Lecture 6.3: quality estimation

Bruno Pouliquen

(with Jeevanthi Liyanapathirana)

What is quality estimation?

- How good/bad is a MT output…
- Estimate a priori the quality of a MT sentence
- Why?
  - For dissemination:
    - A measure for post editing effort
    - Discard very bad MT
  - For assimilation:
    - Show some warning when the translation is supposed to be bad

Challenging…

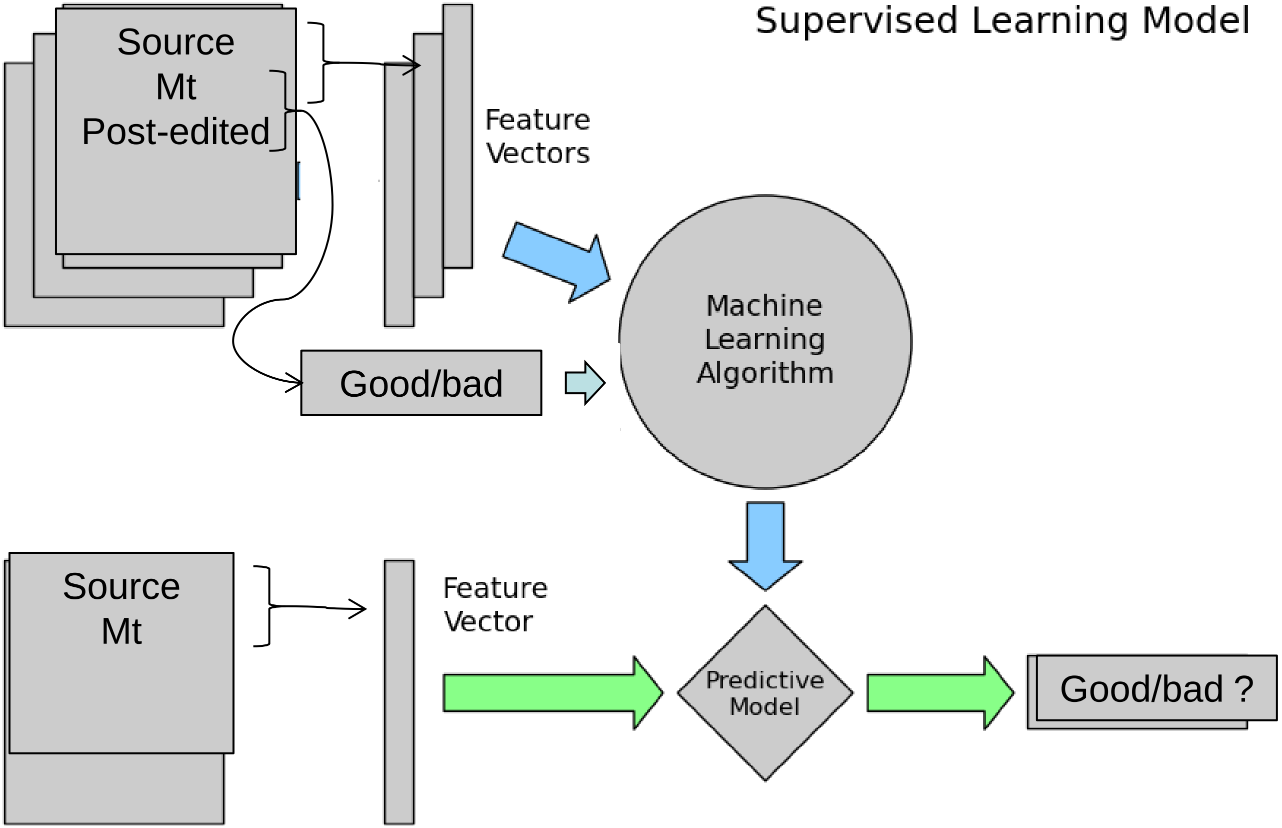
- A machine has difficulties to translate
- A machine has difficulties to evaluates quality of MT translation
- Human reference are the gold standard
  - Many ways to translate
- QE must decide on good or bad?
- QE must decide on quality scale (1-5)?
- QE as regression classification (between 0 and 1)?
- QE for ranking? Decide between various MTs, between n-best-lists?

How?

For measuring post-editing effort:

- “Learn” from post-edited sentences (PE)
- Use (supervised) machine learning on source/MT/PE
- Measure the post-editing effort for each training example
- Estimate the effort on a new source/MT

How?


Supervised Learning Model
Source
Mt
Post-edited
Feature
Vectors
Good/bad
Machine
Learning
Algorithm
Source
Mt
Feature
Vector
Predictive
Model
Good/bad ?

Examples

Original English	***Machine Translation (French)***	***Post edited***
food processing treatment	traitement de traitement des aliments	traitement d'aliments transformés
processing machines for local specialty products	machines de traitement pour produits locaux spéciaux	machines pour la transformation de produits locaux spéciaux
cooked foods (treatment of -)	aliments cuits (traitement d’)	aliments cuits (traitement d')

Can we “learn” that “processing”/”traitement” is usually badly translated with our MT?

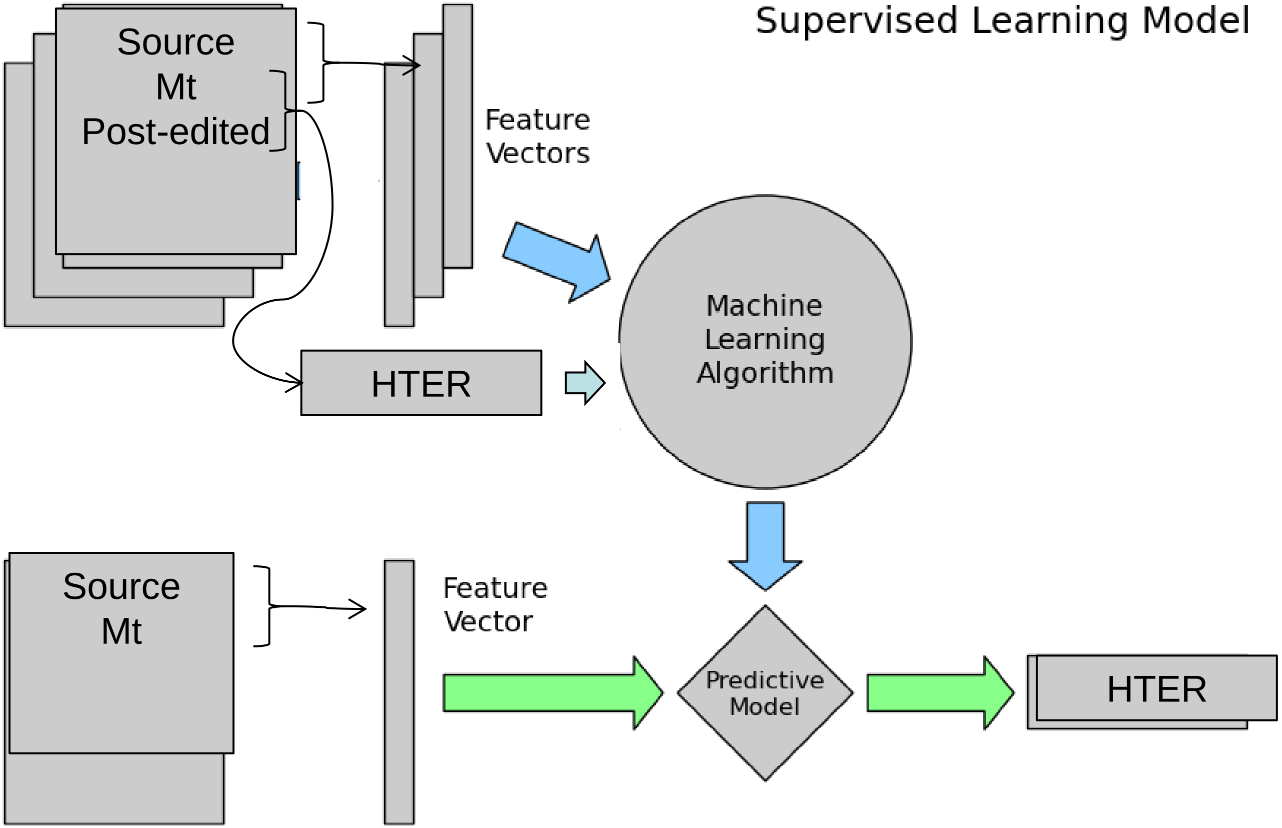
machines for processing of filter material for tobacco products	machines de traitement de matériaux filtrants pour produits du tabac	machines pour la transformation de matières filtrantes pour produits de tabac

How?

Regression model:

- Extract features from source & MT
- Compute HTER between training MT&PE
- Use machine learning to predict HTER

How?


Source
Mt
Post-edited
Supervised Learning Model
Feature
Vectors
HTER
Machine
Learning
Algorithm
Source
Mt
Feature
Vector
Predictive
Model
HTER

Feature extraction

- Extract features out of the English source input and French MT output
- QUEST Framework

  Extracts a number of sentence-level features (and a few word-level features)
- (in our WIPO experiment) 50 features, for the baseline system
- Up to 80 features using linguistic parsers (English and French)

Resources for feature extraction

E.g. we can extract features

- From Moses SMT engine
- French : (big) ngram count file, language model, lexical translation table
- English : Language model, English training corpus
- Syntactic parsers for English and French (available in Quest)

Early experiment example: in WIPO

- Training data: 76620 instances
- Good or bad : HTER > 0.3

Label type	Number of instances
Good Label	46652
Bad Label	29968

- With these data, we trained a classifier and checked accuracy

Experiment example: in WIPO

MT Algorithm	Accuracy
Random Choice	50%
Majority Class	60%
Random Forest (selected features)	**71.5%**
Support Vector Machine	63%
IBK	64.5%
Decision Table	67%

QE: some conclusion

- Hard problem
- It is often as hard to estimate the quality of MT than to produce MT itself
-

bibliography

- QE shared task: (WMT12-WMT16)
  - http://www.statmt.org/wmt16/quality-estimation-task.html
- Lucia Specia and Carolina Scarton, QE tutorial (Jan 2016) http://staffwww.dcs.shef.ac.uk/people/C.Scarton/resources/slides_tutorial.pdf
- Quality estimation for machine translation: some lessons learned (Guillaume Wisniewski, Anil Kumar Singh, François Yvon), In Machine Translation, Springer Netherlands, volume 27, 2013.