Minimum Error Rate Training Semiring

Artem Sokolov & François Yvon

LIMSI-CNRS & LIMSI-CNRS/Univ. Paris Sud
{artem.sokolov,francois.yvon}@limsi.fr

EAMT'2011
31 May 2011

Talk Plan

Probability model and inference in SMT system

Probability of translation **e** given source sentence **f**:

$$p(\mathbf{e}|\mathbf{f}) = Z(\mathbf{f})^{-1} \exp(\bar{\lambda} \cdot \bar{h}(\mathbf{e}, \mathbf{f}))$$

- $\bar{h}(\mathbf{e}, \mathbf{f})$ – feature vector (various compatibility measures of **e** and **f**)
- $\bar{\lambda}$ – parameter vector, λ_i regulates importance of the feature $h_i(\mathbf{e}, \mathbf{f})$

Translating by MAP-inference:

$$\tilde{\mathbf{e}}_{\mathbf{f}}(\bar{\lambda}) = \arg\max_{\mathbf{e} \in E} p(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e} \in E} \bar{\lambda} \cdot \bar{h}(\mathbf{e}, \mathbf{f})$$

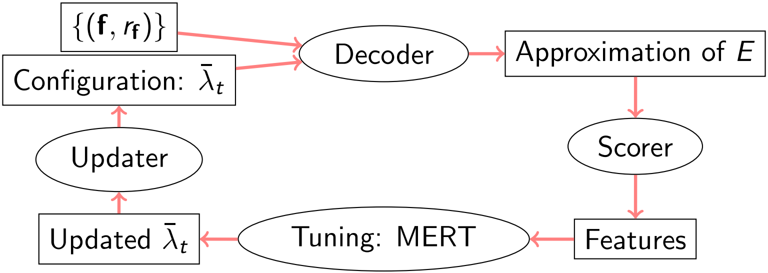
- E – reachable translations (search space), can be approximated by:
  - list of n-best hypotheses
  - word lattice

Tuning SMT system with MERT

Given: development set $\{(\mathbf{f}, r_\mathbf{f})\}$ (source $\mathbf{f}$ & reference $r_\mathbf{f}$ pairs)
Solve:

$$\bar{\lambda}^* = \arg\max_{\bar{\lambda}} BLEU(\{\tilde{\mathbf{e}}_\mathbf{f}(\bar{\lambda}, E(\bar{\lambda})), r_\mathbf{f}\})$$

- BLEU is non-convex and not differentiable, hence heuristics (MERT).
- Search space approximation depends on $\bar{\lambda}$, so iterative tuning:



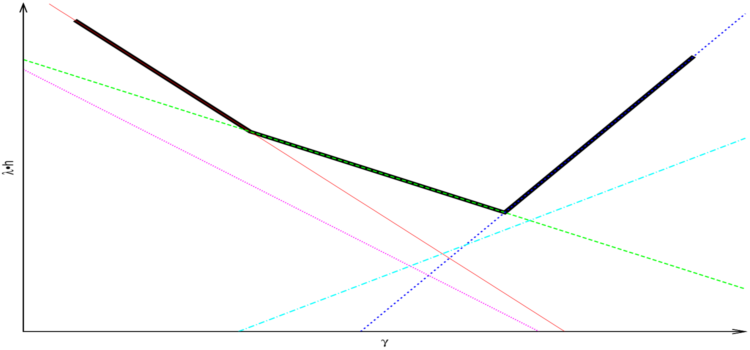
MERT proceeds in series of optimizations along directions $\bar{r}$:

$$\bar{\lambda} = \bar{\lambda}_0 + \gamma \bar{r}$$

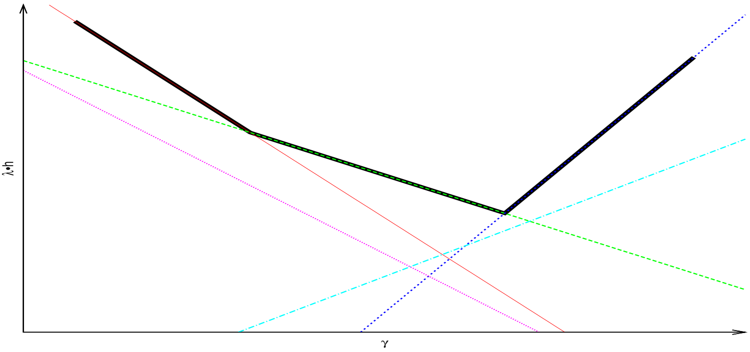
Optimal translation:

$$\tilde{\mathbf{e}}_{\mathbf{f}}(\gamma) = \arg\max_{\mathbf{e} \in E} \bar{\lambda} \cdot \bar{h}(\mathbf{e}, \mathbf{f}) = \arg\max_{\mathbf{e} \in E} \underbrace{\bar{\lambda}_0 \cdot \bar{h}(\mathbf{e}, \mathbf{f})}_{\text{intercept}} + \gamma \underbrace{\bar{r} \cdot \bar{h}(\mathbf{e}, \mathbf{f})}_{\textit{slope}}$$

- each translation hypothesis is associated with a line,
- **upper envelope**: dominating lines when $\bar{\lambda}$ is moved along $\bar{r}$



- γ-projections of intersections give intervals of constant optimal hypothesis
- optimal γ^* found by merging intervals for $\mathbf{f} \in F$ and scoring each
- update $\bar{\lambda} = \lambda_0 + \gamma^*_{i^*} \bar{r}_{i^*}$,
  where i^* is the index of the direction yielding the highest BLEU



MERT problems

- very slow, because of:
  - overall number of iterations
    folklore: number of iterations $\simeq$ number of dimensions
  - slowness of each iteration (dominated by decoding time)
- non-monotonicity/instability of the training process
- sensitivity of the resulting solutions to initial conditions

Ways to tackle the problems

- improve optimization
  - other target function approximations
  - changes into optimization algorithms
- improve search space processing $\longleftarrow$ this presentation
  - use lattices (better approximation of the complete search space)
  - reduce search to standard operations (facilitates implementation)
- reduce number of iteration $\longleftarrow$ this presentation

Contribution

- Recast Lattice MERT algorithm of [Macherey et al., 2008] in a semiring framework
  - has already been hinted to in [Dyer et al., 2010]
  - but was never formally described
  - lack of implementation details
- Reimplement MERT using this reformulation
  - and general-purpose FST toolbox OpenFST

Semirings

Semiring $\mathbb{K} = \langle K, \oplus, \otimes, \bar{0}, \bar{1} \rangle$:

- $\langle K, \oplus, \bar{0} \rangle$ is a commutative monoid with identity element $\bar{0}$:
  - $a \oplus (b \oplus c) = (a \oplus b) \oplus c$
  - $a \oplus b = b \oplus a$
  - $a \oplus \bar{0} = \bar{0} \oplus a = a$
- $\langle K, \otimes, \bar{1} \rangle$ is a monoid with identity element $\bar{1}$
- $\otimes$ distributes over $\oplus$
  - $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$
  - $(b \oplus c) \otimes a = (b \otimes a) \oplus (c \otimes a)$
- element $\bar{0}$ annihilates K
  - $a \otimes \bar{0} = \bar{0} \otimes a = \bar{0}$.

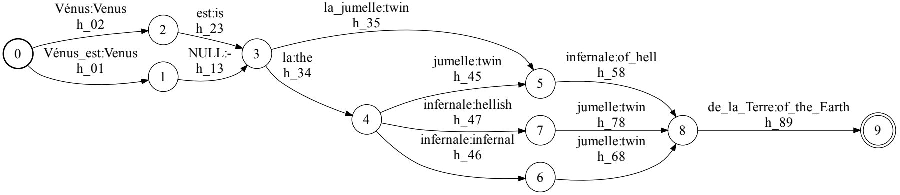
Examples

- $\langle \mathbb{R}, +, \times, 0, 1 \rangle$ – real semiring
- $\langle S, \Delta, \cap, \emptyset, \cup_i S_i \rangle$ – semiring of sets

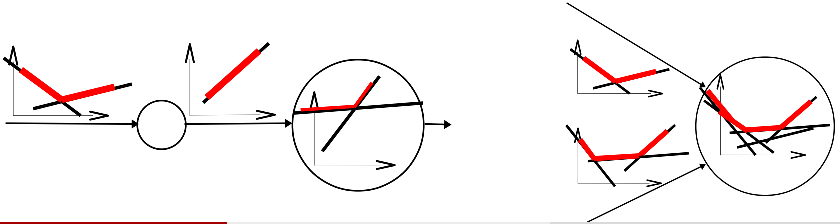
Lattice MERT [Macherey et al., 2008]

source **fr**: Vénus est la jumelle infernale de la Terre

target **en**: Venus is Earth's hellish twin



- Decomposability of $\bar{h}(\mathbf{e}, \mathbf{f})$ into a sum of *local* features $h\_01, h\_02 \ldots$
- Envelopes are distributed over nodes in the lattice



MERT Semiring

$$\mathbb{D} = \langle D, \oplus, \otimes, \bar{0}, \bar{1} \rangle$$

Host set:

- a line: $d_y + d_s \cdot x$ (hypothesis)
- set of lines d_i: $d = \{d_{i,y} + d_{i,s} \cdot x\}$ (set of hypotheses)
- set of sets d^k of lines: $D = \{\{d^k_{i,y} + d^k_{i,s} \cdot x\}\}$

Operations $\oplus$ and $\otimes$:

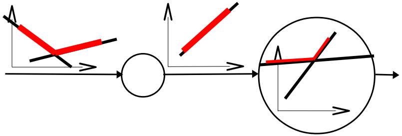
- for $d^1, d^2 \in D$
- $d^1 \oplus d^2 = \text{env}(d^1 \cup d^2)$
- $d^1 \otimes d^2 = \text{env}(\{(d^1_{i.y} + d^2_{j.y}) + (d^1_{i.s} + d^2_{j.s}) \cdot x |\ \forall d^1_i \in d^1, d^2_j \in d^2\})$

Unities:

- $\bar{0} = \emptyset$
- $\bar{1} = \{0 + 0 \cdot x\}$

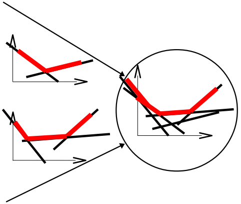
Semiring Operations Illustration

⊗-example



$$d^1 \otimes d^2 = \text{env}(\{(d^1_{i.y} + d^2_{j.y}) + (d^1_{i.s} + d^2_{j.s}) \cdot x |\ \forall d^1_i \in d^1, d^2_j \in d^2\})$$

⊕-example



$$d^1 \oplus d^2 = \text{env}(d^1 \cup d^2)$$

Shortest Paths for MERT Semiring

Each arc in the FST carries:

- target word a
- vector $\bar{h}(a, \mathbf{f})$ of local features associated with a
- singleton set containing line d with
  - slope $d_s = (\bar{r} \cdot \bar{h}(a, \mathbf{f}))$
  - y-intercept $d_y = (\bar{\lambda}_0 \cdot \bar{h}(a, \mathbf{f}))$

Weight of a candidate translation path $\mathbf{e} = e_1 \ldots e_\ell$:

$$w(\mathbf{e}) = \bigotimes_{i=1}^{\ell} w(e_i) = \{\bar{\lambda}_0 \cdot \sum_{i=1}^{\ell} \bar{h}(e_i, \mathbf{f}) + (\bar{r} \cdot \sum_{i=1}^{\ell} \bar{h}(e_i, \mathbf{f})) \cdot x\}$$

Upper envelope of all the lines (hypotheses):

$$\mathrm{env}(\bigcup_{\mathbf{e}} w(\mathbf{e})) = \bigoplus_{\mathbf{e}} w(\mathbf{e}) = \bigoplus_{\mathbf{e}} \bigotimes_{i=1}^{\ell(\mathbf{e})} w(e_i).$$

Generic shortest distance algorithms over acyclic graphs calculate this.

Implementation

- **Basics:** OpenFST toolbox
  - works with **any** semiring
  - proven and well optimized `ShortestPath` algorithms
  - other useful algorithms: `Union`, `Determinize`, etc.
- **Lattice minimization**:
  - `Union` of lattices between decoder runs
  - `Determinize`+`Minimize` to eliminate duplicate hypotheses
    won't work – MERT semiring is not divisible
  - circumvent by performing `Union`+`Determinize` over $(min, +)$ semiring
- **All directions simultaneously**
  - weights as arrays of envelopes
  - 20-30 random direction $\simeq +0.3$-0.5 BLEU
- Random restarts help only for the first iteration

Experiments

Data:

- NewsCommentary (dev: 2051) & WMT10 (dev: 1026), common test
- French to English

FST MERT tuning:

- OpenFST-based multi-threaded implementation
- zero restart points
- axes and additional random directions

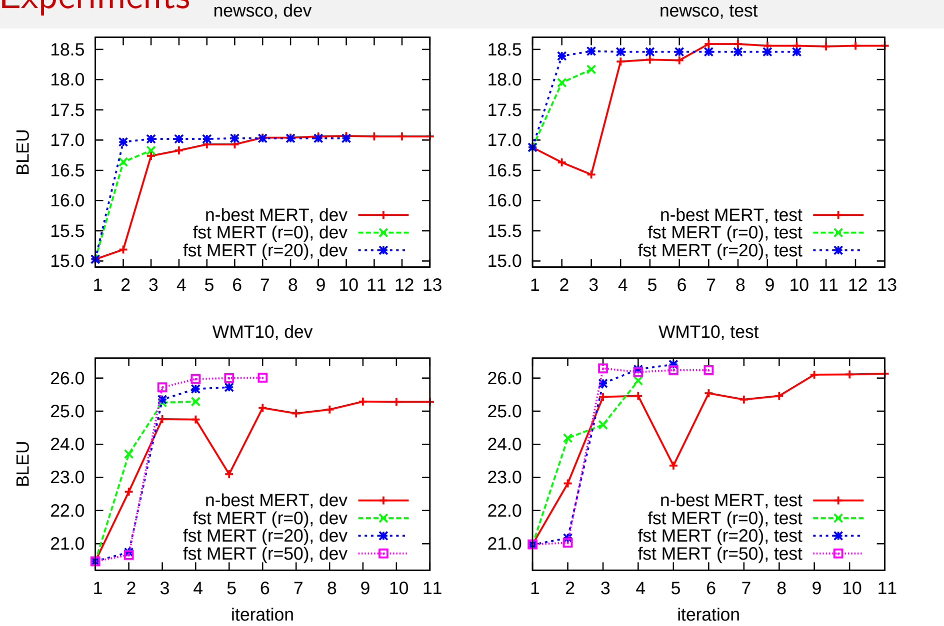
Baseline MERT tuning:

- MERT implementation included in MOSES toolkit
- 100-best list, 20 restart points
- Koehn's coordinate descend (only axis directions)

Decoder: *n*-gram phrase-based SMT system N-code[1], 11 features

[1] Demo on `http://ncode.limsi.fr/`

Experiments



Conclusion & Future Work

Conclusion

- Semiring formalization allows using generic FST toolkits to do MERT
- Convergence in less iterations

Future Work

- Better stopping criteria to detect saturation
- Faster $\oplus$ – should be most helpful for speed up

Thank you for your attention!

Bibliography

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., & Resnik, P. (2010).
cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models.
In *Proc. of the ACL* (pp. 7–12).

Macherey, W., Och, F. J., Thayer, I., & Uszkoreit, J. (2008).
Lattice-based minimum error rate training for statistical machine translation.
In *Proc. of the Conf. on EMNLP* (pp. 725–734).