# Online Learning for CAT applications

### Nicolò Cesa-Bianchi Gabriele Reverberi

Università degli Studi di Milano



# S M A R T

Statistical Multilingual Analysis for Retrieval and Translation

### What is SMART about





### Interactive Machine Translation

- Learning/optimization techniques are used to tune the parameters of SMT systems
- Online learning adjusts parameters incrementally [Lian et al., 2006; Arun and Koehn, 2007; Tillman and Zhang, 2008]
- Especially useful when the system interacts with the user



# Computer Assisted Translation (CAT)

-						
	IT Translation 📃 🗆 🔀					
	Workbench					
:( (*	₹ (Õ (⇔ (∅) ≥) )					
studie techn Prevaj	urpose of this report is to establish the scenarios which will be evaluated in the three case is which the SMR troject and to detail the requirements of the case studies towards the ical work packages (both in terms of required functionality and integration related issues). alm spornin je sestavljen iz delov besedila iz izvornega jezika in njihovih prevodov v enega ali vec jezikov					
	nslation memory consists of text segments in a source language and their translations into one re target languages.					
These segments can be blocks, paragraphs, sentences, or phrases. A translator first supplies a source text (that is, a text to be translated) to the translation memory.						
	The program will then analyze the text, trying to find segment pairs in its translation memory where the text in the new source segment matches the text in the source segment in a previously					
100%	match from Translation Memory:					
Source:	The purpose of this report is to establish the accompany which will be evaluated in the three case studies within the SMART project and to detail the requirements of the case studies towards the technical work packages (both in terms of required functionality and integration related issues).					
Target:	Prevajalni spomin je sestavljen iz delov besedila iz izvomega jezika in njihovih prevodov v enega ali vec ciljnih jezikov.					
	Trades Translator's Workbench - project - English (United Kingdom) -> Slovenian					
The purpose of this report is to establish the scenarios which will be evaluated in the three case						
studies within the SMART project and to detail the requirements of the case studies towards the						
pro fun	purpose of this report is to establish the scenarios which will be evaluated in the three case studies within the SMAPT jest and to detail the requirements of the case studies towards the technical work packages (both in terms of required clinical) and integration related issues).					
Pre	зкајани зротни је зезначуви из ценоч цезецна из изкотнеда језика in njinovin prevodov v enega ali vec ciljnin језикоv.					
<	100% > Exact Match Match 1 of 1					



# CAT meets online learning





### Adaptive decoding [Liang, Bouchard-Côté, Klein, and Taskar, 2006]



# Experimental setup (based on Portage SMT system)

Feature set for online weights

A new feature is created for each phrasetable entry



# Experimental setup (based on Portage SMT system)

### Feature set for online weights

A new feature is created for each phrasetable entry

#### Phase 1 – offline mode

- Building of phrasetable on a training corpus
- Tuning of loglinear weights on a development corpus
- $\rightarrow$  This gives the **baseline** system



# Experimental setup (based on Portage SMT system)

#### Feature set for online weights

A new feature is created for each phrasetable entry

#### Phase 1 – offline mode

- Building of phrasetable on a training corpus
- Tuning of loglinear weights on a development corpus
- $\rightarrow$  This gives the **baseline** system

#### Phase 2 – online mode

Online weights are adapted during CAT process

# Adaptive decoding — basic definitions

- **f**(x<sub>t</sub>, y) is the vector of **phrasetable feature values** when considering y as candidate translation for the source sentence x<sub>t</sub>
- The vector *w* contains the decoder online weights
- The decoder builds a N-best list  $Y_t$  of candidate translations y by ranking them according to margin

 $\mathbf{w}^{\top}\mathbf{f}(\mathbf{x}_{t},\mathbf{y})$ 



# Adaptive decoding — basic definitions

- **f**(x<sub>t</sub>, y) is the vector of **phrasetable feature values** when considering y as candidate translation for the source sentence x<sub>t</sub>
- The vector *w* contains the decoder online weights
- The decoder builds a N-best list  $Y_t$  of candidate translations y by ranking them according to margin  $w^{\top} f(x_t, u)$

 The 1-best translation is ŷ<sub>t</sub> = argmax w<sup>T</sup>f(x<sub>t</sub>, y) y∈Y<sub>t</sub>
The pseudo-target translation is y<sup>\*</sup><sub>t</sub> = argmax BLEU(y<sub>t</sub>, y) y∈Y<sub>t</sub>



#### Recall:

Decoder ranks translations y according to  $w^{\top} f(x_t, y)$ 

• Margin difference for weight w when y is chosen instead of  $y^*$  $_{MARGIN_t}(y^*, y) = w^{\top} (f(x_t, y^*) - f(x_t, y))$ 

• Linear constraints the learner tries to enforce at each step t  $\begin{array}{l} \text{Margin}_t(y^*,y) \geqslant \text{Bleu}(y_t,y_t^*) - \text{Bleu}(y_t,y) \quad \forall y \in Y_t \end{array}$ 

• Constraints are approximately enforced by projecting current *w* onto (some of the) hyperplanes defined by constraints



### Cost-sensitive margin condition





# Update of parameters

### Recall:

y = reference translation

y\* = pseudo-target translation (highest BLEU in N-best)

- $\hat{y}$  = guessed translation (1-best)
- w = current value of online weights

Enforce margin difference between pseudo-target  $y^*$  and 1-best  $\widehat{y}$ 

$$\min_{\boldsymbol{w}',\boldsymbol{\xi}} \left\| \boldsymbol{w} - \boldsymbol{w}' \right\|^2 + C \,\boldsymbol{\xi}$$

such that  $\operatorname{Margin}(y^*, \widehat{y}) \ge (\operatorname{Bleu}(y, y^*) - \operatorname{Bleu}(y, \widehat{y})) - \xi$ 

#### Passive-aggressive update

[Crammer et al., 2006]

 $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta_t \big( \texttt{Bleu}(\boldsymbol{y}_t, \boldsymbol{y}_t^*) - \texttt{Bleu}(\boldsymbol{y}_t, \widehat{\boldsymbol{y}}_t) \big)$ 

# Theoretical guarantees

### For any sequence $(x_1, y_1), (x_2, y_2), \dots$ of source/reference pairs

• If there exists choice **u** for the parameters that satisfies all constraints at each step, then

$$\sum_{t} \text{Bleu}(\mathbf{y}_{t}, \widehat{\mathbf{y}}_{t}) \geq \sum_{t} \text{Bleu}(\mathbf{y}_{t}, \mathbf{y}_{t}^{*}) - \|\mathbf{u}\|^{2}$$

• If no such **u** exists, then  $\sum_{BLEU}(y_t, \hat{y}_t)$  is at least

$$\sum_{t} \text{Bleu}(y_t, y_t^*) - \inf_{u} \left(1 + \frac{1}{C}\right) \left( \|u\|^2 + C \sum_{t} H_t(u) \right)$$

- C is the aggressiveness parameter associated with the constraints
- H<sub>t</sub>(**u**) measures by how much the margin of **u** fails the worst constraint at time t

### Performance measure

- Learning algorithms and their analysis do not require **BLEU**
- For robustness reasons, we train and test the system using **BLEUMIX**, an average of different sentence-level measures (1 **BLEUMIX**  $\approx 0.65$  **BLEU**)



### Performance measure

- Learning algorithms and their analysis do not require **BLEU**
- For robustness reasons, we train and test the system using **BLEUMIX**, an average of different sentence-level measures (1 **BLEUMIX**  $\approx 0.65$  **BLEU**)

#### Cumulative **BLEUMIX** difference

The cumulative difference in sentence-level BLEUMIX points between online system translations  $\hat{y}_t$  and Portage baseline translations  $y'_t$  with respect to the common reference translation  $y_t$ 

$$\sum_{t=1}^{T} \left( \text{Bleumix}(y_t, \widehat{y}_t) - \text{Bleumix}(y_t, y_t') \right)$$



- Corpus: English  $\rightarrow$  Spanish section of Europarl
- Training set: 165,000 sentences
- Dev set: (used to tune Portage) 6,000 sentences
- Test set: (used for online learning) Five adjacent nonoverlapping blocks of 1,000 sentences each



- Online learner attempts to improve on tuned Portage performance by a *single run* over 1,000 sentences
   → less than 0.6% of Portage training set!
- Learner does so by simultaneously tuning 1,7M parameters associated with the phrasetable entries
  → about 1,700 parameters per observed sentence!
- We get an improvement of about 0.4 BLEUMIX points per observed sentence



# Weight adaptation





N. Cesa-Bianchi and G. Reverberi (UNIMI)

# Oracolar phrasetable adaptation

Dynamic growth of phrasetable

Problem: on-the-fly alignment of new segments



### Dynamic growth of phrasetable

Problem: on-the-fly alignment of new segments

#### Oracolar PT

- Fake alignment by building an oracolar PT on train + test corpora
- After translating each new sentence, the relevant segments are moved from the oracolar PT to the working PT
- The weights associated with new segments are incrementally learned



### Weight adaptation + PT adaptation





### Nonparametric randomized test

### [Riezler and Maxwell III, 2005]

- We estimate the probability p that the performance difference increases when each translation in turn is obtained from a random system (adaptive or baseline)
- This is a p-value for the null hypothesis that baseline and adaptive have the same performance

p-values					
0.01	0.28	0.33	0.18	0.45	
0.01	0.40	0.20	0.13	0.41	



### Weight adaptation — 5 runs





N. Cesa-Bianchi and G. Reverberi (UNIMI)

20/22

# Weight adaptation + PT adaptation — 5 runs





- More stable learning curves
- On-the-fly alignment to replace oracolar PTT
- TM's crippling effect

