Université de Montréal

Environnement et Changement climatique Canada | Environment and Climate Change Canada

ENVIRONMENT & CLIMATE CHANGE CANADA

Development of a weather text generator

11th Montreal Industrial Problem Solving Workshop

Guy Lapalme, Prof. emeritus, Université de Montréal
Fabrizio Gotti, Researcher, Université de Montréal
Jérémy Rieussec, Ph.D. student, Université de Montréal

DIRO

rali

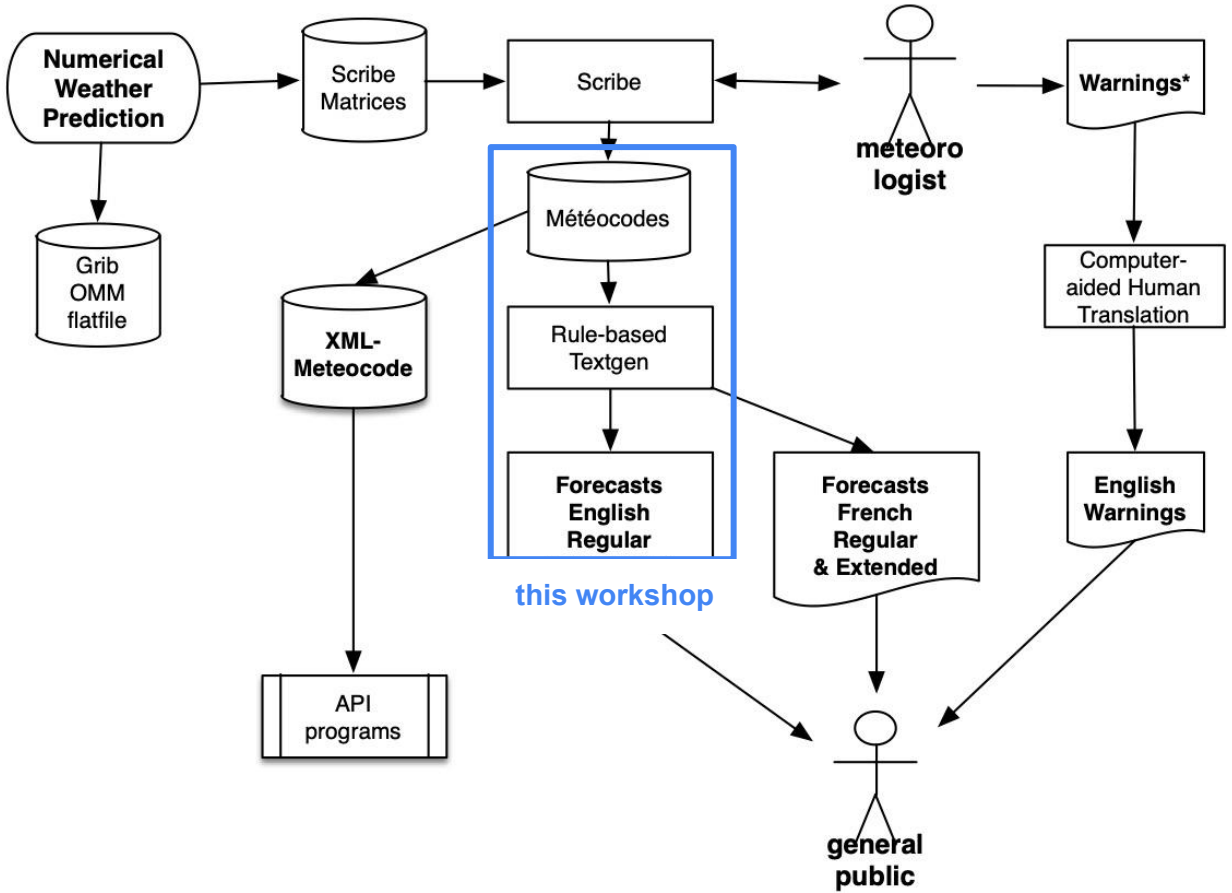
Natural Language Generation (NLG)

- Create readable text from data
- Input : numbers, texts, logical formulas,...
- Output : text, graphics, conversations
- Two classic processes usually pipelined:
  - What to say ?
  - How to say it ?
- Contrarily to parsing:
  - Fuzzy starting point
  - Easy for a human to appreciate the result

Weather Bulletins in Canada

- Canada is a *wide* country
- Thousands of forecast bulletins each day
  - 3 times a day : ~5h, ~11h, ~16h00
  - Regular (today, tomorrow), Extended (7 days)
  - French and English
- All times in data are in *Coordinated Universal Time* (UTC)
  - 6 time zones in Canada
  - daylight saving or standard

Context of the data



Data sources

- Meteocode
  - possibly human edited
- Bulletins
  - generated by Scribe in English and French
- Data for the workshop
  - Ontario and Québec English bulletins
  - 2018 and 2019
  - Timezone shift is given (-5h EST, -6 EDT)

Météocode for a bulletin

{*Field*: {"("start_h end_h ...other infos...")")}";"} "."

Documentation of the fields

```
entete: (FPCN71 CWUL EST5EDT regulier  2018 06 03 0900 00
         prochaine_prevision 2018 06 03 1530 30);

regions: (r71.1);
...
regions: (r71.15);
accum:  (-49 -46  pluie totale pres_de 4 6)
...      (69 72  pluie totale pres_de 2);
ciel:  (-53 -24 9 9 9) (-24 -21 9 1 9) (-21 24 1 1)
...    (180 186 5 5 9) (186 213 5 5 9) (213 228 1 1);
climat_temp:  (-52 -38 min 5.0) (-38 -28 max 18.0)
       (-28 -14 min 5.0) (-14 -4 max 18.0)
...    (212 226 min 7.0) (226 236 max 20.0);
indice_qa:  (-39 -38 1.6) (-38 -37 1.3) (-37 -36 1.0)
...    (33 34 1.2) (34 35 1.1) (35 36 1.1);
indice_uv:  (-32 -30 7.0) (-8 -6 6.9)
            (16 18 7.2) (40 42 7.2);
pcpn:  (-52 -44 certain debut_fin pluie nil continuel)
..     (135 144 possible debut_fin averses nil frequent);
prob: (seuil 0.2  (-54 -31 80) (-31 32 0) (32 46 80)
       (46 48 100) (48 60 90) (60 72 100)
..     (216 228 10)     );
rosee:  (-51 -48 point_intermediaire 18) (-48 -45 point_inter
...     (225 228 point_intermediaire 7);
temp:  (-51 -48 point_intermediaire 20) (-48 -45 point_interm
...    (225 228 point_intermediaire 14);
vents:  (-49 -42 sw  vitesse 20 (-49 -42 rafales 40)) (-42 -3
40))
...     (216 228 w  vitesse 10);

regions: (r71.16);
```

JSON


```
{"header":["FPCN71","CWUL","EST5EDT","regulier",2018,6,3,900,0,
  "prochaine_prevision",2018,6,3,1530,30],
 "names-en":["Matagami"],
 "names-fr":["Matagami"],
 "regions":["r71.15"],
 "accum":[[-49,-46,"pluie","totale","pres_de",4,6],
          [69,72,"pluie","totale","pres_de",2]],

 "ciel":[[-53,-24,9,9,9],
         [213,228,1,1]],

 "climat_temp":[[-52,-38,"min",5.0],
                [226,236,"max",20.0]],

 "indice_uv":[[-32,-30,7.0],
              [40,42,7.2]],

 "pcpn":[[-52,-44,"certain","debut_fin","pluie","nil","continuel"],

         [135,144,"possible","debut_fin","averses","nil","frequent"]],
 "prob":[["seuil", 0.2, -54,-31,80],
          [216,228,10]]],

 "rosee":[[-51,-48,"pi",18],
          [225,228,"pi",7]],

 "temp":[[-51,-48,"pi",20],
         [-42,-39,"min",3],
         [-30,-27,"max",5],
         [225,228,"pi",14]],

 "vents":[[-49,-42, "sw", "vitesse",20,
           [-49,-42,"rafales",40]],
          [216,228,"w","vitesse",10]],
```

JSON


```
{"header":["FPCN71","CWUL","EST5EDT","regulier",2018,6,3,900,0,
  "prochaine_prevision",2018,6,3,1530,30],
 "names-en":["Matagami"],
 "names-fr":["Matagami"],
 "regions":["r71.15"],
 "accum":[[-49,-46,"pluie","totale","pres_de",4,6],

          [69,72,"pluie","totale","pres_de",2]],
 "ciel":[[-53,-24,9,9,9],

         [213,228,1,1]],
 "climat_temp":[[-52,-38,"min",5.0],

                [226,236,"max",20.0]],
 "indice_uv":[[-32,-30,7.0],

              [40,42,7.2]],
 "pcpn":[[-52,-44,"certain","debut_fin","pluie","nil","continu

         [135,144,"possible","debut_fin","averses","nil","freq
 "prob":[["seuil",
          0.2,
          [-54,-31,80],

          [216,228,10]]],
 "rosee":[[-51,-48,"pi",18],

          [225,228,"pi",7]],
 "temp":[[-51,-48,"pi",20],

         [-42,-39,"min",3],
         [-30,-27,"max",5],
         [225,228,"pi",14]],
 "vents":[[-49,-42, "sw", "vitesse",20,
           [-49,-42,"rafales",40]],
          [216,228,"w","vitesse",10]],
```


Generated English bulletin

FPCN11 CWUL 030900
Forecasts for Western Quebec issued by Environment Canada at 5:00 a.m. EDT Sunday 3 June 2018 for today and Monday.
The next scheduled forecast will be issued at 11:30 a.m. EDT.

Metro Montréal - Laval.
...
Vaudreuil - Soulanges - Huntingdon.
...

Matagami.
Today..Sunny. Wind becoming southeast 20 km/h this morning. High 23. UV index 8 or very high.
Tonight..Increasing cloudiness. Rain beginning before morning. Wind east 20 km/h gusting to 40. Low 9.
Monday..Rain. Wind east 20 km/h gusting to 40. Temperature steady near 10. UV index 2 or low.

Waskaganish.
....

Some challenges

- Identify the appropriate data for a given *period* (today, tonight, tomorrow)
  - e.g. **today** : between 5h and 18h local time
  - data is in Universal Time (i.e. Greenwich) so must subtract 4 or 5 depending on the date
  - data is given as ranges [start end values…]
  - must find the data ranges that intersect with the start and end time of the period
  - possibly expand the data values for each hour
- Some values (e.g UV index or Wind chill) depend on many types of values

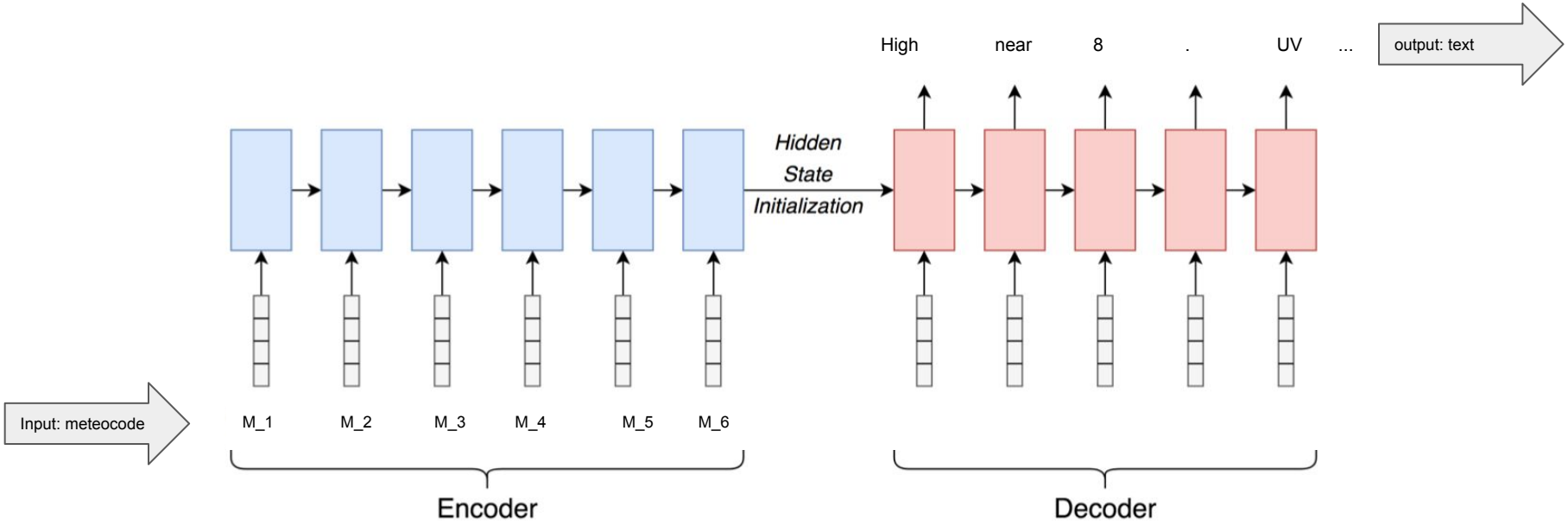
Remarks

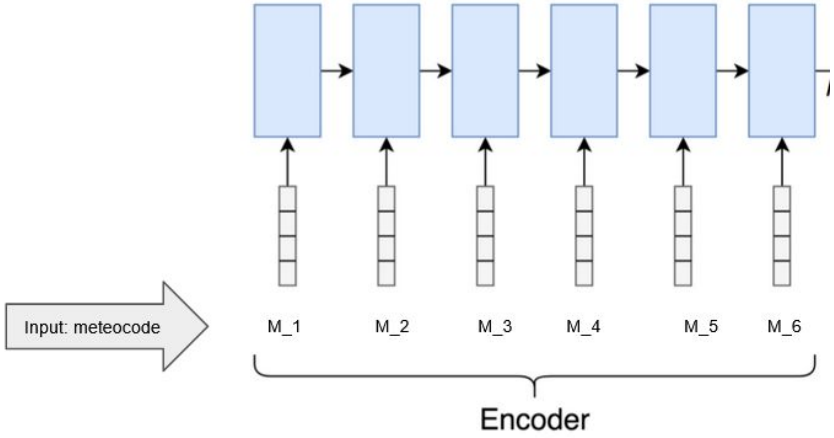
- Input is data generated by a mathematical process, not data created by humans (e.g. sales, click rate…)
- The job is to try to reproduce with a neural approach the output of a symbolic system

Compromises given the time allotted and the difficulty encountered:

- temperatures only (`temp` field)
- separate the generation by period of the day
- train: ~200k, valid: ~10k, test: ~10k bulletins, but subset of train used

Strategy: seq2seq architecture






```
temp:
  0:
    0:  -51
    1:  -48
    2:  "pi"
    3:  -17
  1:
    0:  -48
    1:  -45
    2:  "pi"
    3:  -17
  2:
    0:  -45
    1:  -42
    2:  "pi"
    3:  -18
```

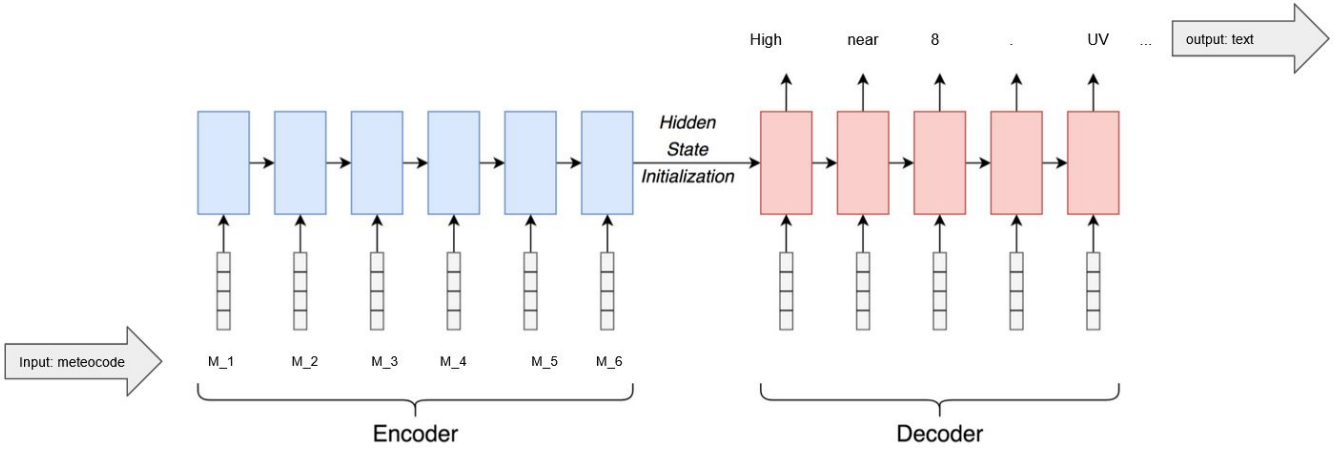

rules

M_0: <start_hour , one-hot(category), temperature>

M_1: <start_hour + 1, one-hot(category), temperature>

M_2: <start_hour + 2, one-hot(category), temperature>

...



- no word embeddings in input, we build meteocode embeddings instead
- small vocabulary of ~100 words for temperature
- meteocode temperature embeddings of size 8 only

- … so modest architecture (38k parameters), faster to train, 1 epoch = 10 minutes on GPU, 3 epochs give good results

Results: convincing

- BLEU score of 76% (100% = perfect)

reference	generation
Low 10 .	Low 10 .
High 26 .	High 26 .
Low 9 .	Low 8 .
Low 12 .	Low 12 .
High 28 .	High 27 .
Low 12 .	Low 12 .
Low 14 .	Low 14 .
High 28 .	High 26 .
Low 14 .	Low 14 .
Temperature steady near minus 1 .	Temperature steady near zero .
Low minus 8 .	Low minus 8 .
High minus 5 .	High minus 5 .
High minus 8 .	High minus 8 .
Temperature rising to minus 3 by morning .	Temperature rising to minus 3 by morning .
High plus 1 .	High plus 1 .
High minus 1 .	High minus 1 .
Low minus 2 .	Low minus 2 .
High plus 2 .	High plus 2 .
High minus 3 .	High minus 3 .
Temperature steady near minus 3 .	Temperature steady near minus 3 .
High minus 1 .	High minus 1 .
Temperature steady near minus 1 .	Temperature steady near minus 1 .
Low minus 9 .	Low minus 9 .
High minus 7 .	High minus 7 .

… but beware of “revisionist” seq2seq

reference	generation
High 28 .	High 27 .
Low 12 .	Low 12 .
Low 14 .	Low 14 .
High 28 .	High 26 .
Low 14 .	Low 14 .
Temperature steady near minus 1 .	Temperature steady near zero .
Low minus 8 .	Low minus 8 .
High minus 5 .	High minus 5 .
High minus 8 .	High minus 8 .
Temperature rising to minus 3 by morning .	Temperature rising to minus 3 by morning .
High plus 1 .	High plus 1 .
High minus 1 .	High minus 1 .
Low minus 2 .	Low minus 2 .

Materials & methods

- All written in Python
- Data processing: custom-made parser, JSON technologies
- The usual deep learning libraries: pytorch, torchtext
- Interesting starting point: NLP From Scratch: Translation with a Sequence to Sequence Network and Attention — PyTorch Tutorials 1.9.0+cu102 documentation
- About 3 person-weeks of work, ~2k lines of (sometimes hurried) code
- GPU for fast training and inference

Conclusions

- As usual, data is a very time-consuming element (extraction, format, documentation, interpretation, etc.)
- seq2seq seems to be a good start for temp field
- Whether it works with other fields remains to be seen...
- Rarer meteocode phenomena will create difficulties, because of scarcity of examples
- In NLG, evaluation is a recurring challenge, and BLEU is a very coarse metric