

Incremental Learning for Statistical Machine Translation

Daniel Ortiz Martínez

Pattern Recognition and Human Language Technologies Group

Universitat Politècnica de València

December 3, 2012



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
4. Experiments
5. Conclusions
6. Demo System

1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
4. Experiments
5. Conclusions
6. Demo System

Motivation

- ▶ Translation needs have dramatically increased during the last years
- ▶ The high demand of translations cannot be satisfied by linguists
- ▶ Statistical machine translation (SMT) can help to alleviate the problem
- ▶ SMT output is not error free, but can be supervised by the user
- ▶ *User feedback can be used to update the SMT system*

Statistical Machine Translation

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

$$\hat{\mathbf{y}} = \arg \max_y \{Pr(\mathbf{y}|\mathbf{x})\}$$

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

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

(\mathbf{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 \mathbf{x} and \mathbf{y} is determined: $(\tilde{\mathbf{x}}_1^K, \tilde{\mathbf{y}}_1^K, \tilde{\mathbf{a}}_1^K)$

Post-Editing and Interactive Machine Translation

- ▶ SMT allows us to translate a given source text without human intervention
- ▶ Unfortunately, SMT results are not error-free
- ▶ Output of SMT systems 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 applications are oriented 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 \mathbf{s} for a given prefix \mathbf{p} plus the next key-stroke k introduced by the user:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} \{p(\mathbf{s} | \mathbf{x}, \mathbf{p}, k)\}$$

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

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

(note that $\mathbf{y} \equiv \mathbf{pks}$)

IMT Example

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

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

1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
4. Experiments
5. Conclusions
6. Demo System

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
- ▶ Incremental learning is opposed to batch learning, where there is a finite set of examples that are available a priori

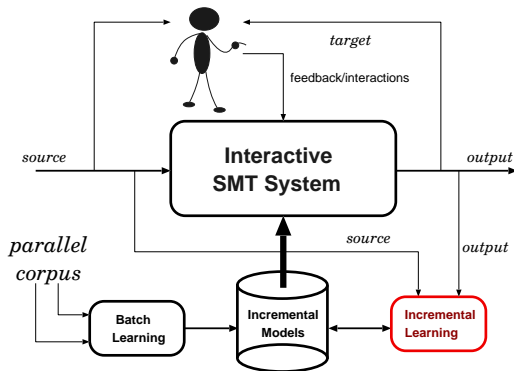
Incremental Learning Algorithms

- ▶ Main Features [Giraud-Carrier, 2000]:
 - 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
- ▶ Important issues [Giraud-Carrier, 2000]:
 - **Ordering effects:** Chronology is an inherent aspect of incrementality
 - **Learning curve:** An incremental system may start from scratch and gain knowledge from examples given one at a time over time. As a result, the system experiences a sort of learning curve
 - **Open-world assumption:** All the data relevant to the problem at hand is not available a priori, there is a need for special learning mechanisms that invalidate portions of knowledge

1. Introduction
2. Incremental Learning
- 3. Incremental Learning for SMT**
4. Experiments
5. Conclusions
6. Demo System

SMT and Incremental Learning

- ▶ Incremental learning is not appropriate in a fully-automatic SMT scenario
- ▶ By contrast, incremental learning fits naturally in PE and IMT



Basic SMT System

- ▶ We use a log-linear model composed of seven feature functions:
 - $h_1(\mathbf{y}) = \log(\prod_{i=1}^{|\mathbf{y}|+1} p(y_i | y_{i-n+1}^{i-1}))$ **Language model**
 - $h_2(\mathbf{y}, \mathbf{x}) = \log(p(|\mathbf{x}| | |\mathbf{y}|))$ **Sentence length model**
 - $h_3(\mathbf{y}, \mathbf{a}, \mathbf{x}) = \log(\prod_{k=1}^K p(\tilde{x}_k | \tilde{y}_{\tilde{a}_k}))$ **Inverse translation model**
 - $h_4(\mathbf{y}, \mathbf{a}, \mathbf{x}) = \log(\prod_{k=1}^K p(\tilde{y}_{\tilde{a}_k} | \tilde{x}_k))$ **Direct translation model**
 - $h_5(\mathbf{y}, \mathbf{a}, \mathbf{x}) = \log(\prod_{k=1}^K p(|\tilde{y}_k|))$ **Target phrase length model**
 - $h_6(\mathbf{y}, \mathbf{a}, \mathbf{x}) = \log(\prod_{k=1}^K p(|\tilde{x}_k| | |\tilde{y}_{\tilde{a}_k}|))$ **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 (\mathbf{x}, \mathbf{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 $h_i(\mathbf{y}, \mathbf{a}, \mathbf{x})$
- ▶ In this presentation we will focus on the sufficient statistics for the language (h_1) and translation models (h_3 and h_4):

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

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

(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 inverse phrase-based translation model:

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

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

- ▶ Inverse 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

Incremental Learning of Log-Linear Weights

- ▶ Log-linear weights λ_m can also be updated using incremental 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 $\check{\lambda}_t$ such that $\mathbf{R}_x \cdot \check{\lambda}_t \propto \mathbf{I}_x$, with
 - \mathbf{R}_x difference of values in \mathbf{h} between every $\mathbf{y} \in n$ -best and best hypothesis
 - \mathbf{I}_x difference in quality between every $\mathbf{y} \in n$ -best and best hypothesis

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

1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
- 4. Experiments**
5. Conclusions
6. Demo System

Corpora (I)

- ▶ Experiments were carried out using Xerox and EU-TT2 corpora (only Xerox results are shown here)
- ▶ The Xerox task consists on the translation of a set of printer manuals from English to Spanish, French and German

		Spanish	English	French	English	German	English
Training	Sentences	55 761		52 844		49 376	
	Running words	657 172	571 960	573 170	542 762	440 682	506 877
Development	Sentences	1 012		994		964	
	Running words	13 808	12 111	9 801	9 480	8 283	9 162
Test	Sentences	1125		984		996	
	Running words	9 358	7 634	9 805	9 572	9 823	10 792

Corpora (II)

- ▶ The EU-TT2 corpus is extracted from the proceedings of the European Parliament in the same language pairs as the Xerox task
- ▶ EU-TT2 is the corpus used in the demo system

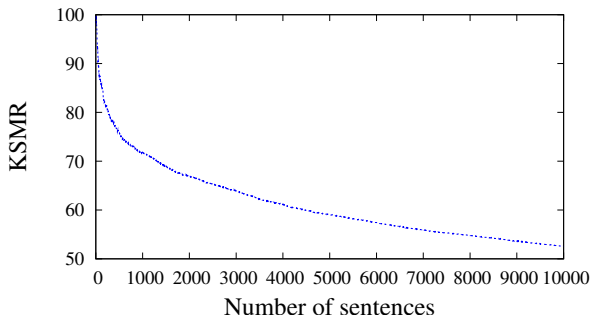
		Spanish	English	French	English	German	English
Training	Sentences	214 473		215 216		222 644	
	Running words	5.8M	5.2M	6.5M	5.9M	6.1M	6.4M
Development	Sentences	400		400		400	
	Running words	11 471	10 080	12 250	11 106	10 730	11 106
Test	Sentences	800		800		800	
	Running words	22 631	19 944	23 945	21 906	20 791	21 906

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]
- ▶ We carried out experiments in two different scenarios:
 1. Experiment using a system without any preexistent model stored in memory (learning from scratch)
 2. Experiment comparing the performance of a batch system with that of an online system. Both systems were initialized with a log-linear model trained in batch mode by means of the XEROX training corpora
- ▶ The search engine is based on the use of partial statistical phrase alignments [Ortiz-Martínez et al., 2009]

Results: Learning from Scratch

- ▶ 10 000 sentences randomly extracted from the English-Spanish Xerox corpus were interactively translated
- ▶ User effort measured in terms of KSMR decreases as the number of interactively translated sentences increases



Results: Learning from Previously Estimated Models

- ▶ Experiments with the English-French Xerox corpus are shown
- ▶ Both systems were initialized with a log-linear model trained in batch mode by means of the Xerox training corpora

	ITP system	BLEU	KSMR	LT (s)
English-French	batch	33.7± 2.0	33.9± 1.3	-
	incremental	42.2± 2.2	27.9± 1.3	0.09

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

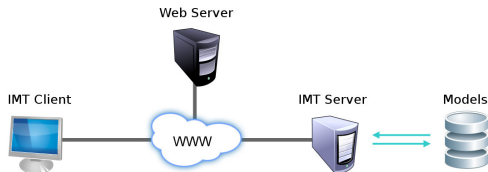
1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
4. Experiments
- 5. Conclusions**
6. Demo System

- ▶ An SMT system with incremental learning have been proposed
- ▶ Such system is able to incrementally update the parameters of the statistical models
- ▶ Training times allow the system to be used in a real time scenario
- ▶ Empirical results clearly show the utility of incremental learning in PE and IMT

1. Introduction
2. Incremental Learning
3. Incremental Learning for SMT
4. Experiments
5. Conclusions
- 6. Demo System**

Main Features

- ▶ Developed within the EU CASMACAT project (www.casmacat.eu)
- ▶ Corpora extracted from the proceedings of the European Parliament was used (EU-TT2)
- ▶ Response times of fractions of a second
- ▶ IMT engine developed in C++ and user interface developed in HTML 5
- ▶ Client-server application:



Demo Description

- ▶ Three modes of operation: PE, standard IMT and incremental IMT
- ▶ Source sentence can be selected from a list
- ▶ Translations are generated using the “Translate” button
- ▶ Initial output sentence can be interactively edited
- ▶ Models can be updated using the “Update” button

Demo System





[Options](#)

Mode: **IMT**. Suggestions: **false**. Confidences: **false [3/40]**. Alignments: **false [matrix: false]**.

Source: * participación de los países candidatos en los programas comunitarios .

* participación de los países candidatos en los programas comunitarios .

* participation of the candidate countries in Community programmes .

-  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.
-  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.



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 Politècnica de Valencia.

Advisors: Ismael García Varea and Francisco Casacuberta.



Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2009).

Interactive machine translation based on partial statistical phrase-based alignments.

In Proceedings of the International Conference Recent Advances in Natural Language Processing.



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!