Ensemble Decoding for Statistical Machine Translation

Anoop Sarkar
Joint work with ***Majid Razmara*** *and* ***Baskaran Sankaran***
SFU Natural Language Lab
Simon Fraser University, Vancouver, Canada
http://natlang.cs.sfu.ca

Ensemble Decoding

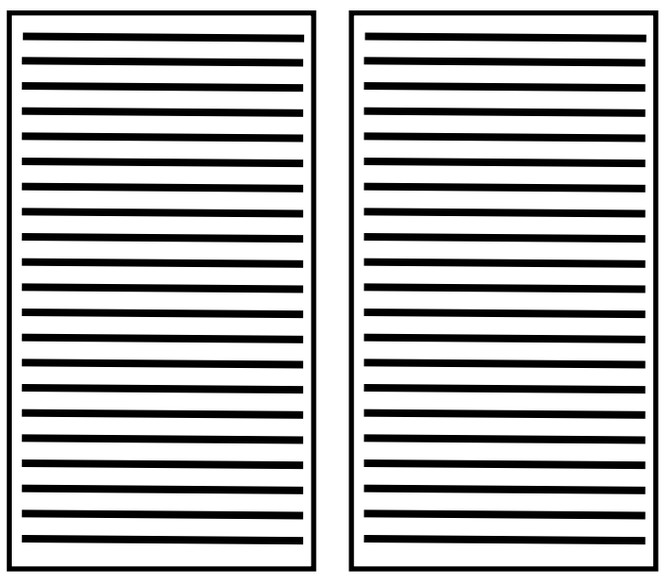
Domain Adaptation

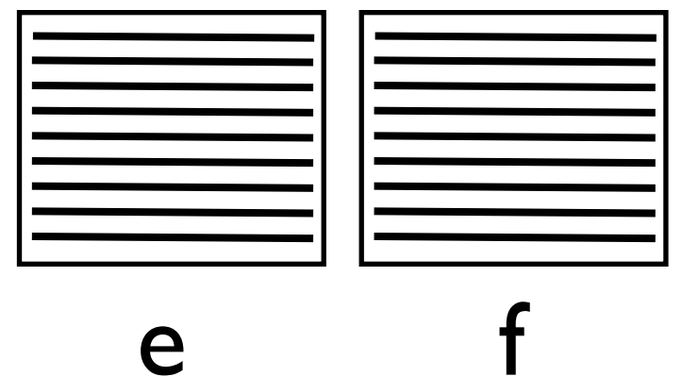
Multi-metric optimization

Pivot language triangulation

Ensemble Decoding

But first, Mixtures





Translation models: $m = 1 \ldots M$

Log-linear mixture: $$p(\bar{e}|\bar{f}) \propto \exp\left(\sum_{m}^{M} \lambda_m \log p_m(\bar{e}|\bar{f})\right)$$

- Each (phrase-table) component in the usual discriminative SMT model is a mixture.
- The mixture weights are tuned on a dev set.

Linear Mixtures

$$p(\bar{e}|\bar{f}) = \sum_{m}^{M} \lambda_m p_m(\bar{e}|\bar{f})$$

- Extract joint phrase pair distribution $p(e, f)$
- Find the weights that minimize the cross-entropy of the mixture $p(e \mid f)$ with respect to $p(e, f)$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \sum_{\bar{e},\bar{f}} \tilde{p}(\bar{e}, \bar{f}) \log \sum_{m}^{M} \lambda_m p_m(\bar{e}|\bar{f})$$

Linear Mixtures

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} \sum_{\bar{e},\bar{f}} \tilde{p}(\bar{e},\bar{f}) \log \sum_{m}^{M} \lambda_m p_m(\bar{e}|\bar{f})$$

- Train the weights on the dev set using any optimization technique (L-BFGS).
- Linear mixtures are used as feature functions in standard discriminative SMT.
- State of the art for domain adaptation in SMT (Foster et al, EMNLP 2010).

Ensemble Decoding

- Previous mixtures of translation models were pre-processing steps.
- This work: Explore mixtures of translation models in the decoder.
- On the fly combination of models in Hiero

0.5 yu X_1 you X_2 / have X_2 with X_1

$\otimes$

1.5 Beihan / North Korea

$\otimes$

3.0 bangjiao / diplomatic relations

$\oplus$

1.5 yu X_1 you X_2 / with X_1 have X_2

Semi-ring

$\oplus$ $\otimes$

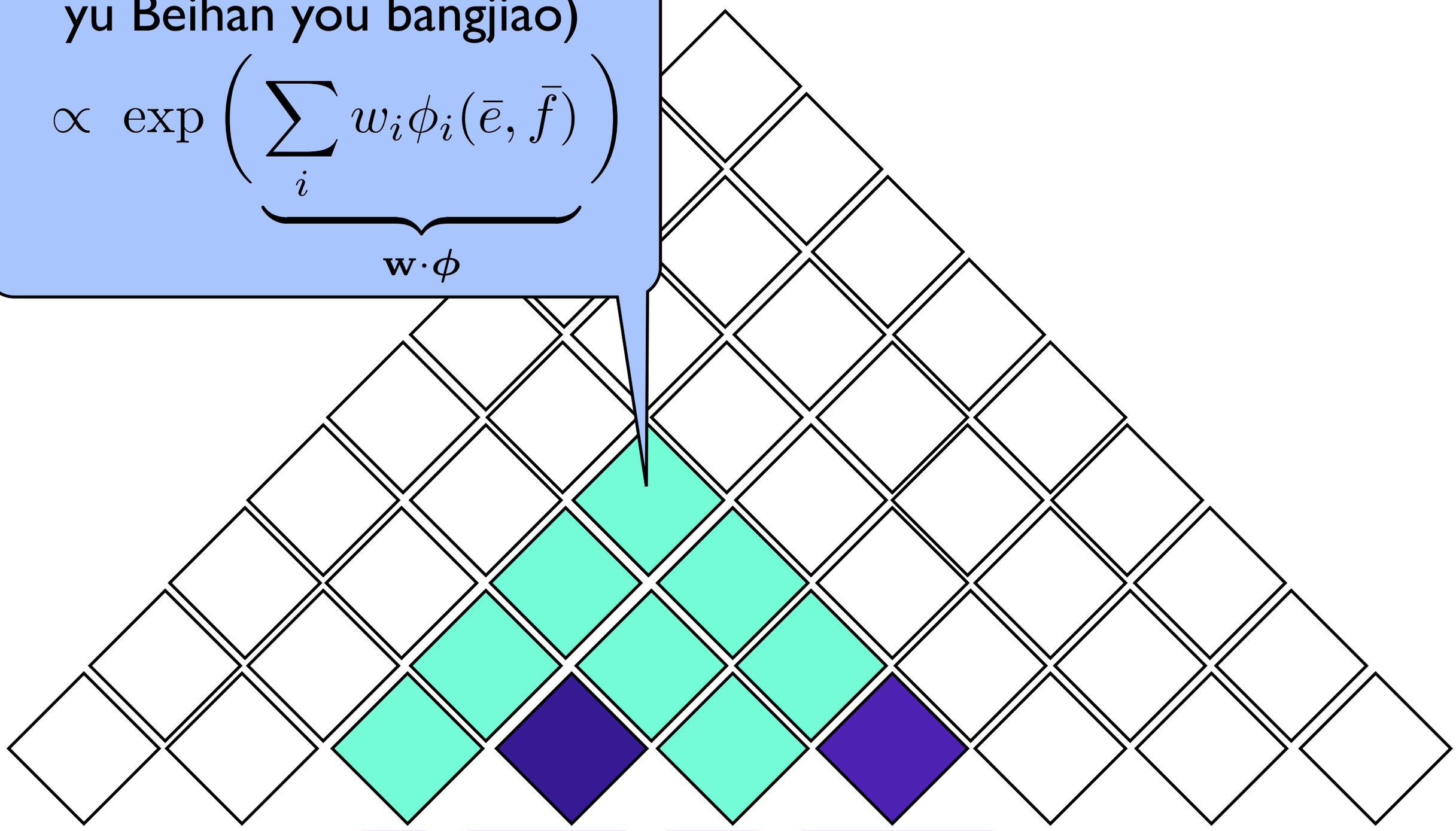
$(\mathbb{R} \cup \{\infty\}, \min, +, \infty, 0)$

1.have dipl. relns. with N. K = 4.0
2.with N.K. have dipl. relns. = 6.0
•min(4.0, 6.0) option #1 wins

Aozhou	shi	yu	Beihan	you	bangjiao	de	guojia	shiyi
澳洲	是	与	北韩	有	邦交	的	国家	之一

P(have dipl. relns. with N.K | yu Beihan you bangjiao)

$$\propto \exp\left(\underbrace{\sum_i w_i \phi_i(\bar{e}, \bar{f})}_{\mathbf{w}\cdot\boldsymbol{\phi}}\right)$$

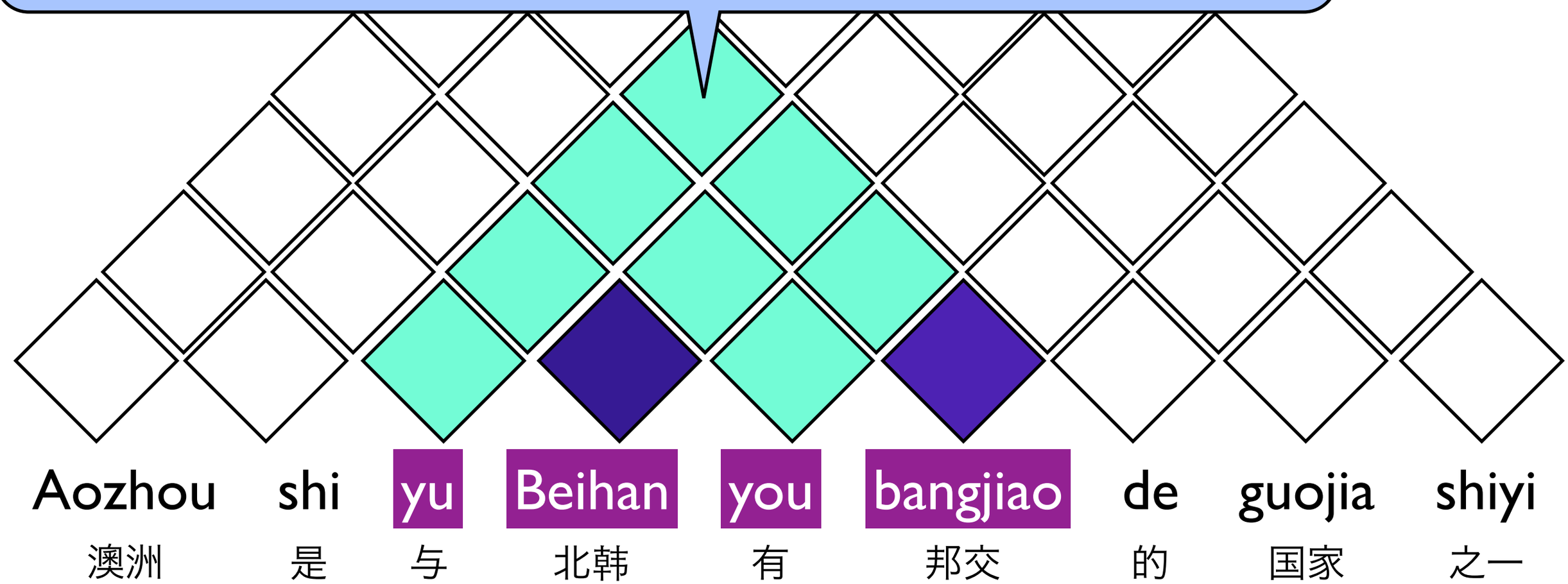


Aozhou	shi	yu	Beihan	you	bangjiao	de	guojia	shiyi
澳洲	是	与	北韩	有	邦交	的	国家	之一

Ensemble Decoding

P(have dipl. relns. with N.K | yu Beihan you bangjiao)

$$\propto \exp\left(\underbrace{\mathbf{w_1} \cdot \boldsymbol{\phi_1}}_{1^{st}\ \text{model}} \otimes \underbrace{\mathbf{w_2} \cdot \boldsymbol{\phi_2}}_{2^{nd}\ \text{model}} \otimes \cdots \right)$$



Ensemble Operations

Weighted Sum (wsum)

$$p(\bar{e} \mid \bar{f}) \quad \propto \quad \sum_m^M \lambda_m \, \exp\left(\mathbf{w}_m \cdot \boldsymbol{\phi}_m\right)$$

- Ensemble score is the weighted sum of individual model scores
- *m* is each component model, a total of *M* components in the ensemble.

fr une autre maladie métabolique héréditaire

en or other disease hereditary metabolic

or another hereditary metabolic disease

m1	4.5
m2	10.5
ens	7.25

m1	16.5
m2	3.5
ens	10

Weighted Max (wmax)

$$p(\bar{e} \mid \bar{f}) \quad \propto \quad \max_m \left(\lambda_m \, \exp\left(\mathbf{w}_m \cdot \boldsymbol{\phi}_m\right)\right)$$

- Ensemble score is the weighted max of all the model scores
- The n-best list can contain entries from different models

fr une autre maladie métabolique héréditaire

en or other disease hereditary metabolic

or another hereditary metabolic disease

m1	4.5
m2	10.5
ens	4.5

m1	16.5
m2	3.5
ens	3.5

Model Switching (Switch)

$$p(\bar{e} \mid \bar{f}) = \sum_m^M \delta(\bar{f}, m)\, p_m(\bar{e} \mid \bar{f})$$

- Switch models in each CKY cell. Possibly picking a different model from the ensemble.
- The n-best list can contain entries from only one model.

$$\delta(\bar{f}, m) = \begin{cases} 1, & m = \operatorname*{argmax}_{n \in M} \psi(\bar{f}, n) \\ 0, & \text{otherwise} \end{cases}$$

Model Switching (Switch)

$$\delta(\bar{f}, m) = \begin{cases} 1, & m = \operatorname*{argmax}_{n \in M} \psi(\bar{f}, n) \\ 0, & \text{otherwise} \end{cases}$$

For each cell the model that has the highest weighted score wins:

$$\psi(\bar{f}, n) = \lambda_n \max_{\bar{e}} (\mathbf{w}_n \cdot \boldsymbol{\phi}_n(\bar{\mathbf{e}}, \bar{\mathbf{f}}))$$

For each cell, the model with highest weighted sum of scores wins:

$$\psi(\bar{f}, n) = \lambda_n \sum_{\bar{e}} \exp\left(\mathbf{w}_n \cdot \boldsymbol{\phi}_n(\bar{\mathbf{e}}, \bar{\mathbf{f}})\right)$$

Product (prod)

$$p(\bar{e} \mid \bar{f}) \quad \propto \quad \exp\left(\sum_m^M \lambda_m \left(\mathbf{w}_m \cdot \boldsymbol{\phi}_m \right) \right)$$

- Compute the product of all the probabilities in the ensemble (sum of log-probs).
- A Logarithmic Opinion Pool (LOP).
- LOPs work best when the ensemble is used to down-vote a highly confident but incorrect candidate.

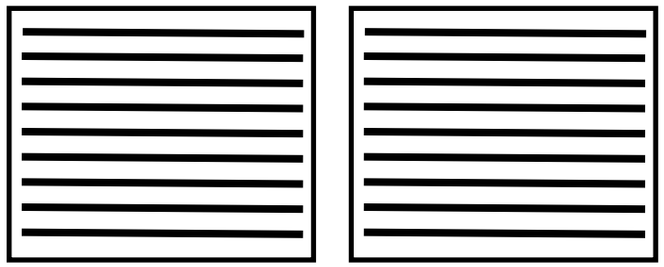
fr une autre maladie métabolique héréditaire

en or other disease hereditary metabolic

m1	1.5
m2	10.5
m3	12.0
ens	24.0

Domain Adaptation

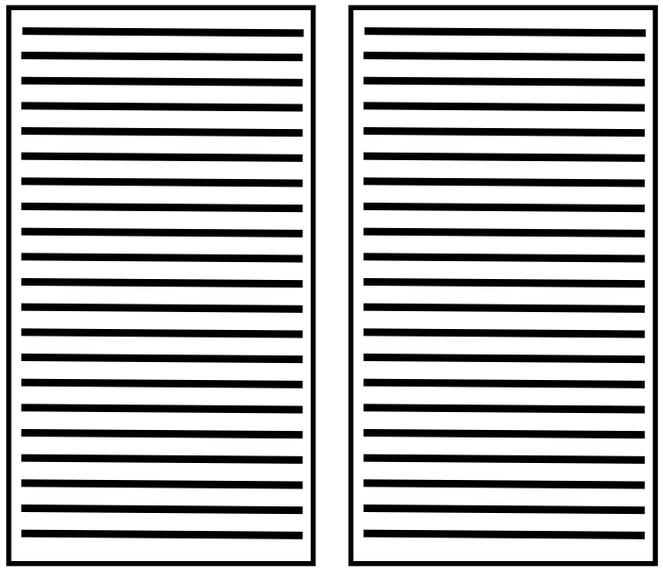
Domain Adaptation



IN-domain

EMEA (Medical)

Train	11770
Dev	1533
Test	1522



OUT-of-domain

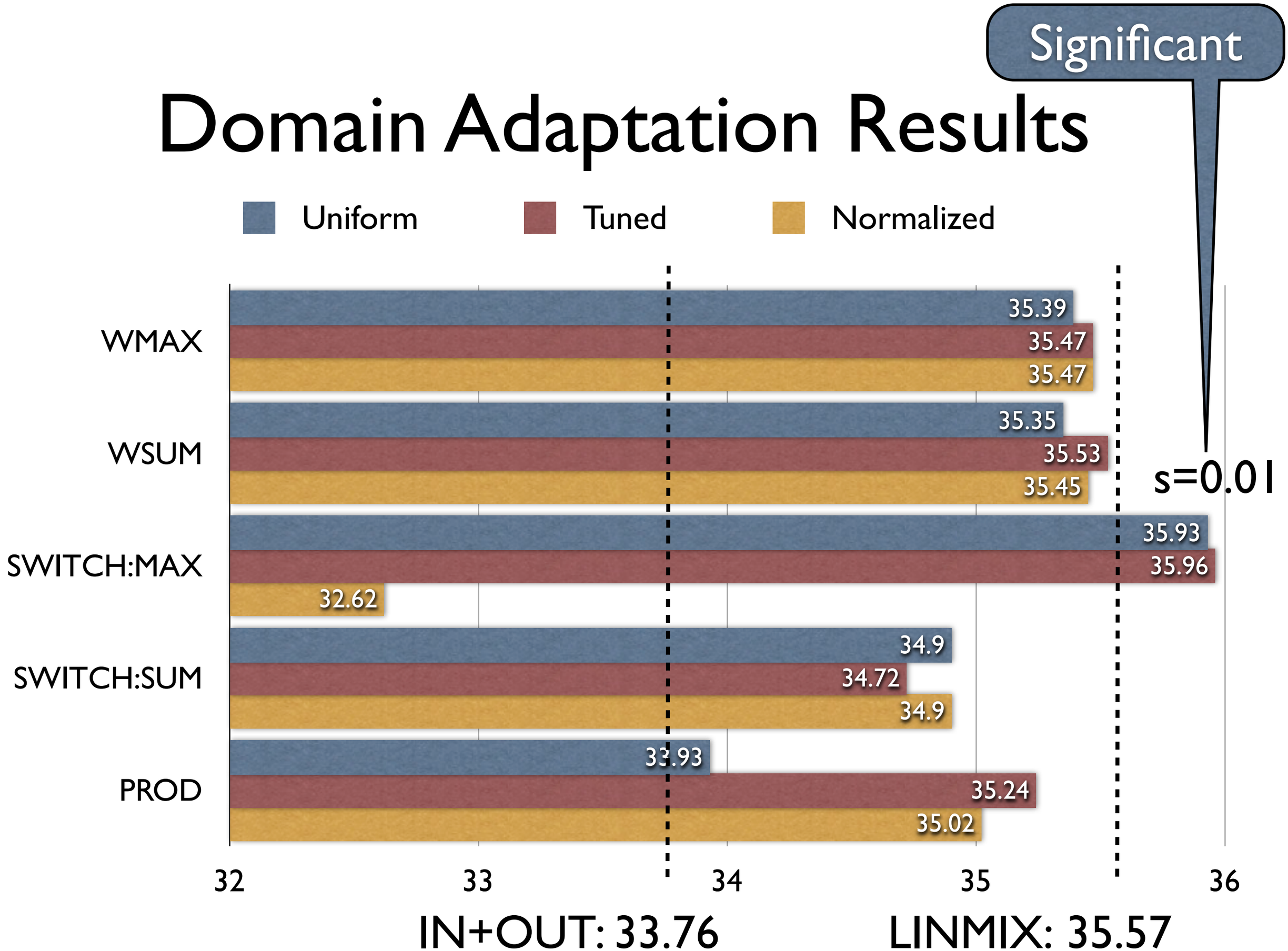
EuroParl (fr-en)

Train	1.3M

Domain Adaptation

- Scaling the model scores using ensemble weights:
  - Find the appropriate model scores that can participate in an ensemble.
  - We use CONDOR (Vanden Berghen and Bersini, 2005) which uses Powell's algorithm and no gradient information.
  - Component weights for each mixture operation is tuned on the dev set.

Domain Adaptation Results



Example

SOURCE	aménorrhée , menstruations irrégulières
REF	amenorrhoea , irregular menstruation
IN	amenorrhoea , menstruations irrégulières
OUT	aménorrhée , irregular menstruation
ENSEMBLE	amenorrhoea , irregular menstruation

Example

SOURCE	le traitement par naglazyme doit être supervisé par un médecin ayant l' expérience de la prise en charge des patients atteints de mps vi ou d' une autre maladie métabolique héréditaire .
REF	naglazyme treatment should be supervised by a physician experienced in the management of patients with mps vi or other inherited metabolic diseases .
IN	naglazyme treatment should be supervisé by a doctor the with in the management of patients with mps vi or other hereditary metabolic disease .
OUT	naglazyme 's treatment must be supervised by a doctor with the experience of the care of patients with mps vi. or another disease hereditary metabolic .
ENSEMBLE	naglazyme treatment should be supervised by a physician experienced in the management of patients with mps vi or other hereditary metabolic disease .

Multi-metric optimization

Joint work with Baskaran Sankaran and Kevin Duh

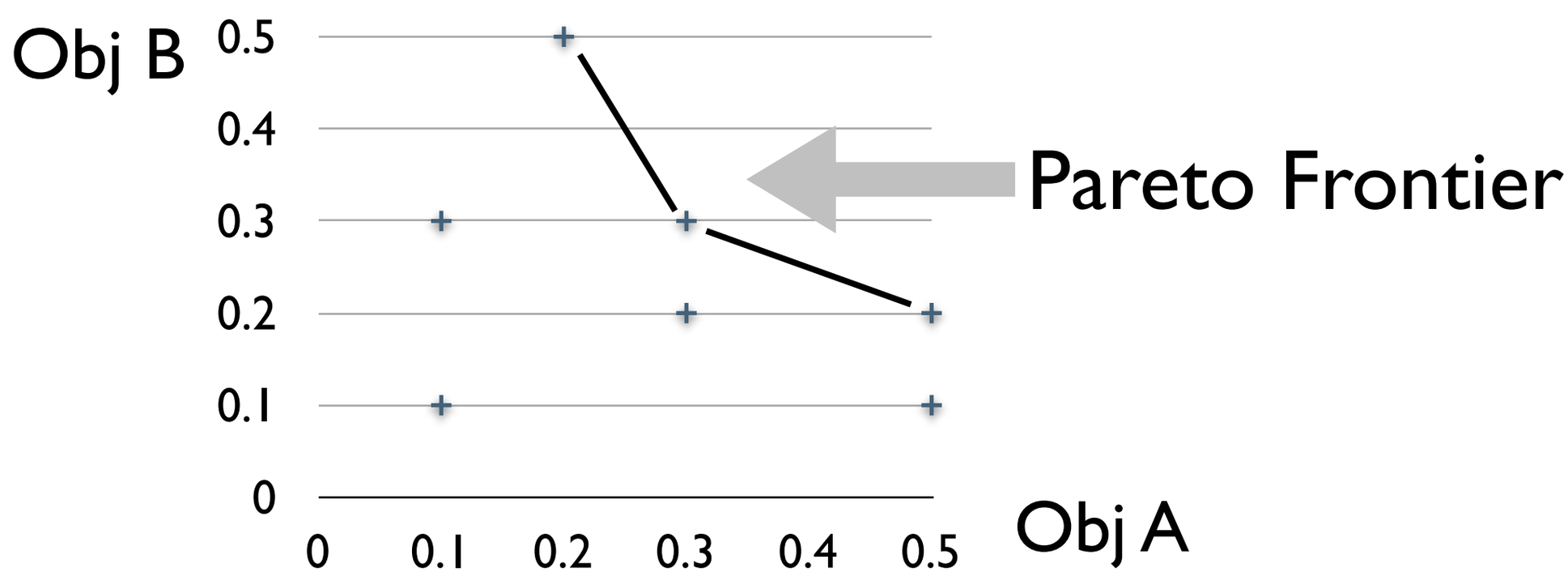
Multi-metric optimization

- Quite a few proposals for MT evaluation.
- In this talk, the focus is on BLEU, RIBES, TER, METEOR.
- Can other metrics be useful as a loss function for training SMT systems. (be useful how?)
- Most systems tune towards BLEU and test on BLEU. Can other metrics provide a second opinion?

Multi-objective optimization

$$\max_w \left(F_1(w), F_2(w), \ldots, F_k(w)\right)$$

- Find one w that simultaneously optimizes k objectives.
- A well formed notion of optimality wrt multiple objectives: Pareto optimality



Finding Pareto points

- Duh et al (ACL 2012) give an algorithm called PMO-PRO that finds Pareto optimal points as part of the tuning step.
- PRO (May and Hopkins, EMNLP 2011) show a pairwise ranking classifier can be used to train an SMT log-linear model.
- PMO-PRO puts Pareto points as positive examples and low scoring non-Pareto points as negative examples.
- This can be used to find Pareto points in the dev set.

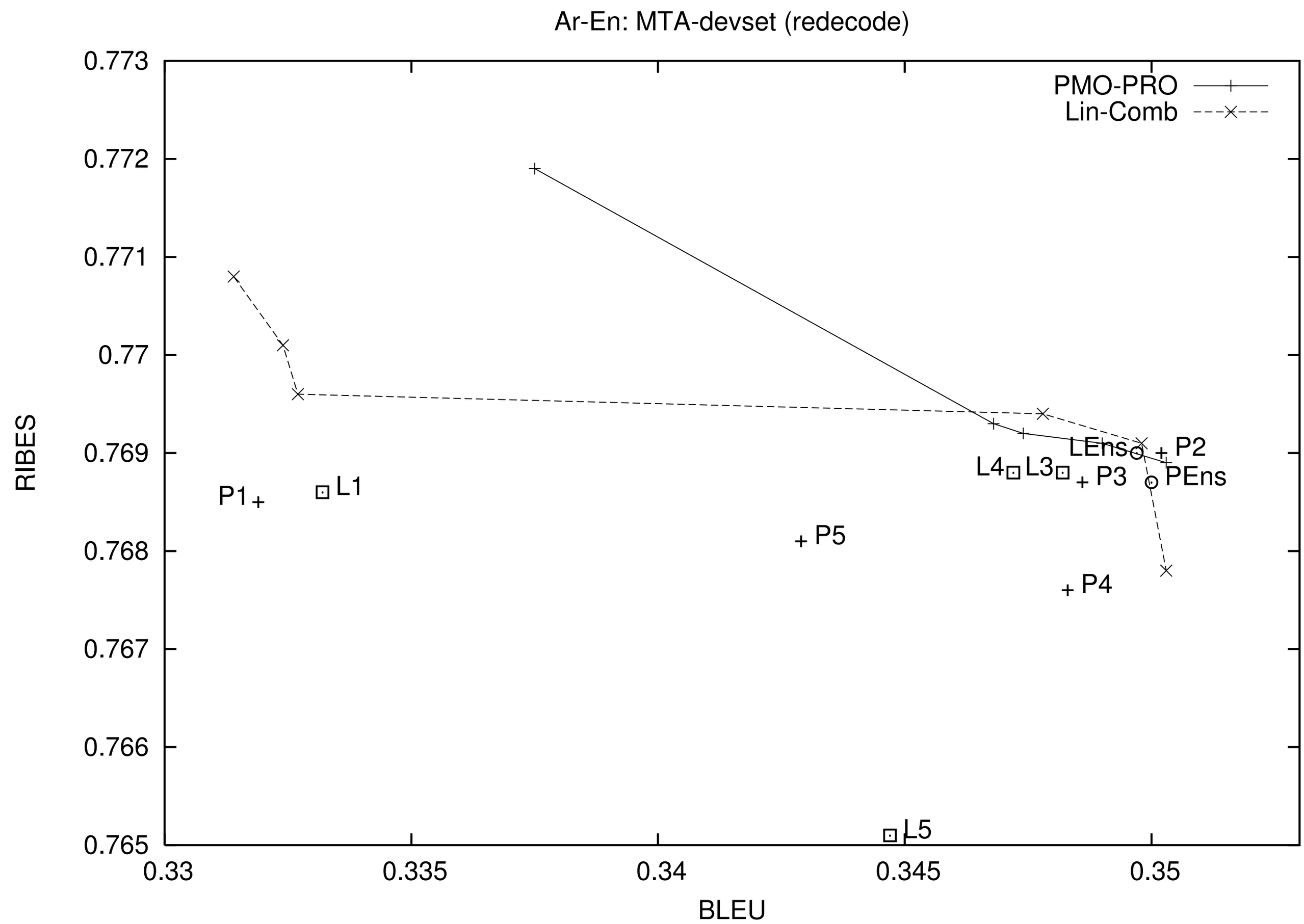
Using the Pareto Points

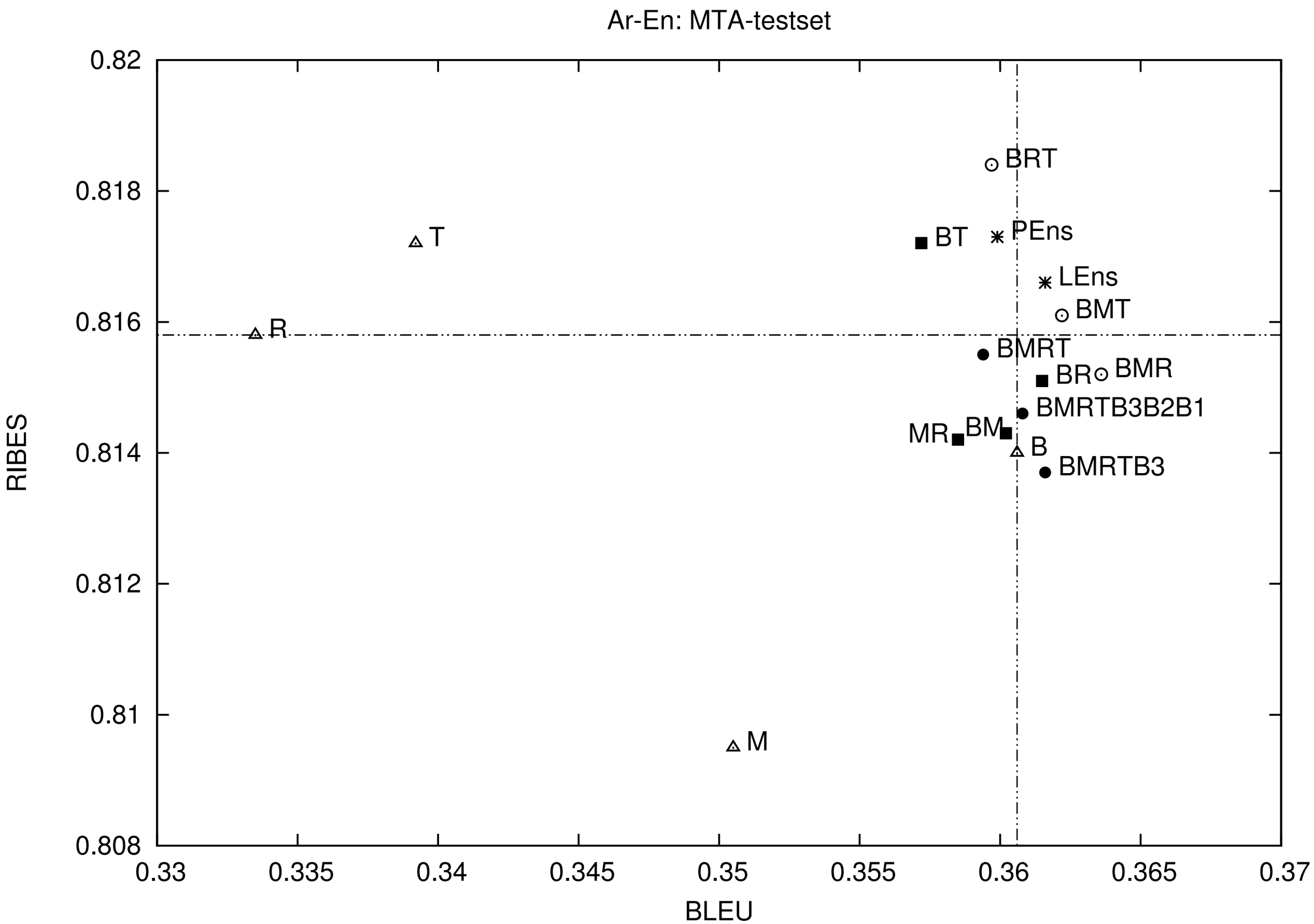
- Each Pareto point in the dev set is a weight vector that produced that point.
- **PMO-Ensemble**: Each of these weight vectors is a model and we can simply combine them using an ensemble model.
- **Union**: Take the union of "good" points wrt multiple objectives as positive examples and vice versa for negative examples. Simpler version of PMO-PRO.

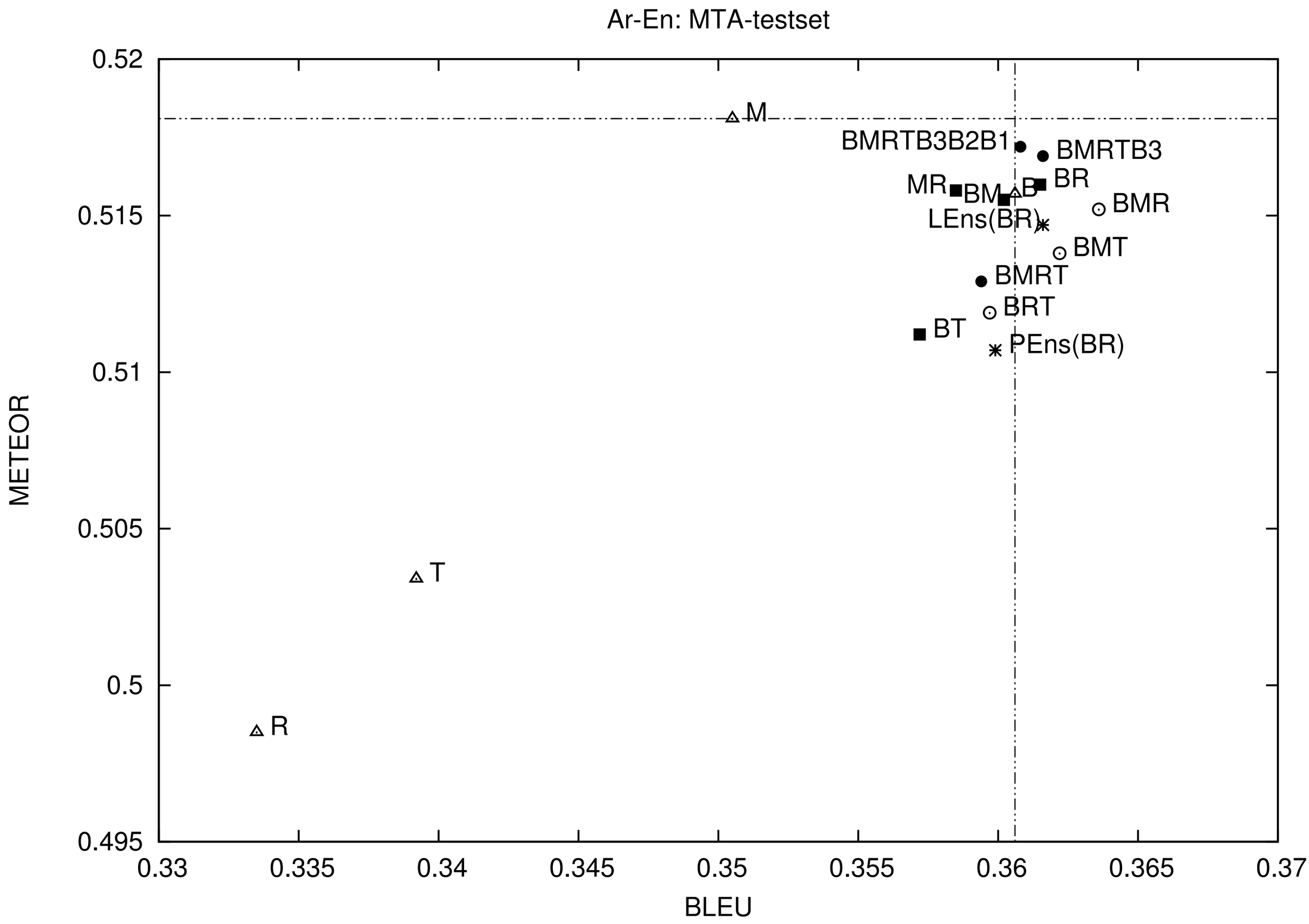
Using the Pareto Points

$$\max_w (F_1(w), F_2(w), \ldots, F_k(w))$$

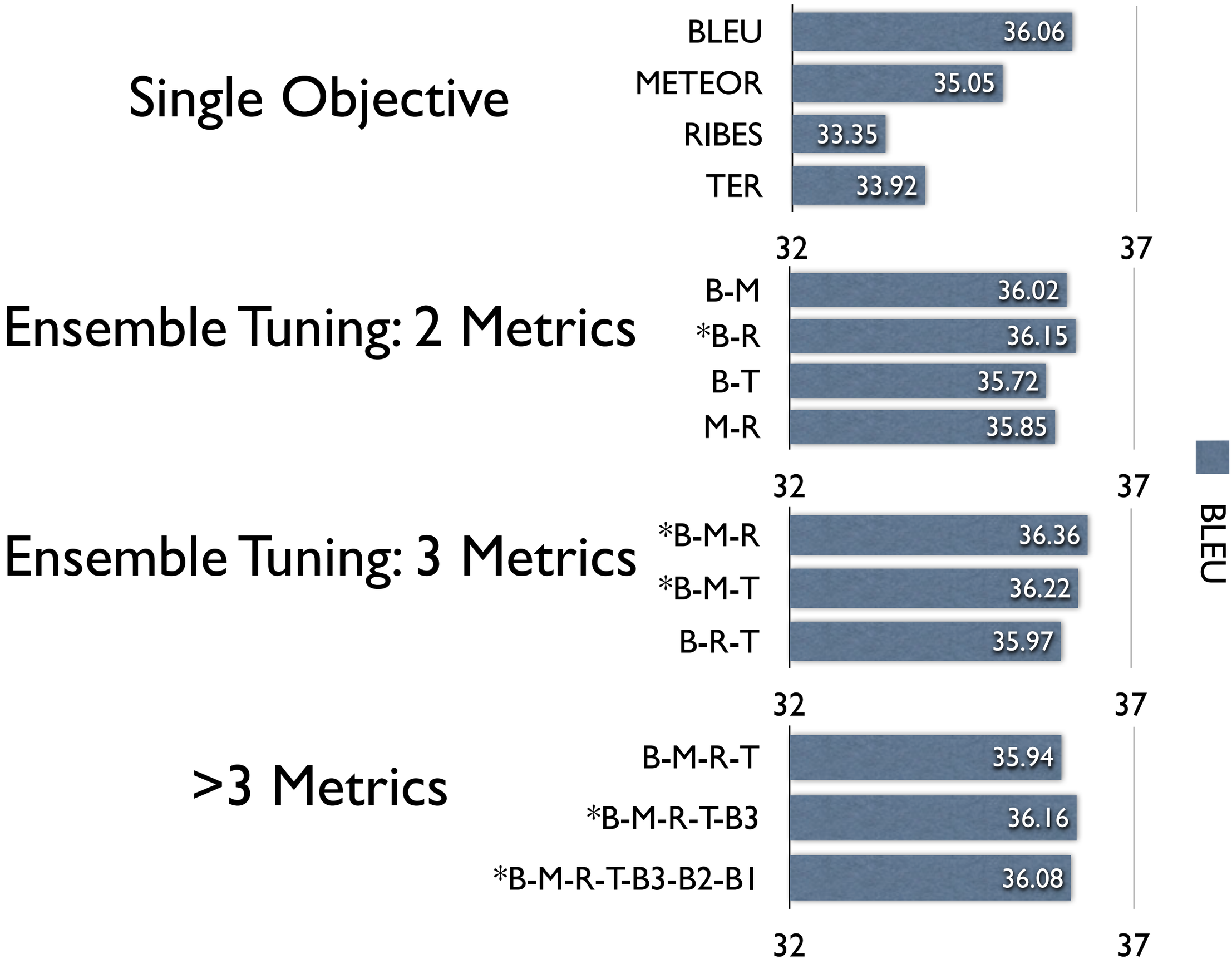
- **Ensemble Tuning**:
  - For each $F_i(w)$ perform error rate tuning (using PRO) to obtain the best w_i according to F_i in each iteration of tuning.
  - When decoding the dev set for the next search for w use an ensemble model with the same features but weights: $w_1, \ldots, w_k$
  - Tune the ensemble model hyperparameters using PMO-PRO to get Pareto points in the ensemble.
  - Repeat.


Ar-En: MTA-devset (redecode)
PMO-PRO
Lin-Comb
RIBES
BLEU
0.773
0.772
0.771
0.77
0.769
0.768
0.767
0.766
0.765
0.33
0.335
0.34
0.345
0.35
P1
L1
P5
L5
L4
L3
LEns
P2
P3
PEns
P4

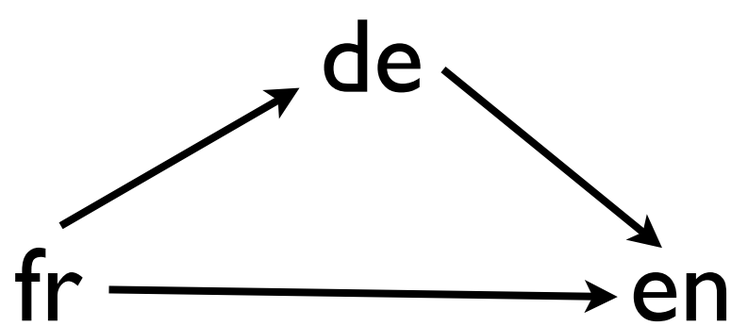

Ar-En: MTA-testset
RIBES
BLEU
0.82
0.818
0.816
0.814
0.812
0.81
0.808
0.33
0.335
0.34
0.345
0.35
0.355
0.36
0.365
0.37
BRT
T
BT
PEns
LEns
BMT
R
BMRT
BR
BMR
BMRTB3B2B1
MR
BM
B
BMRTB3
M


Ar-En: MTA-testset
METEOR
BLEU
0.52
0.515
0.51
0.505
0.5
0.495
0.33
0.335
0.34
0.345
0.35
0.355
0.36
0.365
0.37
M
BMRTB3B2B1
BMRTB3
MR
BM
B
BR
BMR
LEns(BR)
BMT
BMRT
BRT
BT
PEns(BR)
T
R

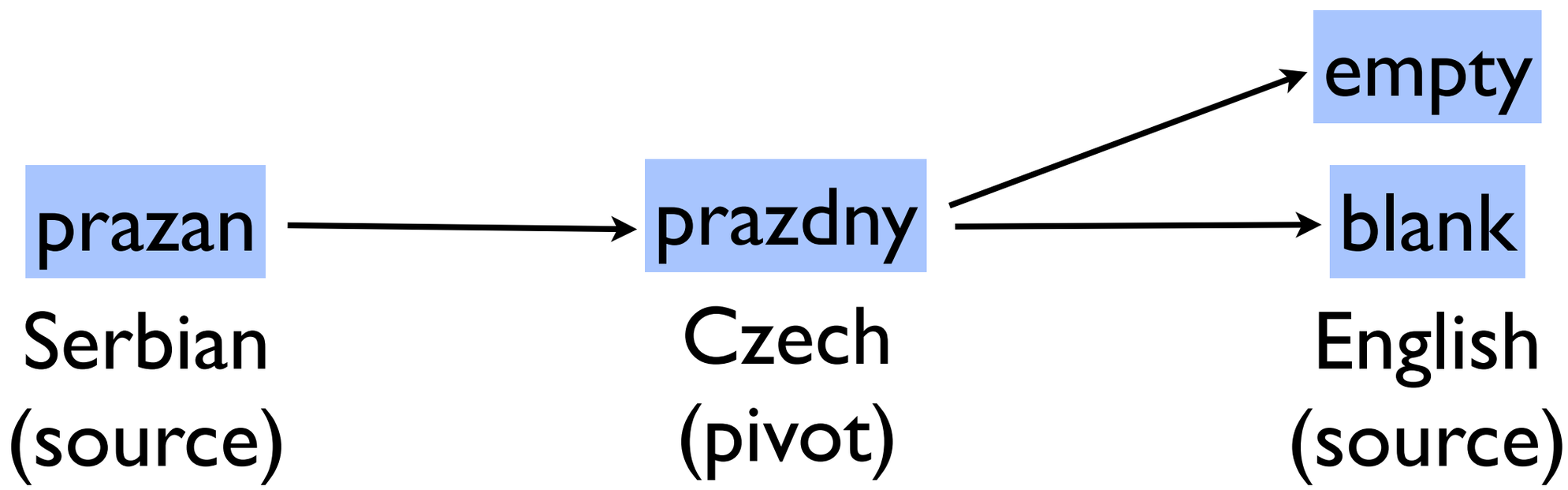
Can multi-metric tuning help a single metric?



Pivot language triangulation



Triangulation



- **Direct**: From source to target using available data
- Phrase-based triangulation (Cohn and Lapata, ACL 2007)

Phrase-based Triangulation

- Create a triangulated phrase table using source, pivot and target language data.

$$
\begin{aligned}
p(e|f) &= \textstyle\sum_i p(e, i|f) \\
&= \textstyle\sum_i p(e|i, f)p(i|f) \\
&\approx \textstyle\sum_i p(e|i)p(i|f)
\end{aligned}
$$

- **Mixture**: Interpolate the triangulated model with the direct source to target model

Ensemble-based Triangulation

- Typically, one pivot language does not provide an improvement.
- More pivot languages used, the better.
- Ensemble-based Triangulation: an ensemble of different pivot models.
- Each one goes from source to pivot_k to target for pivot_1 to pivot_M
- **Ensemble**: Finally add the direct source to target model to the ensemble as well.

Experiment

- Compare Direct, Mixture and Ensemble
- Use EuroParl (en, fr, de, es, it)
- Each source language is translated to a target language through 3 pivot languages.
- For example, en to fr goes through de, es, it
- 10K sentence pairs (as in Cohn and Lapata, ACL 2007)
- To be done: 700K EuroParl corpus.

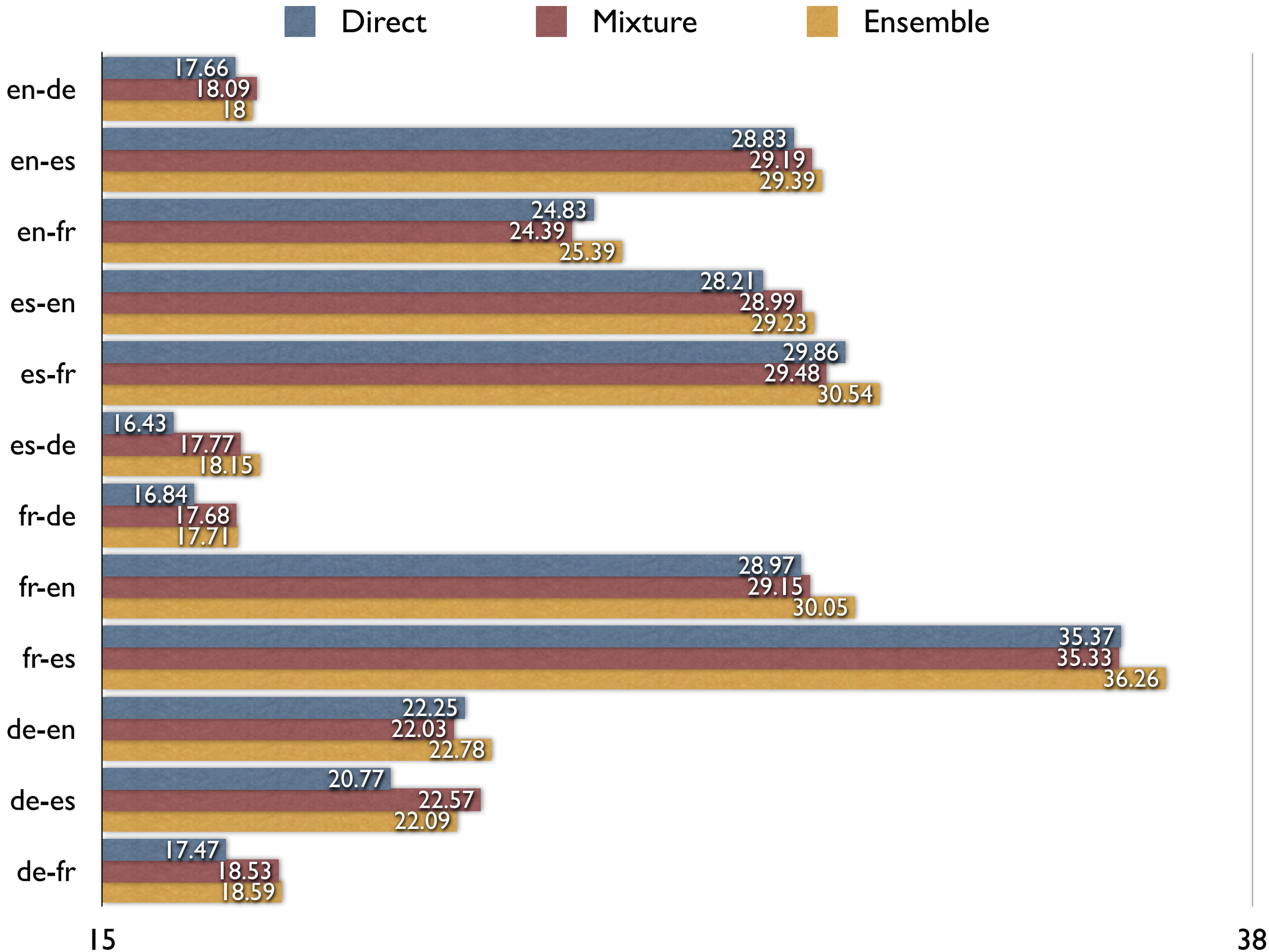
Ensemble Triangulation

	en	es	fr	de
en	-	+0.2	+1.0	-0.09
es	+0.24	-	+1.06	+0.38
fr	+0.9	+0.93	-	+0.03
de	+0.75	-0.48	+0.06	-

Comparison of Ensemble Model and Mixture model
(Cohn and Lapata, ACL 2007)

Summary

- Ensemble models combine translation models during SMT decoding.
  - Allows more dynamic combination methods.
  - Do not need to be tuned (with uniform weights)
- Applied to:
  - Domain adaptation
  - Multi-metric optimization
  - Pivot language triangulation


Direct
Mixture
Ensemble
en-de
17.66
18.09
18
en-es
28.83
29.19
29.39
en-fr
24.83
24.39
25.39
es-en
28.21
28.99
29.23
es-fr
29.86
29.48
30.54
es-de
16.43
17.77
18.15
fr-de
16.84
17.68
17.71
fr-en
28.97
29.15
30.05
fr-es
35.37
35.33
36.26
de-en
22.25
22.03
22.78
de-es
20.77
22.57
22.09
de-fr
17.47
18.53
18.59
15
38