Domain Adaptation in Statistical MT

Barry Haddow

University of Edinburgh

3 October, 2013

SMT at Edinburgh: EU Projects

ACCEPT

EU★BRIDGE

MOSES CORE

SMT at Edinburgh: People

Faculty

Philipp Koehn

Lexi Birch

Researchers

Hieu Hoang

Students

Christian Buck

Maria Nadejde

Miles Osborne

Nadir Durrani

Matthias Huck

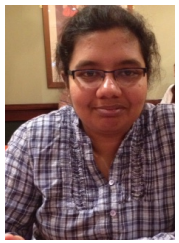
Herve Saint-Amand

Liane Guillou

Phil Williams

Bonnie Webber

Ulrich Germann



Annie Louis

Eva Hasler

Dominikus Wetzel

What is Domain?

The minutes of yesterday's sitting have been distributed. Are there any comments? The next item is the statements by the Council and the Commission on the situation in Palestine.

What is Domain?

The minutes of yesterday's sitting have been distributed. Are there any comments? The next item is the statements by the Council and the Commission on the situation in Palestine.

Parliamentary

What is Domain?

Re: Norton Ghost 15 Install errors and “error in Main configeration.
nope was during first back-up but I can’t even start Ghost process without it crashing with taht eror on startup.

What is Domain?

Re: Norton Ghost 15 Install errors and "error in Main configeration.
nope was during first back-up but I can't even start Ghost process without it crashing with taht eror on startup.

Tech. Support Forum

What is Domain?

Yetta, any chicken left from last night?
I'il make you a sandwich.
Make it to go.
Is Shayna coming or what?
Come on, give it to me.

What is Domain?

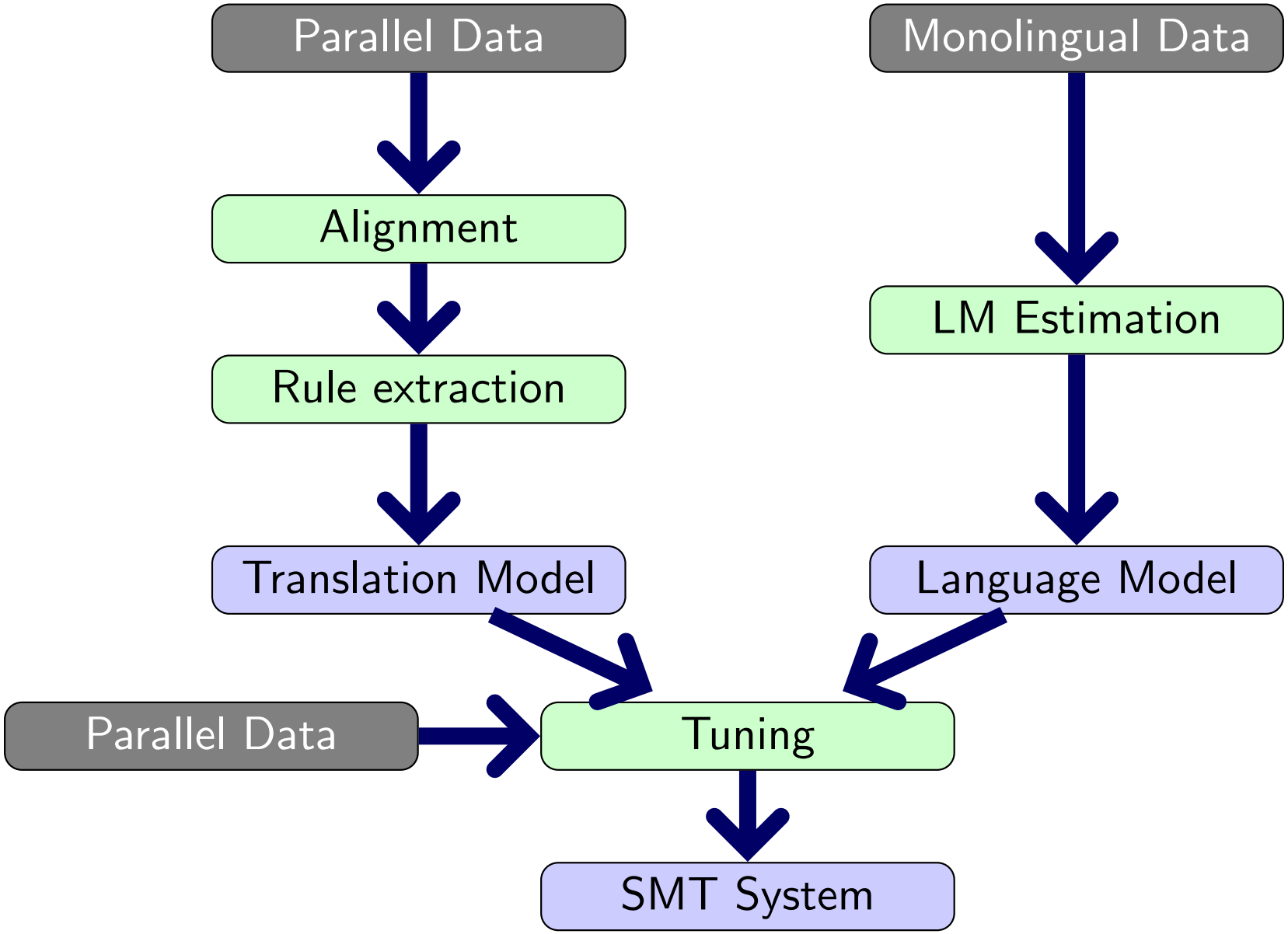
Yetta, any chicken left from last night?
I'il make you a sandwich.
Make it to go.
Is Shayna coming or what?
Come on, give it to me.

Subtitles

Definition of Domain?

- Encompasses notions of genre and topic.
- Practical definition as subcorpus
- But
  - Domains such as news cover many topics
  - Some domains more different than others
- Main issue is difference between training and test data

The SMT Training Pipeline



The Standard (Phrase-Based) SMT Model

Best translation given by:

$$e^* = \arg\max_e \sum_i \lambda_i h_i(e, f)$$

Typical features:

- log of forward and reverse translation probabilities
- forward and reverse lexical scores
- log of language model probability
- phrase and word penalty
- reordering model scores

Domain Mismatch: What goes wrong?

unadapted	Europarl $\approx$ 2 million sentences
adapted	Europarl plus news-commentary ($\approx$ 140k sentences)
test	news commentary

What goes wrong? (I)

Source	Als gelernter Arabist , der tief in die arabische und mohammedanische Kultur eingedrungen war ...
Unadapted	As a skilled Arabist , deeply in the Arab and mohammedanische culture ...
Adapted	As a skilled Arabist steeped in the Arab and Muslim culture ...
Reference	As an Arabist by training , immersed in Arab and Muslim culture ...

What goes wrong? (I)

Source	Als gelernter Arabist , der tief in die arabische und mohammedanische Kultur eingedrungen war ...
Unadapted	As a skilled Arabist , deeply in the Arab and mohammedanische culture ...
Adapted	As a skilled Arabist steeped in the Arab and Muslim culture ...
Reference	As an Arabist by training , immersed in Arab and Muslim culture ...

Occurrences of mohammedanische:

europarl: 0

news-commentary: 13

Seen Error

What goes wrong? (II)

Source	die indische Zentralbank schwimmt im Geld
Unadapted	the Indian Central Bank in the money they swim
Adapted	India's central bank is awash in money
Reference	India's central bank is rolling in cash

What goes wrong? (II)

Source	die indische Zentralbank schwimmt im Geld
Unadapted	the Indian Central Bank in the money they swim
Adapted	India's central bank is awash in money
Reference	India's central bank is rolling in cash

Translations of schwimmt

Unadapted		Adapted	
swim	2/5	swim	2/11
floats	1/5	is awash	2/11
then something	1/5	awash	1/11
then	1/5	rows	1/11

Sense Error

What goes wrong? (III)

Source	... suchte Japan auf breiter Front nach Hilfsmitteln und Innovationen ...
Unadapted	... Japan on a broad front sought after aids and innovation ...
Adapted	... Japan on a broad front sought after tools and innovations ...
Reference	... Japan searched broadly for tools and innovations ...

What goes wrong? (III)

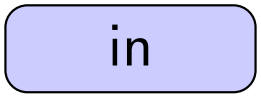
Source	... suchte Japan auf breiter Front nach Hilfsmitteln und Innovationen ...
Unadapted	... Japan on a broad front sought after aids and innovation ...
Adapted	... Japan on a broad front sought after tools and innovations ...
Reference	... Japan searched broadly for tools and innovations ...

Translations of Hilfsmitteln

Unadapted		Adapted	
aid	8/38	tools	10/46
tools	6/38	aid	8/46
aids	4/38	aids	4/46
devices	3/38	resources	3/46

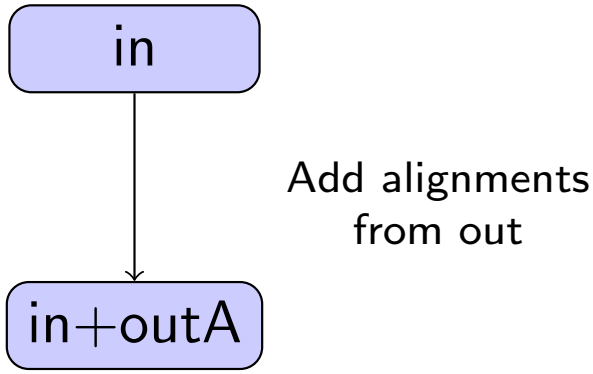
Score Error

Quantify Effect of Domain on Translation Model



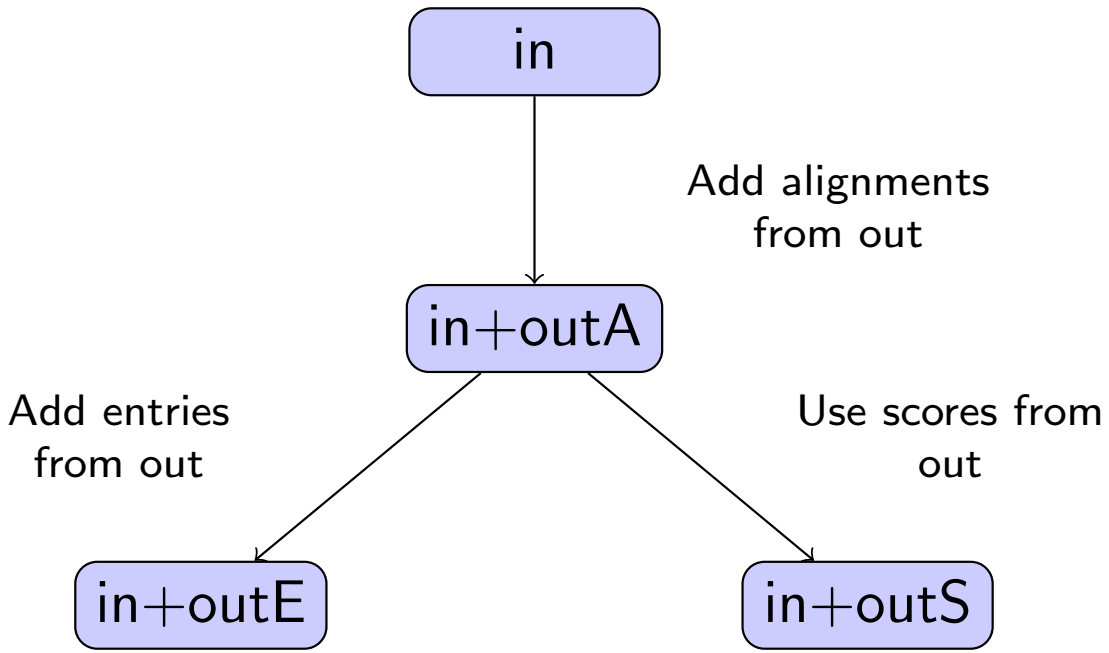
Assume two corpora (in and out) build 5 different phrase tables

Quantify Effect of Domain on Translation Model



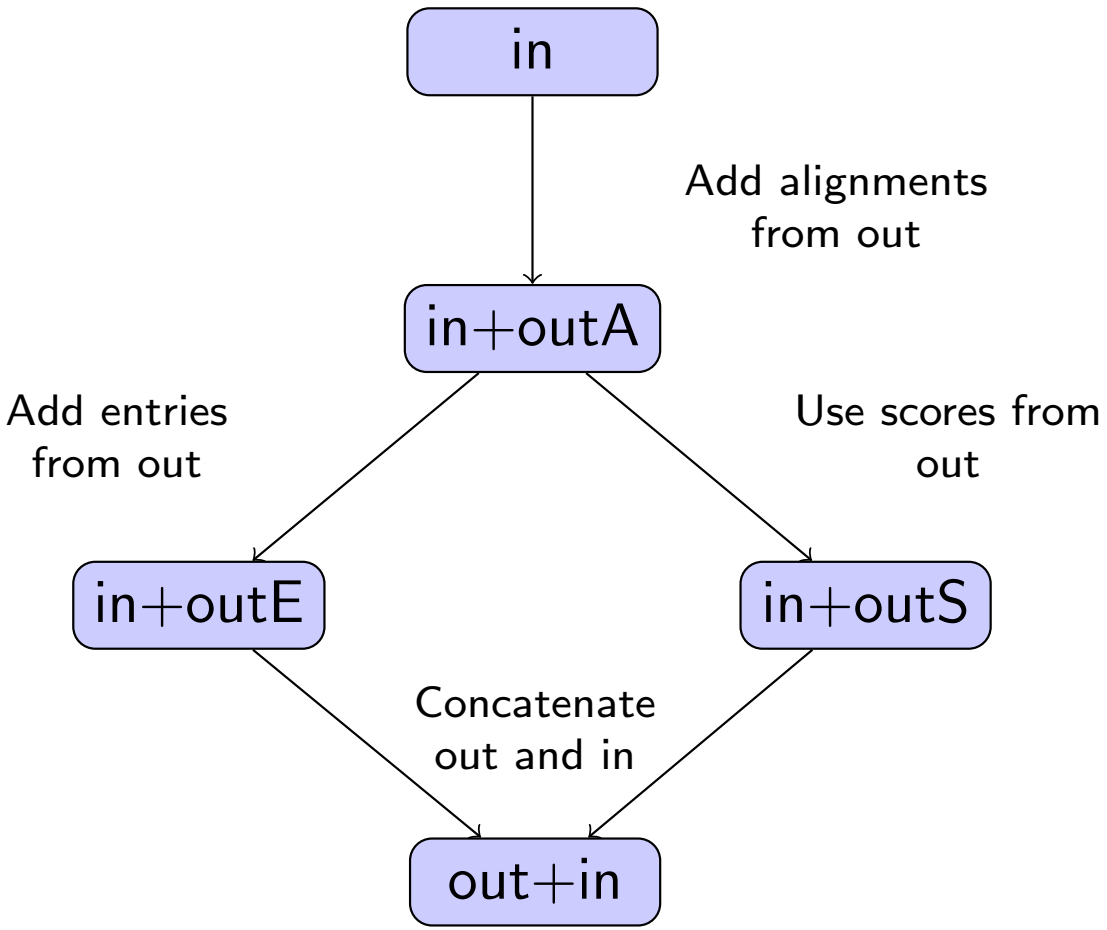
Assume two corpora (in and out) build 5 different phrase tables

Quantify Effect of Domain on Translation Model



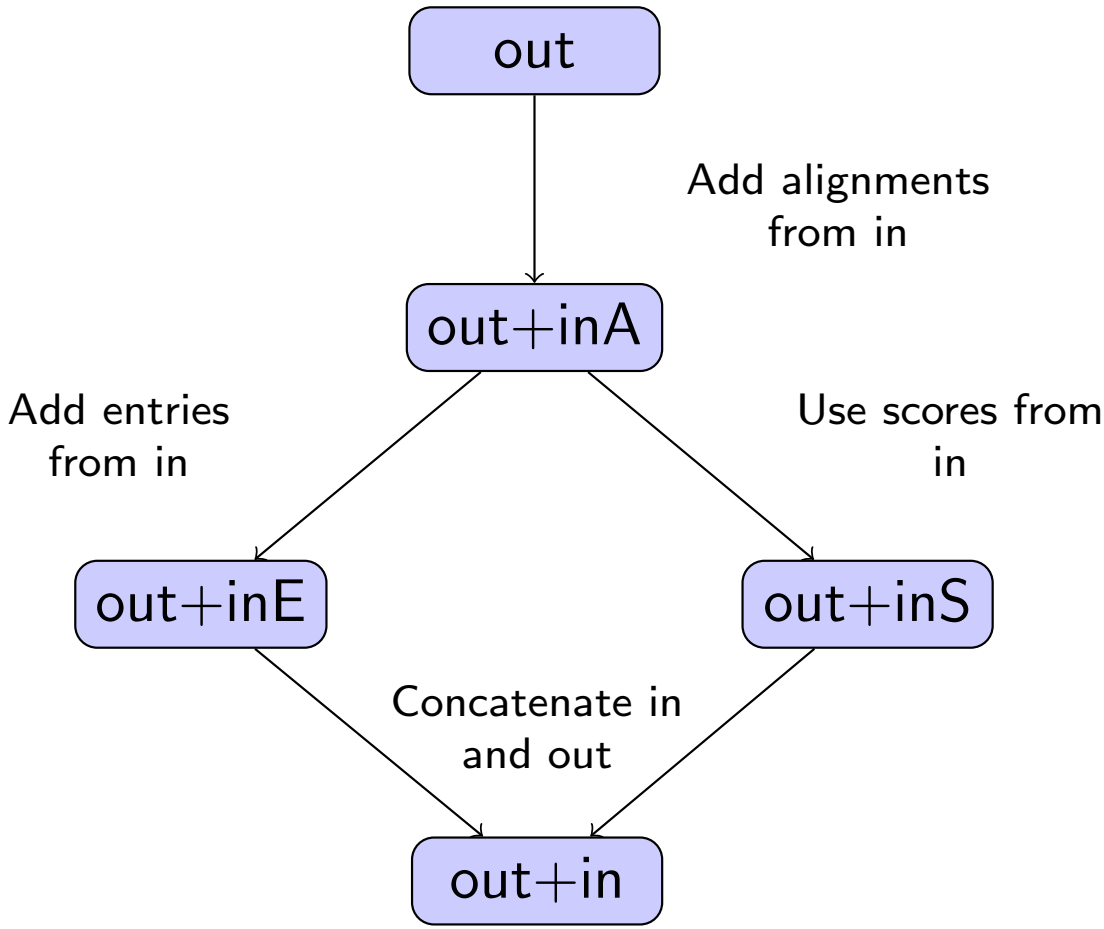
Assume two corpora (in and out) build 5 different phrase tables

Quantify Effect of Domain on Translation Model



Assume two corpora (in and out) build 5 different phrase tables

Quantify Effect of Domain on Translation Model



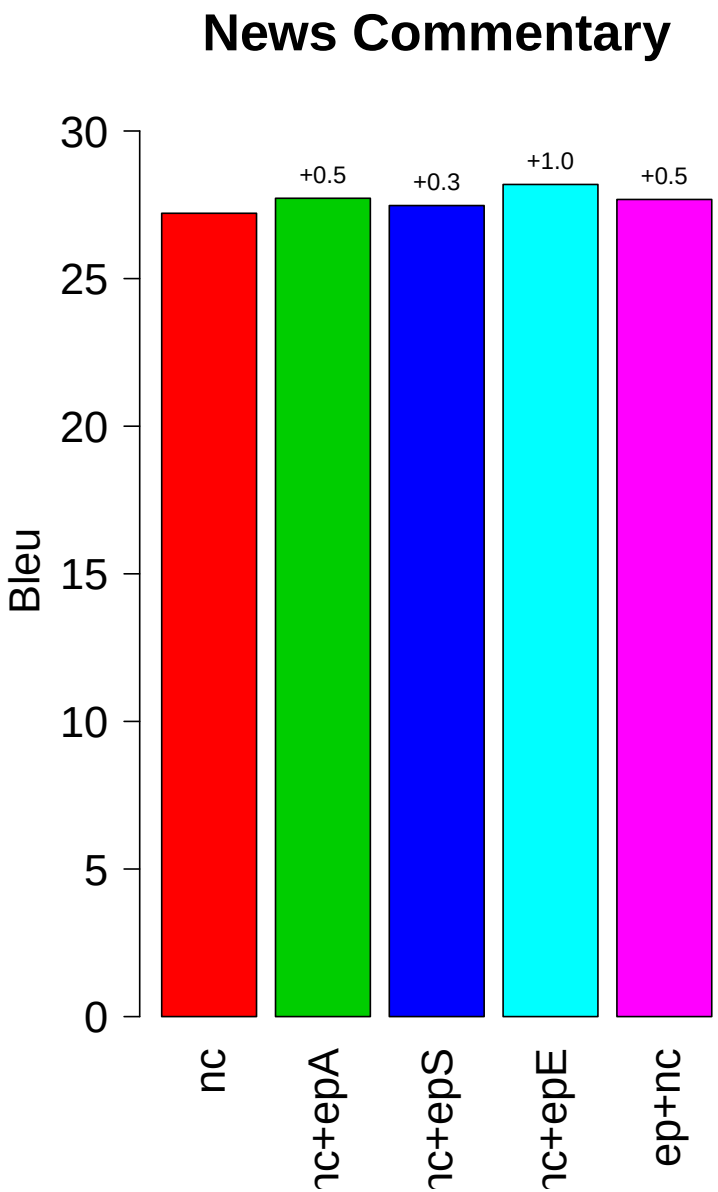
Assume two corpora (in and out) build 5 different phrase tables

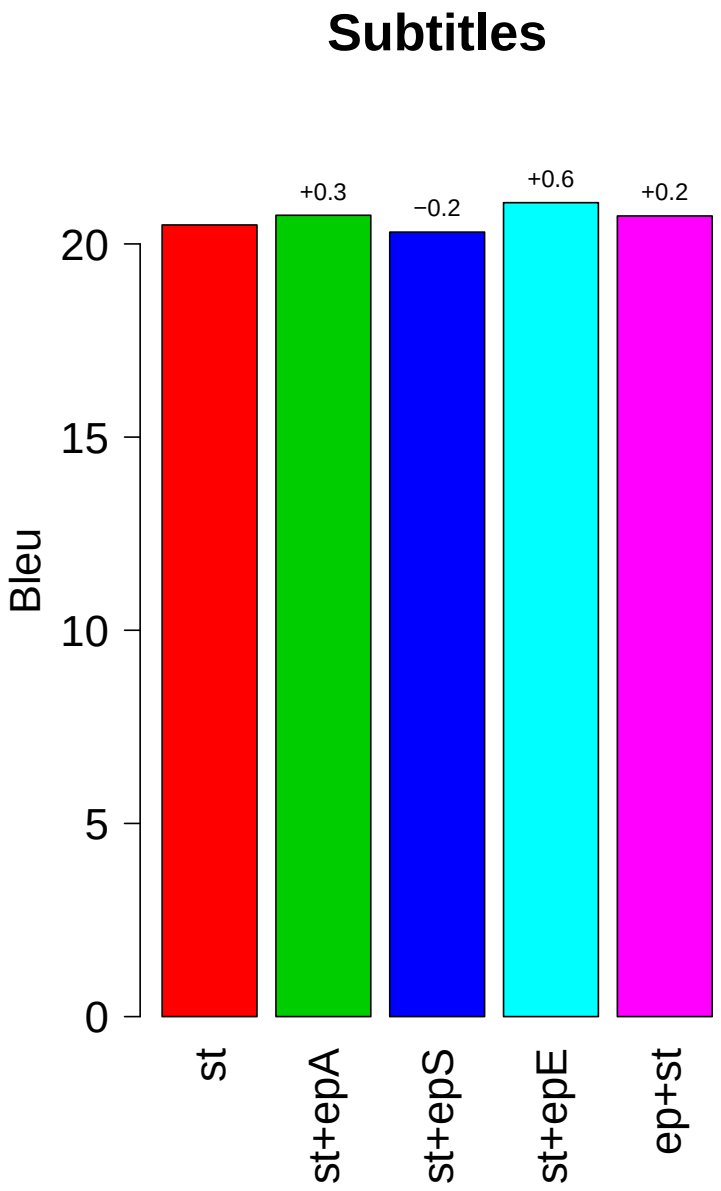
Experimental Setup

out-of-domain	**ep**	European parliament
in-domain	**nc**	News commentary
	st	Subtitles

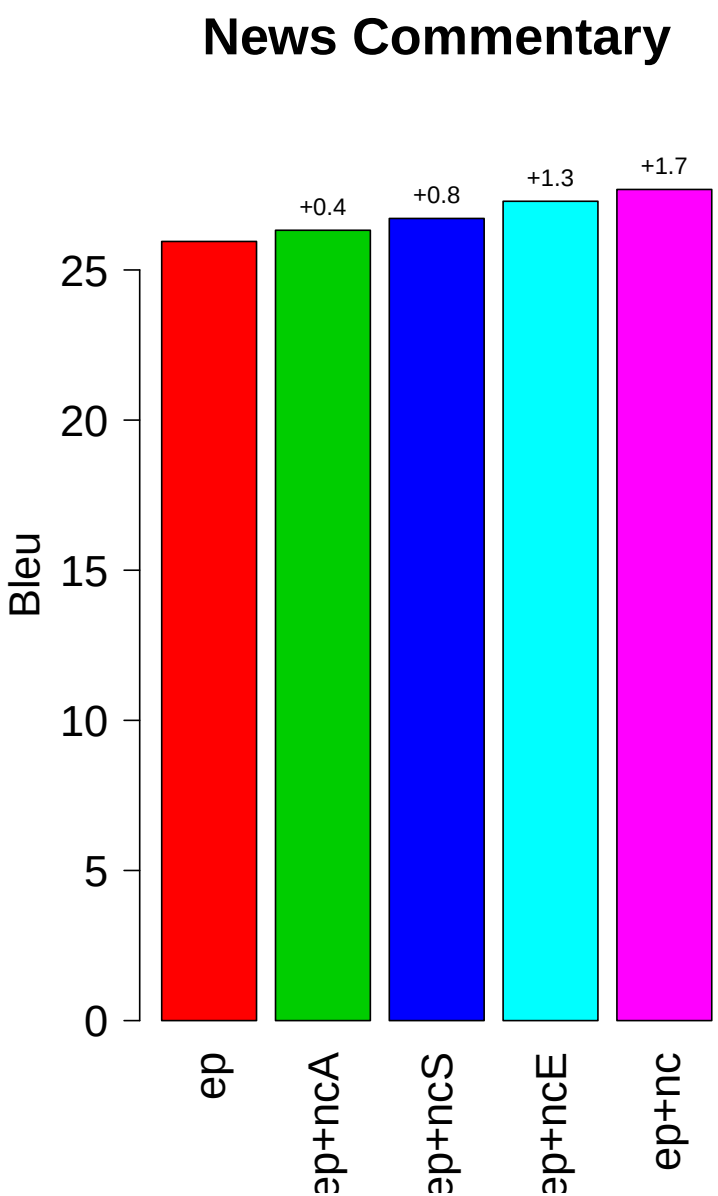
- Language and reordering model built from all data
- Tune and test on in-domain
- Average across 8 language pairs

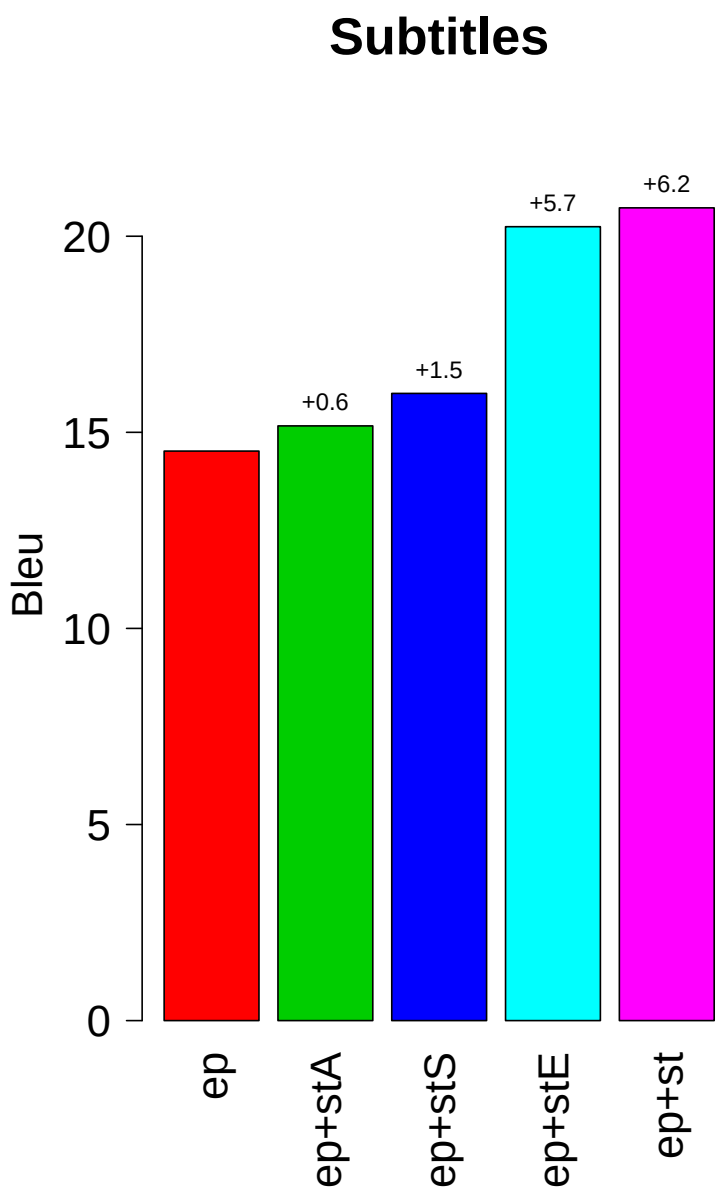
Supplementing In-Domain Data



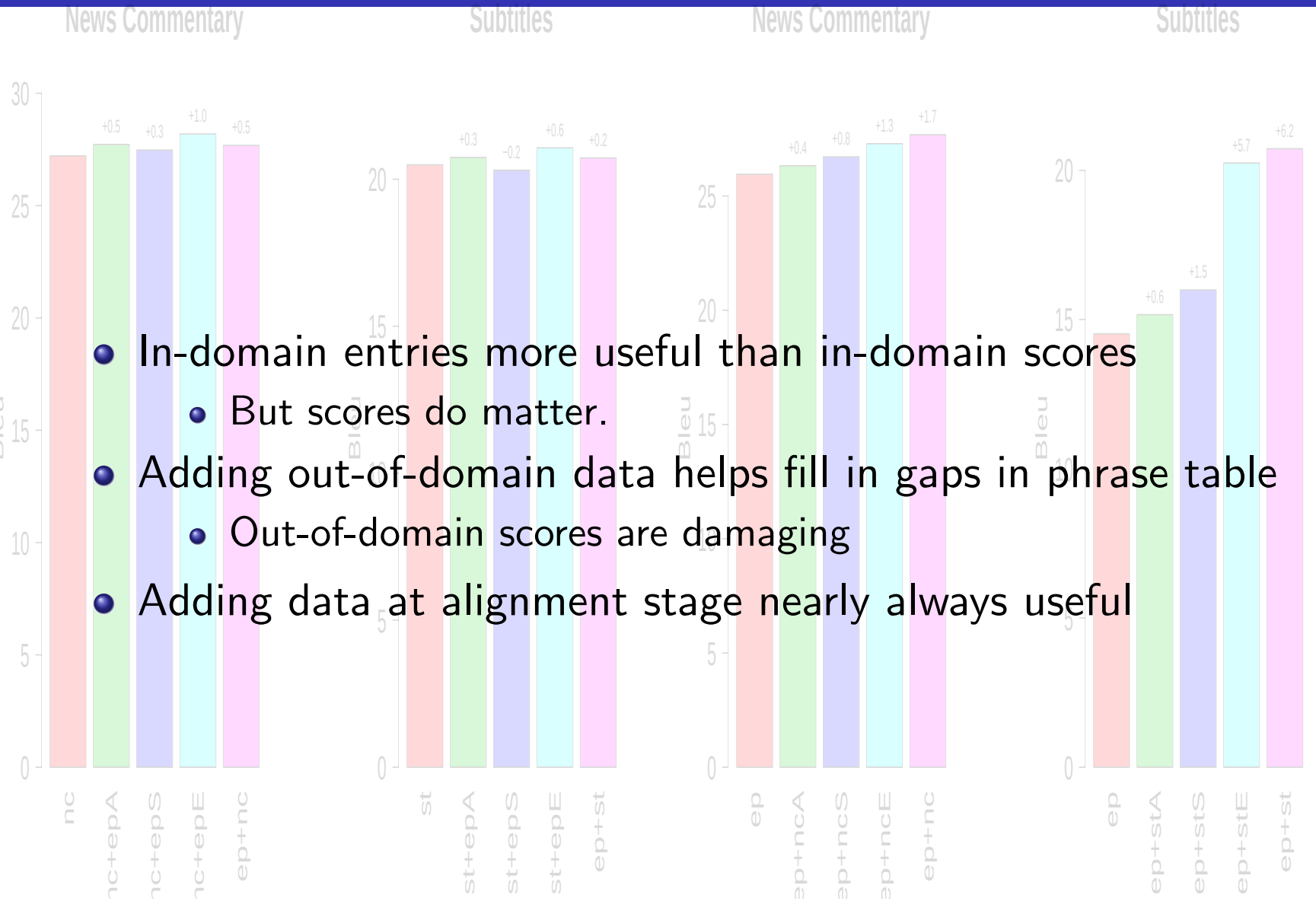


Supplementing Out-of-Domain Data





Effect of Adding Data - Conclusions

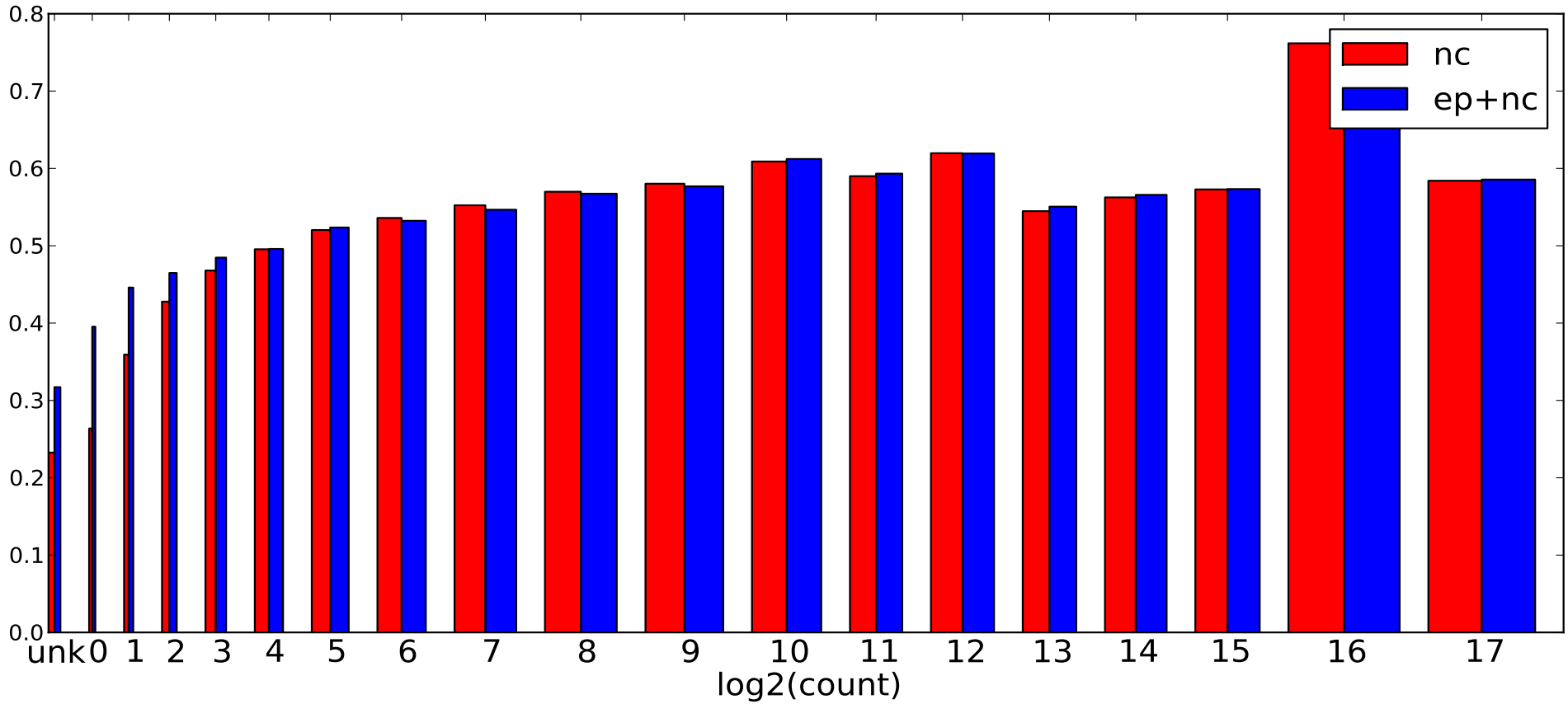


- In-domain entries more useful than in-domain scores
  - But scores do matter.
- Adding out-of-domain data helps fill in gaps in phrase table
  - Out-of-domain scores are damaging
- Adding data at alignment stage nearly always useful

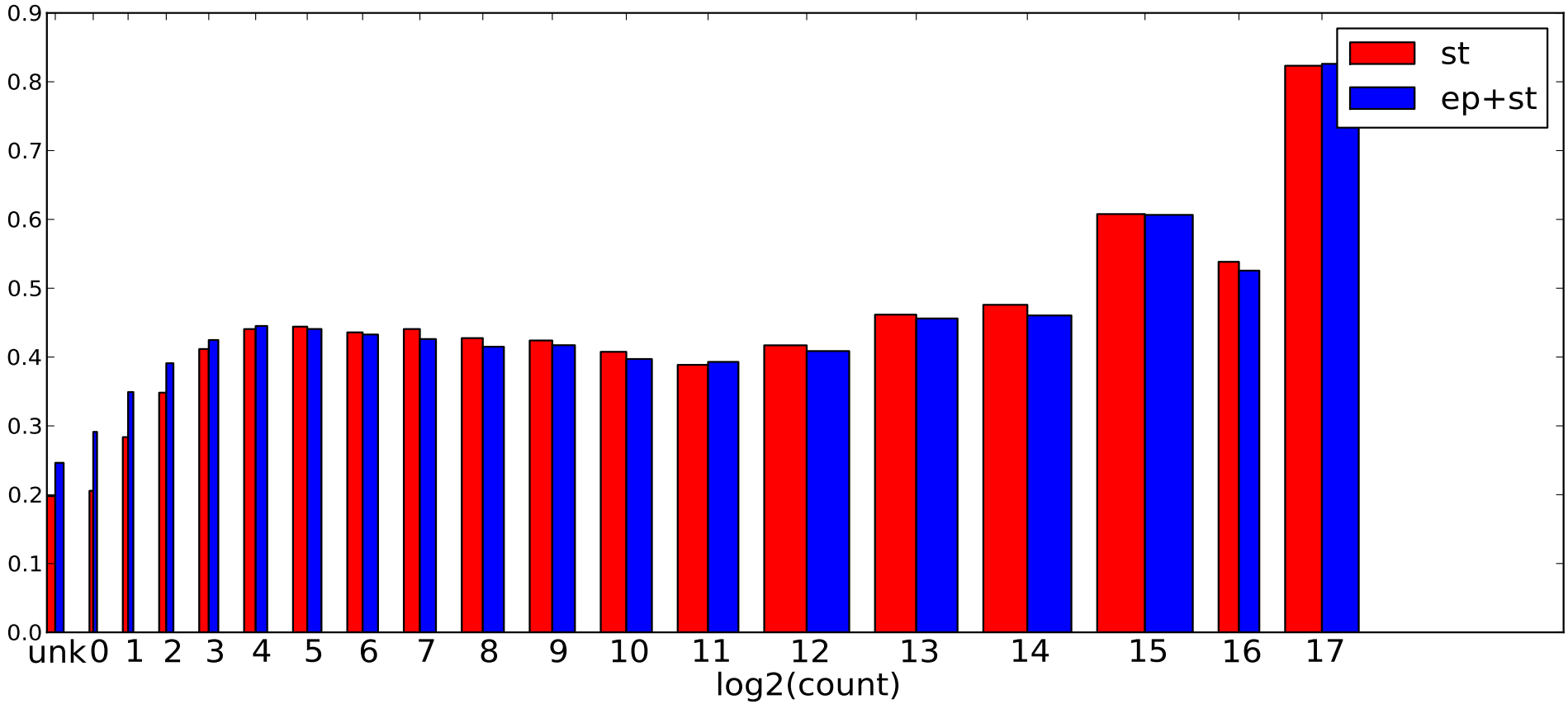
Which Type of Words are Improved?

- Adding out-of-domain appears to help with OOVs?
  - What about other frequencies?
- Measure source-word precision by tracking word translation
- Calculate average precision, binned by log frequency (in training).

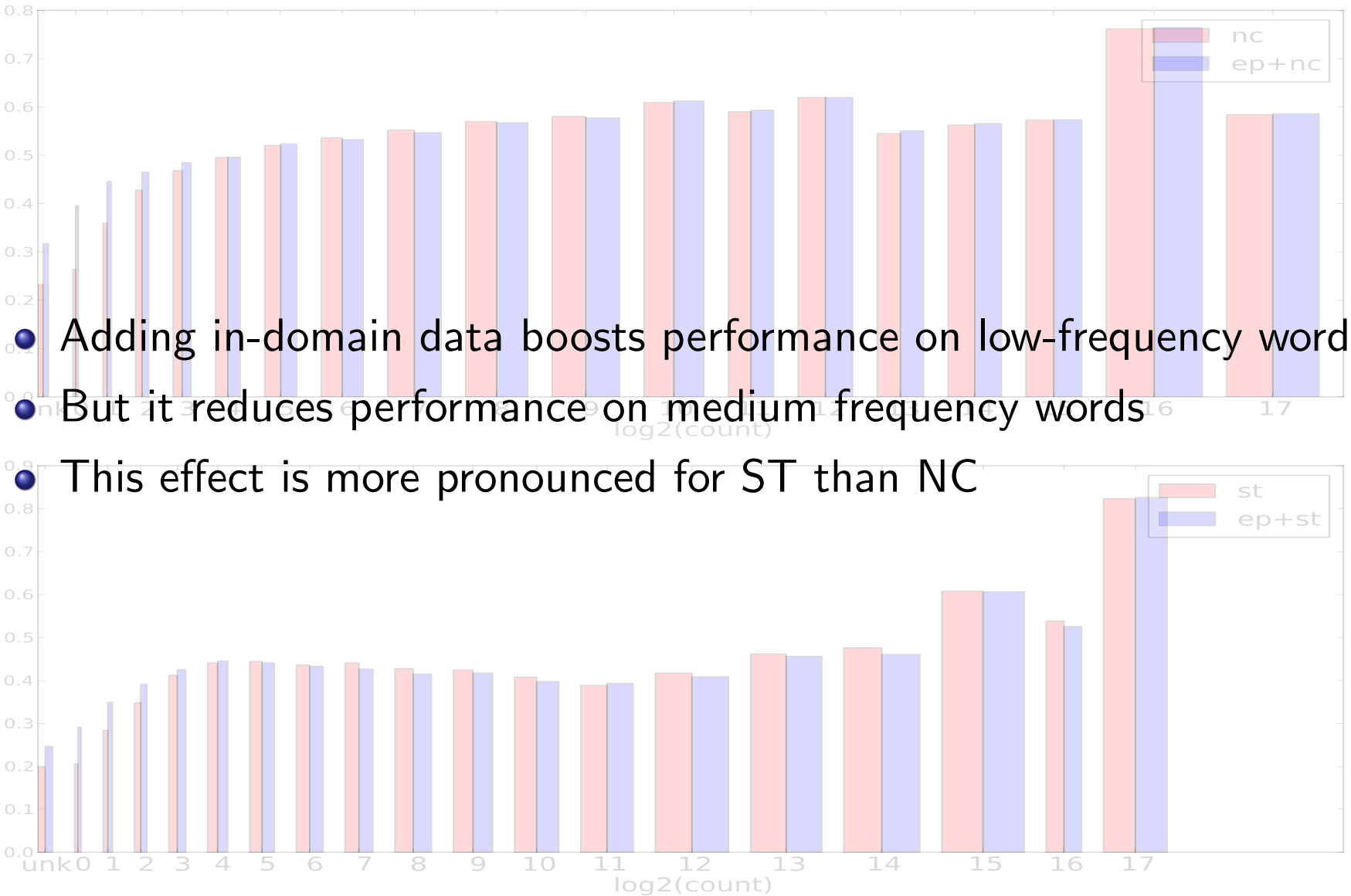
Precision vs Frequency (NC)



Precision vs Frequency (ST)

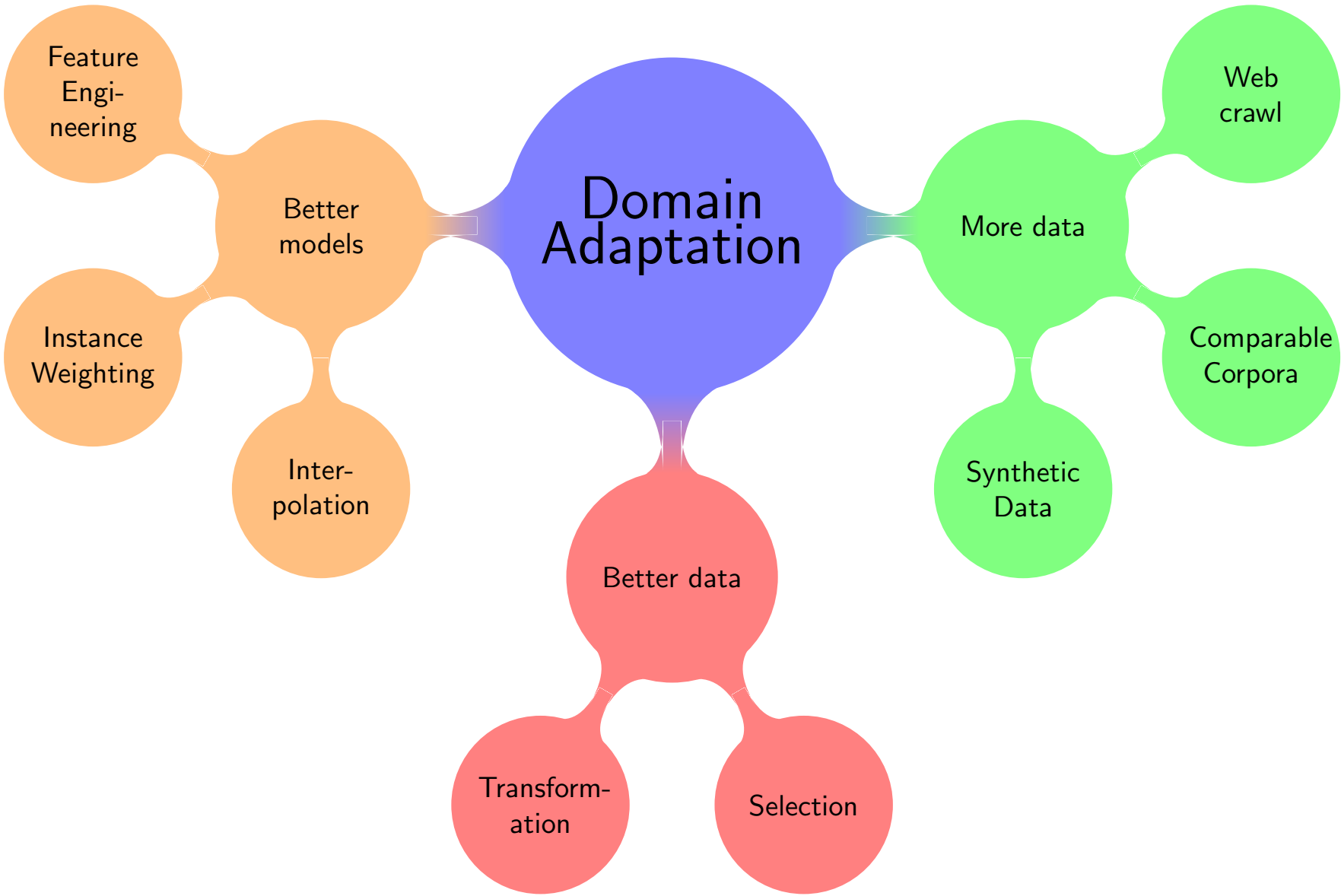


Precision vs Frequency - Conclusions

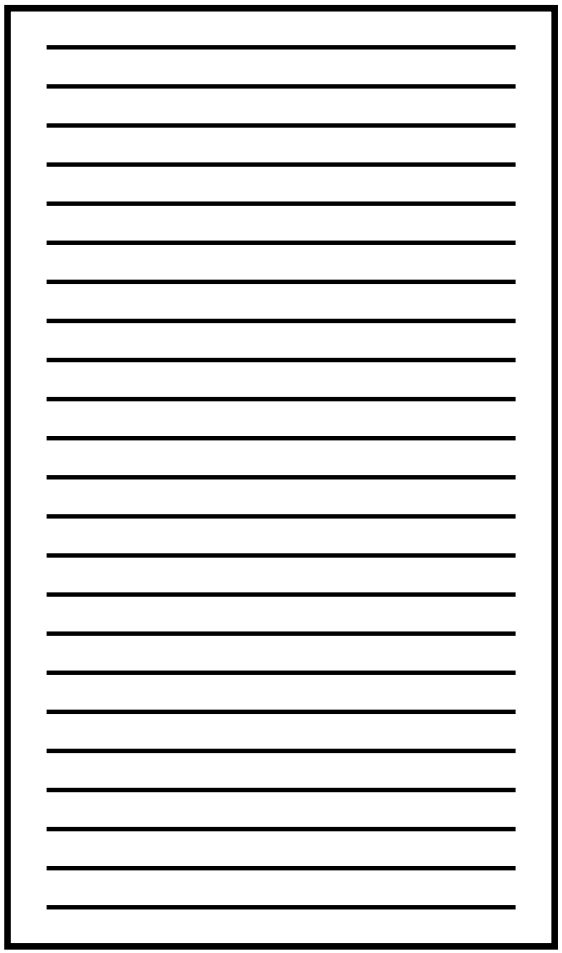


- Adding in-domain data boosts performance on low-frequency words
- But it reduces performance on medium frequency words
- This effect is more pronounced for ST than NC

Methods for Domain Adaptation in SMT

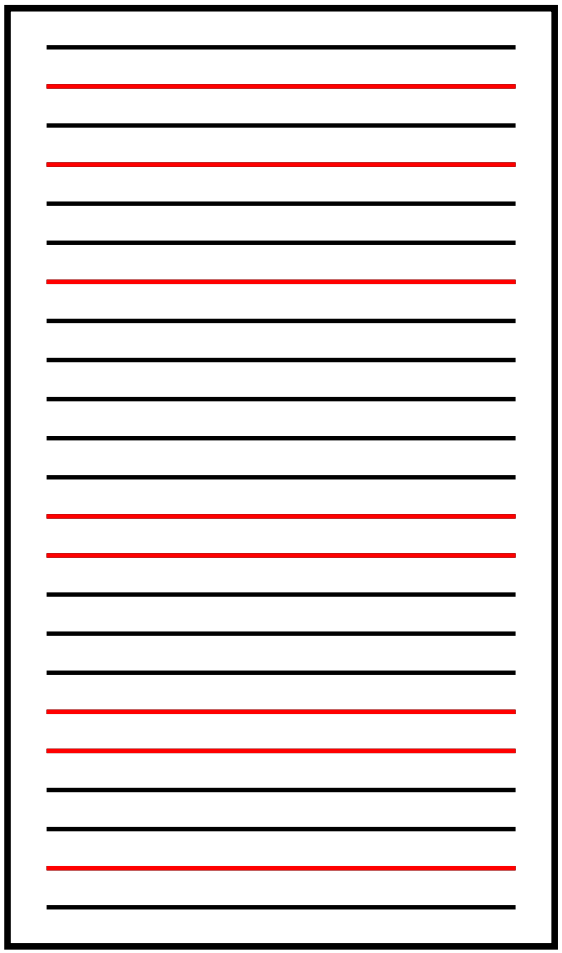


Instance Selection



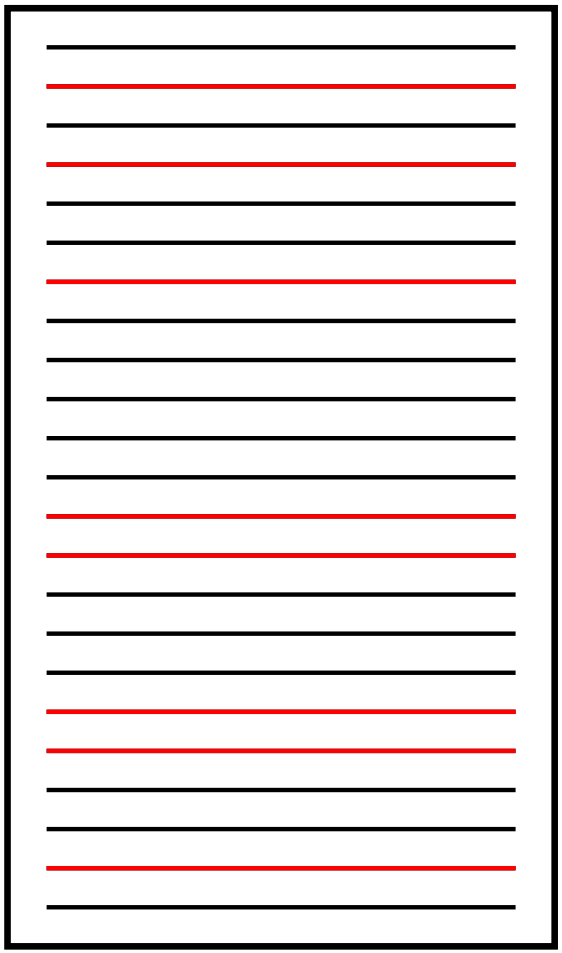
- Instead of using all training data ...

Instance Selection



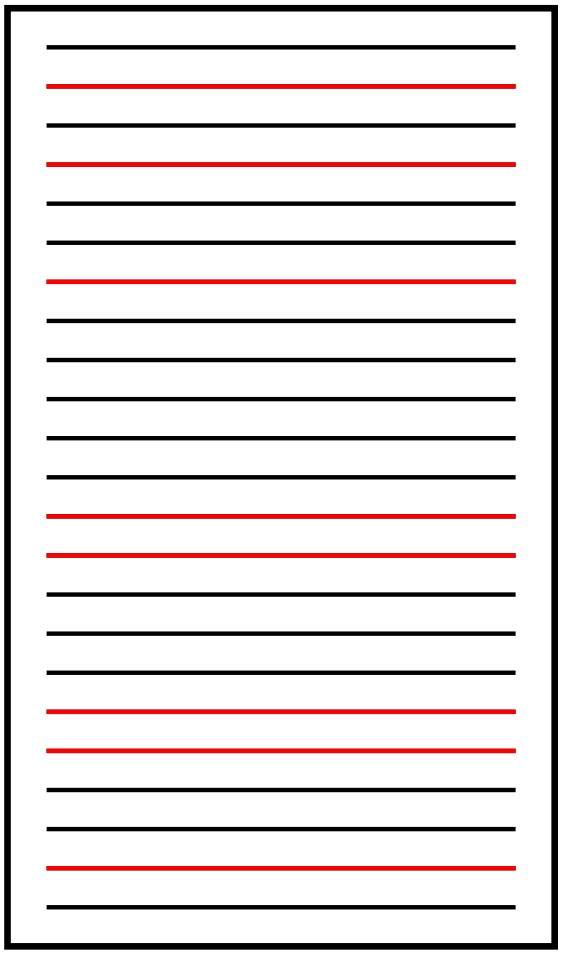
- Instead of using all training data ...
- Use only selected data
- But how to select?

Instance Selection



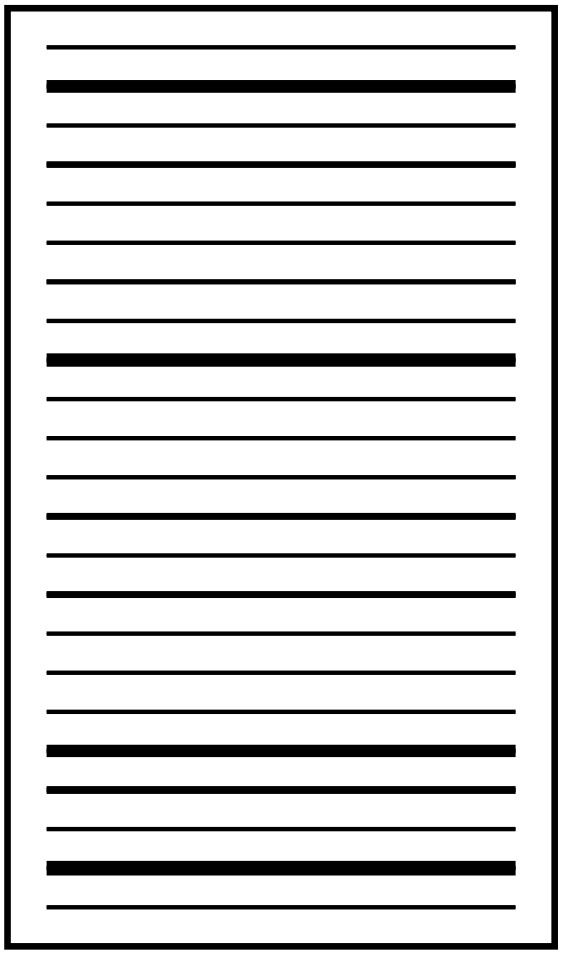
- Instead of using all training data ...
- Use only selected data
- But how to select?
  - Modified Moore-Lewis
  - Select similar to in-domain ...
  - ...but different from out
  - Use LM perplexity for similarity

Instance Weighting



- Instance selection is a 1-0 weighting

Instance Weighting



- Instance selection is a 1-0 weighting
- Can we improve by a allowing variable weights?
  - Use MML scores for weighting
  - .. really should learn weights

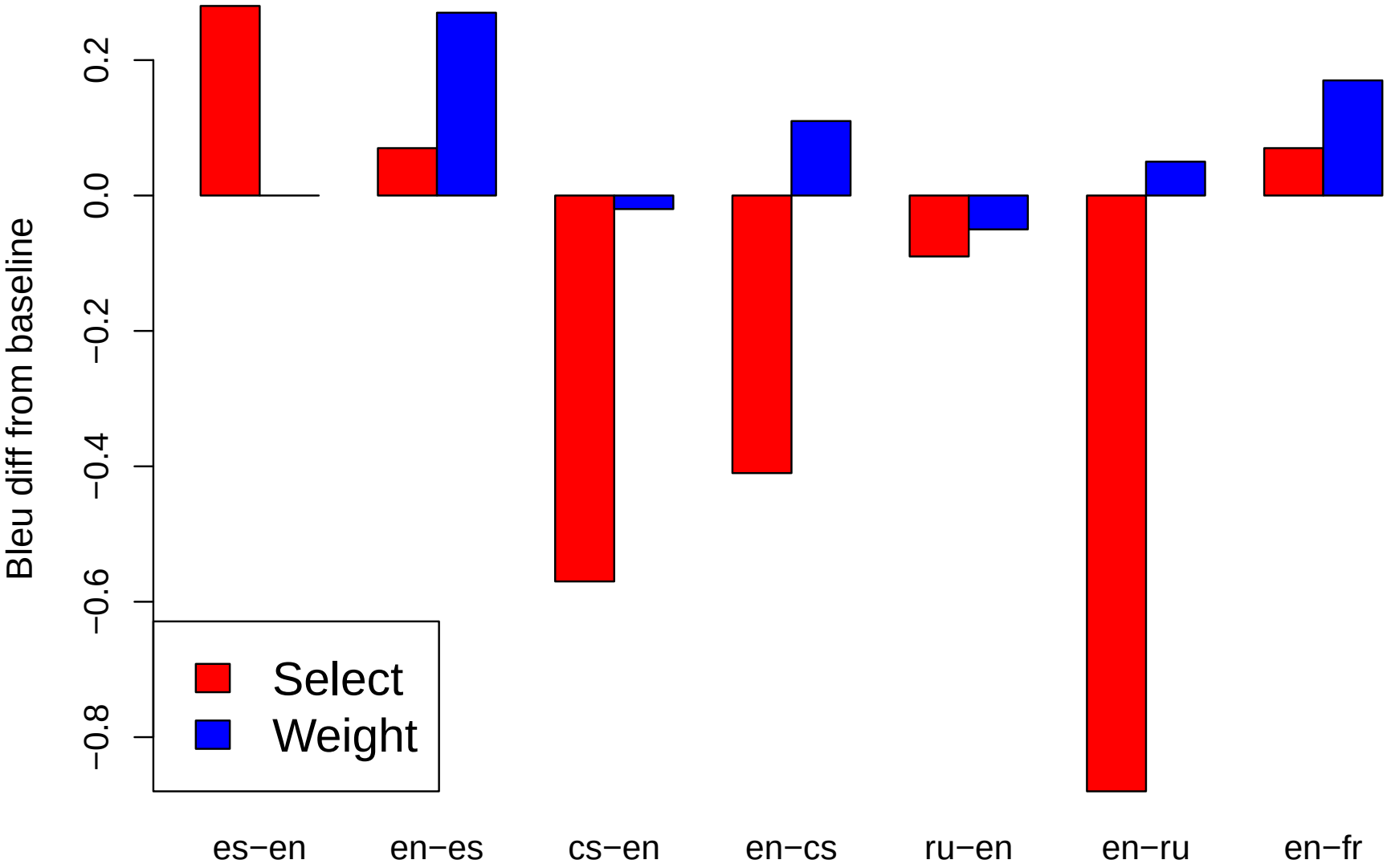
WMT13 Experiments

Parallel Training Data

	en↔cs	en↔es	en↔fr	en↔ru
europarl	✓	✓	✓	
news-comm	✓	✓	✓	✓
UN		✓	✓	
GigaFrEn			✓	
CommonCrawl	✓	✓	✓	✓
Yandex				✓
Czeng	✓			

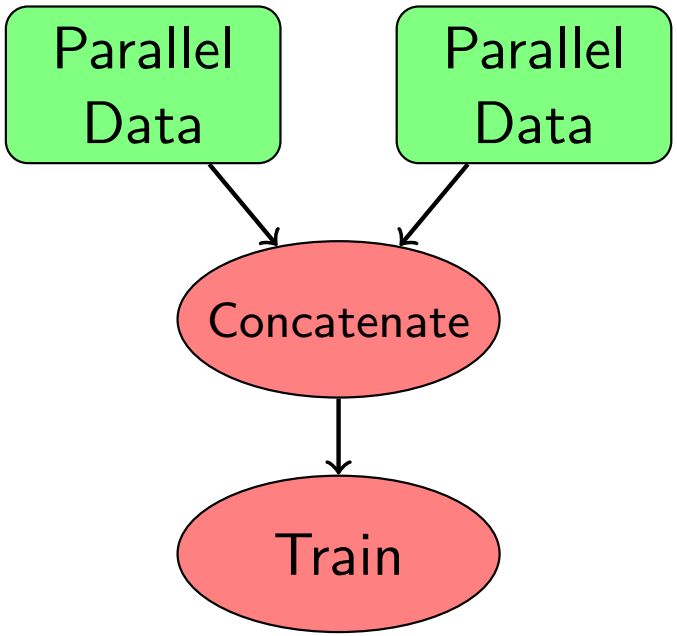
Test data: Extracted from online news.

Instance Selection and Weighting - WMT13 Experiments

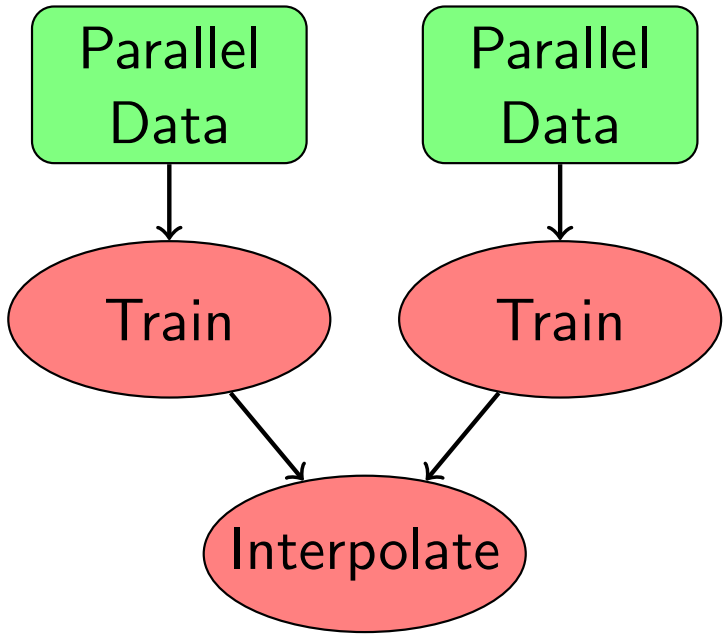


Mixture Models (Interpolation)

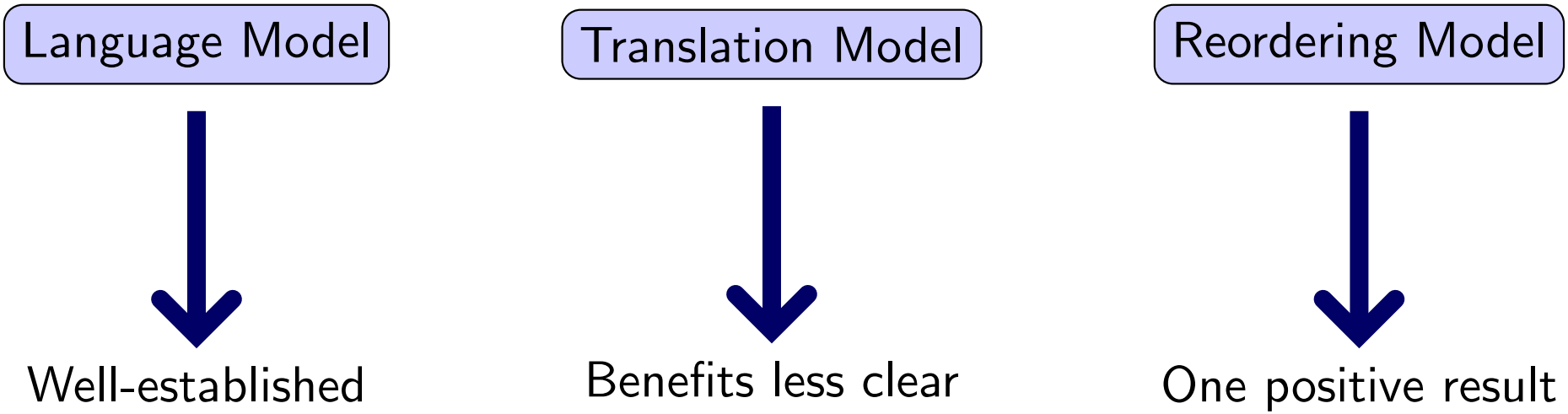
Usual Approach



Mixture Model



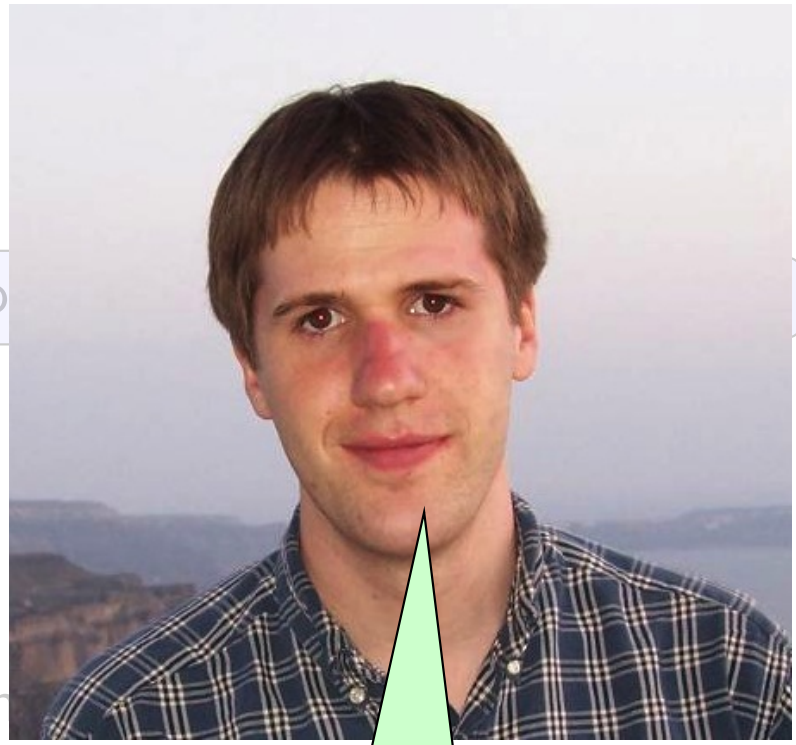
Which models to mix?



Which models to mix?

Language Mo

Reordering Model



Well-establish

One positive result

Warning: SRILM Interpolation is broken!

Mixture Models in Practice

Log-Linear

$$\ldots w_{T1} \log(p_1(e|f)) + w_{T2} \log(p_2(e|f)) \ldots$$

- Optimise with MERT etc.
- Problems with zeros

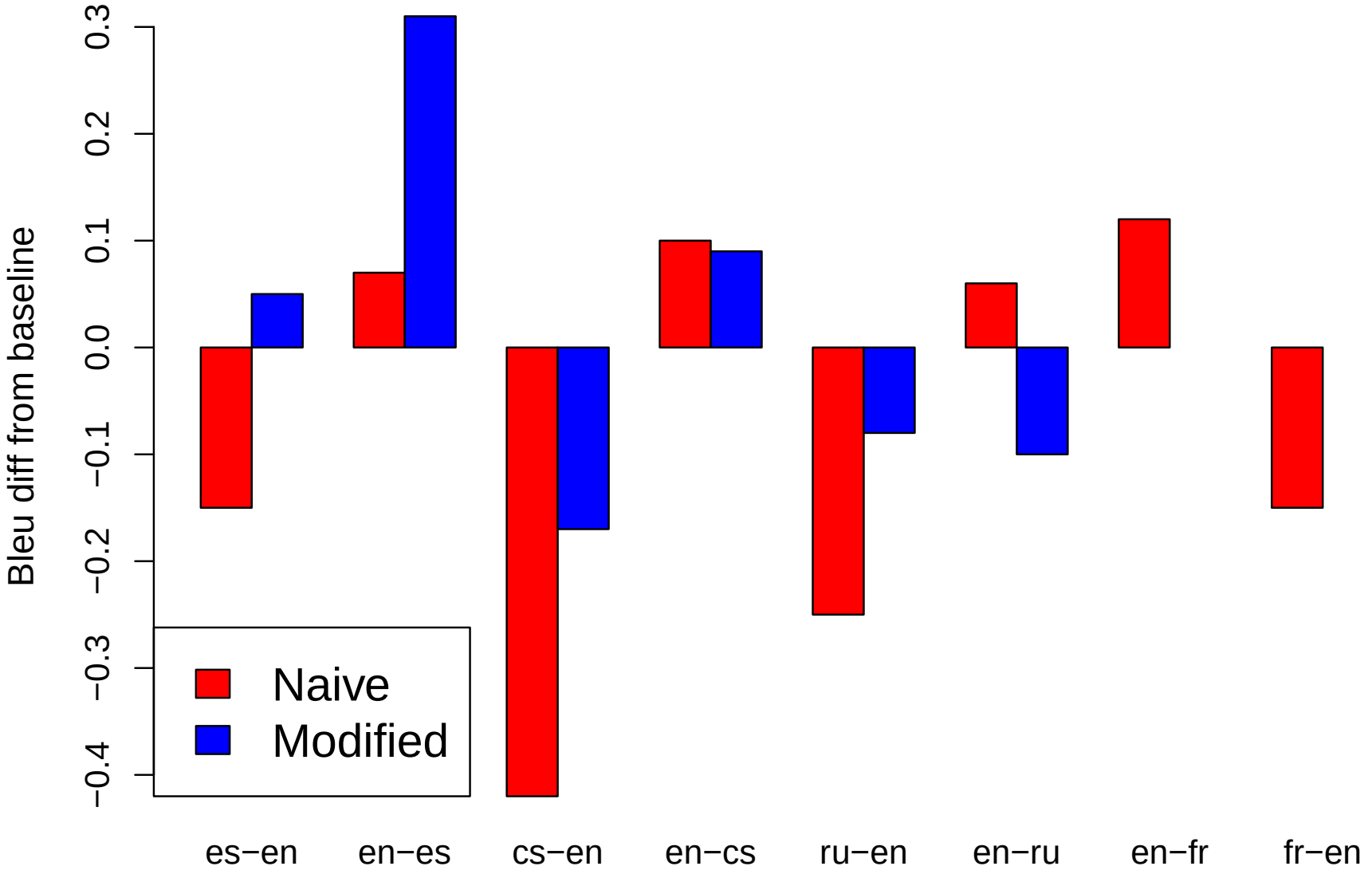
Vs.

Linear

$$\ldots w_T \log\left(\lambda p_1(e|f) + (1-\lambda) p_2(e|f)\right) \ldots$$

- More natural
- Separate optimisation

Linear Mixture Models - WMT13 Experiments



Improving Translation Model Mixture Training

- Linear mixture better than log linear
  - → But perplexity is indirect and ill-defined objective
- What if we could optimise directly for translation performance?
  - e.g. BLEU
- This can be done with *Pairwise Ranked Optimisation (PRO)*

PRO: Pairwise Ranked Optimisation

- Batch tuning algorithm – optimises standard linear model
- Sample pairs of hypotheses from n-best lists

$$S = \{(e_1^1, e_2^1), \ldots (e_1^n, e_2^n)\}$$

- Feature weights optimised by:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \log\left(\sigma\left(y_i \cdot (score_{\mathbf{w}}(e_1^i) - score_{\mathbf{w}}(e_2^i))\right)\right)$$

$$\text{where} \quad \sigma(x) = 1/(1+\exp(-x))$$

$$\text{and} \quad y_i = \text{sgn}\left(\text{bleu}(e_1^i) - \text{bleu}(e_2^i)\right)$$

PRO for Mixture Model Training

- Standard linear model

$$\text{score}_{\mathbf{w}}(e) = \mathbf{h}(e) \cdot \mathbf{w}$$

- Interpolated TM makes *score*() function of mixture weights

$$\sum_{j=1}^{m} \left(w^j \cdot \log \left(\lambda^j p_A^j + (1 - \lambda^j) p_B^j \right) \right) + \sum_{j=m+1}^{n} w^j h^j$$

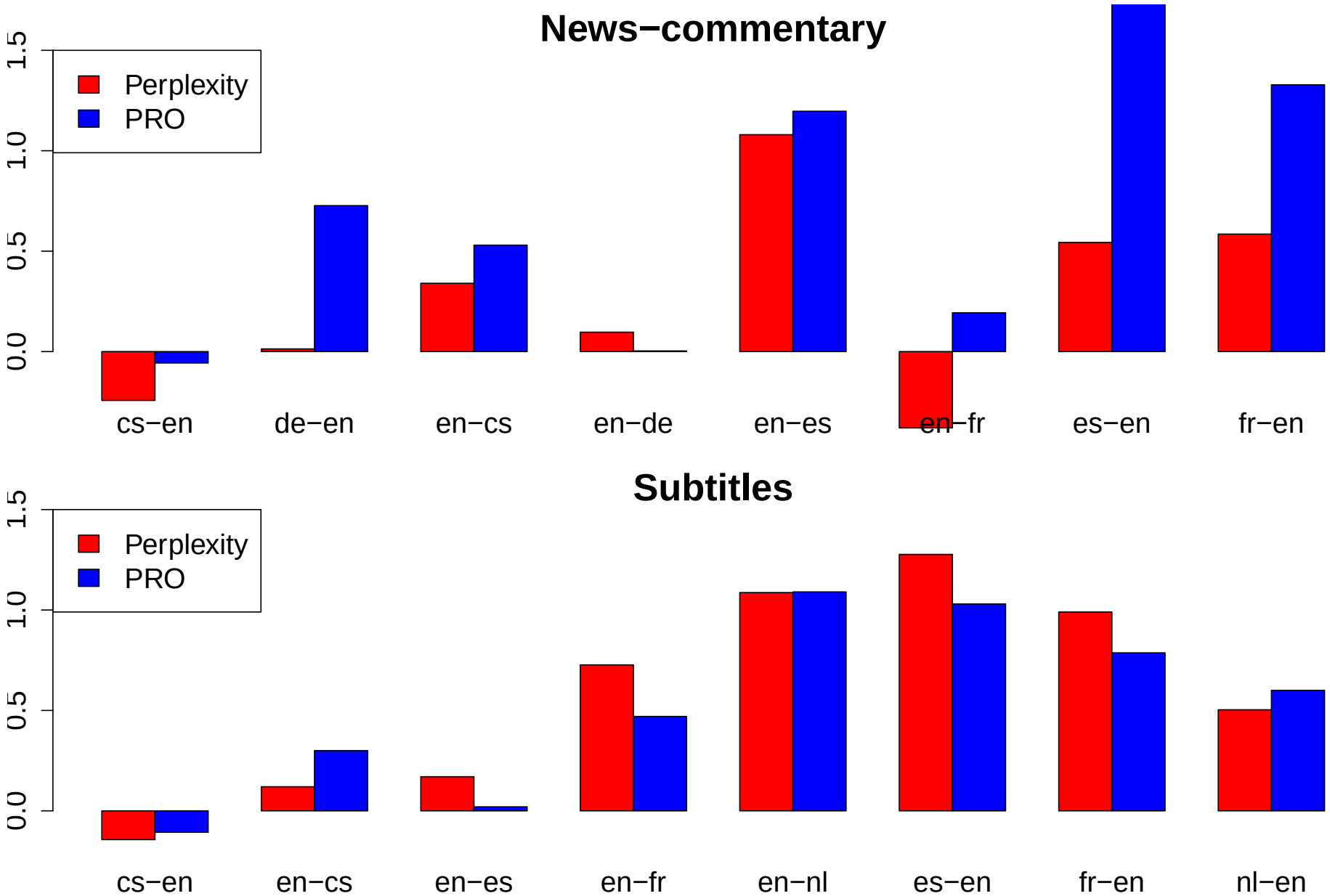
- So just optimise PRO objective for both $\mathbf{w}$ and $\boldsymbol{\lambda}$!

Experimental Setup

out-of-domain	**ep**	European parliament
in-domain	**nc**	News commentary
	st	Subtitles

- Language and reordering model on all data
- Tune and test on in-domain
- Average across 8 language pairs
- Baseline is concatenation

PRO for Mixture Models – Results



Feature Engineering for Domain Adaptation

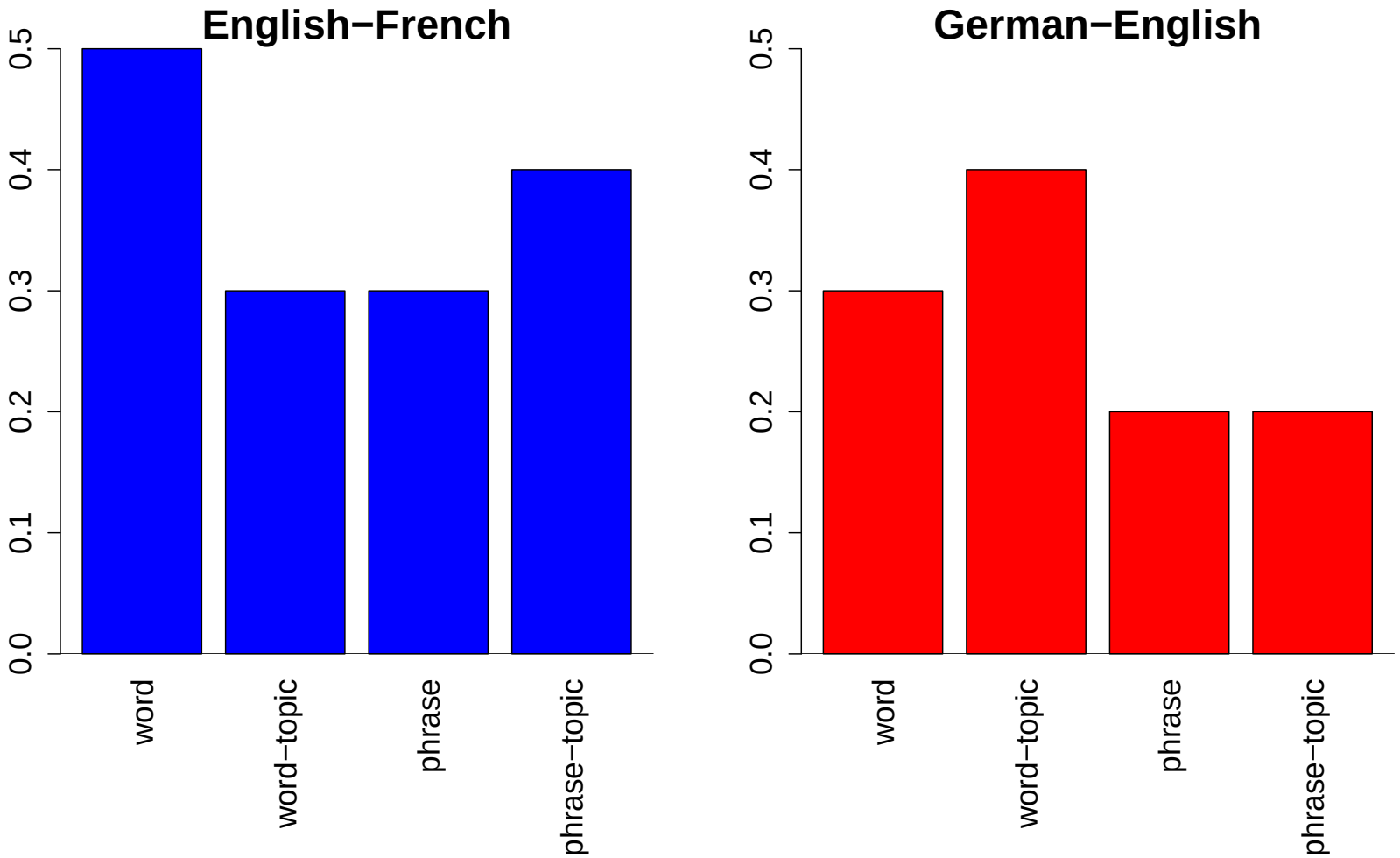
- Can now tune models with many features
- Can this help with domain adaptation?

Feature Engineering for Domain Adaptation

- Can now tune models with many features
- Can this help with domain adaptation?

	Standard	Word	Word-Topic
matter ↕ important	p_{dir}=0.2, p_{inv}=0.4, l_{dir}=0.5, l_{inv}=0.1	wp_matter_important=1	wt_matter_important_T1=1 wt_matter_important_T6=1
matter ↕ matière	p_{dir}=0.1, p_{inv}=0.6, l_{dir}=0.4, l_{inv}=0.2	wp_matter_matiere=1	wt_matter_matiere_T2=1 wt_matter_matiere_T7=1

Feature Engineering: Results



Experiments using IWSLT (TED) Data

Transforming Data

Training Data

Madame la Présidente, c'est une motion de procédure. Vous avez probablement appris par la presse et par la télévision que plusieurs attentats à la bombe et crimes ont été perpétrés au Sri Lanka.

Support Forums

Tu ne retrouve pas ton compte Norton, je te conseille de joindre le Support des produits Norton et ils seront à meme via différents moyens, de le retrouver pour toi.

Transforming Data

Training Data

Madame la Présidente, c'est une motion de procédure. Vous avez probablement appris par la presse et par la télévision que plusieurs attentats à la bombe et crimes ont été perpétrés au Sri Lanka.

Support Forums

Tu ne retrouve pas ton compte Norton, je te conseille de joindre le Support des produits Norton et ils seront à meme via différents moyens, de le retrouver pour toi.

How can we deal with register differences?

Data Transformations

Informal vs. Formal French

Tu l'as dit toi-même	$\leftrightarrow$	*Vous l'avez dit vous-même*
Est-ce que tu as des ...	$\leftrightarrow$	*Avez-vous des ...*

- Two Approaches:
  - $\rightarrow$ Transform the training data
  - $\rightarrow$ Transform the test data
- Both methods are effective for *tu/vous*
  - Transform test better, but effect is additive
- *est-ce que* transform not effective – too much variety

Domain Adaptation in SMT: Summary

- Heuristic, multi-stage pipeline makes DA difficult
- OOVs biggest problem, also score errors
- Variety of techniques:
  - Looked mainly at filtering and model interpolation
  - Really only tackling score errors
- Often no clear winner ... use many languages and data sets

Domain Adaptation in SMT: What next?

- Better analysis of why things work
- Move away from data-set as domain
- Deal better with informal text – preprocessing
- Improve handling of OOVs
- Make more use of non-parallel data
- Improved pipeline could make DA easier to analyse

The End

Thank You. Questions?

Collaborators

Pierrette Bouillon, Nadir Durrani, Eva Hasler, Kenneth Heafield, Philipp Koehn, Manny Rayner