Seminar at LIMSI, Spoken Language Processing Group
Paris, 20 June 2014

Online Learning for Statistical Machine Translation

Daniel Ortiz Martínez

Pattern Recognition and Human Language Technologies Group

Universitat Politècnica de València

Outline

Outline

Motivation

- Translation needs have dramatically increased during the last years
- Data scarcity in restricted domains is a major problem for SMT systems
- Training data is inherently generated in many real translation scenarios
- Conventional SMT systems cannot deal with such new data (retraining)
- Online learning can be used to efficiently update the statistical models

Statistical Machine Translation

- The statistical approach to MT [Brown et al., 1993] is as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \{Pr(\boldsymbol{y}|\boldsymbol{x})\}$$

- State-of-the-art SMT systems follow a log-linear approach [Och and Ney, 2002]:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \left\{ \max_{\mathbf{a}} \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) \right\}$$

(**a** is the hidden alignment variable introduced by the translation models)

Phrase-Based Translation

- Current SMT systems are focused on phrase models
- Generative process of phrase-based translation:
  1. Segment the source sentence into K phrases
  2. Translate each source phrase into a target phrase
  3. Reorder the translated target phrases
- A bisegmentation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is determined: $(\tilde{x}_1^K, \tilde{y}_1^K, \tilde{a}_1^K)$

Post-Editing and Interactive Machine Translation

- SMT allows us to translate a source text without human intervention
- Unfortunately, SMT results are not error-free
- SMT system output can be supervised to obtain high-quality translations
- Two SMT applications allow users to collaborate with the system:
  - Post-editing (PE): sequential collaboration
  - Interactive Machine Translation (IMT): interactive collaboration
- PE and IMT try to increase the productivity of translation companies

Interactive Machine Translation

- IMT can be seen as an evolution of the SMT framework
- In the IMT scenario, we have to find a suffix $\boldsymbol{s}$ for a given prefix $\boldsymbol{p}$ plus the next key-stroke k introduced by the user:

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} \left\{ p(\boldsymbol{s} \mid \boldsymbol{x}, \boldsymbol{p}, k) \right\}$$

- Search is restricted to those sentences $\boldsymbol{y}$ containing $\boldsymbol{p}$ plus k as prefix
- Following the log-linear approach we obtain:

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} \left\{ \max_{\mathbf{a}} \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) \right\}$$

(note that $\boldsymbol{y} \equiv \boldsymbol{p}k\boldsymbol{s}$)

IMT Example

source($\boldsymbol{x}$): Para ver la lista de recursos
reference($\hat{\boldsymbol{y}}$): To view a listing of resources

interaction-0	***p***						
	s	*To*	*view*	*the*	*resources*	*list*	
interaction-1	***p***	To	view				
	k			**a**			
	s				*list*	*of*	*resources*
interaction-2	***p***	To	view	a	list		
	k				**i**		
	s				*ng*	*resources*	
interaction-3	***p***	To	view	a	listing		
	k					**o**	
	s					*f*	*resources*
acceptance	***p***	To	view	a	listing	of	resources

Outline

Online Learning

Concept

- Appropriate in those learning tasks in which learning must take place over time
- Examples are not available a priori but become available over time, usually one at a time
- Learning may need to go on indefinitely
- Online learning is opposed to batch learning, where there is a finite set of examples that are available a priori

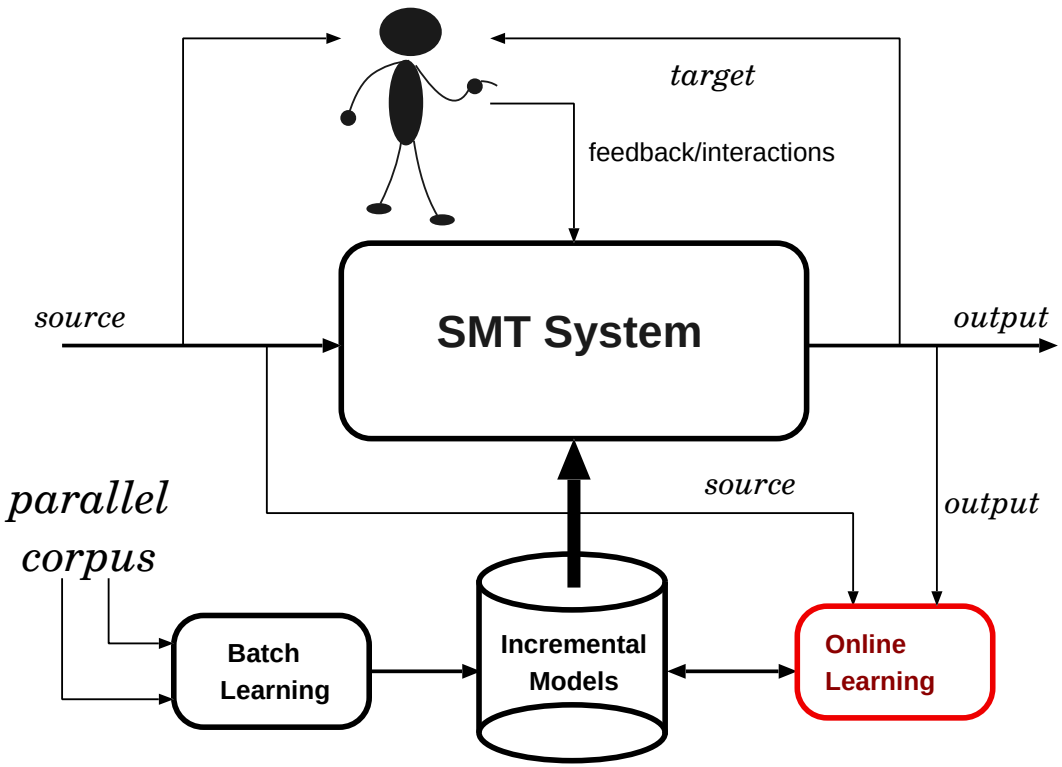
Main Features of Online Learning

- No re-processing of previous samples is required.
- The learner can, at any time, produce an answer to a query
- The quality of the answers improves over time
- All the data relevant to the problem is not available a priori

Outline

SMT and Online Learning

- Online learning fits naturally in (but is not restricted to) PE and IMT applications



Basic SMT System

- We use a log-linear model composed of seven feature functions:
  - $h_1(\boldsymbol{y}) = \log(\prod_{i=1}^{|\boldsymbol{y}|+1} p(y_i|y_{i-n+1}^{i-1}))$ **Language model**
  - $h_2(\boldsymbol{y}, \boldsymbol{x}) = \log(p(|\boldsymbol{x}|\,||\boldsymbol{y}|))$ **Sentence length model**
  - $h_3(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log(\prod_{k=1}^{K} p(\tilde{x}_k|\tilde{y}_{\tilde{a}_k}))$ **Inverse translation model**
  - $h_4(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log(\prod_{k=1}^{K} p(\tilde{y}_{\tilde{a}_k}|\tilde{x}_k))$ **Direct translation model**
  - $h_5(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log(\prod_{k=1}^{K} p(|\tilde{y}_k|))$ **Target phrase length model**
  - $h_6(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log(\prod_{k=1}^{K} p(|\tilde{x}_k|\,||\tilde{y}_{\tilde{a}_k}|))$ **Source phrase length model**
  - $h_7(\mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{a}_k|\tilde{a}_{k-1}))$ **Distortion model**

Learning from New Sentence Pairs

- Given a new sentence pair $(\boldsymbol{x}, \boldsymbol{y})$, the log-linear model is updated
- To do this, a set of *sufficient statistics* that can be incrementally updated is maintained for each feature function $h_i(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x})$
- In this presentation we will focus on the sufficient statistics for the language (h_1) and translation models (h_3 and h_4):

$$h_1(\boldsymbol{y}) = \log\left(\prod_{i=1}^{|\boldsymbol{y}|+1} p(y_i|y_{i-n+1}^{i-1})\right)$$

$$h_3(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log\left(\prod_{k=1}^{K} p(\tilde{x}_k|\tilde{y}_{\tilde{a}_k})\right)$$

(h_4 is defined analogously to h_3)

For a more detailed explanation see [Ortiz-Martínez et al., 2010, Ortiz-Martínez, 2011]

Incremental Language Model (h_1)

- An n-gram language model with interp. Kneser-Ney smoothing is used:

$$p(y_i|y_{i-n+1}^{i-1}) = \frac{\max\{c_X(y_{i-n+1}^{i}) - D_n, 0\}}{c_X(y_{i-n+1}^{i-1})} + \frac{D_n}{c_X(y_{i-n+1}^{i-1})} N_{1+}(y_{i-n+1}^{i-1}\bullet) \cdot p(y_i|y_{i-n+2}^{i-1})$$

  where:
  - $D_n = \frac{c_{n,1}}{c_{n,1}+2c_{n,2}}$ (fixed discount)
  - $N_{1+}(y_{i-n+1}^{i-1}\bullet)$ (number of unique words that follows the history y_{i-n+1}^{i-1})
  - $c_X(y_{i-n+1}^{i})$ ($c_X(\cdot)$ can represent true $c_T(\cdot)$ or modified $c_M(\cdot)$ n-gram counts)
- Sufficient statistics: $c_{k,1}$, $c_{k,2}$, $N_{1+}(\cdot)$, $c_T(\cdot)$, $c_M(\cdot)$
- The set of sufficient statistics are updated using an appropriate algorithm

Incremental Inverse Translation Model (h_3)

- We use a smoothed phrase-based translation model:

$$p(\tilde{x}_k|\tilde{y}_{\tilde{a}_k}) = \beta p_{phr}(\tilde{x}_k|\tilde{y}_{\tilde{a}_k}) + (1-\beta)p_{hmm}(\tilde{x}_k|\tilde{y}_{\tilde{a}_k})$$

$p_{phr}(\cdot) \rightarrow$ statistical phrase-based dictionary
$p_{hmm}(\cdot) \rightarrow$ HMM-based alignment model

- Phrase model probabilities are estimated from phrase counts:

$$p(\tilde{x}|\tilde{y}) = \frac{c(\tilde{x},\tilde{y})}{\sum_{\tilde{x}'} c(\tilde{x}',\tilde{y})}$$

- Standard estimation procedures use word alignment matrices to extract phrase counts

Incremental Inverse Translation Model (h_3)

- HMM models are used here for:
  - smoothing
  - generating word alignment matrices
- Estimation of HMM models is based on the EM algorithm
- **Problem**: standard EM algorithm requires to retrain the whole training set when a new training pair is available
- **Solution**: use incremental EM algorithm to train the HMM models
- The sufficient statistics are a set of expected counts collected after the presentation of a new training pair

Online Learning of Log-Linear Weights

- Log-linear weights λ_m can also be updated using online learning
- Discriminative ridge regression (DRR) technique is used
- Good hypotheses within a *n*-best list score higher, bad hypotheses lower
- Establish correlation between difference in translation quality and difference in score
- Find $\check{\boldsymbol{\lambda}}_m$ such that $\mathbf{R}_{\boldsymbol{x}} \cdot \check{\boldsymbol{\lambda}}_m \propto \mathbf{l}_{\boldsymbol{x}}$, with
  - $\mathbf{R}_{\boldsymbol{x}}$ difference of scores between every $\boldsymbol{y} \in$ *n*-best and best hypothesis
  - $\mathbf{l}_{\boldsymbol{x}}$ difference in quality between every $\boldsymbol{y} \in$ *n*-best and best hypothesis

See [Martínez-Gómez et al., 2012] for more details

Outline

The Open-Source Thot Toolkit

- The Thot toolkit [Ortiz-Martínez and Casacuberta, 2014] implements a fully fledged phrase-based SMT system
- Released under LGPL license
- Can be downloaded from `https://github.com/daormar/thot/`
- Implements incremental estimation of:
  - *n*-gram language models
  - HMM-based single-word alignment models
  - Phrase-based models
- Used to carry out the experiments reported here
- One of the CASMACAT project SMT engines (`www.casmacat.eu`)

Outline

Corpus

- We will show experiments using the Xerox and the Europarl corpus
- The Xerox task consists on the translation of a set of printer manuals:

		English	Spanish	English	French	English	German
Training	Sentences	55 761		52 844		49 376	
	Running words	571 960	657 172	542 762	573 170	506 877	440 682
	Vocabulary	25 627	29 565	24 958	27 399	24 899	37 338
Dev	Sentences	1 012		994		964	
	Running words	12 111	13 808	9 480	9 801	9 162	8 283
Test	Sentences	1125		984		996	
	Running words	7 634	9 358	9 572	9 805	10 792	9 823

- The Europarl task consists on the translation of parliamentary proceedings:

		English	Spanish	English	French	English	German
Training	Sentences	1 547 596		1 525 315		1 601 936	
	Running words	33 125 K	34 116 K	31 923 K	34 571 K	36 185 K	34 070 K
	Vocabulary	96 741	146 288	95 232	114 417	101 113	318 475
Dev	Sentences	3 003		3 003		3 003	
	Running words	72 988	78 888	72 988	81 800	72 988	72 603
Test	Sentences	3 000		3 000		3 000	
	Running words	64 809	70 562	64 809	73 664	64 809	63 411

Experiments

Evaluation Methodology

- Our proposals were evaluated using:
  - BLEU score [Papineni et al., 2002]
  - Key-stroke and mouse-action ratio (KSMR) measure: effort required from the user to generate the target translations [Barrachina et al., 2009]

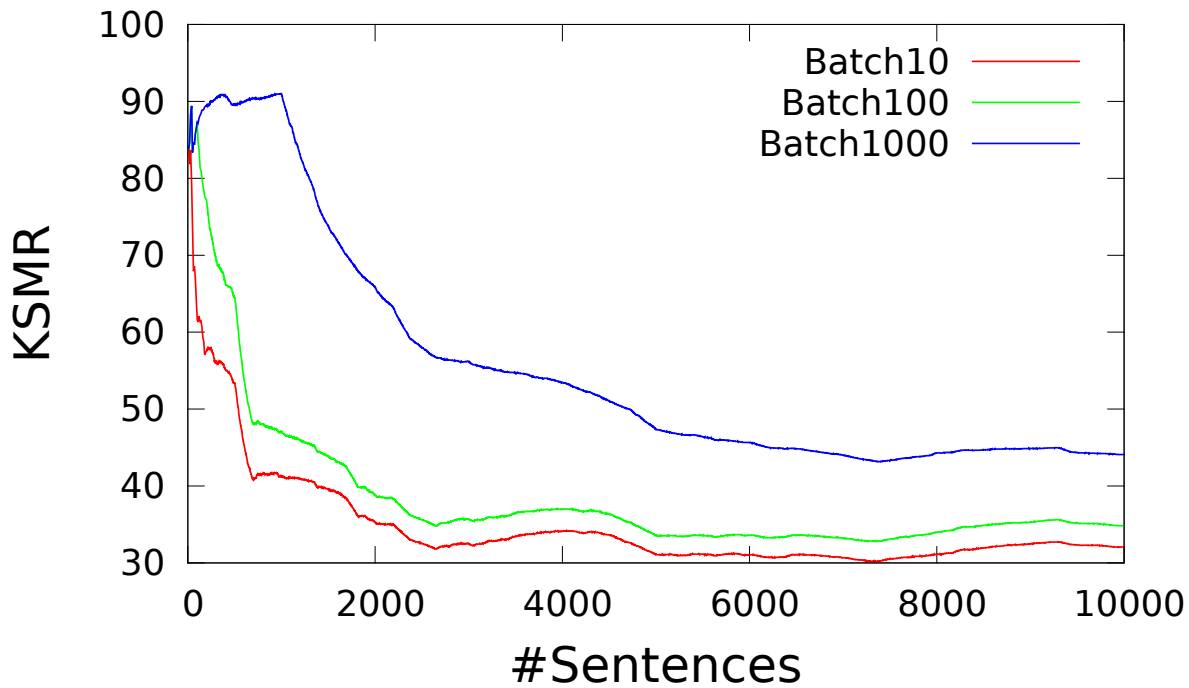
- Three different experimentation scenarios:
  1. Measuring the importance of frequent updates
  2. Online versus batch learning performance
  3. Learning from previously estimated models

Results: Measuring the Importance of Frequent Updates (I)

- Is it really important to perform frequent model updates?
  - If the answer is no, the training process can be executed overnight
  - Otherwise, online learning techniques are required
- A conventional (batch) SMT system without any preexistent model stored in memory (learning from scratch) was used
- After translating a certain number of sentences, the system is retrained using the whole corpus
- Here we only show the obtained IMT results

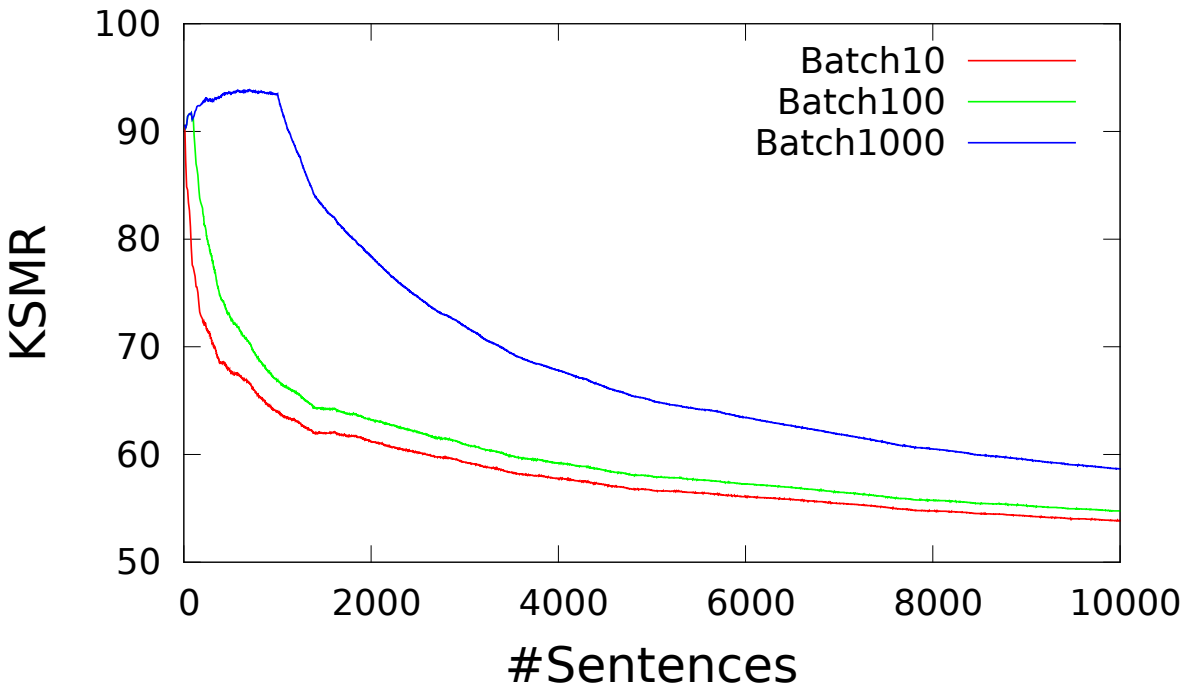
Results: Measuring the Importance of Frequent Updates (II)

- The first 10 000 sentences of the English-Spanish Xerox training corpus were interactively translated
- Three different retraining frequencies were tested (10, 100 or 1000 sentences)



Results: Measuring the Importance of Frequent Updates (III)

- English-Spanish Europarl results:

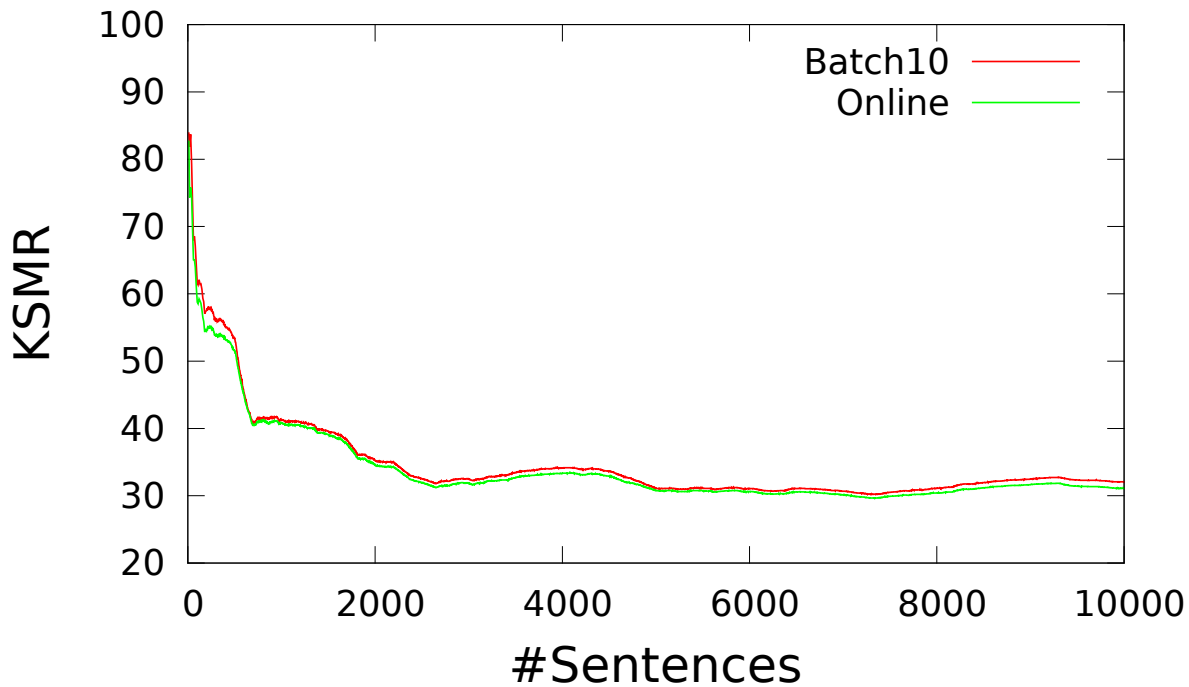


Results: Online versus Batch Learning Performance (I)

- Is online learning able to deliver the same performance as batch learning?
  - If the answer is no, then we need to reduce the time cost of batch learning
  - Otherwise, online learning should be used
- When processing a new sample:
  - Batch learning processes all previously seen samples (non-constant time)
  - Online learning processes only the last one (constant time)

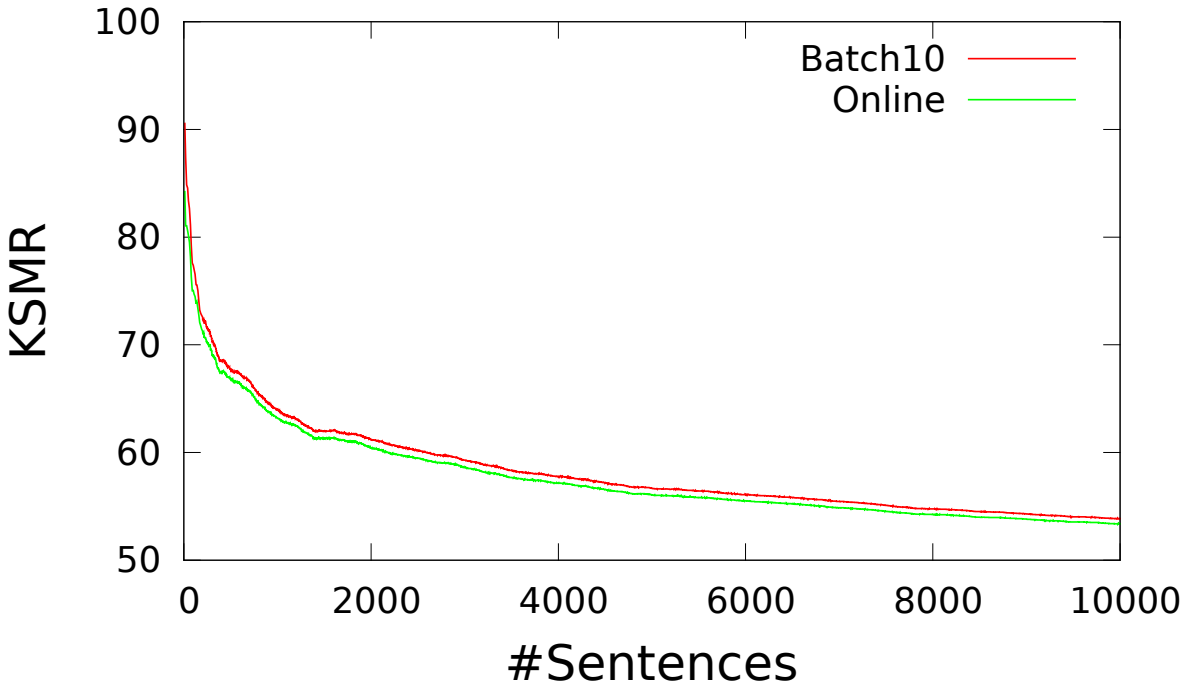
Results: Online versus Batch Learning Performance (II)

- The first 10 000 sentences of the English-Spanish Xerox training corpus were interactively translated
- Online learning system was compared with a batch system retrained every ten sentences



Results: Online versus Batch Learning Performance (III)

- English-Spanish Europarl results:



Results: Learning from Previously Estimated Models

- English-Spanish Xerox and Europarl results are shown
- Both systems were initialized with a log-linear model trained in batch mode by means of the Xerox and Europarl training corpora

	SMT system	BLEU	KSMR	LT (s)
Xerox English-Spanish	conventional	58.3± 2.4	19.3± 1.2	-
	online	64.0± 2.4	16.6± 1.1	0.06
Europarl English-Spanish	conventional	21.0± 0.5	45.9± 0.4	-
	online	22.5± 0.6	44.7± 0.4	0.2

- The improvements were statistically significant
- Learning times (LT) allow the system to be used in a real-time scenario

Outline

Conclusions

- An SMT system with online learning have been proposed
- The greater the update frequency of the system, the better the results
- Online learning performance is comparable to that of batch learning
- Training times allow the system to be used in a real time scenario
- Empirical results clearly show the utility of online learning in PE and IMT
- Strong potential when new training data becomes periodically available

Future Work

- Current proposal is not able to give more importance to in-domain data
- One possible solution is to combine a static out-of-domain translation table with an online in-domain model
- Incorporate bounds to data structures (incoming data is unbounded)
- Application of online sequential monte-carlo algorithm [Cappé, 2009]
- Continuous space models are particularly appropriate for online learning

References I


Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., and Vidal, E. (2009).
Statistical approaches to computer-assisted translation.
Computational Linguistics, 35(1):3–28.


Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., and Mercer, R. L. (1993).
The mathematics of statistical machine translation.
Computational Linguistics, 19(2):263–313.


Cappé, O. (2009).
Online sequential monte carlo em algorithm.


Giraud-Carrier, C. (2000).
A note on the utility of incremental learning.
AI Communications, 13(4):215–223.


Martínez-Gómez, P., Sanchis-Trilles, G., and Casacuberta, F. (2012).
Online adaptation strategies for statistical machine translation in post-editing scenarios.
Pattern Recognition, 45(9):3193–3203.

References II

Och, F. J. and Ney, H. (2002).
Discriminative training and maximum entropy models for statistical machine translation.
In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Ortiz-Martínez, D. (2011).
Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation.
PhD thesis, Universidad Politécnica de Valencia.
Advisors: Ismael García Varea and Francisco Casacuberta.

Ortiz-Martínez, D. and Casacuberta, F. (2014).
The new thot toolkit for fully automatic and interactive statistical machine translation.
In *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, pages 45–48, Gothenburg, Sweden.

References III

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010).
Online learning for interactive statistical machine translation.
In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554, Los Angeles, California. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
BLEU: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Thank you for your attention!