Generative and Discriminative Methods for Online Adaptation in SMT

K. Wäschle †, P. Simianer †, N. Bertoldi ‡, S. Riezler †,
M. Federico‡

Department of Computational Linguistics, Heidelberg University, Germany †
FBK, Trento, Italy ‡

Outline

Outline

Motivation

- SMT systems usually translate each sentence in a document in isolation $\rightarrow$ context information is lost, translations might be inconsistent
- MT systems in a Computer-Assisted Translation (CAT) framework can benefit from user feedback from the same document $\rightarrow$ confirmed translations should be integrated into the MT engine as soon as they become available

Online learning protocol

Train global model M_g
for all documents d of $|d|$ sentences **do**
 Reset local model $M_d = \emptyset$
 for all examples $t = 1, \ldots, |d|$ **do**
  Combine M_g and M_d into M_{g+d}
  Receive input sentence x_t
  Output translation $\hat{y}_t$ from M_{g+d}
  Receive user translation y_t
  Refine M_d on pair (x_t, y_t)
 end for
end for

M_d has only knowledge of the previous $t-1$ sentences!

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	
8	*Sistemi Informativi* **SpA**	
21	*This document is Sistemi Informativi* **SpA** *'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri**
8	*Sistemi Informativi* **SpA**	
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	*Sistemi Informativi* **SpA**
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	*Sistemi Informativi* **SpA** *Sistemi Informativi* **S.p.A.**
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	*Sistemi Informativi* **SpA** *Sistemi Informativi* **S.p.A.**
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	

- MT hypothesis
- user translation

Example

id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	*Sistemi Informativi* **SpA** *Sistemi Informativi* **S.p.A.**
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	*Questo documento è Sistemi Informativi* **SpA** *di* **Tecnica Offri**

- MT hypothesis
- user translation

Example

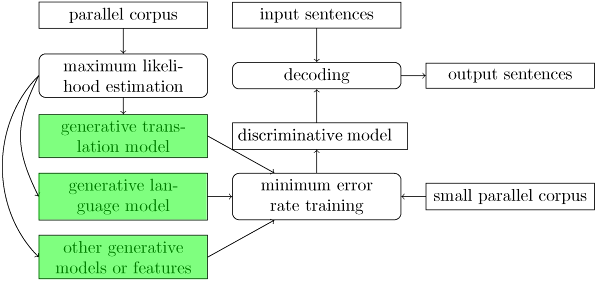
id	source sentence	translation
7	*Annex to the* **Technical Offer**	*Annex all'***Tecnica Offri** *Allegato all'***Offerta Tecnica**
8	*Sistemi Informativi* **SpA**	*Sistemi Informativi* **SpA** *Sistemi Informativi* **S.p.A.**
21	*This document is Sistemi Informativi* **SpA***'s* **Technical Offer**	*Questo documento è Sistemi Informativi* **SpA** *di* **Tecnica Offri** *Il presente documento rappresenta l'***Offerta Tecnica** *di Sistemi Informativi* **S.p.A.**

- MT hypothesis
- user translation

Goals

- integrate user feedback into an SMT system on a per-sentence basis
- enable translation consistency, learn new, document-specific translations
- focus on simple, easily integrable solutions as proof of concept that can serve as a baseline for enhanced approaches

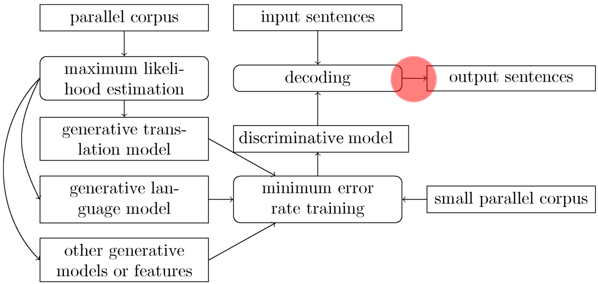
Approaches



Generative: Interacting with the decoder

Adapt language and translation models locally by passing information to the Moses decoder through XML markup and a cache feature.

Approaches



Discriminative: Reranking decoder output

Train an external reranking model of sparse phrase pair and target n-gram features on the k-best output of the decoder; let reranker determine 1best translations.

Related work

- incremental learning for domain adaptation (Koehn and Schroeder, 2007; Bisazza et al., 2011; Liu et al., 2012)
- translation consistency (Carpuat and Simard, 2012)
- online learning for interactive machine translation (Nepveu et al., 2004; Ortiz-Martínez et al., 2010; Cesa-Bianchi et al., 2008)

Outline

Exploiting user feedback

- align source and user translation
- extract

  phrase table (generative approach)
  features (reranking approach)

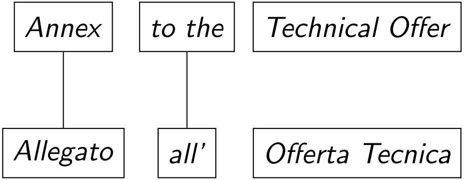
  from the alignment

Constrained search for phrase alignment

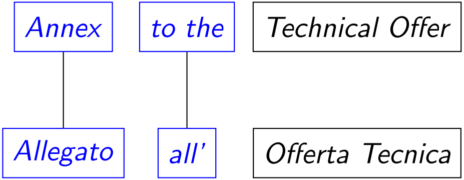
Tool by Cettolo et al. (2010)

- produces an alignment at phrase level
- given a set of translation options, constrained search optimizes the coverage of both source and target sentences
- search produces exactly one phrase segmentation and alignment
- target does not have to be reachable, i.e. gaps are allowed

Phrase extraction



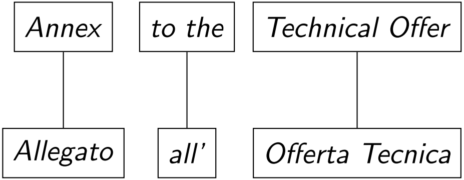
Phrase extraction



known phrase pairs

Annex → *Allegato*,
to the → *all'*

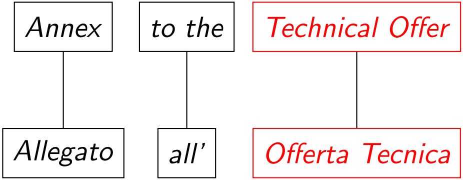
Phrase extraction



known phrase pairs

Annex $\rightarrow$ *Allegato*,
to the $\rightarrow$ *all'*

Phrase extraction



known phrase pairs   new phrase pairs

Annex → *Allegato*, *to the* → *all'*   *Technical Offer* → *Offerta Tecnica*

Phrase extraction

Annex to the Technical Offer

Allegato all' Offerta Tecnica

known phrase pairs　new phrase pairs

Annex → *Allegato*,　*Technical Offer* → *Offerta Tecnica*
to the → *all'*

full sentence
Annex to the Technical Offer → *Allegato all' Offerta Tecnica*

Reranking features

two sparse feature templates are used:

1. phrase pairs used by the decoder (hypotheses); phrase pair features on the **user translation** given by the **alignment** output of the constrained search
2. target *n*-gram features (*n* upto 4)

these are indicator features, but we use source side token count (phrase pairs) or *n* (target *n*-grams) as feature values

Outline

Discriminative reranking

we do reranking using a structured perceptron algorithm

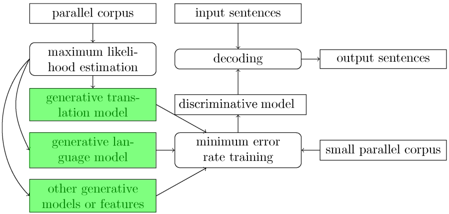
1. decoder produces k-best list of hypotheses
2. for each hypothesis x we build a feature vector $\phi(x)$ and calculate model scores using the current reranking model $\mathbf{w}$
3. output 1best according to current model

$$\text{1best} = \arg\max_{x \in k\text{best}} \mathbf{w} \cdot \phi(x) = \sum_{i=0}^{d} \mathbf{w}_i \phi(x)_i$$

4. update model if reranking prediction is not equal to the user translation (by string comparison)

$$\mathbf{w} \leftarrow \mathbf{w} + (\phi(\text{user translation}) - \phi(\text{1best}))$$

Generative: Adding local models in Moses



TM adaptation suggest phrase pairs from the feedback exploitation step to the decoder at run time using XML input

LM adaptation use an *n*-gram cache feature in Moses that rewards *n*-grams seen in user translations

Moses XML input option

- Moses allows input to be annotated with translation options for phrases:

  Annex to the `<p translation= "`*Offerta Tecnica*`||`*Proposta Tecnica*`" prob="0.75||0.25">`*Technical Offer*`</p>`

  inclusive mode given phrase translations compete with existing phrase table entries

  exclusive mode decoder is forced to choose from the given translations

- probabilities are estimated based on the relative frequency of the target phrase given the source phrase within the local phrase table

Outline

Baseline system

- Moses decoder and tools
- 5-gram Kneser-Ney smoothed LM learned with IRSTLM
- case-sensitive models
- default log-linear weights optimized with MERT

Data

	IT		*patent*	
	doc	sentences	doc	sentences
	train	1,167 K	**train**	4,199 K
dev1	prj1	420	pat1	300
	prj2	931	pat2	227
	prj3	375	pat3	239
dev2	prj4	289		
	prj5	1,183		
	prj6	864		
test	prj7A	176	pat4	232
	prj7B	176	pat5	230
	prj7C	176	pat6	225
			pat7	231

Development: TM adaptation

	IT **dev1**		*IT* **dev2**	
	BLEU	$\Delta[\sigma]$	BLEU	$\Delta[\sigma]$
baseline	22.59		21.49	
new	23.11	+0.52 [±0.57]	21.64	+0.15 [±0.06]
known	23.73	+1.14 [±0.70]	22.24	+0.75 [±0.15]
full	24.22	+1.63 [±1.73]	23.07	+1.58 [±0.91]
new&known&full	**25.49**	+**2.90** [±2.18]	**23.91**	+**2.42** [±0.83]

new&know&full = tm

Development: Reranking & comparison

	IT **dev1**		*IT* **dev2**	
	BLEU	$\Delta[\sigma]$	BLEU	$\Delta[\sigma]$
baseline	22.59		21.49	
rerank	23.74	+1.15 [±0.82]	22.85	+1.36 [±0.65]
known &lm	**25.78**	+**2.69** [±1.68]	**23.43**	+**1.94** [±1.41]

Reranking: Top features

- the baseline translates

  $$\textit{DLI and IBM} \rightarrow \textit{DLI } \mathbf{e} \textit{ IBM}$$

  in consequence the reranker learns that *and* should be translated with *ed* in this case (following vowel)
- the *IT* data contains a lot of title-cased text which is incorrectly translated to lowercase by the baseline system $\rightarrow$ top phrase pairs include corrections for this

Test set results: IT domain

	prj7A		prj7B		prj7C	
	BLEU	Δ	BLEU	Δ	BLEU	Δ
baseline	41.10		39.68		30.68	
tm+lm	**42.97**	**+1.87**	39.72	+0.04	**33.76**	**+3.08**

Test set results: patent domain

	patent **test**	
	BLEU	$\Delta[\sigma]$
baseline	30.26	
rerank	32.54	+2.28 [±1.47]
tm+lm	33.24	+2.98 [±2.03]
tm+lm+rerank	**34.02**	+**3.76** [±2.08]

Conclusions

- simple approaches to integrate new information in an SMT system on each sentence
- significant improvements over baseline
- starting point for advanced incremental approaches

Bibliography

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *IWSLT'11*.

Carpuat, M. and Simard, M. (2012). The trouble with SMT consistency. In *WMT'12*.

Cesa-Bianchi, N., Reverberi, G., and Szedmak, S. (2008). Online learning algorithms for computer-assisted translation. Technical report, SMART (`www.smart-project.eu`).

Cettolo, M., Federico, M., and Bertoldi, N. (2010). Mining parallel fragments from comparable texts. In *IWSLT'10*.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *WMT'07*.

Liu, L., Cao, H., Watanabe, T., Zhao, T., Yu, M., and Zhu, C. (2012). Locally training the log-linear model for SMT. In *EMNLP'12*.

Nepveu, L., Lapalme, G., Langlais, P., and Foster, G. (2004). Adaptive language and translation models for interactive machine translation. In *EMNLP'04*.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *HLT-NAACL'10*.