Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

Translation memory hit ratios in a best-case scenario for WATT

Fabrizio Gotti – gottif@iro.umontreal.ca
Guy Lapalme – lapalme@iro.umontreal.ca

13 June 2016
Recherche Appliquée en Linguistique Informatique

Introduction

WATT[1] uses a translation memory (TM) populated with some of the past translations of weather warning sentences. Even though these sentence pairs (source and translation) are produced by humans, not all of them make it to the TM, mainly because there are doubts on their quality (source and target material)[2].

This report considers a theoretical best-case scenario where WATT's TM is populated with *all* the material available to us, regardless of its quality. The experiments reported here measure the recall of that ideal TM when presented with a new weather warning. The goal is to assess how much better WATT's TM performs when it has access to all past translations, compared to the current situation where only a subset is used.

Methodology

Corpora

We populate a TM with serialized (__DAY__, __NUM__) versions of as many past human translations as we have access to. We use the training corpus described in our article[1] published in NLE, as well as most of the CAP archive shared by EC and described in a previous report[3]. We keep the rest of the CAP archive as a test corpus.

[1] "*Designing a machine translation system for Canadian weather warnings: A case study*" – http://rali.iro.umontreal.ca/rali/sites/default/files/publis/WATT-NLE-Final.pdf
[2] See our previous reports on source and target language quality.
[3] "Updating WATT's translation memory, Part 1: Findings on the CAP data"

In the TM:

- NLE corpus (warnings from 2000 to 2011): **4.9M** sentences
- CAP (warnings from Jan. 2013 to July 2015 inclusively): **0.4M** sentences

Test corpus:

- CAP (WO and WW warnings, Aug. 2015 to Dec. 2015 inclusively): **53k** sentences

Measuring hit ratios: Two settings

We measure how many of the sentences in the test corpus are present in the translation memory. We call this the *hit ratio*, a percentage. We do this for each language separately.

We use two different settings.

The "**traditional**" setting is the most straightforward, where we submit all the sentences in the test corpus and compute the hit ratio from the TM.

The "**chronological**" setting corresponds to a hit ratio where all test warnings are submitted in chronological order[4] to the TM, with an important difference from the previous scenario: as soon as a warning has been submitted to the TM and its hit ratio has been computed, its sentences are immediately added to the TM. The next warning in chronological order is then submitted to the memory, which has therefore access to all past translations, including those from the previous warning. This corresponds to an ideal setting where a TM shared across all human translators is assiduously updated as soon as new material is produced.

Results

Metric	English	French
Nb different sentences in TM	622K	675K
Nb sentences in test	53K	53K
Hit ratio for traditional	67%	67%
Hit ratio for chronological	83%	77%

The TM hit ratio for the current implementation of WATT is 50%.

Interpretation

As it stands, WATT's TM has a hit ratio of 50%, according to the figures we published in previous reports.

The "traditional" setting

Should we add all available training material to the TM, we could potentially reach a hit ratio of 67%, a significant improvement of almost 20% in absolute percentage points. This could translate into potential savings of 20% on translation cost.

[4] All CAP warnings have a timestamp and can therefore be sorted in the order they were issued.

WATT does not benefit from all this material for the time being, mainly because there are issues with the quality of the source and target material produced by humans. To compensate for this quality problem, RALI decided early on that a sentence pair can only be added to the TM if it appears a certain number of times in the training material. This rudimentary filter improves the quality of the TM, but at the cost of rejecting many sentence pairs.

An interesting way to enrich the TM with more of the past translations would be to devise smarter filters, perhaps in the form of classifiers. More subtle features of a given translation pair would then help determine whether it is valid or not.

The "chronological" setting

The chronological setting's results highlight the importance of a well-curated TM shared by all translators. If a translator is to immediately add a warning pair when he finishes a translation, significant improvements to the hit ratios of the TM are to be expected, about 15% in absolute percentage points (less for French).

This result is expected, but our study shows an upper bound on how much of an improvement this strategy could yield.

The figures reported here suppose that the timestamps used in the CAP files accurately reflect the time at which these warnings were issued. When two warnings are issued at the same time (same timestamp), an arbitrary order is used between them.

Some examples of sentences not found in the TM

We show here some of the sentences from the test set that could not be found in the TM using the "chronological" setting explained above. For each one, we give a short comment explaining in part the TM's silence.

English

Sentence not found in TM	Comment
THIS SYSTEM IS GIVING SNOW , AT TIMES HEAVY , TO SOUTHERNMOST NEW BRUNSWICK WHICH WILL CONTINUE INTO THE EVENING .	Unique sentence, quite long. Removing the place name (SOUTHERNMOST NEW BRUNSWICK) does not help.
AFTER A MILD AND BENIGN WEATHER PATTERN , A LOW PRESSURE SYSTEM ORIGINATING FROM THE AMERICAN SOUTHWEST IS EXPECTED TO PROPEL A MIXED BAG OF WINTERY WEATHER INTO SOUTHERN ONTARIO __DAY__ NIGHT INTO __DAY__ .	Unique sentence, quite long. Unusual phrasing 'PROPEL A MIXED BAG OF WINTERY'.
SNOW IS EXPECTED TO PERSIST OVER NORTHERN LABRADOR ON __DAY__ WHERE SIGNIFICANT ACCUMULATIONS ARE POSSIBLE .	First half of the sentence can be found in TM (up to WHERE). As a whole, not in the TM.

RADAR SHOWS THAT AN AREA OF FREEZING RAIN HAS DEVELOPED .	Surprisingly, this is the single occurrence of this sentence. There are plenty of mentions of RADAR and WEATHER RADAR in the TM, but not this particular phrasing.
AS THE WINDS STRENGTHEN BLIZZARD CONDITIONS ARE EXPECTED .	Closest matching sentence in the TM is : AS THE WINDS STRENGTHEN BLIZZARD TO NEAR BLIZZARD CONDITIONS ARE EXPECTED TO DEVELOP .
CONDITIONS ARE EXPECTED TO IMPROVE IN THE AFTERNOON AS THE WINDS DIMINISH .	Closest hit is CONDITIONS ARE EXPECTED TO IMPROVE IN THE AFTERNOON AS WINDS DIMINISH . without the THE.

French

Sentence not found in TM	Comment
EN PLUS DE CETTE BANDE , UNE AUTRE ZONE DE NEIGE FORTE S EST FORMEE LE LONG DES CONTREFORTS SUD ET PERSISTERA PROBABLEMENT JUSQU EN MILIEU DE MATINEE __DAY__ .	Unique sentence, quite long. Removing the place name does not help.
UN SYSTEME FRONTAL DU PACIFIQUE S APPROCHERA DU YUKON AUJOURD HUI .	Unique sentence, even if it is short. Removing the place name (YUKON) does not help the TM. Closest sentence is UN SYSTEME FRONTAL DU PACIFIQUE S APPROCHERA DE LA COTE NORD __DAY__ .
DES VENTS FORTS DU SUD-EST DE __NUM__ A __NUM__ KM / H SE SONT LEVES SUR L ILE DE VANCOUVER NORD ET LA COTE CENTRALE - SECTEURS COTIERS .	Unique sentence, but a close sentence can be found in the TM when ignoring the place names. The closest match is DES VENTS FORTS DU SUD-EST DE __NUM__ A __NUM__ KM / H SE SONT LEVES SUR LES REGIONS DE L ILE DE VANCOUVER NORD VERS LE NORD .
LES AVERSES DE PLUIE DEVRAIENT SE CHANGER EN AVERSES DE NEIGE ET EN BOURRASQUES DE NEIGE PAR ENDROITS TARD CETTE NUIT ET __DAY__ MATIN A MESURE QUE LES TEMPERATURES CHUTERONT DERRIERE UN FRONT FROID .	Not even the segment LES AVERSES DE PLUIE DEVRAIENT SE CHANGER EN AVERSES DE NEIGE is found in the TM.
POUR DE PLUS AMPLES RENSEIGNEMENTS , VISITEZ WWWBCAIRQUALITY.CA.	Typo in the URL prevents TM hit.
UN IMPORTANT CHANGEMENT DU REGIME METEOROLOGIQUE S AMORCERA __DAY__ .	Phrase UN IMPORTANT CHANGEMENT DU REGIME METEOROLOGIQUE is extremely rare. More frequent are UN IMPORTANT CHANGEMENT DES CONDITIONS METEOROLOGIQUES .