Learning to Translate with Multiple Objectives

Kevin Duh (NAIST)
Katsuhito Sudoh (NTT)
Xianchao Wu (Baidu)
Hajime Tsukada (NTT)
Masaaki Nagata (NTT)

How many metrics have been proposed for MT evaluation?

RIBES

DepOverlap

IMPACT

RTE

NIST

TER

BLEU

WER

RED

ParaEval

GTM

PER

TESLA

METEOR

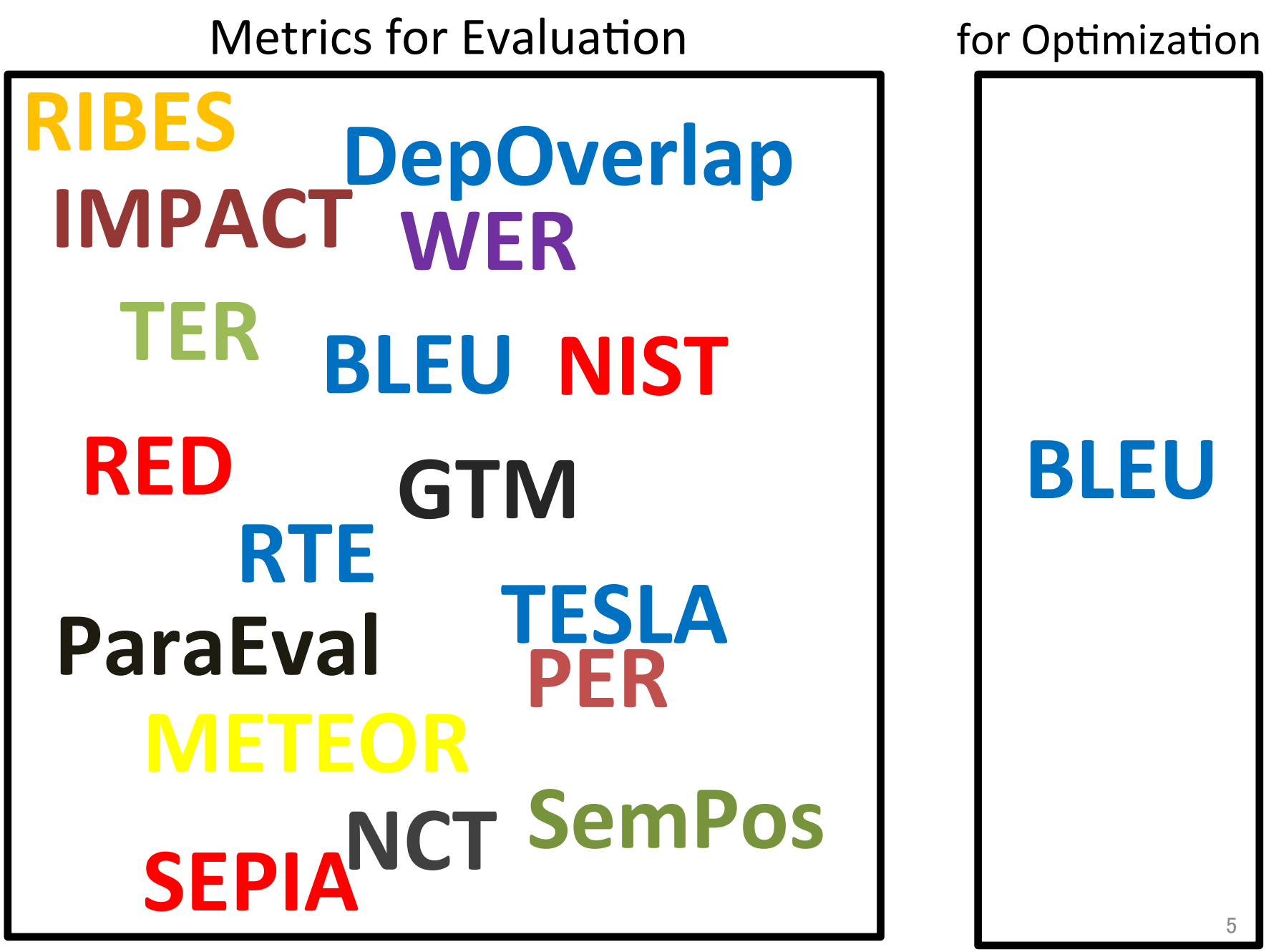
SemPos

NCT

SEPIA

How many metrics are used for MT optimization?

BLEU


Metrics for Evaluation
for Optimization
RIBES
DepOverlap
IMPACT
WER
TER
BLEU
NIST
RED
GTM
RTE
TESLA
ParaEval
PER
METEOR
NCT
SemPos
SEPIA
BLEU

Each metric has its strengths.

→ Optimize with multiple metrics

Outline

1. Motivation
2. Basic Concepts: Pareto optimality
3. Multiobjective optimization in MT
4. Experiments

Outline

1. Motivation
2. Basic Concepts: Pareto optimality
3. Multiobjective optimization in MT
4. Experiments

Multiobjective optimization

$$\max_{w}[F_1(w), F_2(w), ..., F_K(w)]$$

Find one w that simultaneously optimizes K objectives

But what does it mean to be "optimum"?

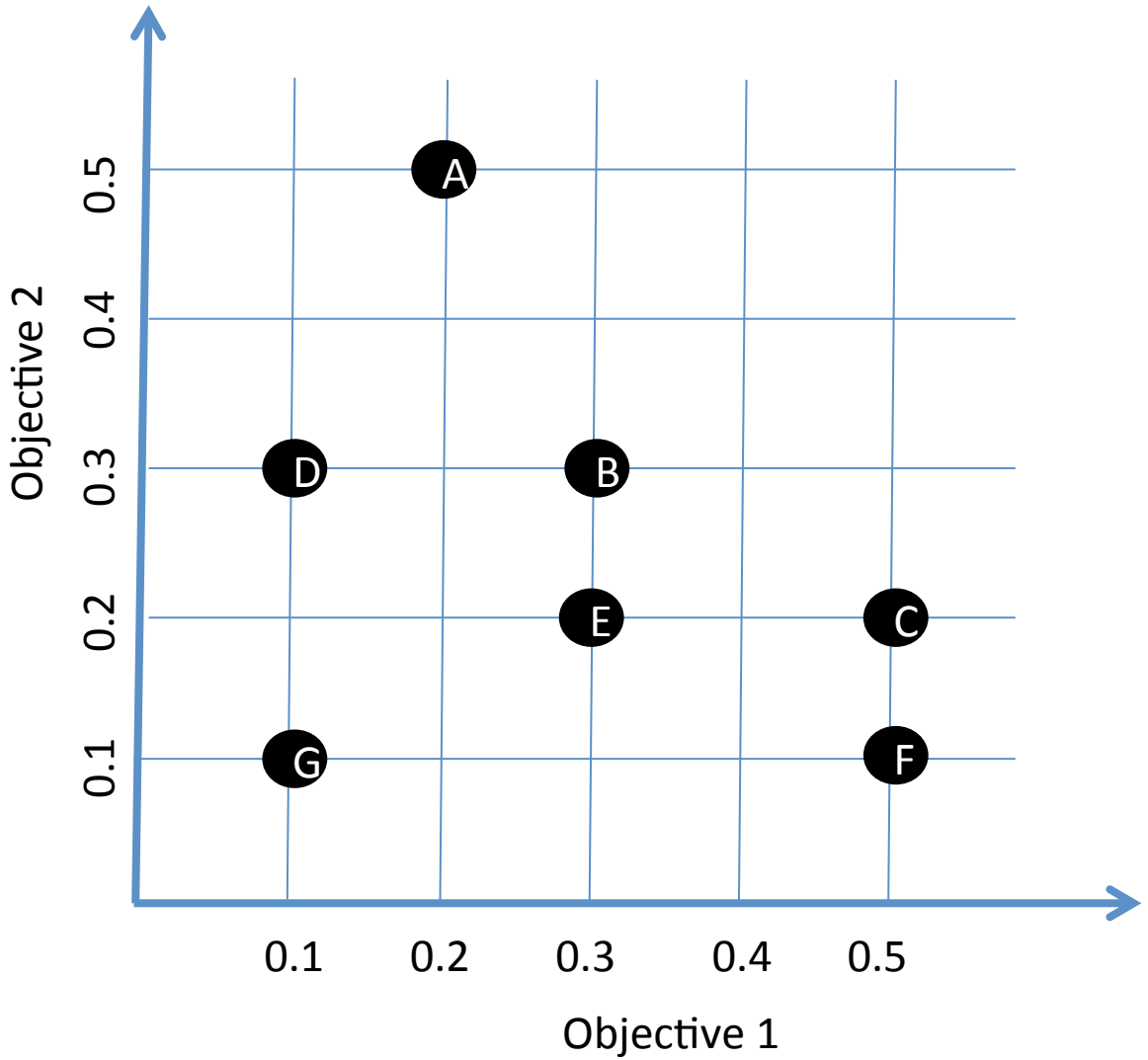
Multiobjective optimization of your ACL Hotel

Hotel	Customer Reviews	Distance to Conference Center	Price (KRW)
The Shilla Jeju		utes	230,000
Hotel Lotte Jeju			200,000
Poonglim Resort	3 stars	utes	
Hana Hotel	3 stars	5 minutes	
Gyulhyanngi Pension	2 stars	10 minutes	

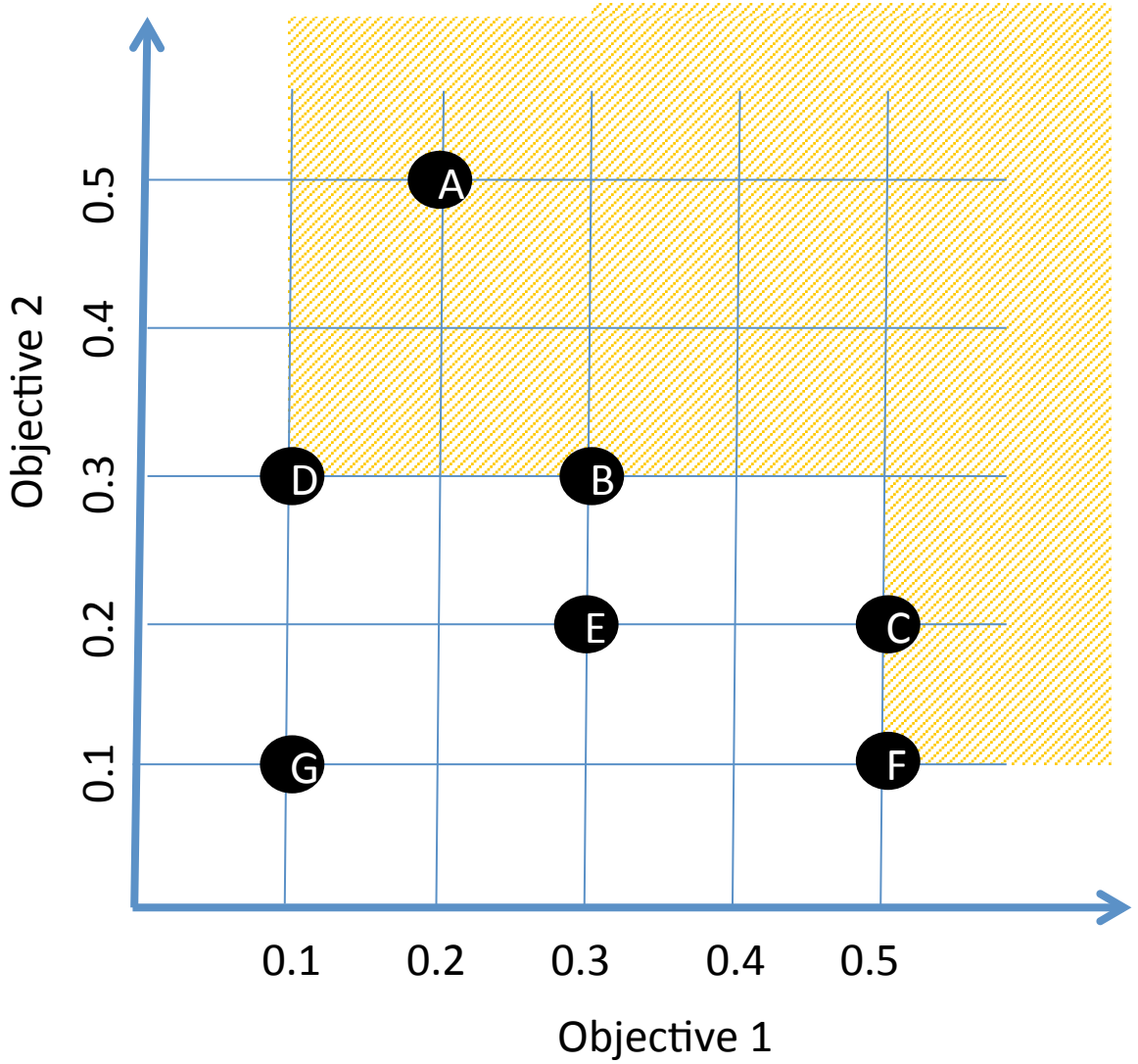
You're irrational! That choice is not Pareto Optimal!

Vilfredo Pareto, Economist (1848-1923)

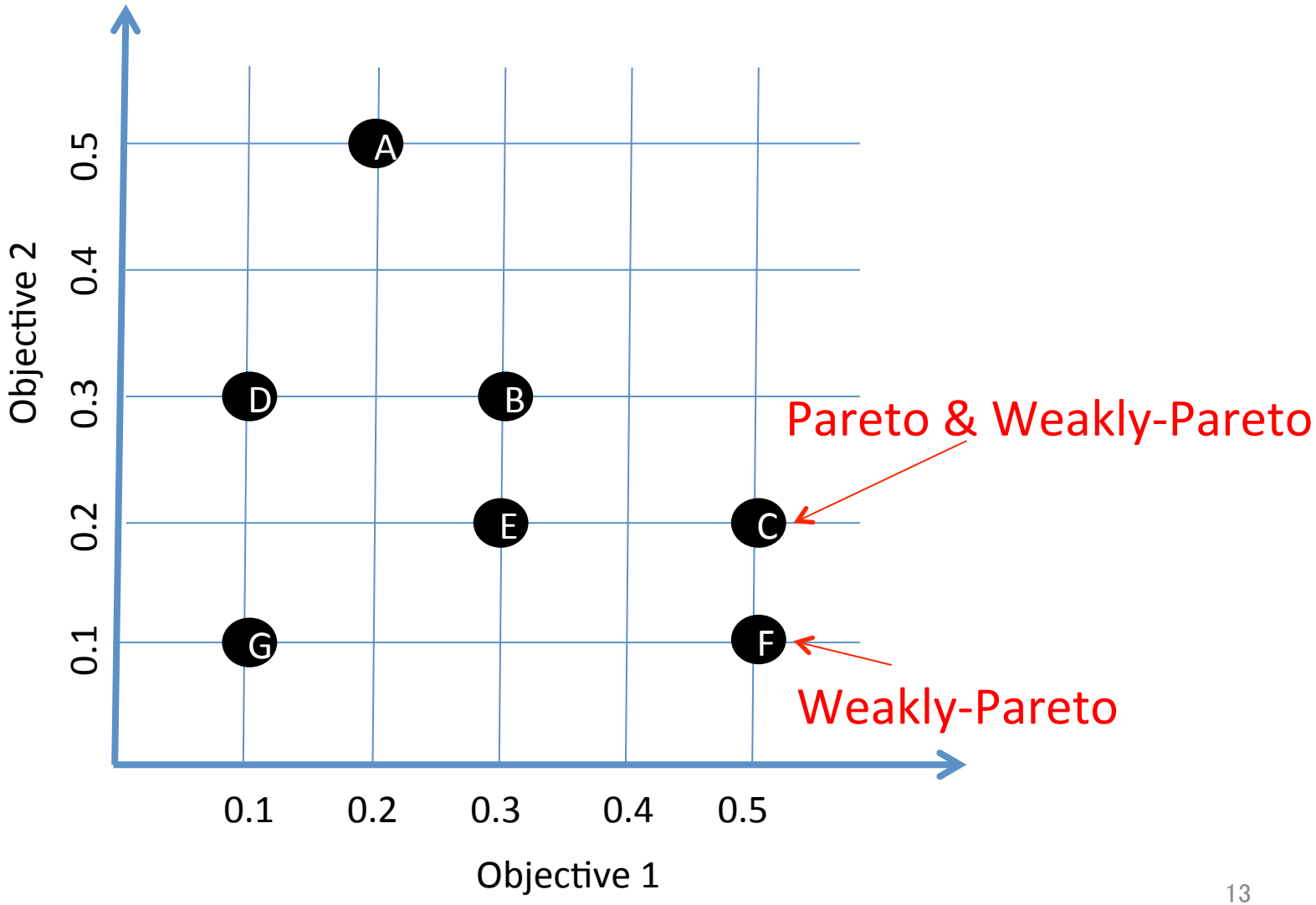
How to define optimality



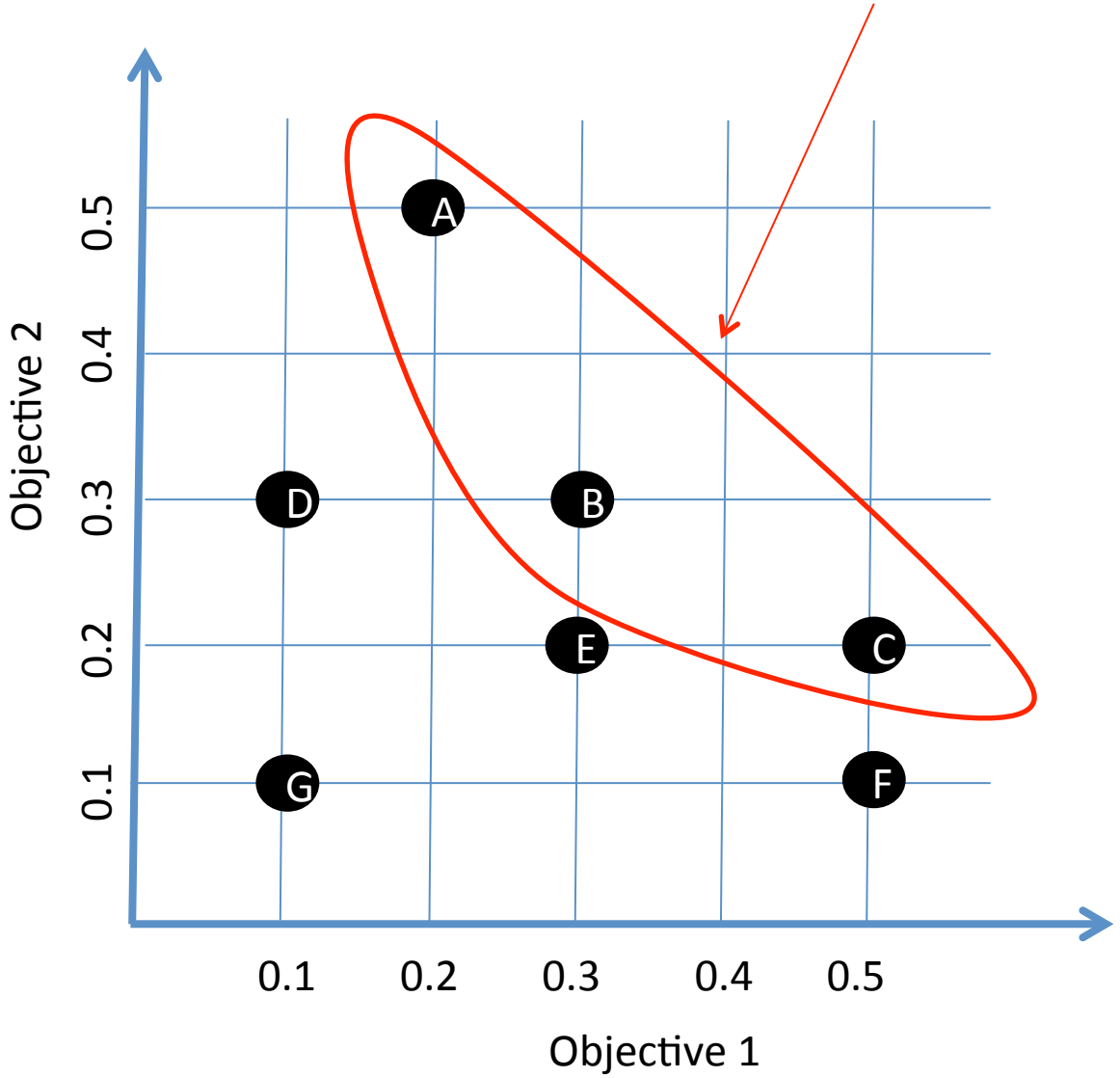
A point p is **weakly pareto-optimal** iff there does not exist another point q such that $F_k(q) > F_k(p)$ for all k



A point p is **pareto-optimal** iff there does not exist a q such that $F_k(q) >= F_k(p)$ for all k and $F_k(q) > F_k(p)$ for at least one k



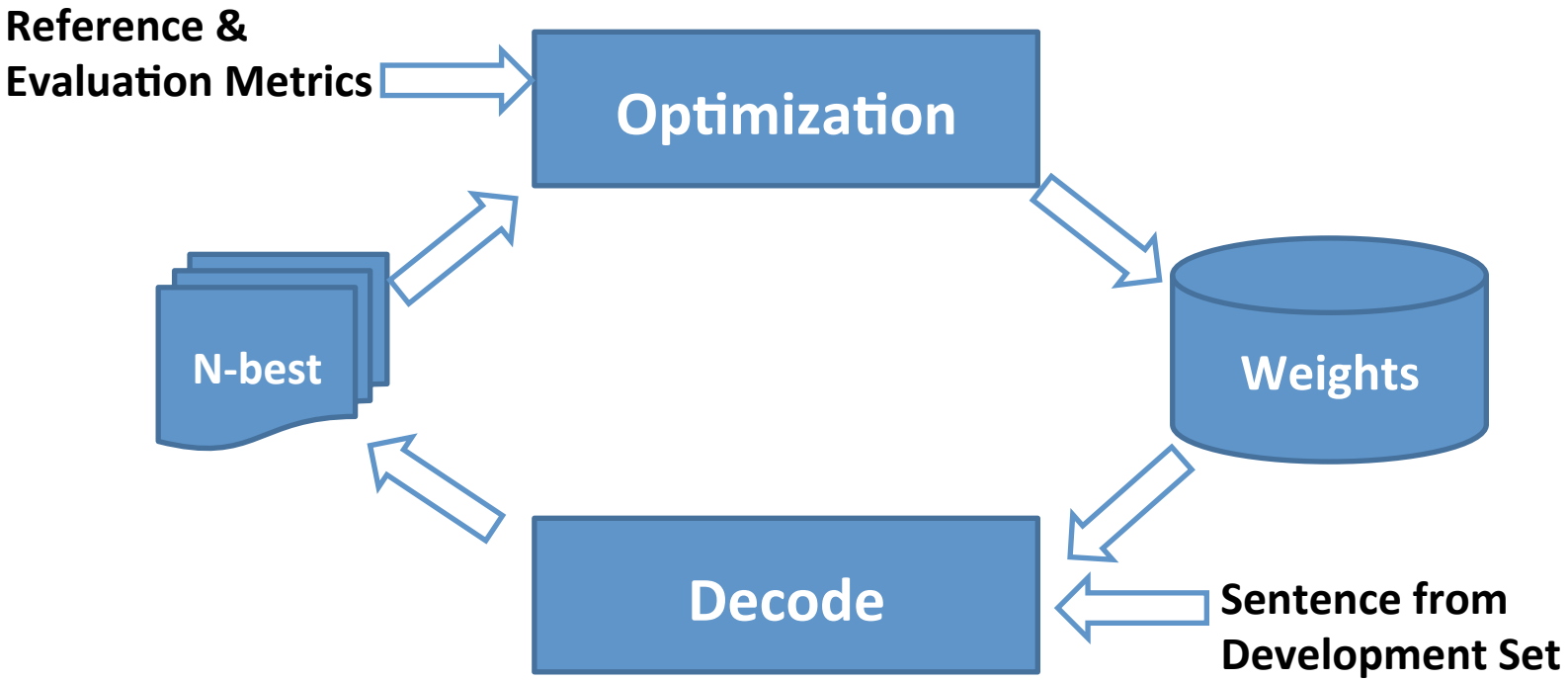
Given a set of points, the subset of pareto-optimal points form the **Pareto Frontier**



Outline

1. Motivation
2. Basic Concepts: Pareto optimality
3. Multiobjective optimization in MT
4. Experiments

Optimization in Machine Translation



Baseline:
Linear Combination

$$\max_{w} \sum_{k=1}^{K} \alpha_k F_k(w)$$

Importance of each objective

$$\alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$$

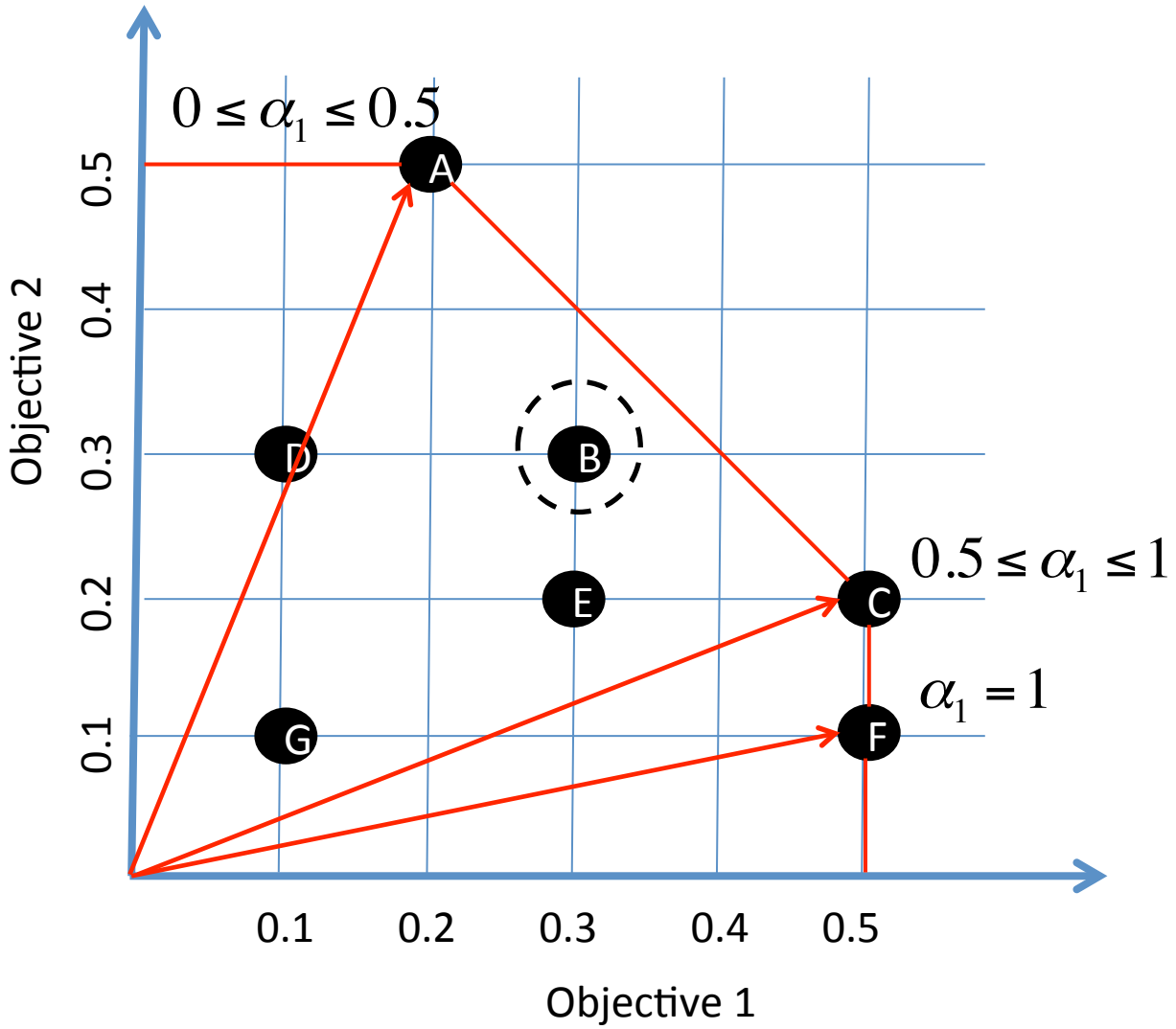
Advantages:

1. Single-objective tools can be used
2. Sufficiency: If w* is a solution, then it's Weakly Pareto

Disadvantages:

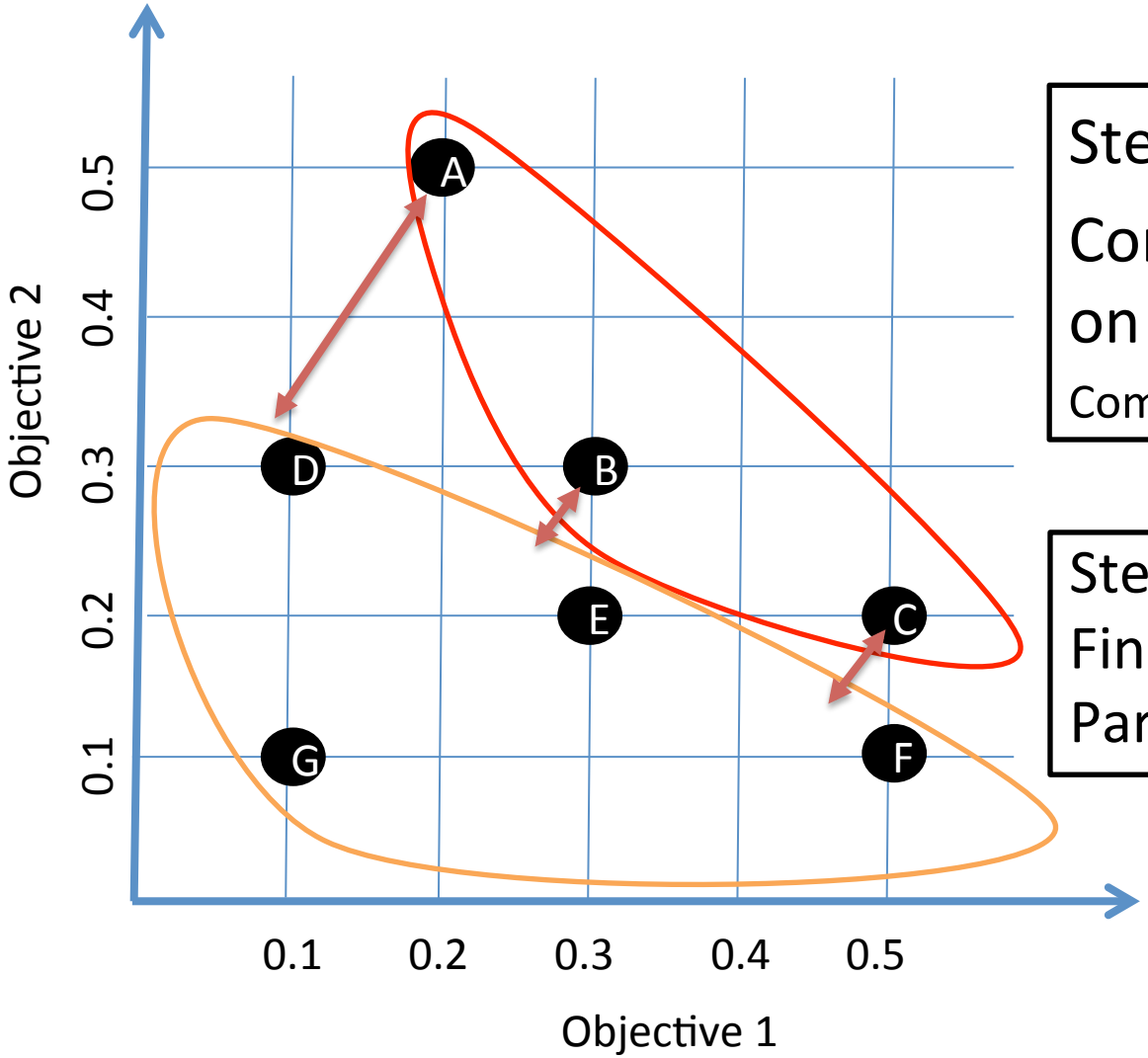
1. How to set α?
2. No Necessary Conditions: Some Pareto points can never been obtained, whatever setting of α.

Pareto points not on Convex Hull are missed



New method: Directly optimize Pareto Front

New method: Directly optimize Pareto Front



Step 1:

Compute Pareto Frontier on N-best List

Complexity O(#objective * N^2)

Step 2:

Find w separating Pareto vs. Non-Pareto

Multi-objective Pairwise Ranking Optimization

Regularizer

Slack

$$\min_{w} \| w \|^2 + c \sum_{ij} \xi_{ij}$$

Feature vector

$$\text{s.t. } w^T \Phi(x, y_i) - w^T \Phi(x, y_j) \geq 1 - \xi_{ij}$$

Input sentence

Good hypothesis

Poor hypothesis

$$\forall y_i \in ParetoFront, y_j \notin ParetoFront$$

i.e. score of pareto hypothesis should be higher than non-pareto hypotheses

Outline

1. Motivation
2. Basic Concepts: Pareto optimality
3. Multiobjective optimization in MT
4. Experiments

Experiment Setup

Task 1: NIST Zh-En

Optimize BLEU & NTER
NTER = max(1-TER,0)

Moses decoder, 7M train sentences, 1.6k dev, 8 features

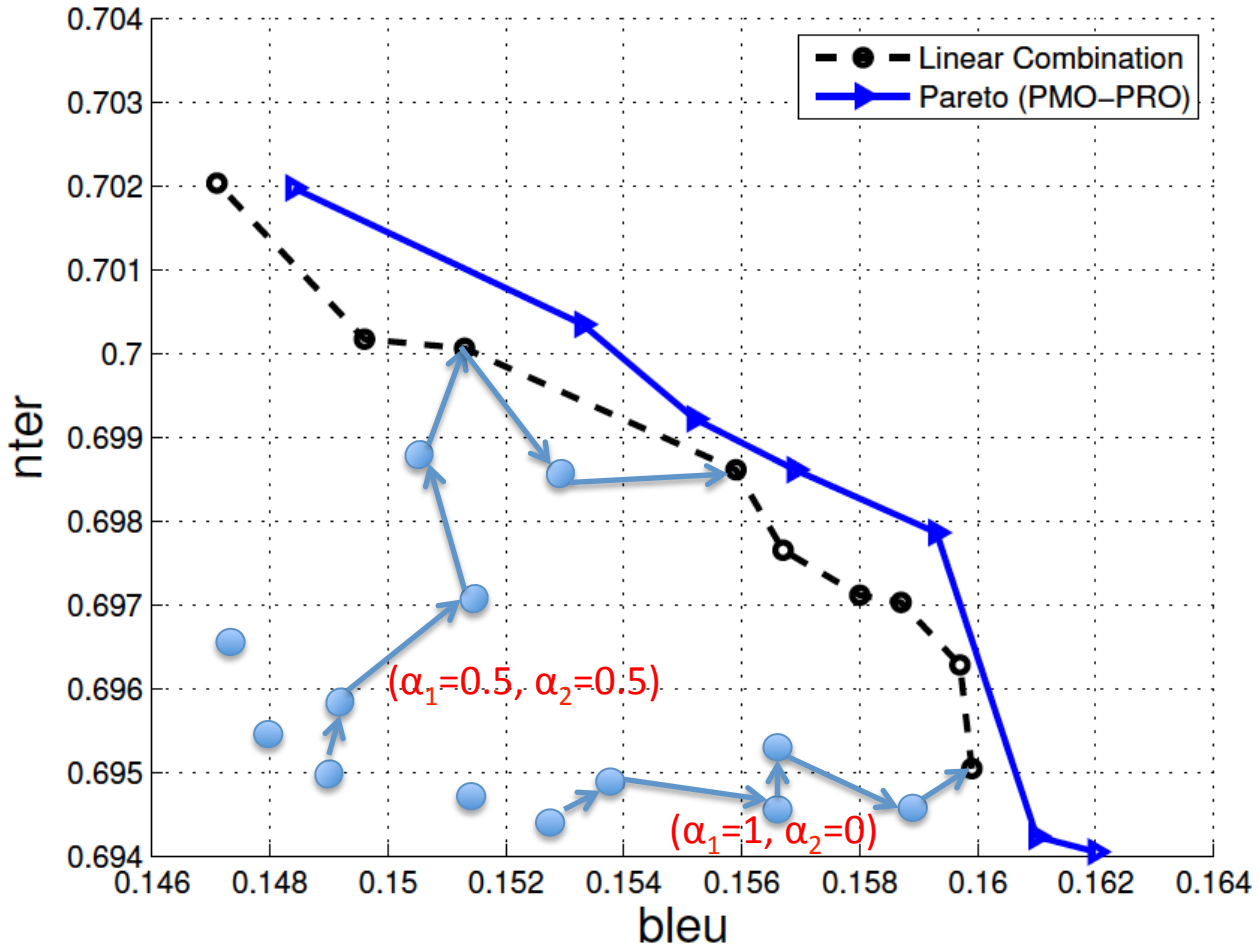
Task 2: PubMed En-Ja

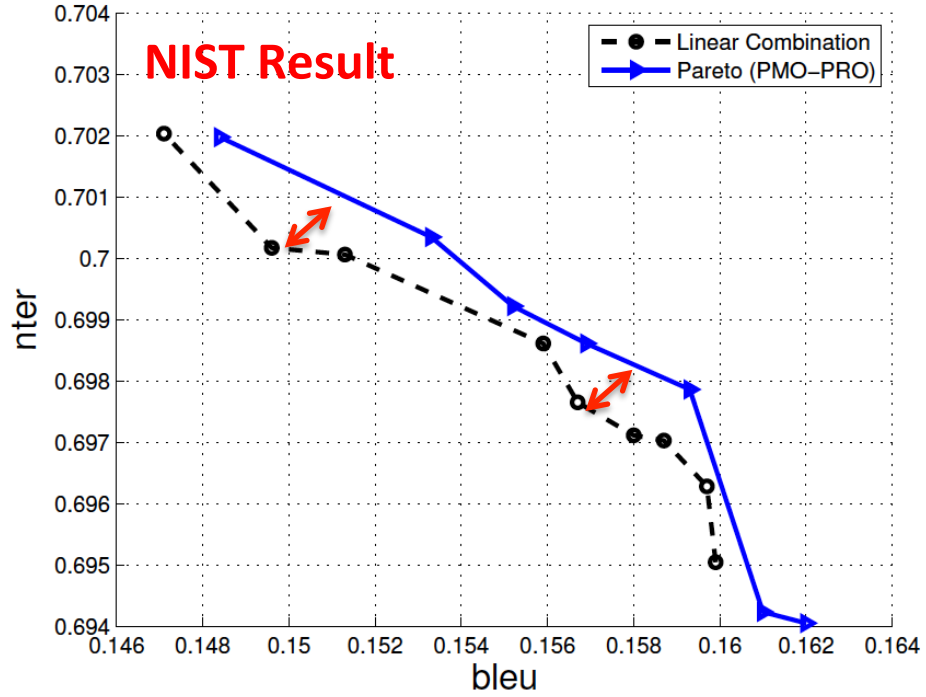
Optimize BLEU & RIBES
RIBES = permutation metric [Isozaki, EMNLP10]

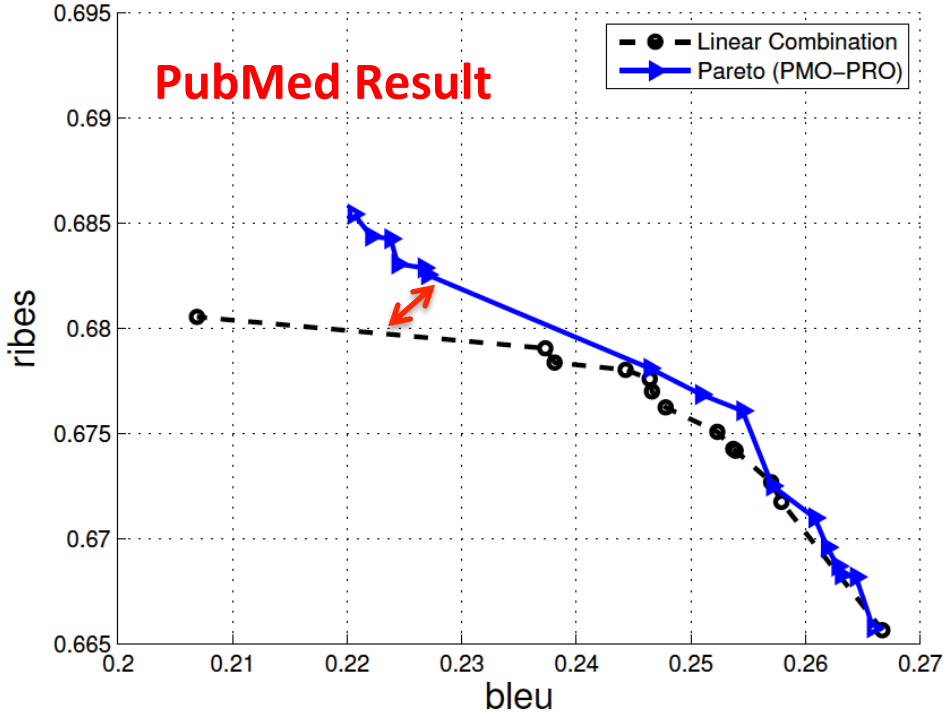
Moses decoder, 0.2M train sentences, 2k dev, 14 features

- Compare Linear Combination vs. Pareto
  - Both use pairwise rank optimization, but different objective.
  - For Linear Combination, multiple α settings (α_1 = {1,0.7,0.5,0.3,0})
  - 5 runs, 20 iterations each. Collect/visualize set of solutions.

Result Visualization

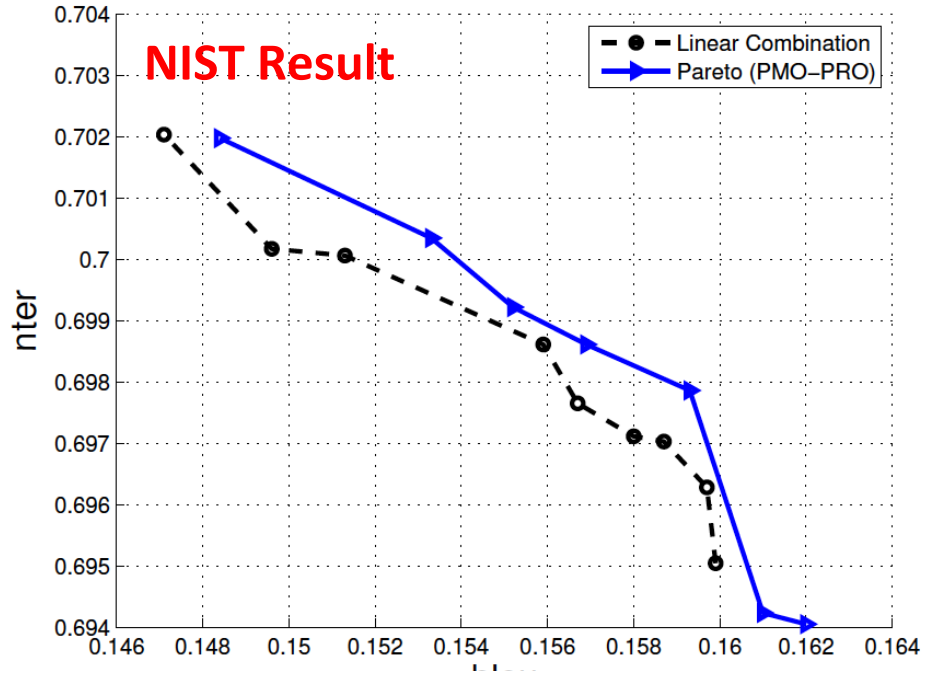


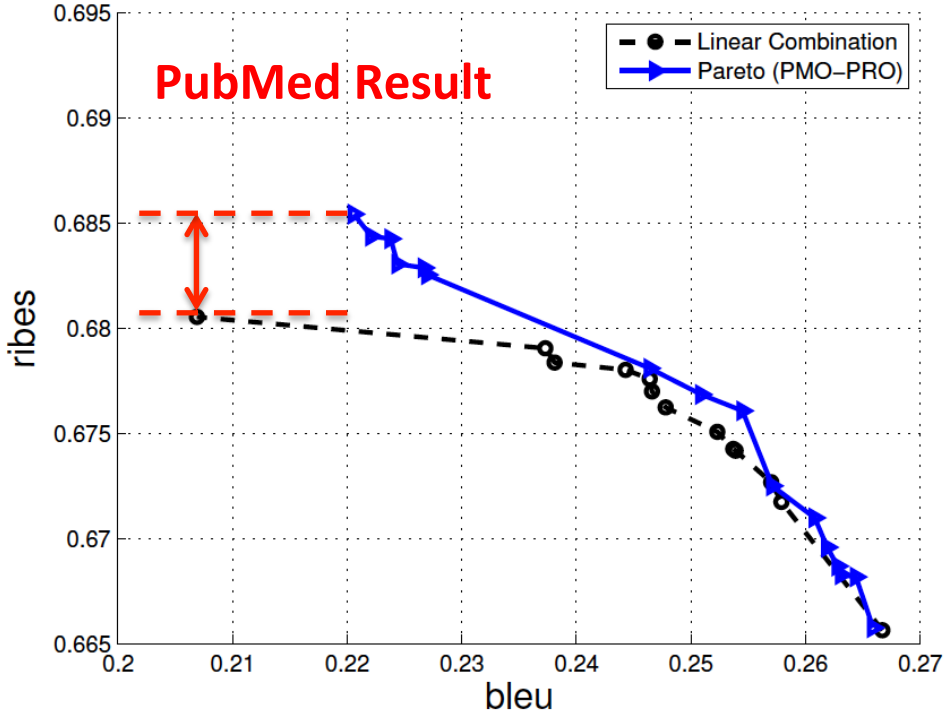




OBSERVARTIONS:

1. Pareto > Linear Combination for any α



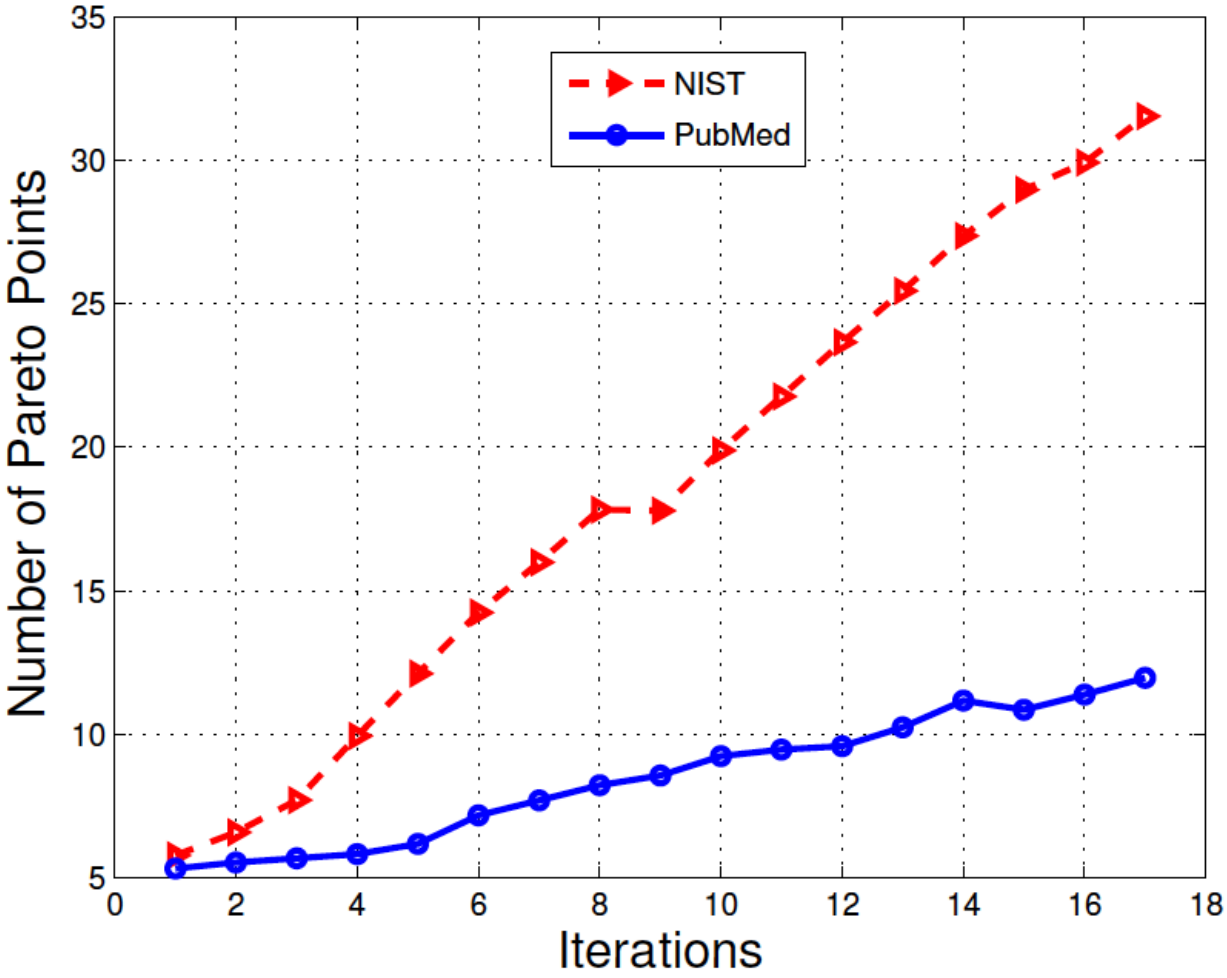


OBSERVARTIONS:

1. Pareto > Linear Combination for any α

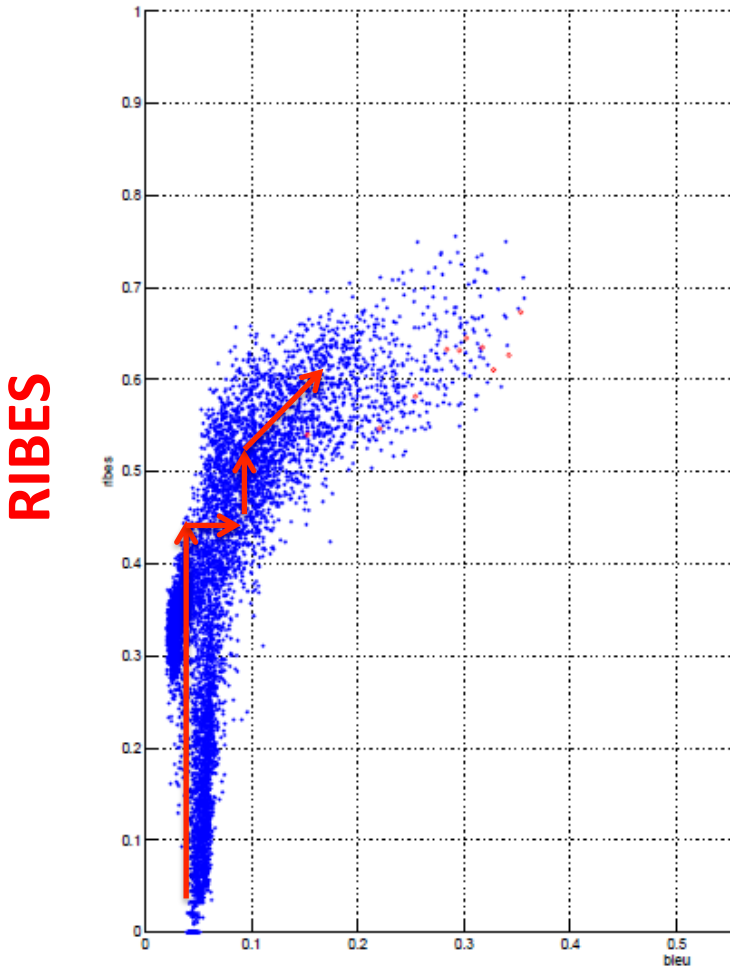
2. Metric tunability: Pareto outperform single-objective optimization of RIBES

Analysis: Number of Pareto Points

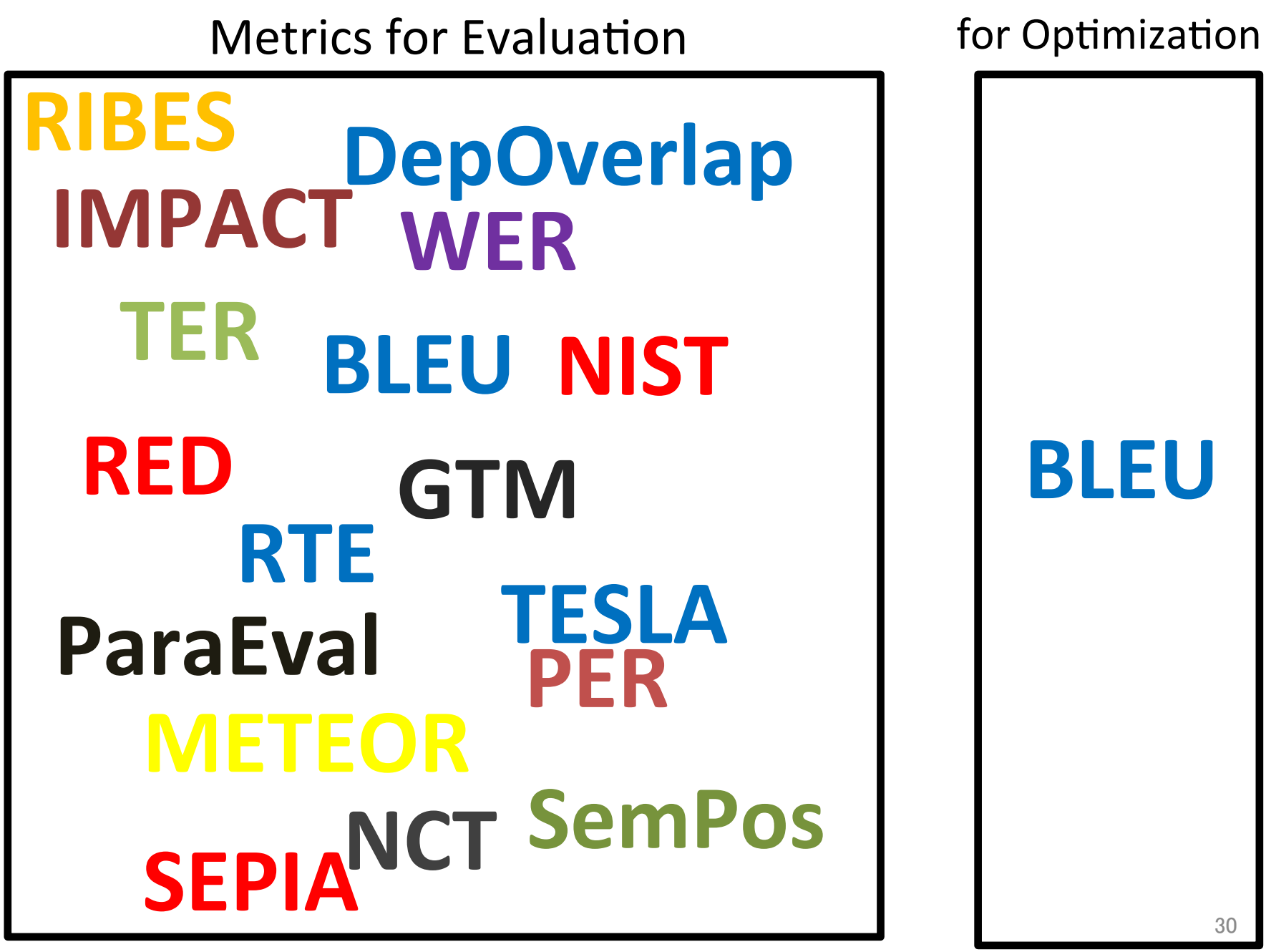


Analysis: Metric Tunability

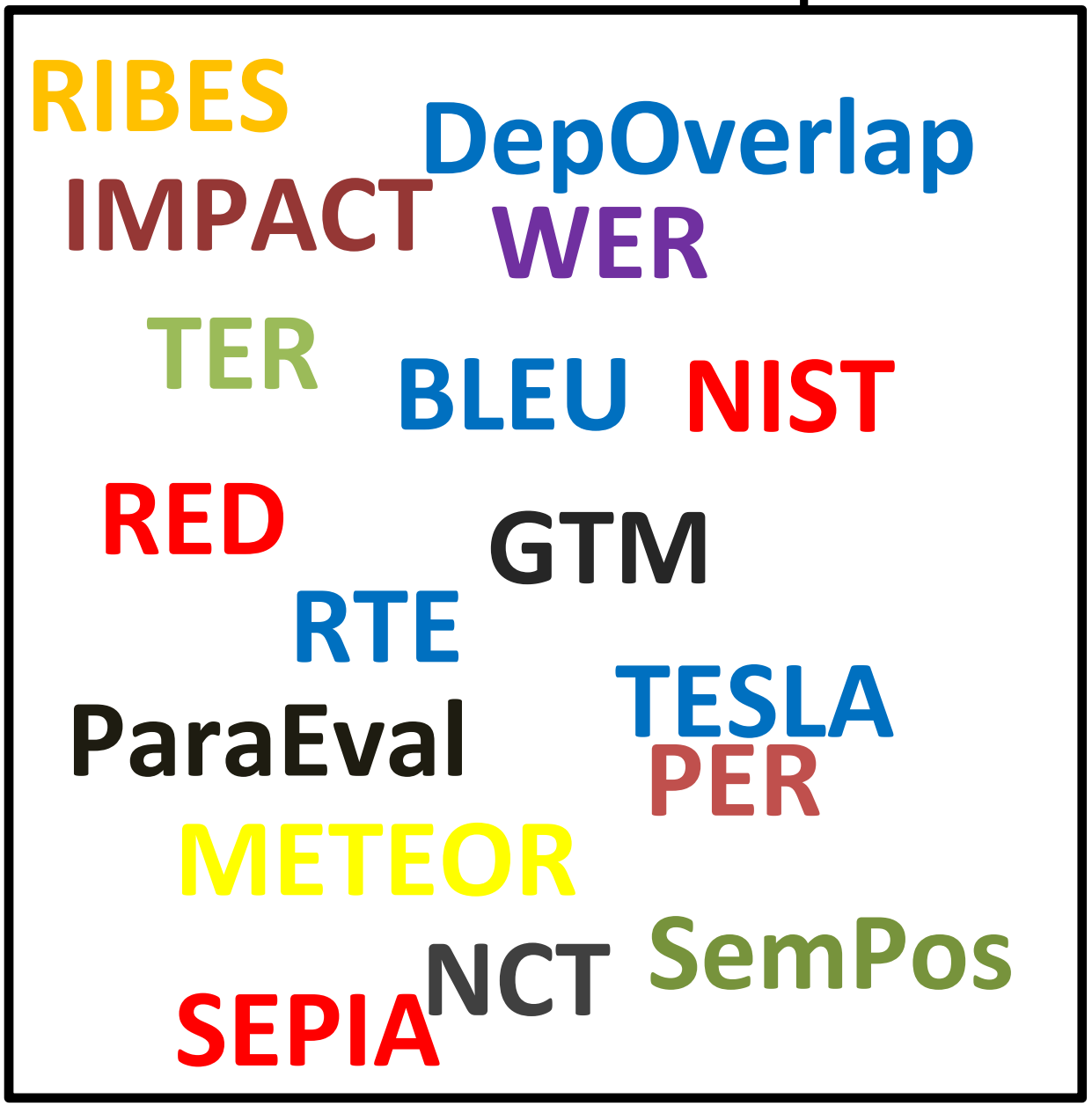
Sampling of 10k random w's



Summary & Final Thoughts


Metrics for Evaluation
for Optimization
RIBES
DepOverlap
IMPACT
WER
TER
BLEU
NIST
RED
GTM
RTE
TESLA
ParaEval
PER
METEOR
NCT
SemPos
SEPIA
BLEU

Metrics for Evaluation and Optimization



Vilfredo Pareto (1848-1923)

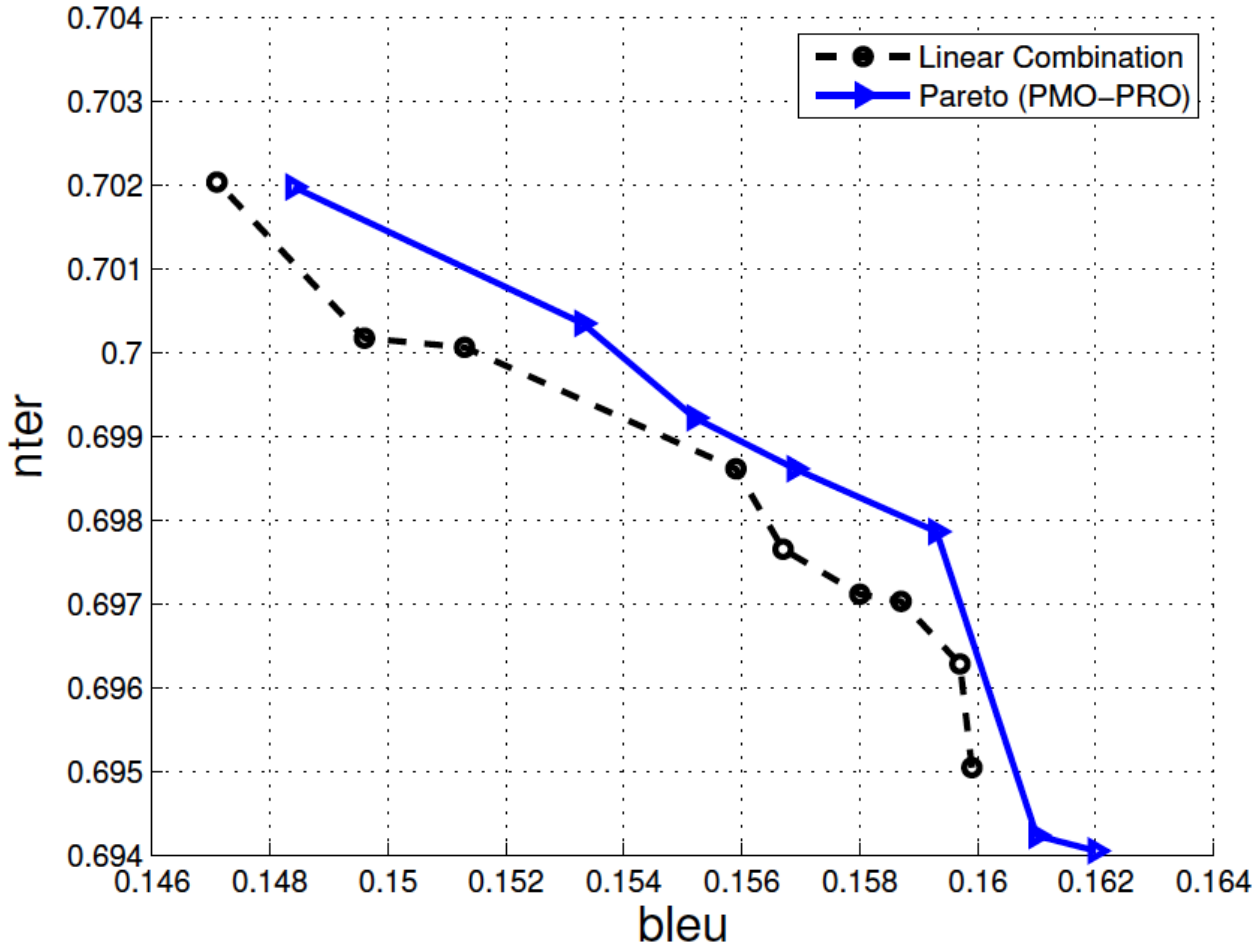
Multi-objective problems are everywhere if we look

- Speed & Accuracy
  - Parsing [Eisner2011]
- Intrinsic & Extrinsic Metrics
  - Parser & downstream Machine Translation [Hall2011]
- Multiple datasets
  - Recommendation system [Agarawl2011]
- Escape local optima
  - Hard & Soft EM in grammar induction [Spitkovsky2011]

Thanks for your attention!

Do you have
a multi-objective problem?

NIST Result


Linear Combination
Pareto (PMO–PRO)
nter
bleu
0.704
0.703
0.702
0.701
0.7
0.699
0.698
0.697
0.696
0.695
0.694
0.146
0.148
0.15
0.152
0.154
0.156
0.158
0.16
0.162
0.164

PubMed Result