Learning to Translate Queries for CLIR

Artem Sokolov   Felix Hieber   Stefan Riezler
Heidelberg University, Germany
{sokolov, hieber, riezler}@cl.uni-heidelberg.de

DFG

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Paper Highlights

Departure from the "SMT as a black-box" paradigm:

- ✔ direct **SMT tuning for CLIR quality**
- ✔ new decomposable proxy for retrieval quality to:
  - ✔ explore full decoder search space instead of k-best lists
  - ✔ train faster than k-best reranking frameworks

Recap: SMT (baseline)

Translation q_f of foreign query f:

- construct q_f from bilingual translation **units** (phrases or grammar rules)
- **units** carry numerical **features** $\mathbf{h}_{u,q,f}$
- decoding: $q_f = \arg\max_q \mathbf{w} \cdot \mathbf{h}_{q,f}$
- features must be **decomposable over units** for efficient arg max: $\mathbf{h}_{q,f} = \sum_u \mathbf{h}_{u,q,f}$
- $\mathbf{w}$ is learned to maximize BLEU on human reference translations r_f

Structural SVM for SMT

Inject task-specific info via margin-rescaling:

1 assume unit-decomposable penalty $\Delta(q, q')$ for producing q instead of q':

- $\Delta(q, q') = 0$, if $q = q'$
- increases as q gets father away from q'

2 closest reachable substitute for reference r_f: $q_f^* = \max_q(-\Delta(q, r_f))$

3 unit-decomposability of Δ is necessary for efficient max in the loss:

$$\mathcal{L} = \sum_f \max_q(\Delta(q, q_f^*) + \mathbf{w} \cdot \mathbf{h}_q) - \mathbf{w} \cdot \mathbf{h}_{q_f^*}$$

4 updates: $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \cdot \nabla_{\mathbf{w}} \mathcal{L}$

In CLIR single r_f does not exist
$\Rightarrow$ a decomposable proxy Δ that reflects retrieval-quality is required

Contribution: Tuning SMT for CLIR

New decomposable penalty Δ:

- let $\mathcal{C}_{f,k}^+$ be docs on k^{th} relevance level for query f
- **relevance score** of a translation q w.r.t. $\mathcal{C}_f^+$
  $S(q, \mathcal{C}_f^+) = \sum_{t \in q} \sum_k \boldsymbol{\omega}_k \sum_{d \in \mathcal{C}_{f,k}^+} \mathbf{bm25}(t, d) \; / \; |\mathcal{C}_{f,k}^+|$
  - ➡ relevance level weights $\boldsymbol{\omega}_k$ are found with grid search
  - ➡ **BM25 decomposes** over terms!
- **novel penalty** for Structural SVM
  $\Delta(q, \mathcal{C}_f^+) = \max_q \left(S_{rel}(q, \mathcal{C}_f^+)\right) - S_{rel}(q, \mathcal{C}_f^+)$

Define hope, fear & oracle [McAllester and Keshet, 2011]:

$$q^{oracle} = \arg\max_{q \in \mathcal{E}_f}(-\Delta(q, \mathcal{C}_f^+)), \quad q^{hope} = \arg\max_{q \in \mathcal{E}_f}(\mathbf{w} \cdot \mathbf{h}_q - \Delta(q, \mathcal{C}_f^+))$$

$$q^{fear} = \arg\max_{q \in \mathcal{E}_f}(\mathbf{w} \cdot \mathbf{h}_q + \Delta(q, \mathcal{C}_f^+))$$

Losses to optimize:

$$\mathcal{L}_{svm} = \sum_f (\mathbf{w} \cdot \mathbf{h}_q^{fear} + \Delta(q^{fear}, \mathcal{C}_f^+)) - \mathbf{w} \cdot \mathbf{h}_q^{oracle}$$

$$\mathcal{L}_{ramp} = \sum_f (\mathbf{w} \cdot \mathbf{h}_q^{fear} + \Delta(q^{fear}, \mathcal{C}_f^+)) - (\mathbf{w} \cdot \mathbf{h}_q^{hope} - \Delta(q^{hope}, \mathcal{C}_f^+))$$

Experiments: patent prior art search

Two baseline SMT systems:

- **Moses** (lattices) & **cdec** (hypergraphs)
- train/dev: 1.8M/2k sentences from NTCIR
- standard dense features and **lexical sparse word-to-word mappings**
- MIRA weight optimization [Chiang et al., 2008]

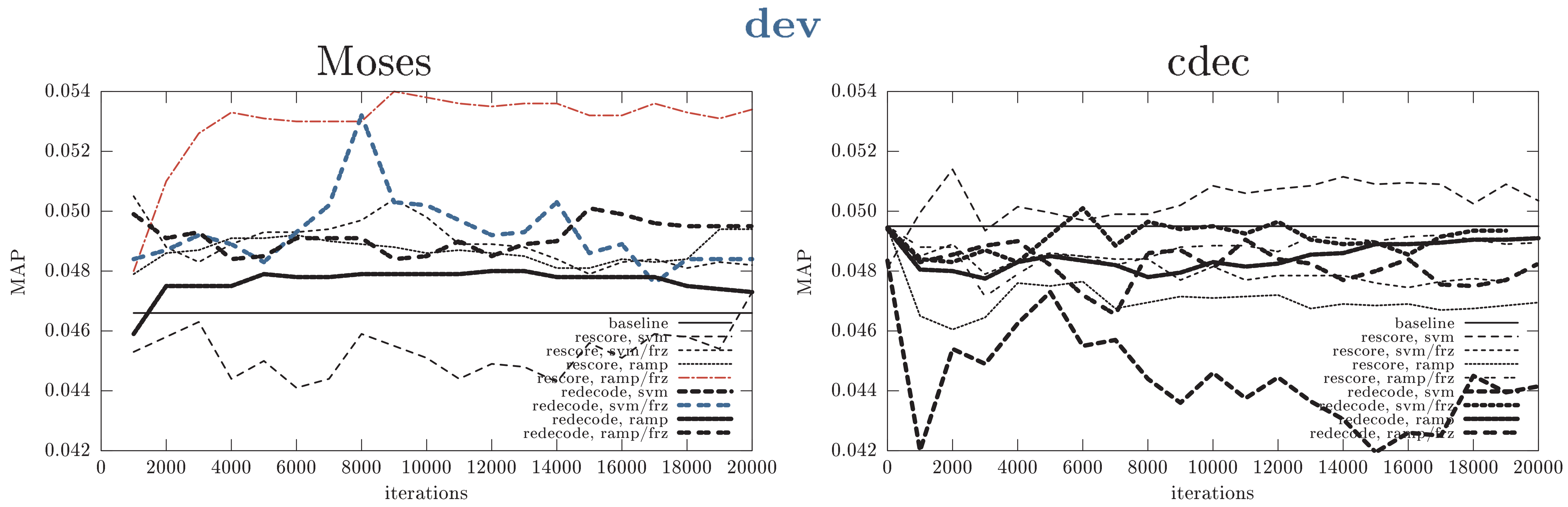
CLIR dataset

- sampled from the BoostCLIR dataset of JP/EN patents
- train: 1k queries (5k sentences)
- dev/test: 400 queries (2k sentences)

Meta-parameters (tuned on dev):

- **rescoring/redecoding**: inference with new $\mathbf{w}$ on **old/rebuilt** MIRA lattices/hypergraphs
- **ramp** or **svm** losses ($\mathcal{L}_{ramp}/\mathcal{L}_{svm}$)
- **freezing** or **learning** dense features
- **#** of iterations

dev



test

config	Moses MAP	Moses NDCG	cdec MAP	cdec NDCG
baseline	0.0438	0.1498	0.0515	0.1600
rescore	$^{0.03}$**0.0498**	$^{0.02}$**0.1575**	$^{0.11}$0.0473	$^{0.08}$0.1548
redecode	$^{0.28}$**0.0463**	$^{0.26}$**0.1532**	$^{0.23}$0.0487	$^{0.27}$0.1571

- Moses (rescore: ramp/frz@9k, redecode: svm/frz@8k)
- cdec (svm/frz: rescore@2k, redecode@6k)

References

[Chiang et al., 2008] Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *EMNLP*.

[McAllester and Keshet, 2011] McAllester, D. A. and Keshet, J. (2011). Generalization bounds and consistency for latent structural probit and ramp loss. In *NIPS*.