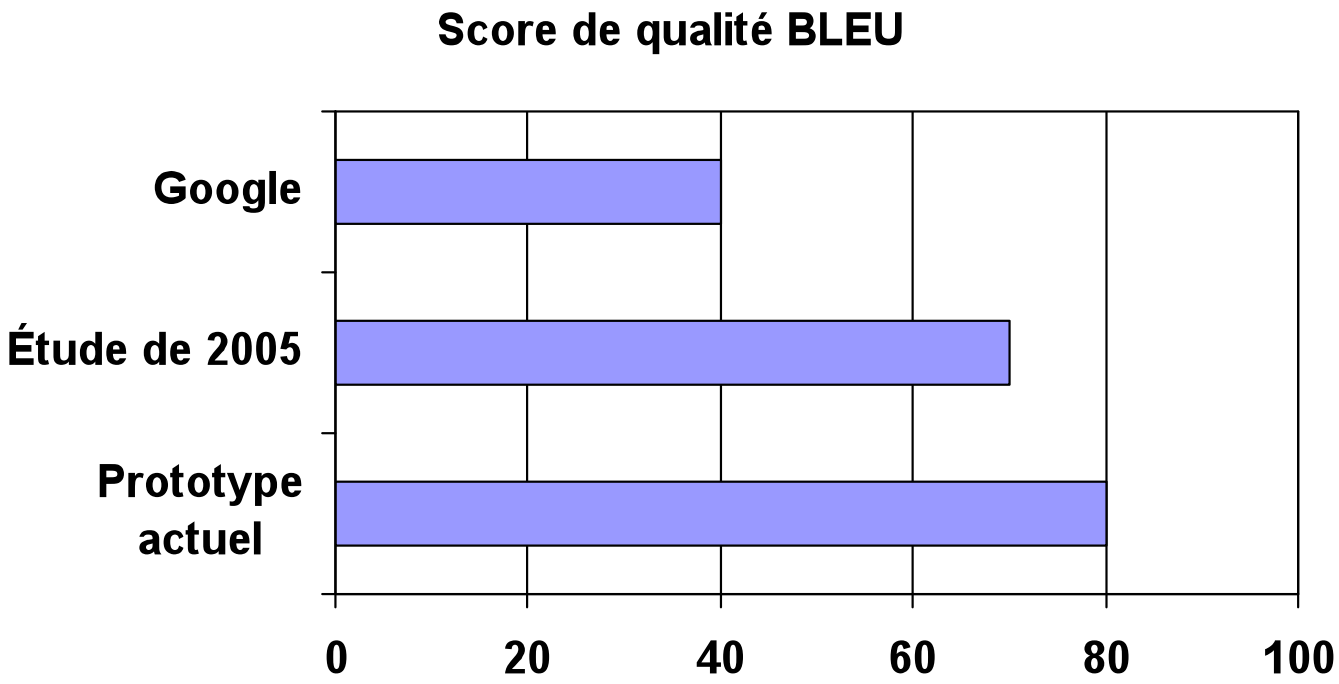
Meanwhile snow will continue across the inland areas and amounts up to 20 cm are forecast. The band of moisture will move out of the region Monday morning as the front moves to the south coast and precipitation will ease off. **Crossing.**

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**..**

Tests du moteur de traduction

- On compare la traduction automatique avec la traduction de référence (humaine)



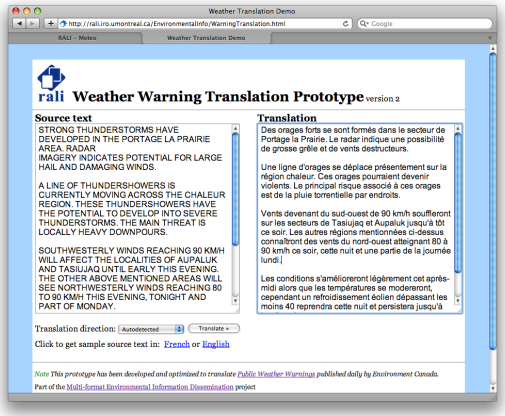
Déploiement chez EC

Serveur Web de traduction

Linux 64 bits

2 Go de mémoire vive

Processeur ~3 GHz



Client Web

Client Python (automatisation)

Pistes intéressantes

- Intégration de nouvelles ressources
  - Les nouvelles alertes (2005-2009)
  - Une liste de noms de lieux canadiens pour améliorer la restauration de la casse
  - Des règles de traduction utilisées par EC (?)
  - Autres
- Traitements d'autres types de bulletins (maritimes, aériens, etc.)
  - Moyennant l'obtention de textes d'exemples en grande quantité, et d'autres ressources

Merci