Workpackage 4
Adaptive Translation Models

Daniel Ortiz-Martínez, Germán Sanchis-Trilles, Francisco Casacuberta

27 November 2012

Index

Introduction

- Tasks in WP4:

Task	Leader	PM	Status
4.1	UPVLC	13	on track
4.2	UPVLC	11	inactive
4.3	UPVLC	13	inactive

- WP4 deals with adaptation in interactive translation prediction (ITP)
- Activity during this first year was limited to Task 4.1
- Task 4.1 studies online learning techniques for ITP
- Initial progress is focused on the online estimation of the parameters of statistical machine translation (SMT) models

Introduction: SMT and ITP

- State-of-the-art SMT systems follow a loglinear approach:

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

  ($\mathbf{a}$ is the hidden alignment variable introduced by the translation models)

- In the ITP 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:

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

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

Online Learning

- Online learning algorithms proceeds in a sequence of trials
- Each trial can be decomposed into three steps:
  1. The learning algorithm receives an instance
  2. The learning algorithm predicts a label for the instance
  3. The true label of the instance is presented to the learning algorithm

SMT and Online Learning

- SMT allows us to translate a given source text without human intervention
- Output of SMT systems can be supervised to obtain high-quality translations
- User feedback can be used to extend the statistical models of the SMT system
- Online learning fits naturally in two well-known SMT applications:
  - Post-editing (PE)
  - ITP
- Online learning can be used to estimate:
  - Parameters of the generative models
  - Scaling factors of the loglinear combination

Task 4.1: Online Learning of SMT Generative Models

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**
  - $h_2(\boldsymbol{y}, \boldsymbol{x}) = \log(p(|\boldsymbol{x}| \mid |\boldsymbol{y}|))$ **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**
  - $h_6(\boldsymbol{y}, \mathbf{a}, \boldsymbol{x}) = \log(\prod_{k=1}^{K} p(|\tilde{x}_k| \mid |\tilde{y}_{\tilde{a}_k}|))$ **Source phrase length model**
  - $h_7(\mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{a}_k|\tilde{a}_{k-1}))$ **Distortion model**

NOTE: Phrase models determine a bisegmentation between $\boldsymbol{x}$ and $\boldsymbol{y}$: $(\tilde{x}_1^K, \tilde{y}_1^K, \tilde{a}_1^K)$

Task 4.1: Online Learning of SMT Generative Models

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 translation models (h_3):

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

Task 4.1: Online Learning of SMT Generative Models

Incremental Inverse Translation Model (h_3)

- We use a smoothed inverse phrase-based translation model:

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

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

Task 4.1: Online Learning of SMT Generative Models

Incremental Inverse Translation Model (h_3)

- HMM models are used here for:
  - smoothing
  - generating word alignment matrices
- We use the 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

Task 4.1: Online Learning of SMT Generative Models

Experiments

- 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
- The EU-TT2 corpus is extracted from the proceedings of the European Parliament in the same language pairs as the Xerox task
- Our proposals were evaluated using:
  - BLEU score
  - *Key-stroke and mouse-action ratio* (KSMR) measure: effort required from the user to generate the target translations

Task 4.1: Online Learning of SMT Generative Models

Results

- **Learning from previously estimated models** (experiments with the English-French Xerox corpus are shown)

	ITP system	BLEU	KSMR	LT (s)
Eng-Fre	batch	33.7± 2.0	33.9± 1.3	-
	online	42.2± 2.2	27.9± 1.3	0.09

- **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

Task 4.1: Online Learning of SMT Scaling Factors
Discriminative Ridge Regression

- Good hypotheses within a n-best list score higher, bad hypotheses lower

- Originally designed for conventional SMT (PE)

- Establish correlation between difference in translation quality and difference in score

- Find $\check{\boldsymbol{\lambda}}_t$ such that $\mathbf{R}_{\boldsymbol{x}} \cdot \check{\boldsymbol{\lambda}}_t \propto \mathbf{l}_{\boldsymbol{x}}$, with
  - $\mathbf{R}_{\boldsymbol{x}}$ difference of values in $\boldsymbol{h}$ between every $\boldsymbol{y} \in n$-best and best hypothesis $\boldsymbol{y}^*$
  - $\mathbf{l}_{\boldsymbol{x}}$ difference in quality between every $\boldsymbol{y} \in n$-best and best hypothesis $\boldsymbol{y}^*$

Task 4.1: Online Learning of SMT Scaling Factors
Modified DRR in ITP

- Several problems need to be tackled

- Metric to be optimised is inherent to a wordgraph
  $\rightarrow$ What are n-best wordgraphs?
  - Consider N random weight-vectors
  - Compute wordgraph associated to each weight-vector
  - Not true list of n-best wordgraphs, but key is to obtain good and bad wordgraphs

- Features should no longer correspond to a single hypothesis, but to a wordgraph
  - Consider the features of the best hypothesis (path) of each wordgraph

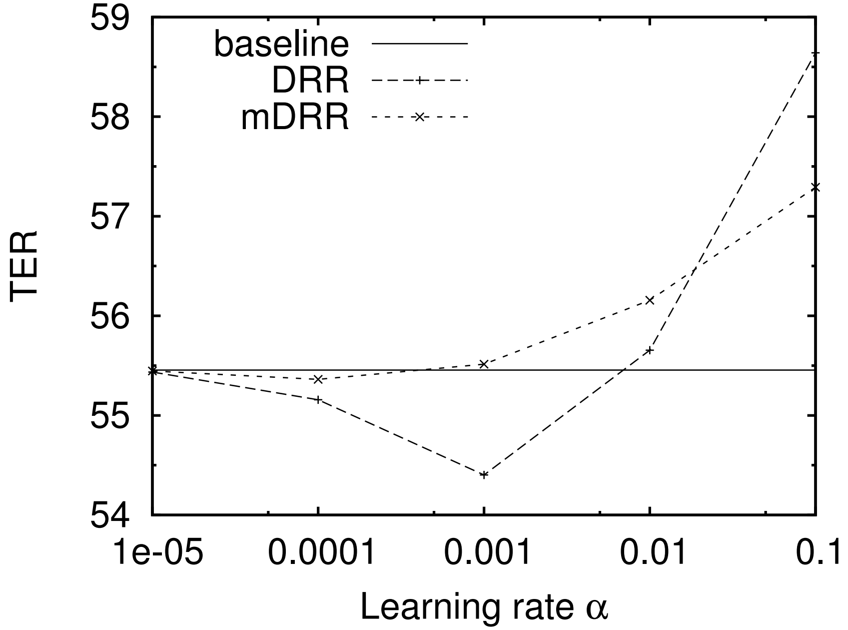
Task 4.1: Online Learning of SMT Scaling Factors

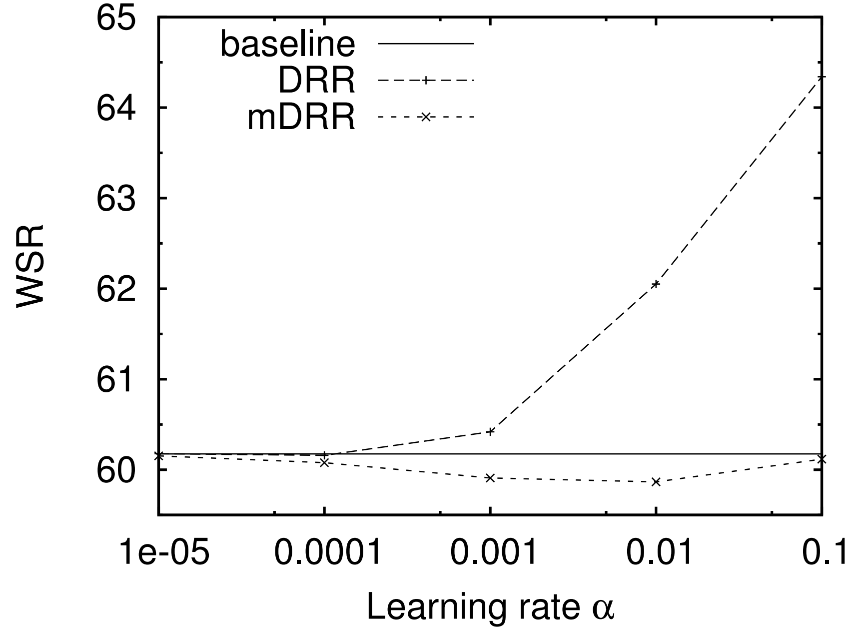
Experiments

- Corpora used:
  - Spanish–English Europarl (training and development)
  - News-Commentary (test)
- Two different scenarios:
  - PE: quality measured with TER
  - ITP: quality measured with WSR
- Baseline with Moses decoder

Task 4.1: Online Learning of SMT Scaling Factors

Results with DRR





- Using TER as translation metric, DRR provides improvements in PE, but not in ITP
- Using modified DRR, improvements in ITP, but not mirrored to PE

Conclusions

- Online learning techniques for SMT have been proposed
- Such techniques allow to incrementally update the parameters of the generative models and their scaling factors
- Estimation of scaling factors in ITP is being studied
- Empirical results clearly show the utility of online learning in SMT and ITP

Future Work

- Test alternative online learning techniques
- Online learning of scaling factors for ITP still poses problems that are to be studied
- Extend experimentation to additional corpora (including the corpora officially used in CASMACAT)

Publications

Publications during the first year of the project:

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

- F. Casacuberta F.J. López-Salcedo, G. Sanchis-Trilles. Online learning of log-linear weights in interactive machine translation. In Proceedings of iberSPEECH, page to be published, 2123 November 2012.

Previous publications:

- Daniel Ortiz-Martínez, Ismael García-Varea, Francisco Casacuberta. Online Learning for Interactive Statistical Machine Translation. Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT), 2010. pp. 546-554.