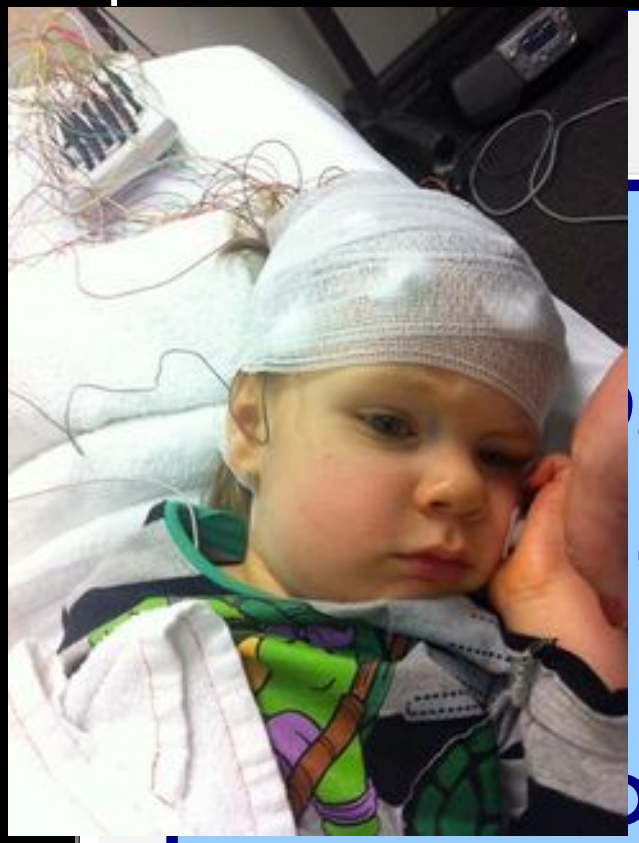
Understanding and adapting statistical models

an exploration in language

Sampling bias is pervasive

- Object detection:
  - Objects don't come segmented in the real world
- Translation:
  - Can't learn to translate tense or colloquial text from news
- Med. Diagnosis
  - Only have data for patients you've seen



Kitchenaid in perfect condition
Anti Obama still perfect

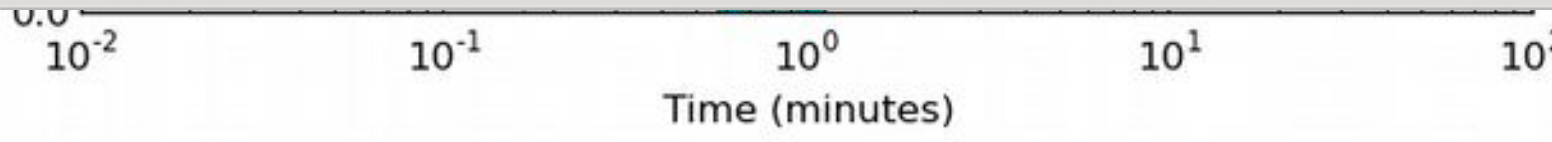
So my little guy had an eeg today to check for seizure activity. He was a champ. He had to be calm while tgey glued 26 wires on his head and lay calmly for 20 minutes for testing. I was so impressed how calm he was. Thank goodness for youtube videos. Hoping the results are negative seizures and show good brain activity. If you have a chance we could use some thoughts and prayers. Thank you.

orne source
is curated
sually want
general
ons/systems

Assumptions that underlie most ML

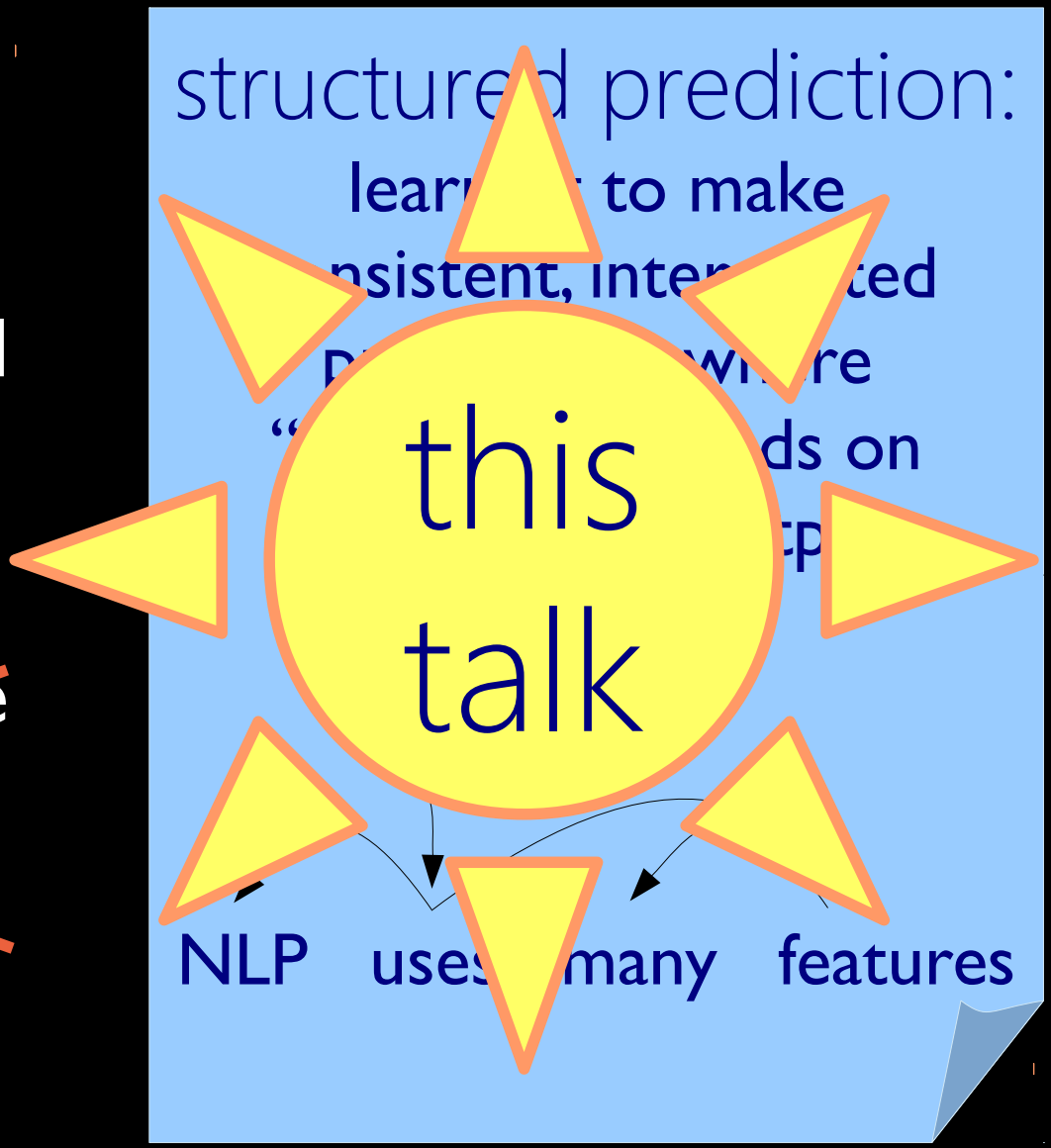
Named entity recognition (200k words, 9 tags)

```
#include "example.h"

namespace SequenceTask {
  using namespace Searn;

  void initialize(searn& srn, size_t& num_actions, std::vector<std::string>&opts, po::variables_map& vm, po::variab
    srn.task_data            = NULL;  // we don't need any of our own data
    srn.auto_history         = true;  // automatically add history features to our examples, please
    srn.auto_hamming_loss    = true;  // please just use hamming loss on individual predictions -- we won't declare
    srn.examples_dont_change = true;  // we don't do any internal example munging
  }

  void finish(searn& srn) { }    // if we had task data, we'd want to free it here

  void structured_predict(searn& srn, example**ec, size_t len, stringstream*output_ss, stringstream*truth_ss) {
    for (size_t i=0; i<len; i++) { //save state for optimization
      srn.snapshot(i, 1, &i, sizeof(i), true);

      OAA::mc_label* y = (OAA::mc_label*)ec[i]->ld;
      size_t prediction = srn.predict(ec[i], NULL, y->label);

      if (output_ss) (*output_ss) << prediction << ' ';
      if (truth_ss ) (*truth_ss ) << (OAA::label_is_test(y) ? '?' : y->label) << ' ';
    }
  }
}

namespace SequenceSpanTask {
```



Assumptions that underlie most ML

- Training data is:
  - ~~independent~~ and
  - ~~identically~~ distributed
- Test data is:
  - ~~drawn from the same distribution as the training data~~



Language does have many flavors!

- Can you guess what domain each of these sentences is drawn from?

News	Many factors contributed to the French and Dutch objections to the proposed EU constitution
Parliament	Please rise, then, for this minute's silence
Medical	Latent diabetes mellitus may become manifest during thiazide therapy
Science	Statistical machine translation is based on sets of text to build a translation model
Step-mother	I forgot to mention in yesterdays post that I also trimmed an overgrown HUGE hedge that spams the entire length of the front of my house and is about 3' accrossed.

An example from clinical text

- Wall Street Journal:

  Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

- Clinical narrati

  LV systolic fn normal with EF 60%.

"is"

"of"

Left ventrical

function

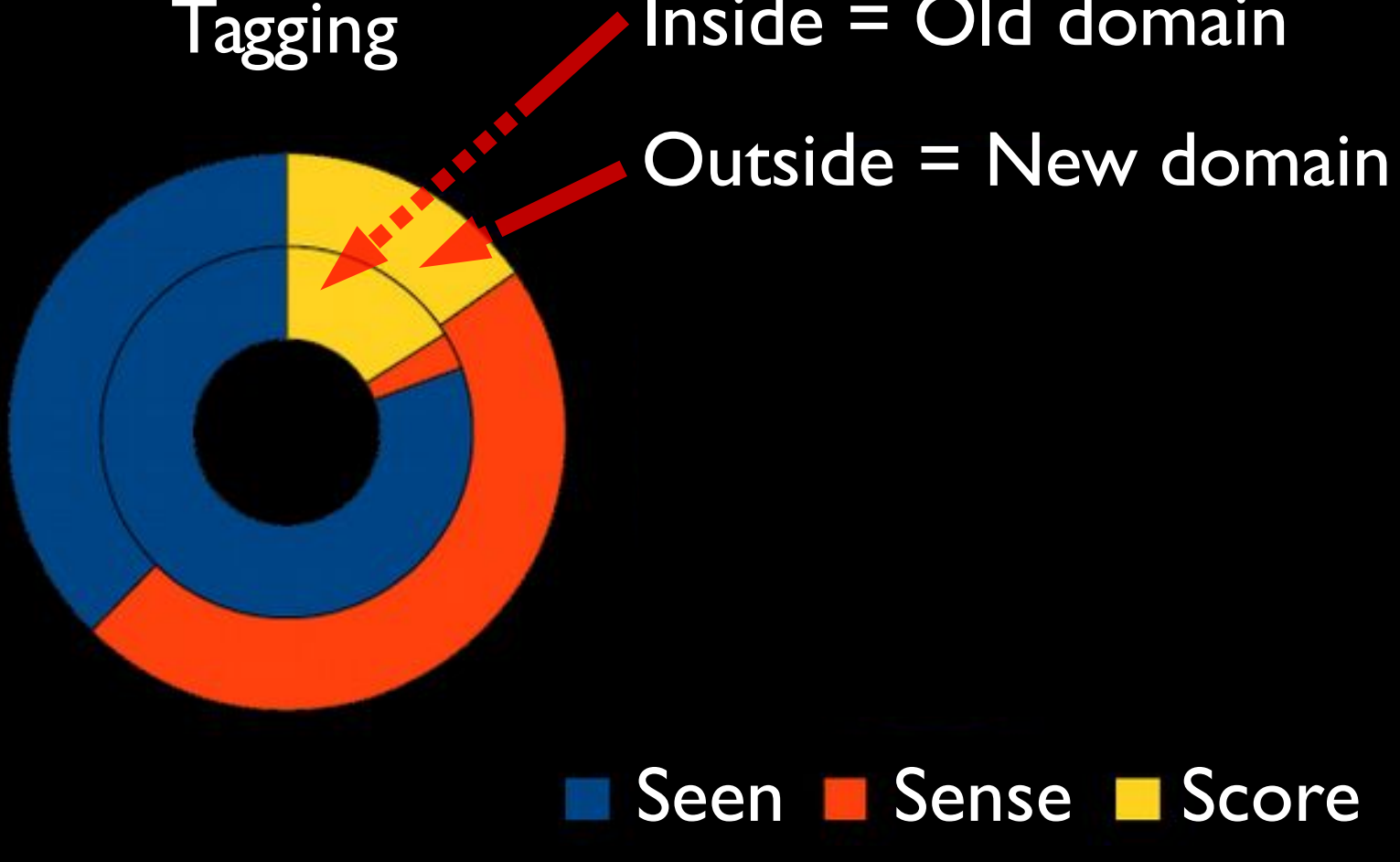
ejection fraction

S4 taxonomy of adaptation effects

- **Seen:** Never seen this word before
  - News to medical: "diabetes mellitus"
- **Sense:** Never seen this word used in this way
  - News to technical: "monitor"
- **Score:** The wrong output is scored higher
  - News to medical: "manifest"
- **Search:** Decoding/search erred

S4 applied to "easy" NLP problems...

Part of Speech Tagging

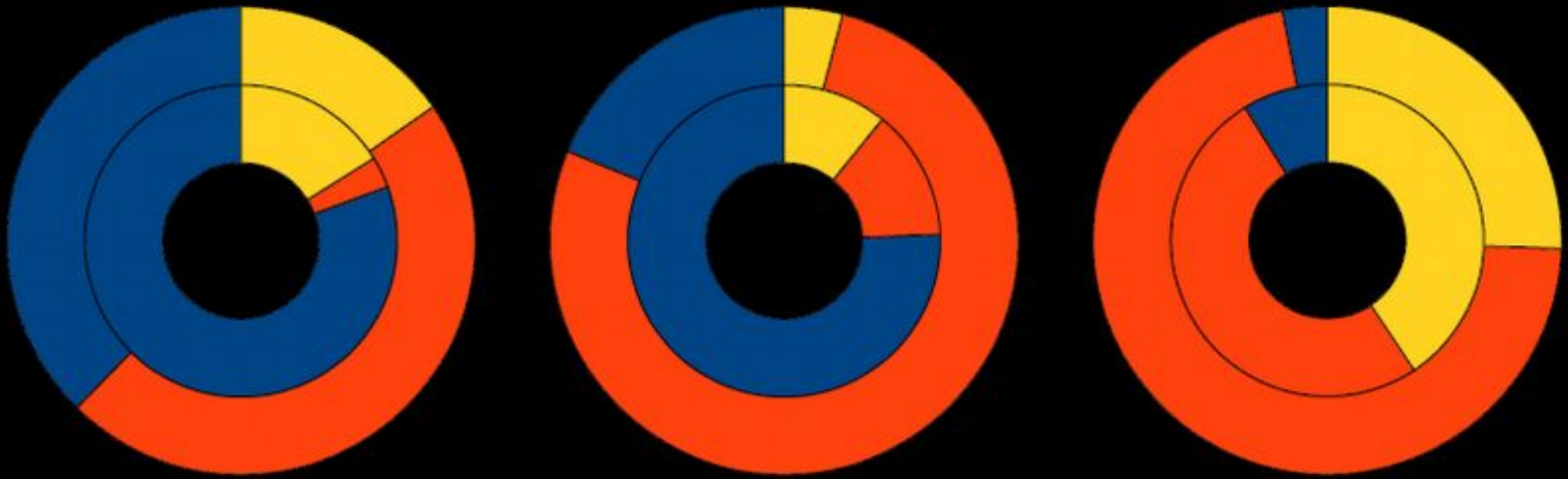


S4 applied to "easy" NLP problems...

Part of Speech Tagging

Shallow Parsing

Named Entity Recognition



Seen Sense Score

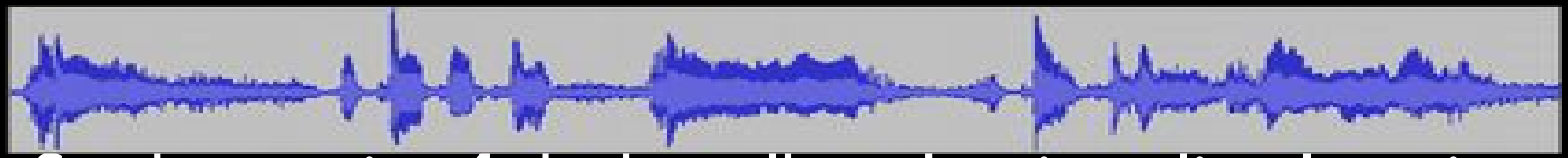
Inside = Old domain Outside = New domain

Classic learning

Predict: $x \rightarrow y$, $(x,y) \sim \Pr[x,y]$

Running with Scissors **Title:** Horrible book, horrible.

This book was horrible. I read half, suffering from a headache the entire time, and eventually i lit it on fire. 1 less copy in the world. Don't waste your money. I wish i had the time spent reading this book back. It wasted my life

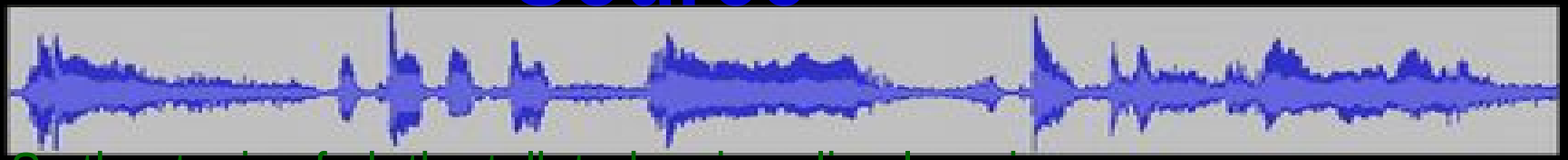


So the topic of ah the talk today is online learning

Domain Adaptation

Training $(x, y) \sim \Pr_S[x, y]$

Source



So the topic of ah the talk today is online learning

Testing $(x, y) \sim \Pr_T[x, y]$

Target

Everything is happening online. Even the slides are produced on-line

You could even get a Nobel prize!

- James Heckman
- Nobel prize in economics (2000)
- *Sample selection bias as specification error.* Econometrica (1979)

"MONITOR" versus "THE"

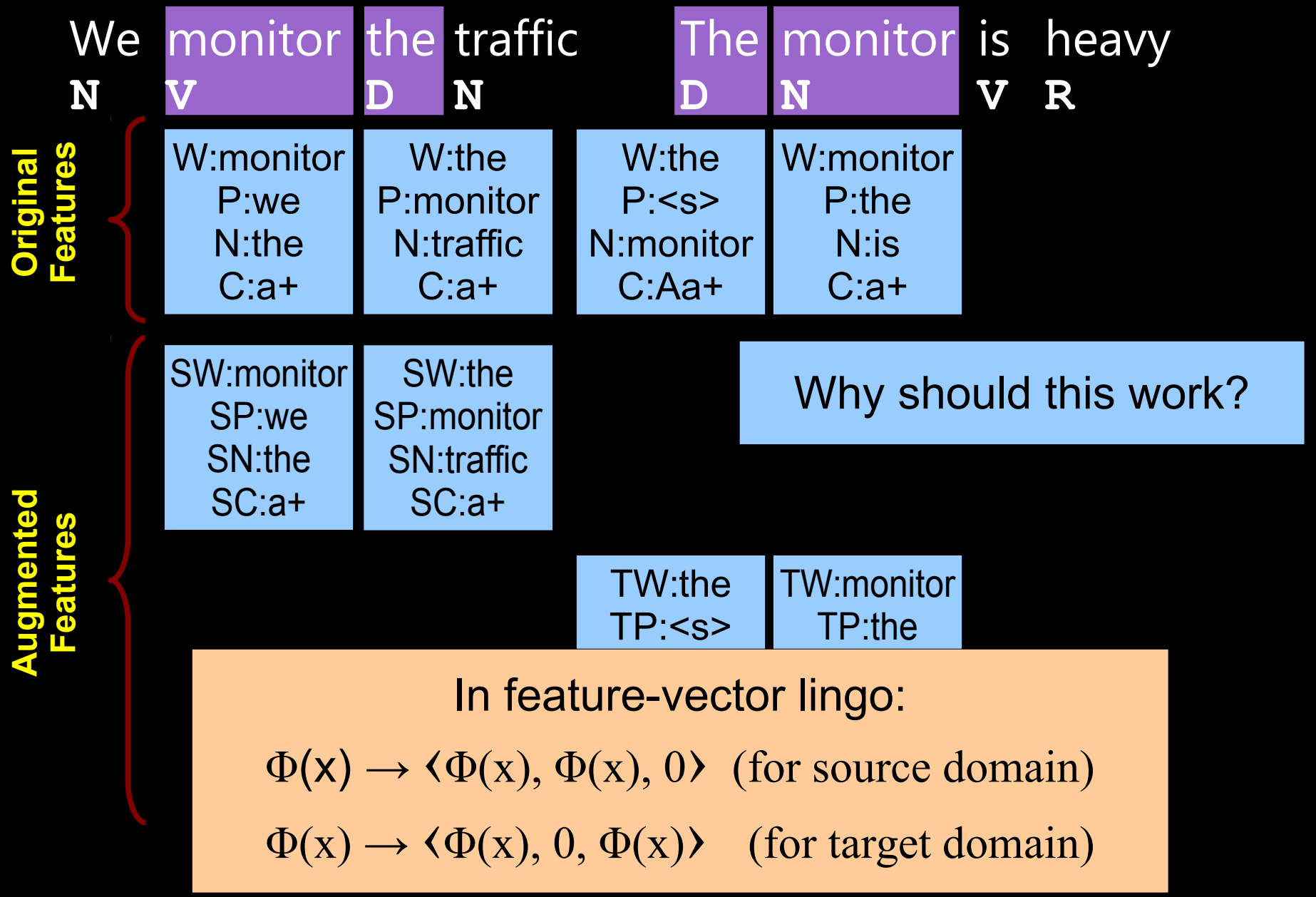
News domain:
"MONITOR" is a **verb**
"THE" is a **determiner**

Technical domain:
"MONITOR" is a **noun**
"THE" is a **determiner**

Key Idea:
Share some features ("the")
Don't share others ("monitor")

(and let the *learner* decide which are which)

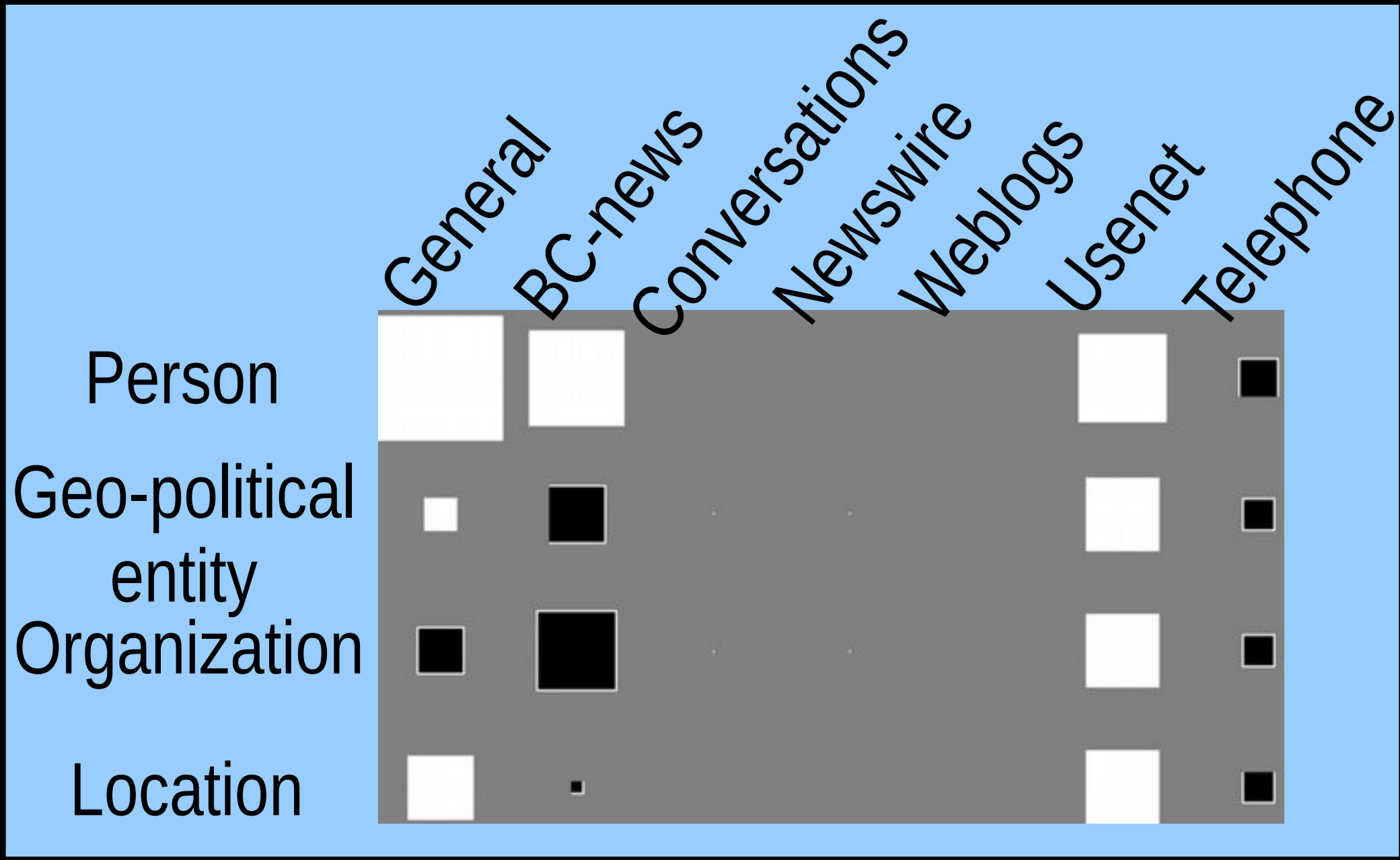
Feature Augmentation



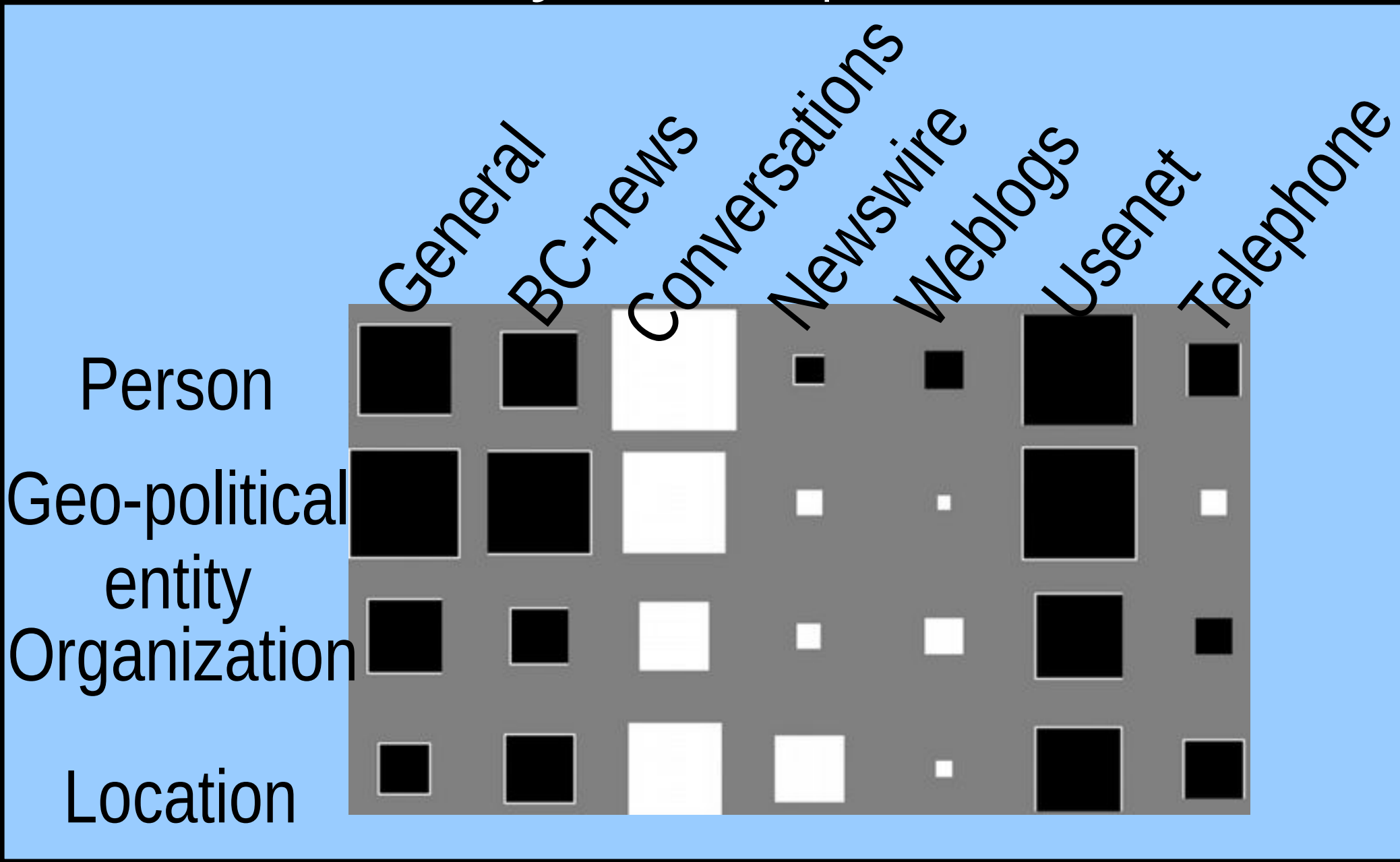
Results – Error Rates

Task	Dom	SrcOnly	TgtOnly	Baseline	Prior	Augment
ACE-NER	bn	4.98	2.37	2.11 (pred)	2.06	**1.98**
	bc	4.54	4.07	3.53 (weight)	**3.47**	**3.47**
	nw	4.78	3.71	3.56 (pred)	3.68	**3.39**
	wl	2.45	2.45	**2.12** (all)	2.41	**2.12**
	un	3.67	2.46	2.10 (linint)	2.03	**1.91**
	cts	2.08	0.46	0.40 (all)	**0.34**	**0.32**
CoNLL	tgt	2.49	2.95	**1.75** (wgt/li)	1.89	**1.76**
PubMed	tgt	12.02	4.15	3.95 (linint)	3.99	**3.61**
CNN	tgt	10.29	3.82	3.44 (linint)	**3.35**	**3.37**
Treebank-Chunk	wsj	6.63	4.35	4.30 (weight)	4.27	**4.11**
	swbd3	15.90	4.15	4.09 (linint)	3.60	**3.51**
	br-cf	5.16	6.27	**4.72** (linint)	5.22	5.15
	br-cg	4.32	5.36	**4.15** (all)	4.25	4.90
	br-ck	5.05	6.32	**5.01** (prd/li)	5.27	5.41
	br-cl	5.66	6.60	**5.39** (wgt/prd)	5.99	5.73
	br-cm	3.57	6.59	**3.11** (all)	4.08	4.89
	br-cn	4.60	5.56	**4.19** (prd/li)	4.48	4.42
	br-cp	4.82	5.62	**4.55** (wgt/prd/li)	4.87	4.78
	br-cr	5.78	9.13	**5.15** (linint)	6.71	6.30
Treebank-brown	brown	6.35	5.75	4.72 (linint)	4.72	**4.65**

Named Entity Rec.: /bush/



Named Entity Rec.: p=/the/



Some Theory

Can bound expected target error:

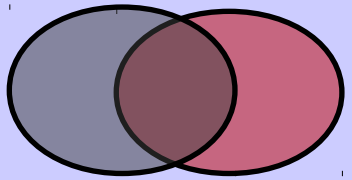
Average training error

$$\epsilon_t \le \frac{1}{2}\left(\hat{\epsilon}_s + \hat{\epsilon}_t\right) + O(complexity)$$

$$+ \left(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}\right) O\left(\frac{1}{\delta}\right) + O\left(disc_H(S,T)\right)$$

source examples

target examples



Semi-supervised Extension

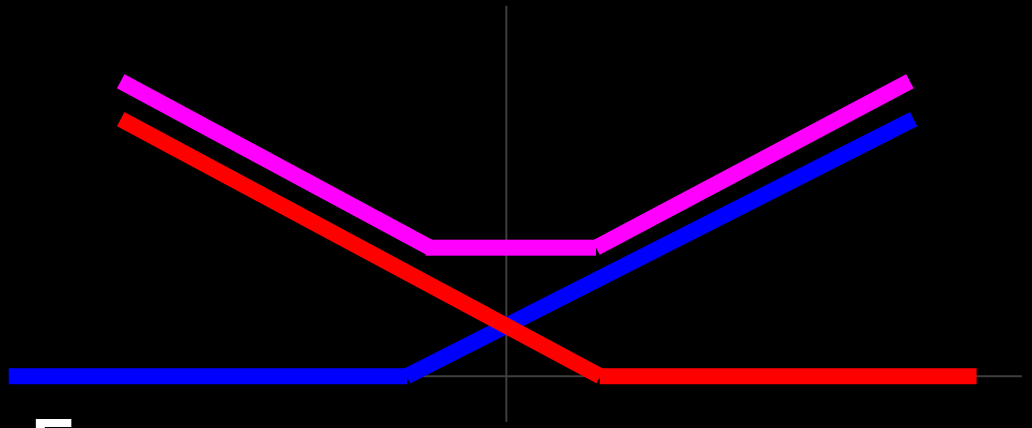
For labeled data:

$(y, x) \rightarrow (y, \langle x, x, 0 \rangle)$ (for source domain)

$(y, x) \rightarrow (y, \langle x, 0, x \rangle)$ (for target domain)

What about unlabeled data?

$(x) \rightarrow \{ (+1, \langle 0, x, -x \rangle) , (-1, \langle 0, x, -x \rangle) \}$



Loss on +ve

Loss on –ve

Loss on unlabeled

Encourage agreement:

$$[h_t(x) = h_s(x)] \Leftrightarrow [w_t \circ x - w_s \circ x = 0]$$

Semi-supervised Bounds

- The complexity drops from $O(tr(B))$ to:

$$O(tr(B) - tr(E(I + kF)^{-1}E'))$$

- Where:

$$K = \begin{matrix} A & C & D \\ C' & B & E \\ D' & E' & F \end{matrix}$$

Source data

Target data

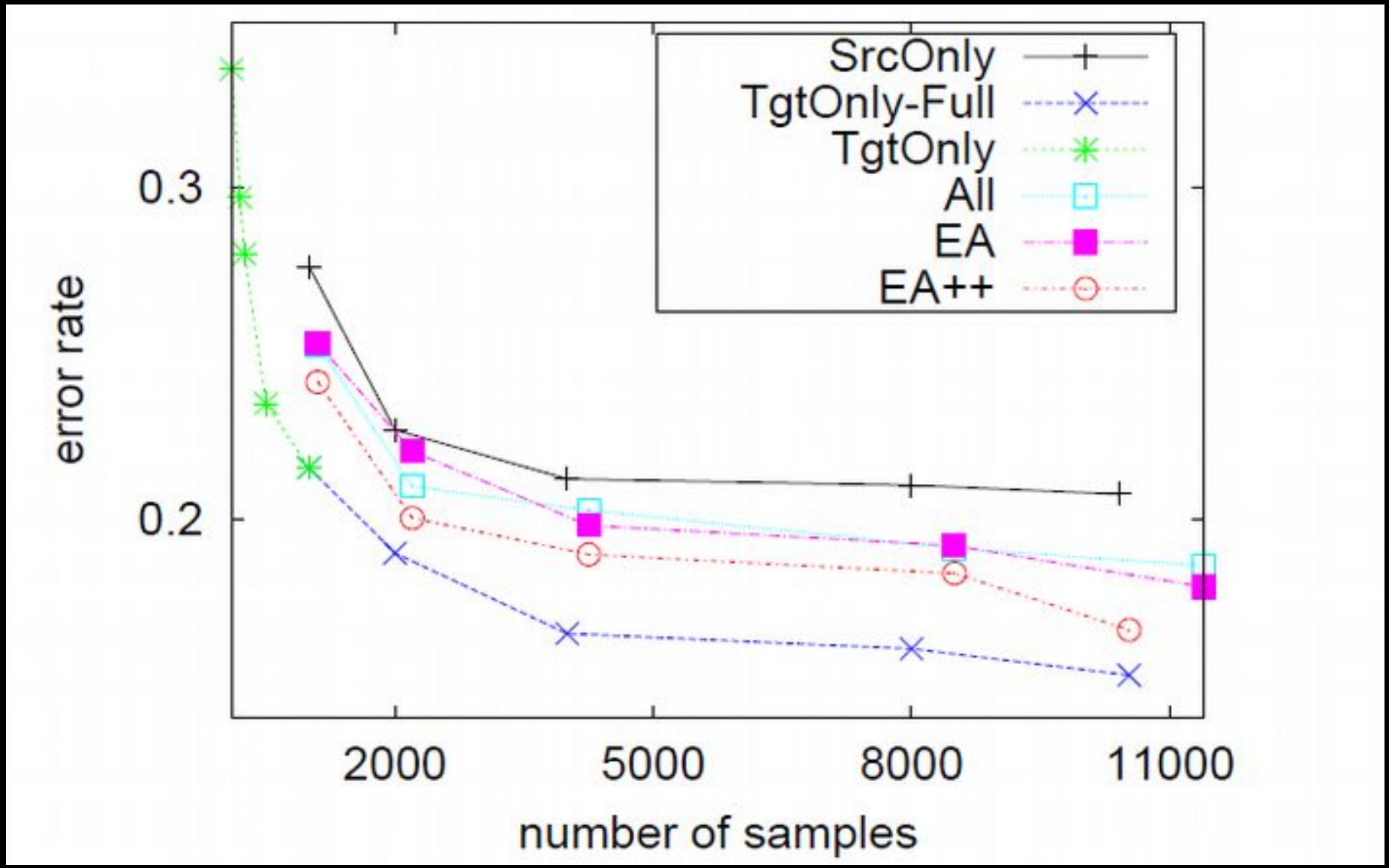
Unlab data

k depends on regularizer

- Proof based on recent results from multi-view learning

D., Kumar & Saha NIPS 2010

Semi-supervised Experiments



(EA++ only gets 50% labeled data!)

D. Kumar & Saha NPS 2010

Is this a problem for harder tasks?

Translating across domains is hard

Old Domain (Parliament)	
Original	monsieur le président, les pêcheurs de homard de la région de l'atlantique sont dans une situation catastrophique.
Reference	mr. speaker, lobster fishers in atlantic canada are facing a disaster.
System	mr. speaker, the lobster fishers in atlantic canada are in a mess.
New Domain	
Original	comprimés pelliculés blancs pour voie orale.
Reference	white film-coated tablets for oral use.
System	white **pelliculés** tablets to oral.
New Domain	
Original	mode et voie(s) d'administration
Reference	method and route(s) of administration
System	**fashion** and **voie(s)** of **directors**

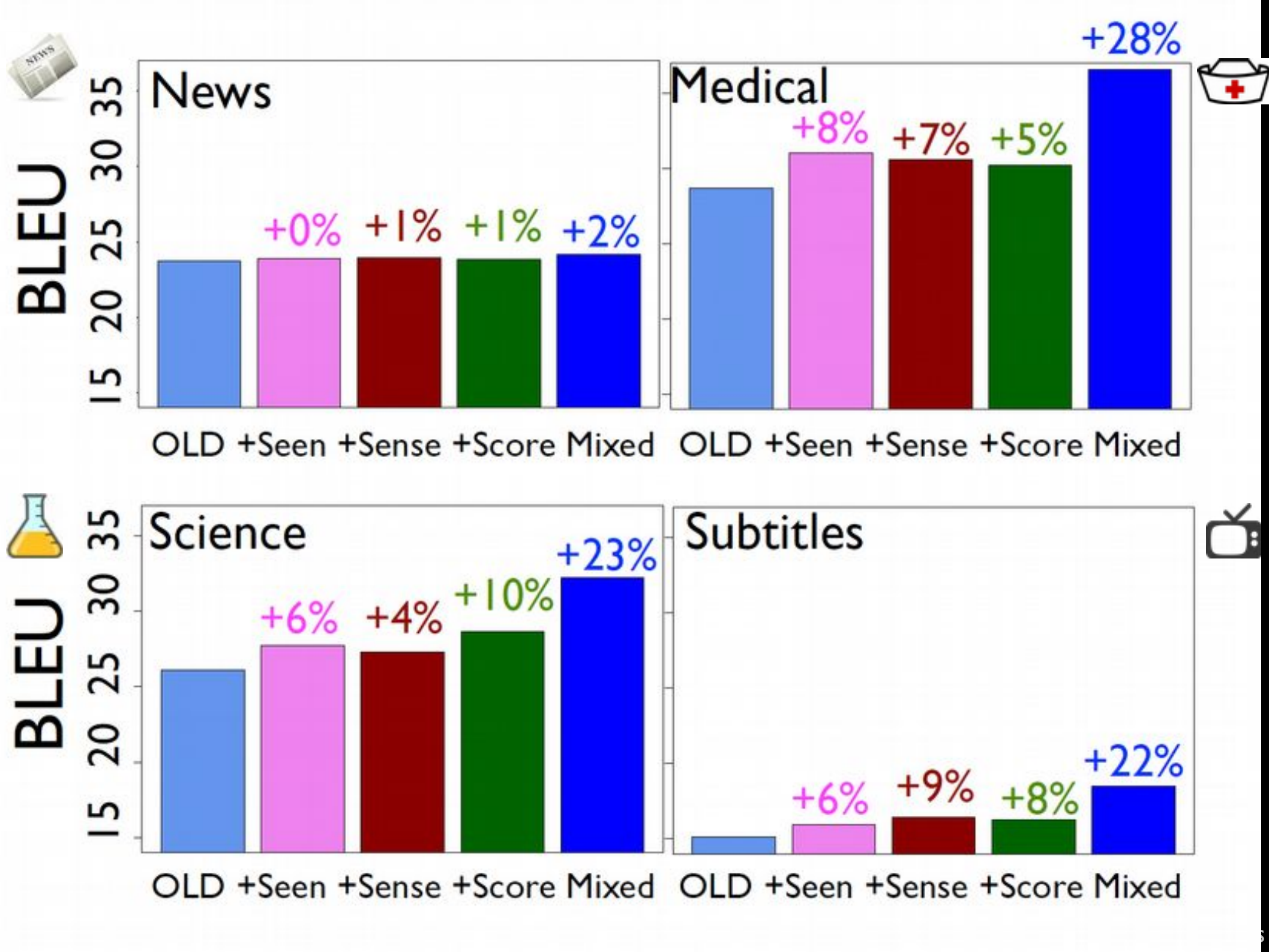
Key Question: What went wrong?

Domain Shift Setting

Old domain: Hansard parliamentary proceedings

New Domain Datasets

Old domain	8,000k	fr	162m	192k
		en	145m	187k


News
Medical
Science
Subtitles
BLEU
15
20
25
30
35
+0%
+1%
+1%
+2%
+8%
+7%
+5%
+28%
+6%
+4%
+10%
+23%
+6%
+9%
+8%
+22%
OLD
+Seen
+Sense
+Score
Mixed

Simultaneously solving seen+sense

- Idea:
  - We have good knowledge of *translations* in the old domain
  - We have good knowledge of raw word frequencies in a new domain in each language individually
  - Can we "nudge" the translation probabilities to match these raw frequencies
- Assumptions:
  - Old domain parallel data
  - New domain comparable data

Marginal matching

	house	place	pregnant	dress	$q^{old}(s)$
enceinte	0.30	0.40	0.10	0	0.80
habiller	0	0	0	0.20	0.20
$q^{old}(t)$	0.30	0.40	0.10	0.20	

(a) OLD-Domain Joint

	house	place	pregnant	dress	girl	$q(s)$
enceinte	?					0.60
habiller						0.20
fille						0.20
$q(t)$	0.12	0.08	0.40	0.20	0.20	

(b) NEW-Domain Marginals

Matched Marginals

	house	place	pregnant	dress	girl	$q^{new}(s)$
enceinte	0.12 ↓	0.08 ↓	0.40 ↑	0	0	0.60
habiller	0	0	0	0.20 =	0	0.20
fille	0	0	0	0	0.20 ↑	0.20
$q^{new}(t)$	0.12	0.08	0.40	0.20	0.20	

(c) Inferred NEW-Domain Joint

Marginal matching details

$\Omega(p)$: regularization term

$f(p)$: edit distance penalty

$$p^{new} = \arg\min_p \left\| p - p^{old} \right\|_1 + \Omega(p) + f(p)$$

subject to:

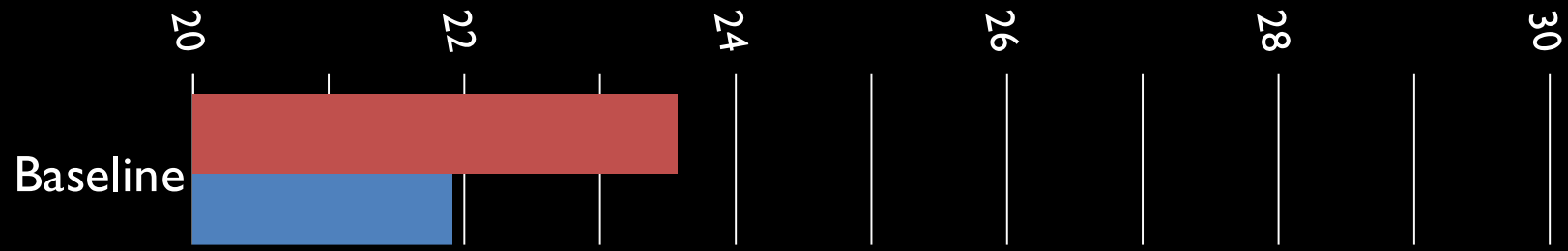
$$\sum_{s,t} p(s,t) = 1, \quad p(s,t) \geq 0$$

$$\sum_s p(s,t) = q(t), \quad \sum_t p(s,t) = q(s)$$

Example Learned Translations

French	Correct	Learned Translations
cisaillement	shear	viscous crack shear
chromosomes	chromosomes	chromosomes chromosome chromosomal
caractérisation	characterization	characterization characteristic π
araignées	spiders	spiders ant spider
tiges	stems	usda centimeters flowering

BLEU Scores



Discussion

- Adaptation effects are real and lead to significant degradation in system performance
- Simple techniques work well in theory + practice
- Understanding source of errors helps address them
- Marginal matching addresses new S4 issues
- How can we combat sampling bias when we know the bias? When we do not?

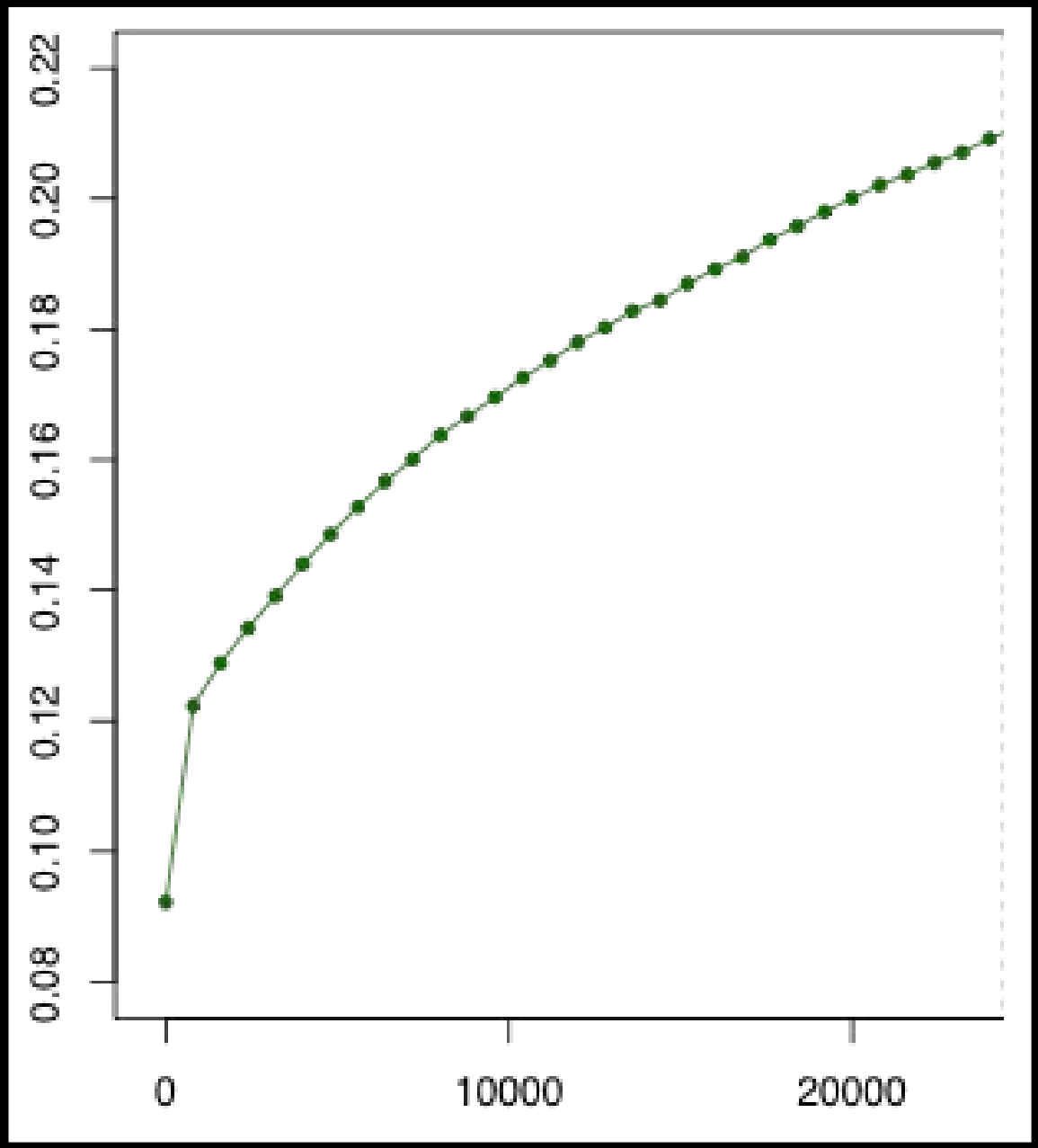
Thanks! Questions?

Thanks to:
John Blitzer
Fabienne Braune
Lyle Campbell
Marine Carpuat
Ann Clifton
Jason Eisner
Alex Fraser
Lise Getoor
Katharine Henry
Ann Irvine
Jagadeesh Jagarlamudi
Abhishek Kumar
John Langford
John Morgan
Jay Pujara
Chris Quirk
Majid Razmara
Rachel Rudinger
Avishek Saha
Aleš Tamchyna

Some example "seen" errors

Dom	Most frequent OOV Words			
News (17%)	behavior	favor	neighbors	fueled
	neighboring	abe	wwii	favored
	favorable	zhao	ahmedinejad	bernanke
	favorite	phelps	ccp	skeptical
Medical (49%)	renal	hepatic	subcutaneous	irbesartan
	ribavirin	olanzapine	serum	patienten
	dl	eine	sie	pharmacokinetics
	ritonavir	hydrochlorothiazide	erythropoietin	efavirenz
Movies (44%)	gonna	yeah	mom	hi
	b****	daddy	s***	later
	f*****g	f***	gotta	wanna
	uh	namely	bye	dude

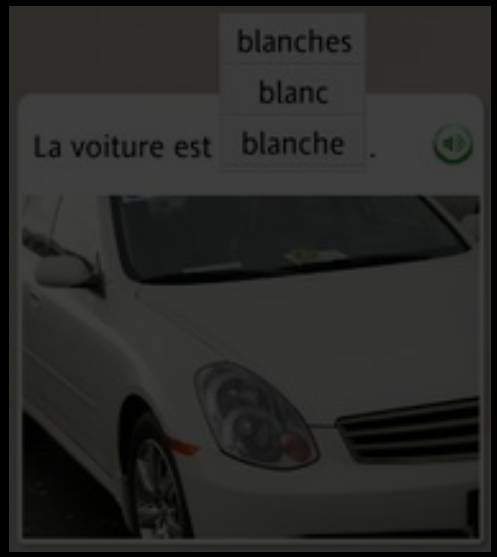
Intrinsic evaluation: MRR



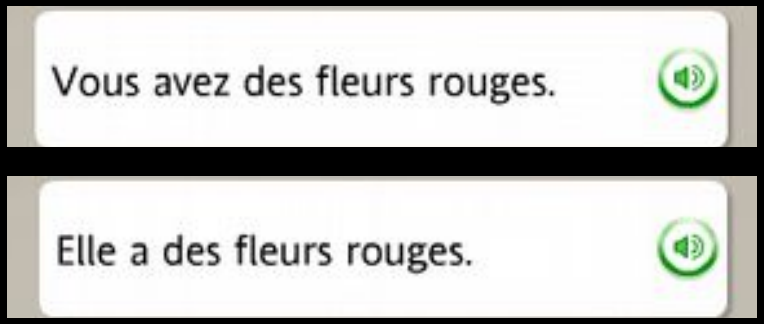
- Ranked document pairs: learning from most science-like first
- Relative gain expected to slow, documents less and less science-y

Aspects of computational 2nd LL

- Very specific linguistic variants
  - Number, case, agreement, etc.
  - *Not enough* to get the majority case

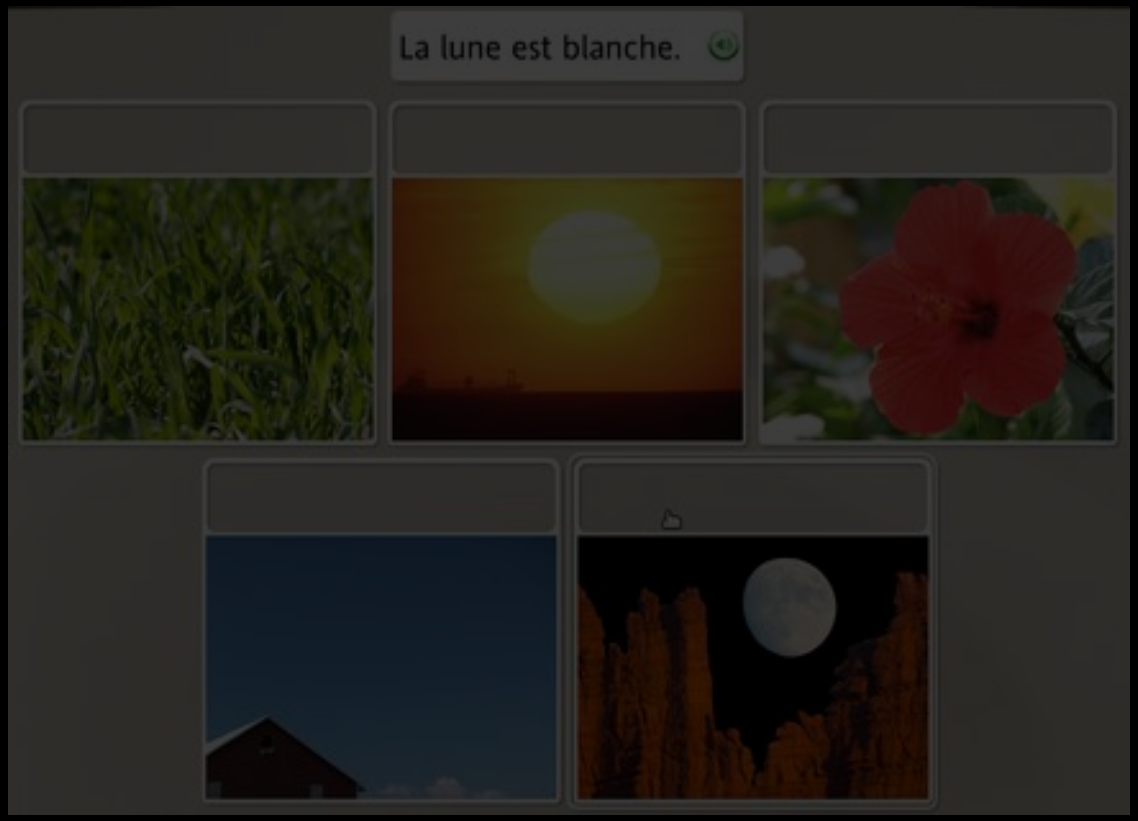


- Focus on subtle visual differences



Aspects of computational 2nd LL

- AI-style reasoning & one-shot learning



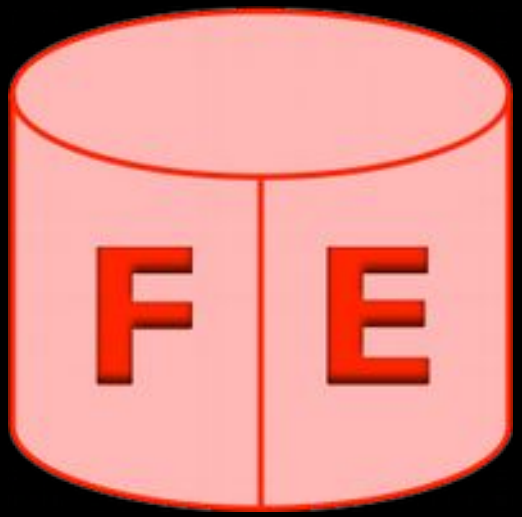
- "It's learnable" proof of concept:

What is NLP?

- Fundamental goal: deep understanding of text
  - Not just string processing or keyword matching
- End systems that we want to build
  - Simple: Spelling correction, text categorization, etc.
  - Complex: Speech recognition, machine translation, information extraction, dialog interfaces, question answering
  - Unknown: human-level comprehension (more than just NLP?)

Macro-Analysis: TETRA

Old-Domain Phrase Table ("*Old*")

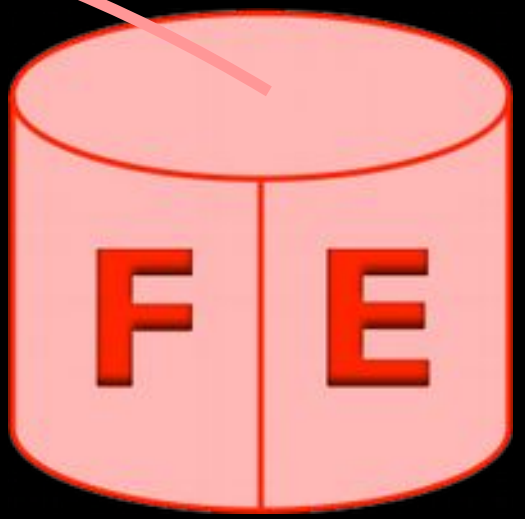


Both-Domains Phrase Table ("*Mixed*")

Macro-Analysis: TETRA

Measuring **SEEN**

Add all phrase pairs with previously unseen F side



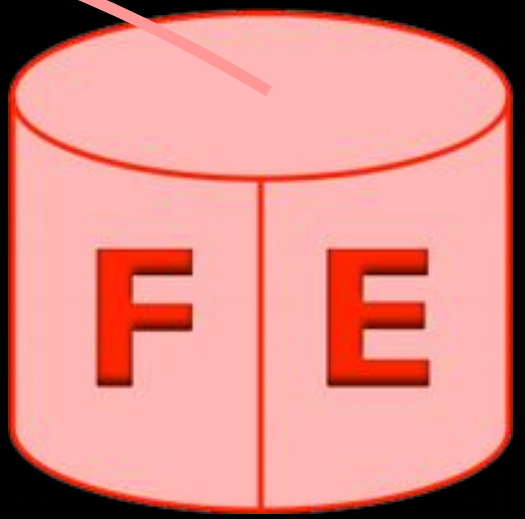
voie(s)	route(s)	0.31 ...

voie(s)	route(s)	0.31 ...
mode	fashion	0.21 ...
mode	method	0.41 ...
administration	directors	0.23 ...
administration	administration	0.15 ...

Macro-Analysis: TETRA

Measuring **SENSE**

Add all phrase pairs with previously seen F side, but unseen translation

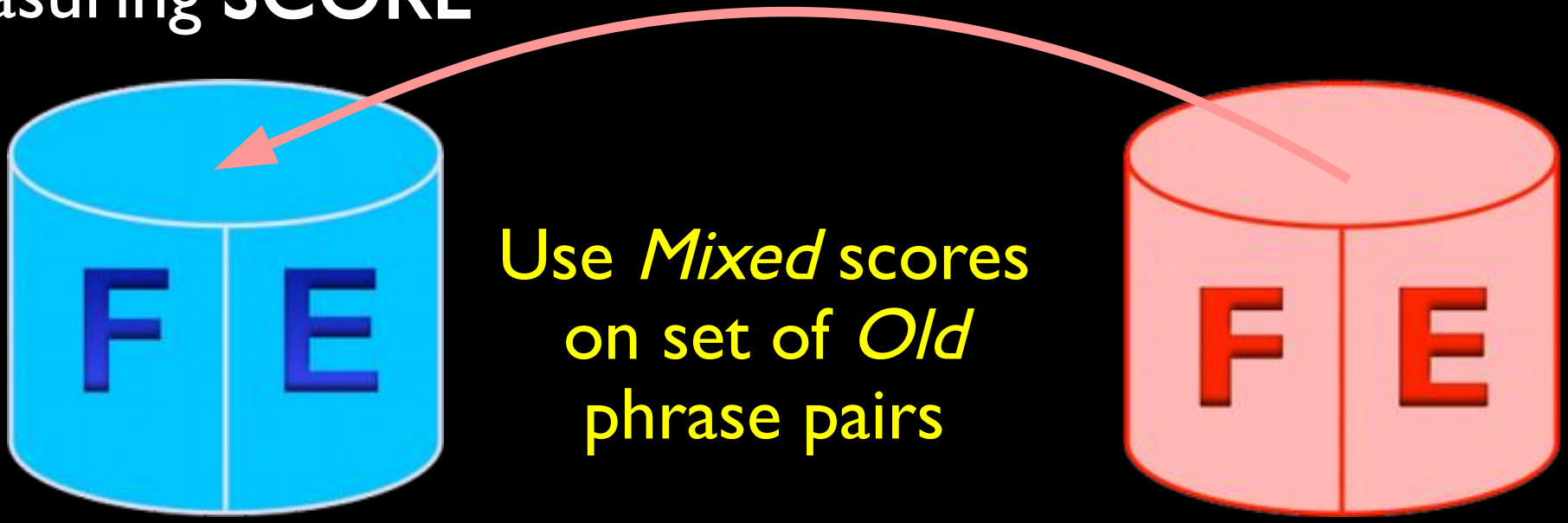


mode	method	0.41 ...

voie(s)	route(s)	0.31 ...
mode	fashion	0.21 ...
mode	method	0.41 ...
administration	directors	0.23 ...
administration	administration	0.15 ...

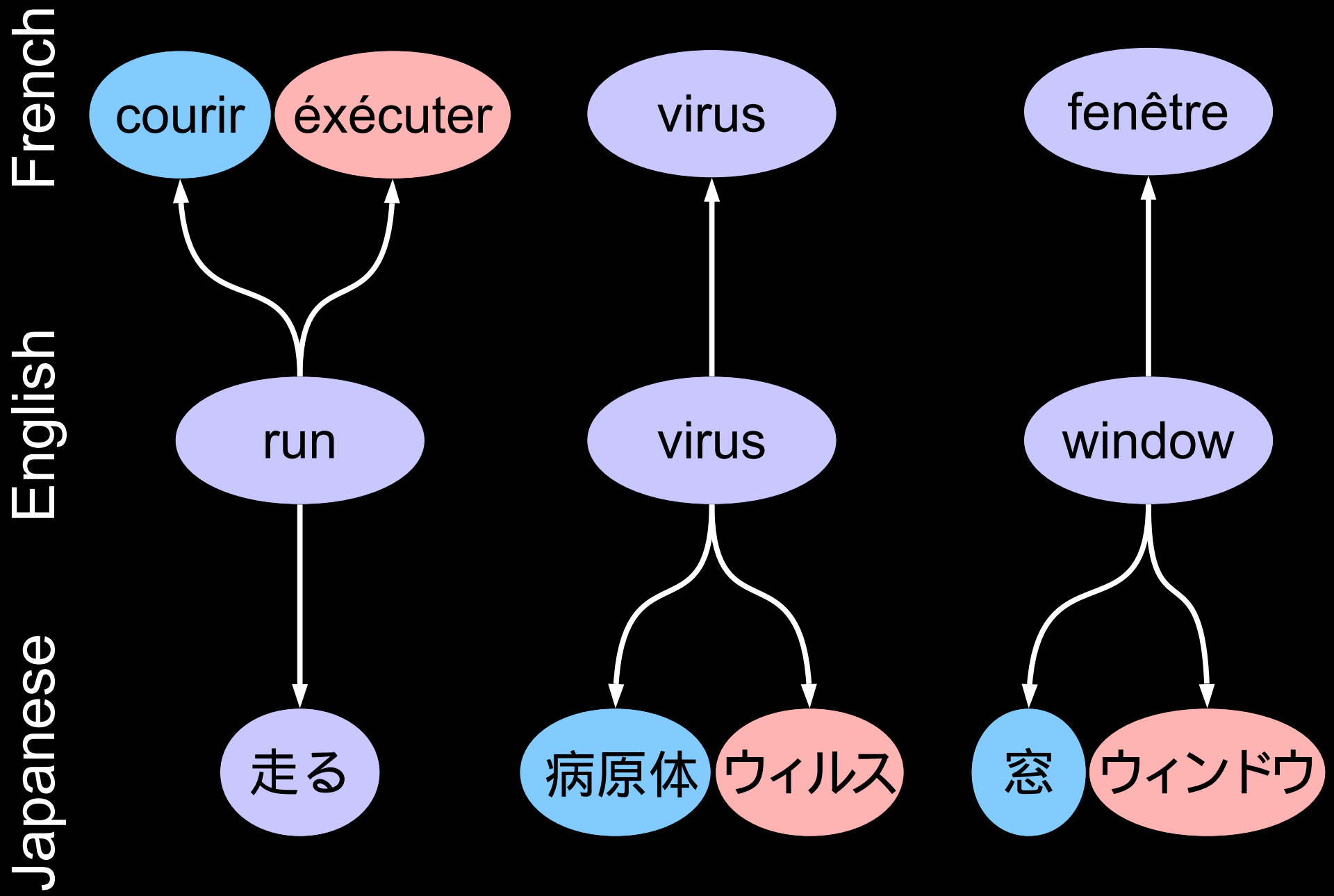
Macro-Analysis: TETRA

Measuring **SCORE**

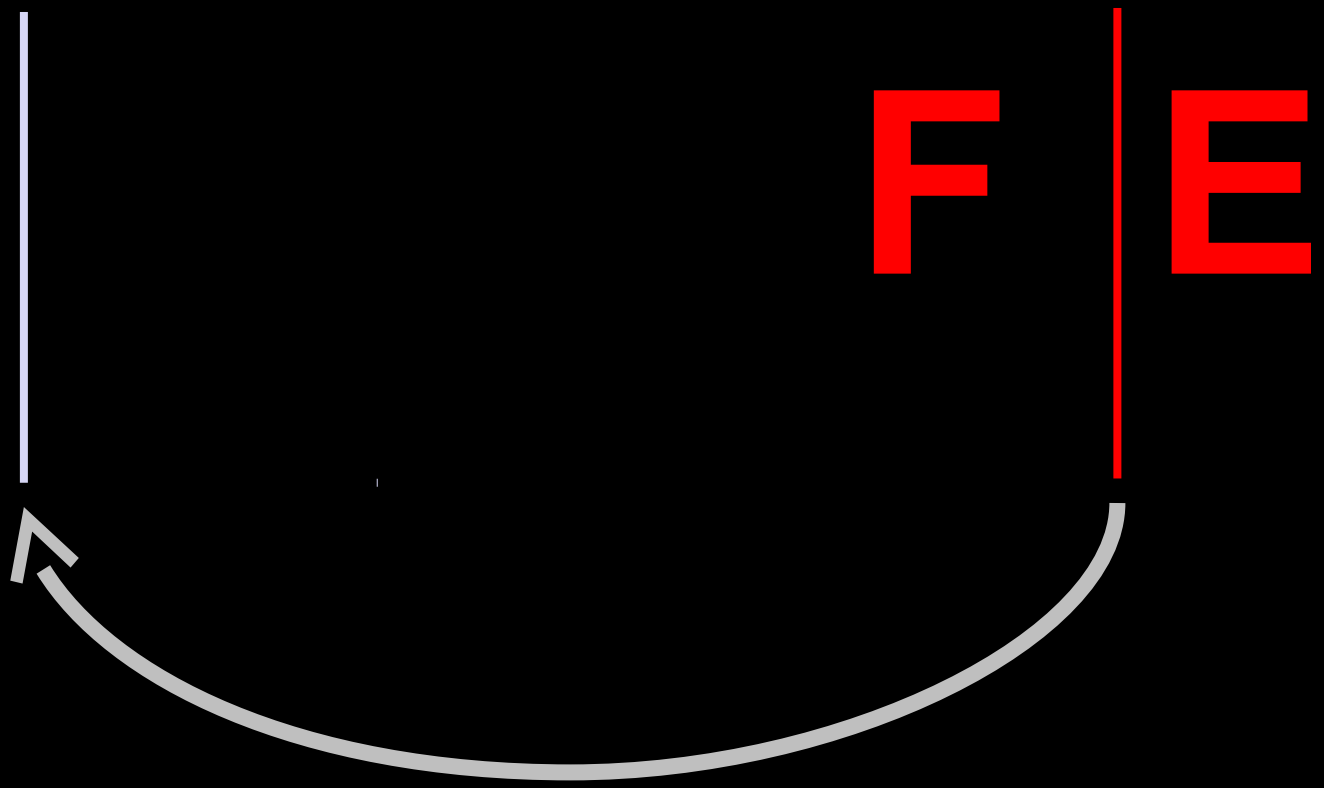


0.21 ...	voie(s)	route(s)	0.31 ...
0.23 ...	mode	fashion	0.21 ...
0.15 ...	mode	method	0.41 ...
	administration	directors	0.23 ...
	administration	administration	0.15 ...

Senses are domain/language specific

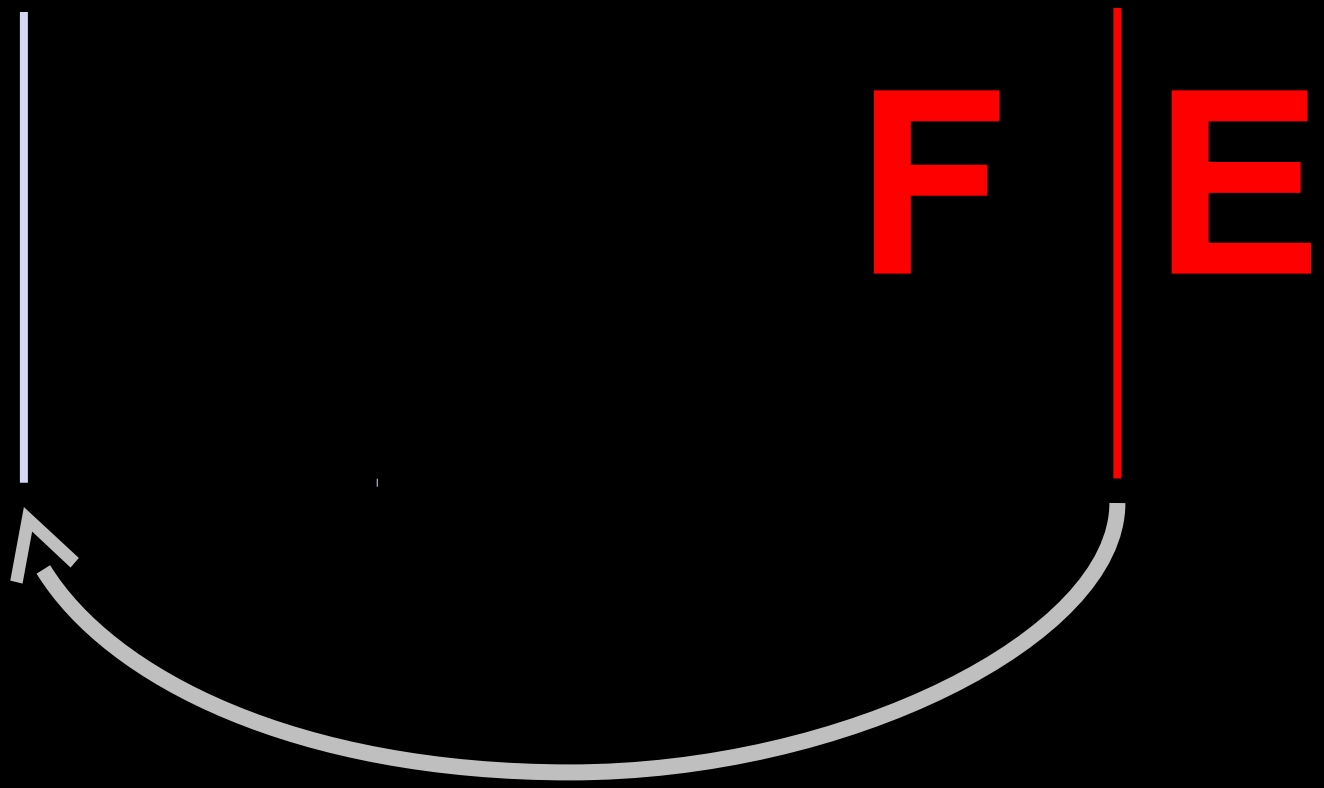


Measuring SEEN effects



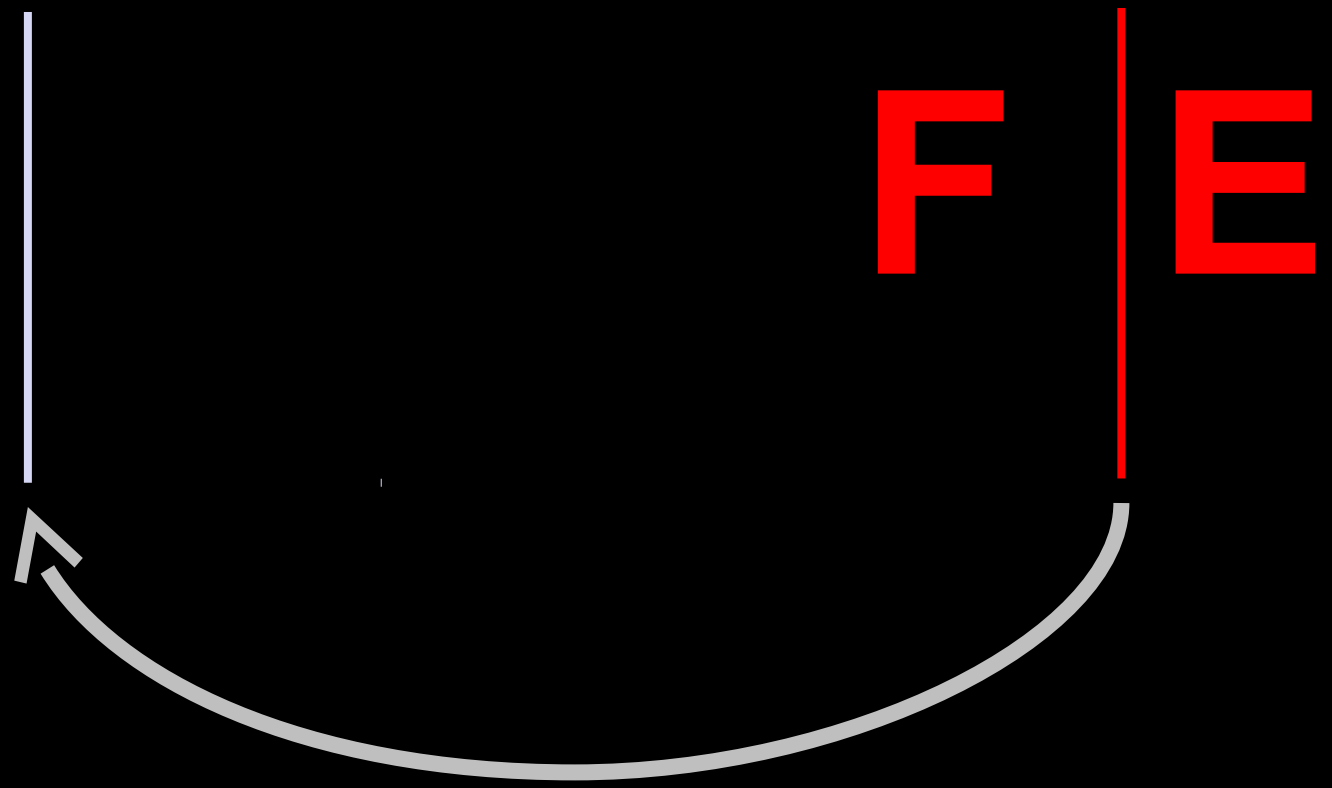
Add all phrase pairs with previously unseen F side

Measuring SENSE effects



Add all phrase pairs with
previously seen F side, but unseen translation

Measuring SCORE effects



Add all phrase pairs, period
(and keep new domain scores)

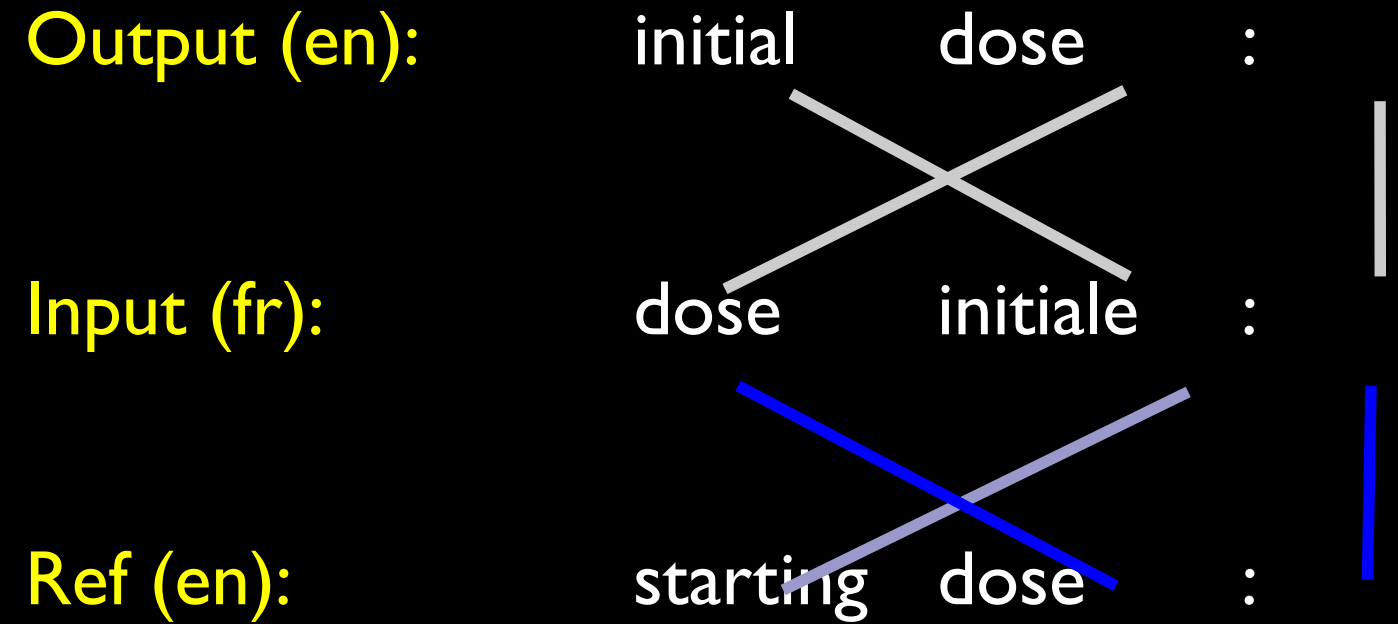
Macro-analysis of S4 effects

- Evaluation using BLEU

	News	Medical	Science	Subtitles
Seen	+0.3%	+8.1%	+6.1%	+5.7%
Sense	+0.6%	+6.6%	+4.4%	+8.7%
Score	+0.6%	+4.5%	+9.9%	+8.4%
Search	+0.0%	+0.0%	+0.0%	+0.0%

- Hansard: 8m sents 161m fr-tokens
- News: 135k sents 3.9m fr-tokens
- Medical: 472k sents 6.5m fr-tokens
- Science: 139k sents 4.3m fr-tokens
- Subtitles: 19m sents 155m fr-tokens

Micro-analysis of S4 effects



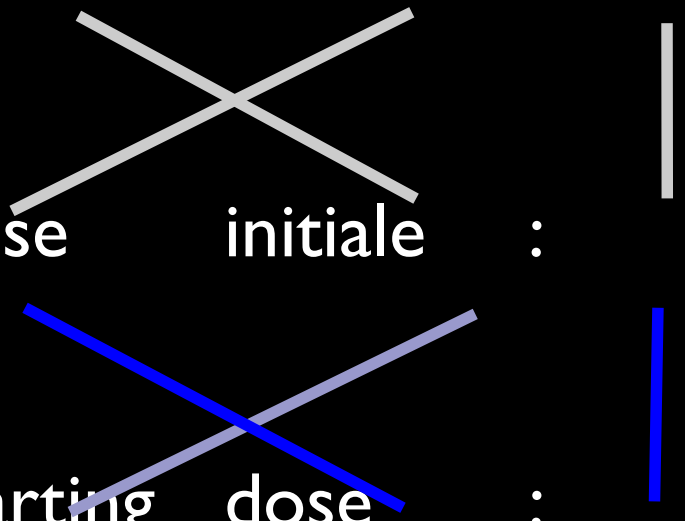
Correct
Seen-freebie
Seen-error
Sense-error
Score-error

Micro-analysis of S4 effects

Output (en): initial dose :

Input (fr): dose initiale :

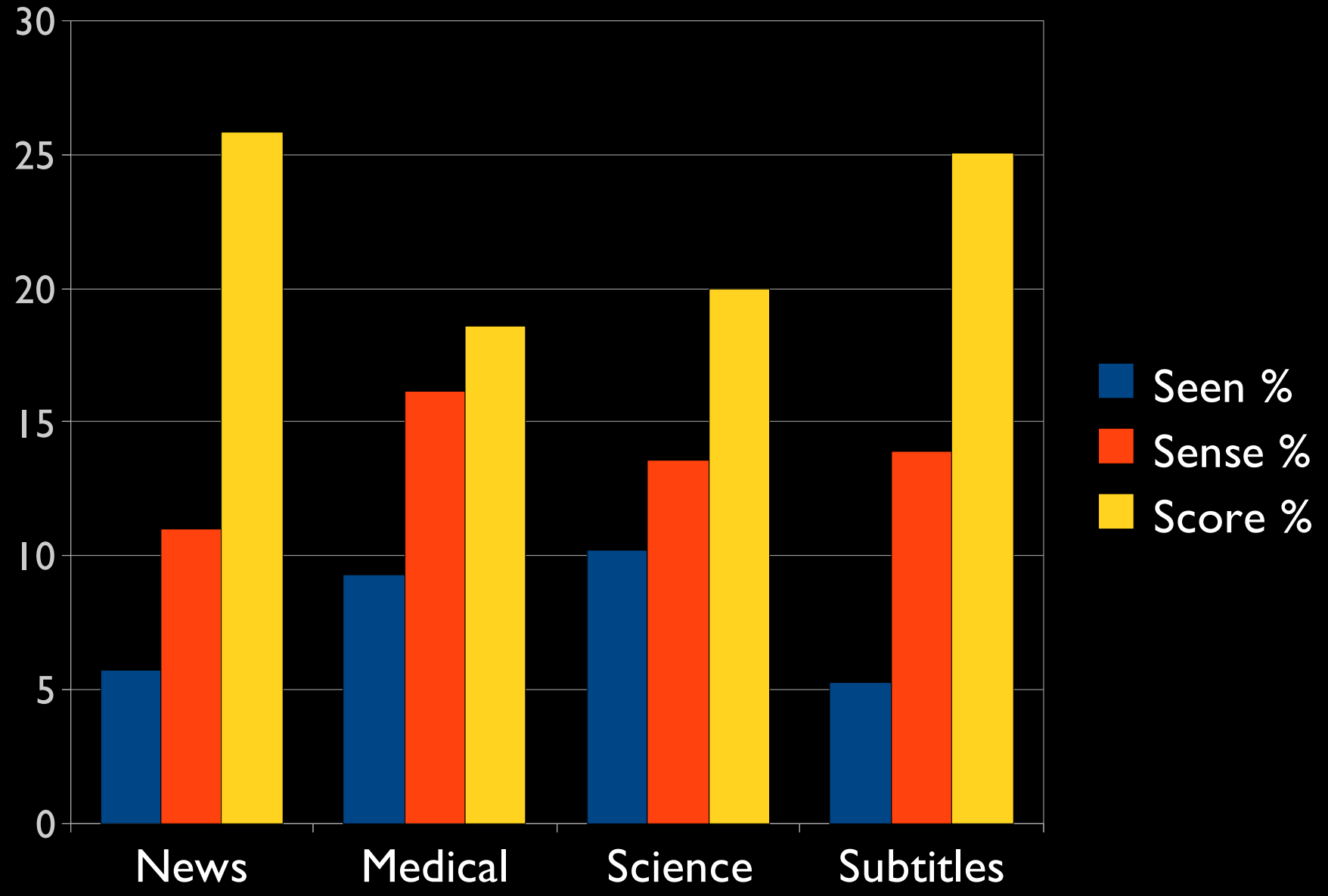
Ref (en): starting dose :



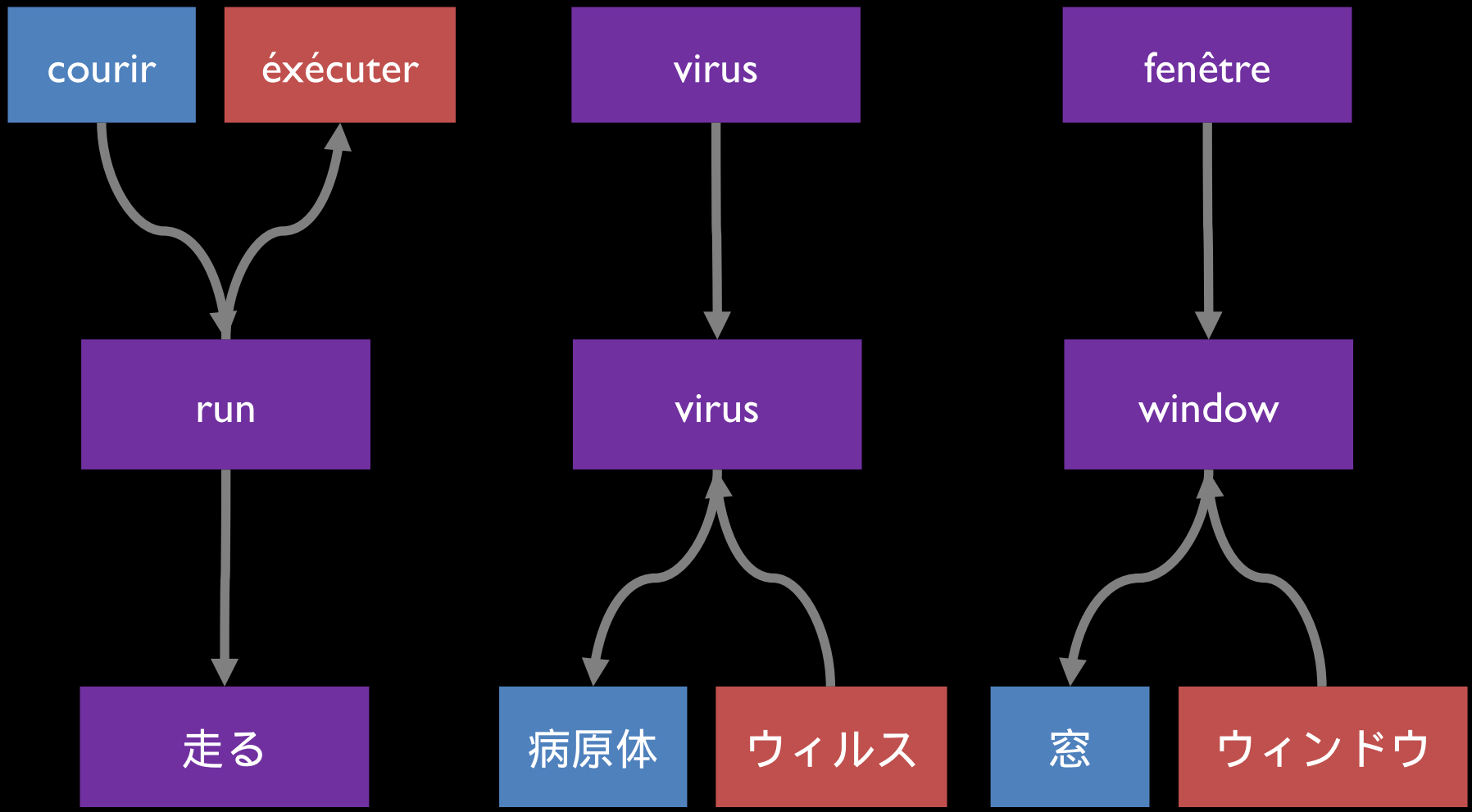
Correct Seen-freebie

Seen-error Sense-error Score-error

Errors found by micro-analysis



Senses are domain/language specific



Case 1: No NEW domain parallel data

- Common situation
  - Lots of data in some OLD domain (e.g., government documents)
  - Need to translate many NEW domain documents
- Acquiring additional NEW domain translations is critical!
- Lots of past work in term mining
  - Distributional similarity [Rapp 1996]
  - Orthographic similarity
  - Temporal similarity

Marginal matching for "sense" errors

Given:

- Joint $p(x, y)$ in old domain
- Marginals $q(x)$ and $q(y)$ in the new domain

Recover:

- Joint $q(x, y)$ in new domain

We formulate as a L1-regularlized linear program

Easier: *many* $q(x)$ and $q(y)$s

	grant	tune	…	
accorder	9	1	…	10+…
…	…	…	…	…
	9+…	1+…	…	

	grant	tune	…	
accorder	???	???	???	5
…	???	???	???	…
	1	5	…	

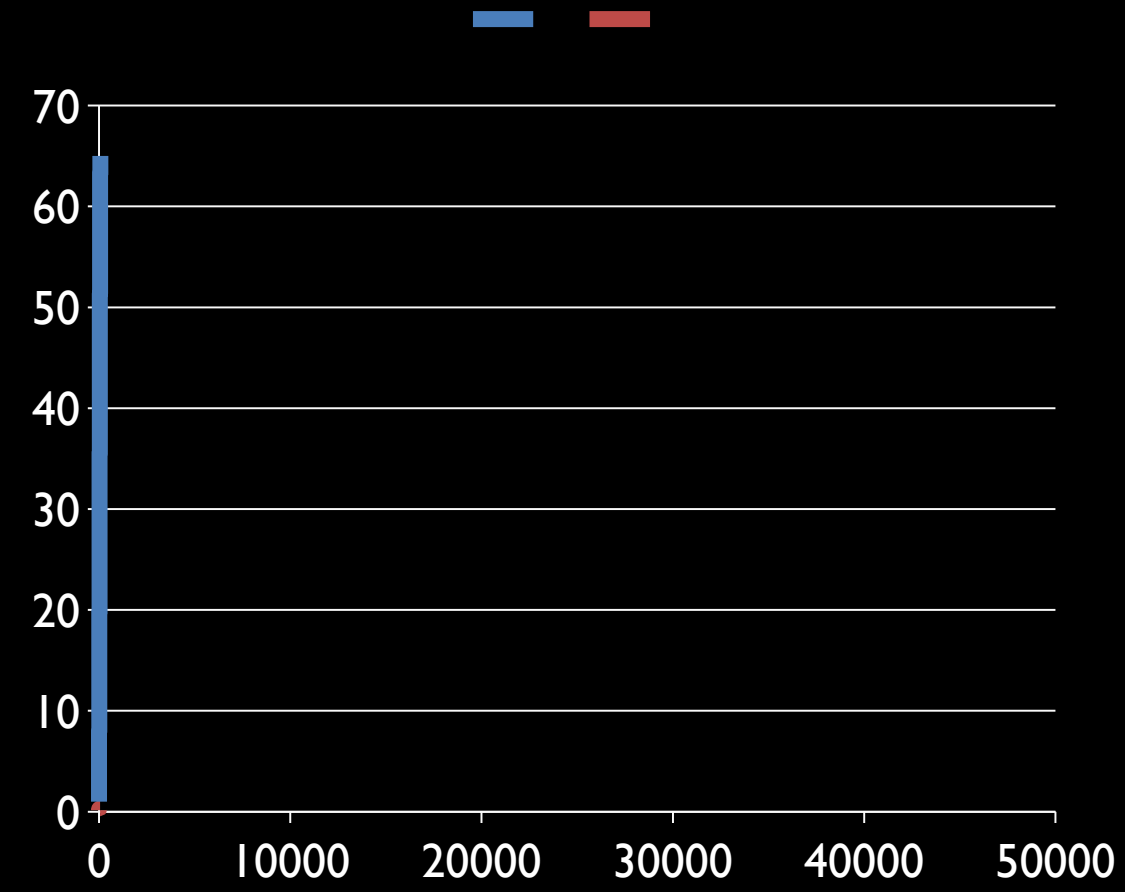
Additional features

- Sparsity: # of non-zero entries should be small
- Distributional: document co-occurrence ⇄ translation pair
- Spelling: Low edit dist ⇄ translation pair
- Frequency: Rare words align to rare words; common words align to common words

c-aractérisation
characterization

E	F
the	le
…	…
spiders	araignées
…	…

Intrinsic evaluation: Mean Reciprocal Rank (MRR)

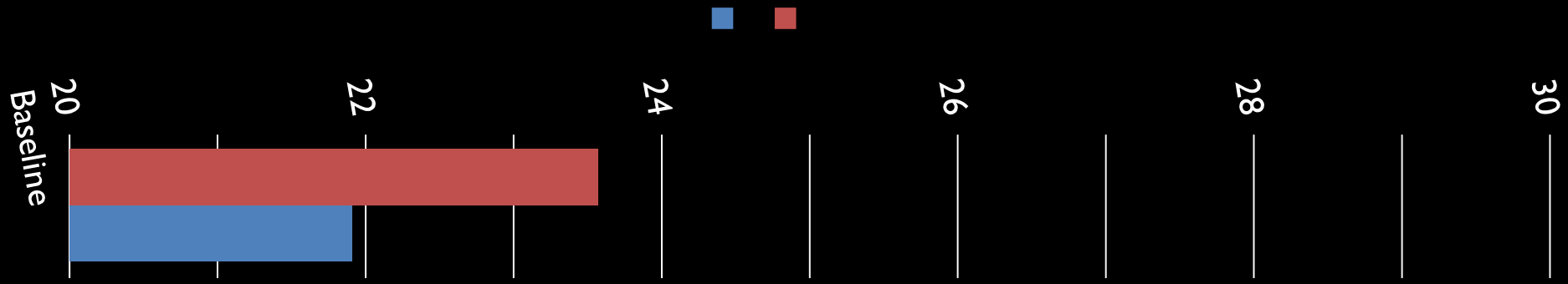


- Ranked Wikipedia document pairs: learning from most science-like first
- Decreasing benefits after ~50,000 document pairs
- Relative gain expected to slow, as documents are less and less science-y

Example learned translations (Science)

French	Correct English	Learned Translations
cisaillement	shear	viscous crack shear
chromosomes	chromosomes	chromosomes chromosome chromosomal
caractérisation	characterization	characterization characteristic
araignées	spiders	spiders ant spider
tiges	stems	usda centimeters flowering

BLEU Scores

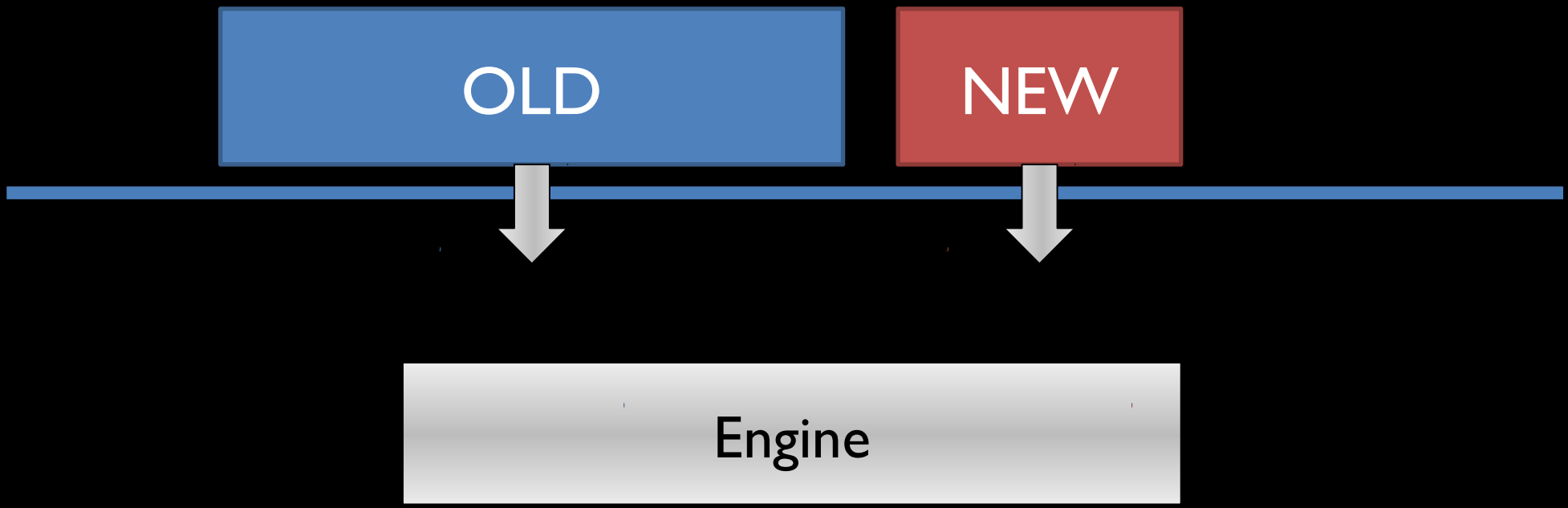


Case 2: Add NEW domain parallel data

- Say we have a NEW domain translation memory
- How can we leverage our OLD domain to achieve the greatest benefit?

Initial adaptation baselines

Use both models (log-linear mixture)



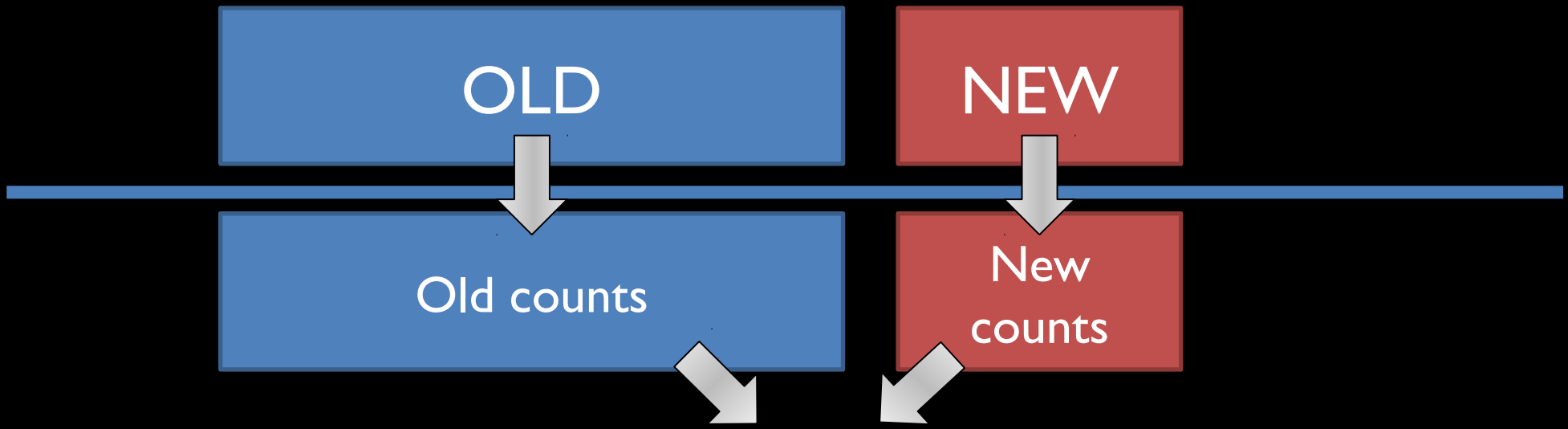
Baseline:

$$\alpha_1 \log p(f|e) + \alpha_2 \log p(e) + \ \ldots$$

New:

$$\alpha_{1OLD} \log p_{OLD}(f|e) + \alpha_{1NEW} \log p_{NEW}(f|e) + \alpha_2 \log p(e) + \cdots$$

Combine models (linear mixture)



Baseline:

$$p(f|e) = \frac{c(f,e)}{c(e)}$$

New – mix with λ picked on dev set:

$$p(f|e) = \lambda \frac{c_{old}(f,e)}{c_{old}(e)} + (1-\lambda) \frac{c_{old}(f,e)}{c_{old}(e)}$$

BLEU results

	OLD	NEW	OLD+ NEW	Use both models	Combine models
News	23.8	21.7	22.0	16.4	21.4
EMEA	28.7	34.8	34.8	32.9	36.6
Science	26.1	32.3	27.5	30.9	32.2
Subtitles	15.1	20.6	20.5	18.4	18.5

Next steps

- These mixtures are simple but coarse
- More fine-grained approaches:
  - Data selection: pick OLD data most like NEW
  - Data reweighting: use fractional counts on OLD data; greater weight to sentence pairs more like NEW
  - Can reweight at the word or phrase level rather than sentence pair [Foster et al., 2010]
- Similar in spirit to statistical domain adaptation
  - but existing machine learning algorithms can't be applied
  - because SMT is not a classification task

Phrase Sense Disambiguation (PSD)

Proposed solution: Phrase Sense Disambiguation

[Carpuat & Wu 2007]

- Incorporate context in lexical choice
  - Yields P(e|f, context) features for phrase pairs
  - Unlike usual P(e|f) relative frequencies
- Turns phrase translation into discriminative classification
  - Just like standard machine learning tasks

[Chan et al. 2007, Stroppa et al. 2007, Gimenez & Màrquez 2008, Jeong et al. 2010, Patry & Langlais 2011, ...]

Why PSD for domain adaptation?

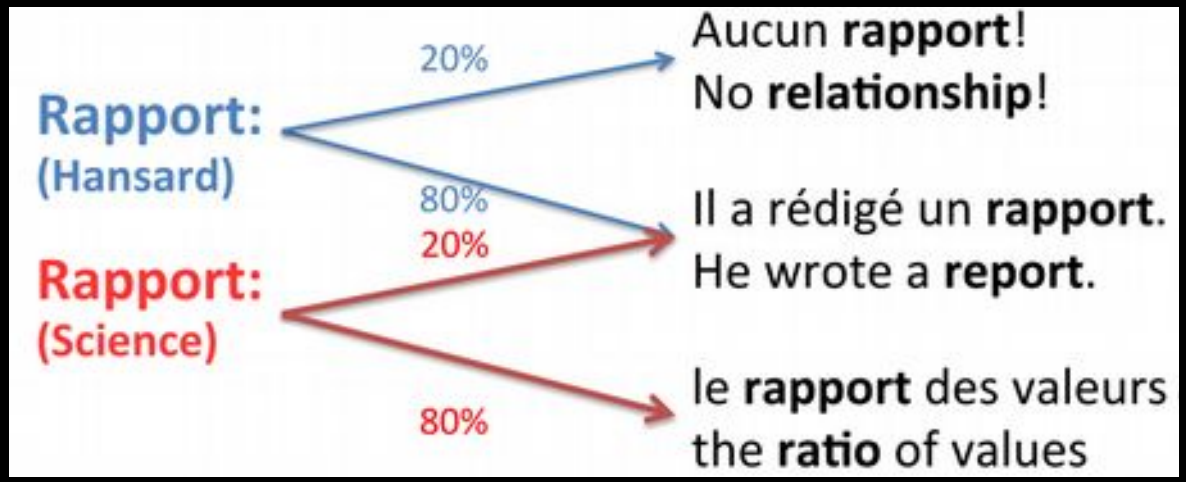
Highest P(e|f) in Science!

New sense in medical domain!

Occurs in new domains but not as often as in Hansard!

Source context can prevent translation errors when shifting domain

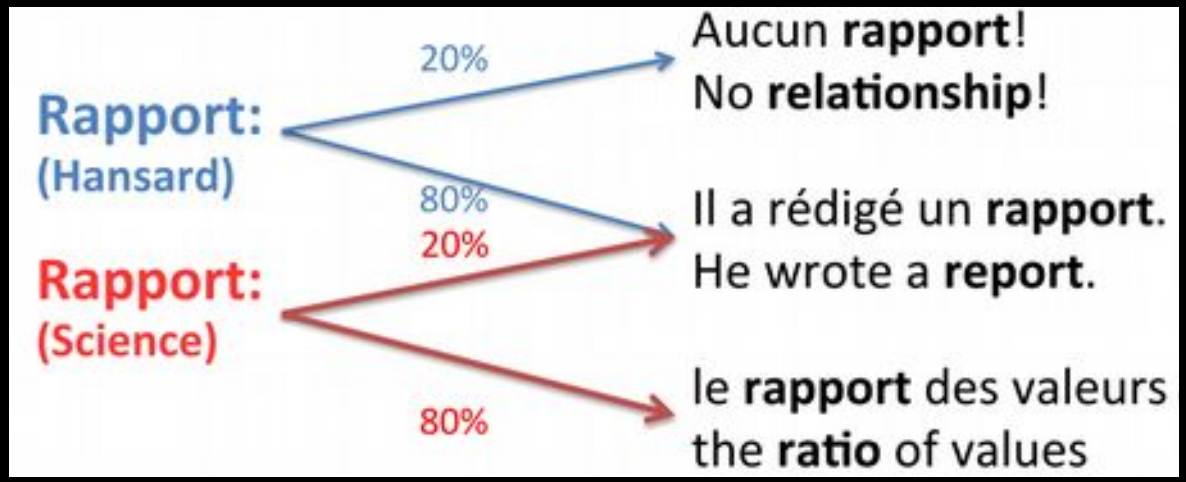
Phrase Sense Disambiguation



- PSD = phrase translation as classification
- PSD at test time
  - use context to predict correct English translation of French phrase
  - local lexical and POS context , global sentence and document context
- PSD at train time
  - extract French phrases with English translations from word alignment
  - throw into off-the-shelf classifier + adaptation techniques

[Blitzer & Daumé 2010]

Domain adaptation in PSD



- Train a classifier over OLD and NEW data
-

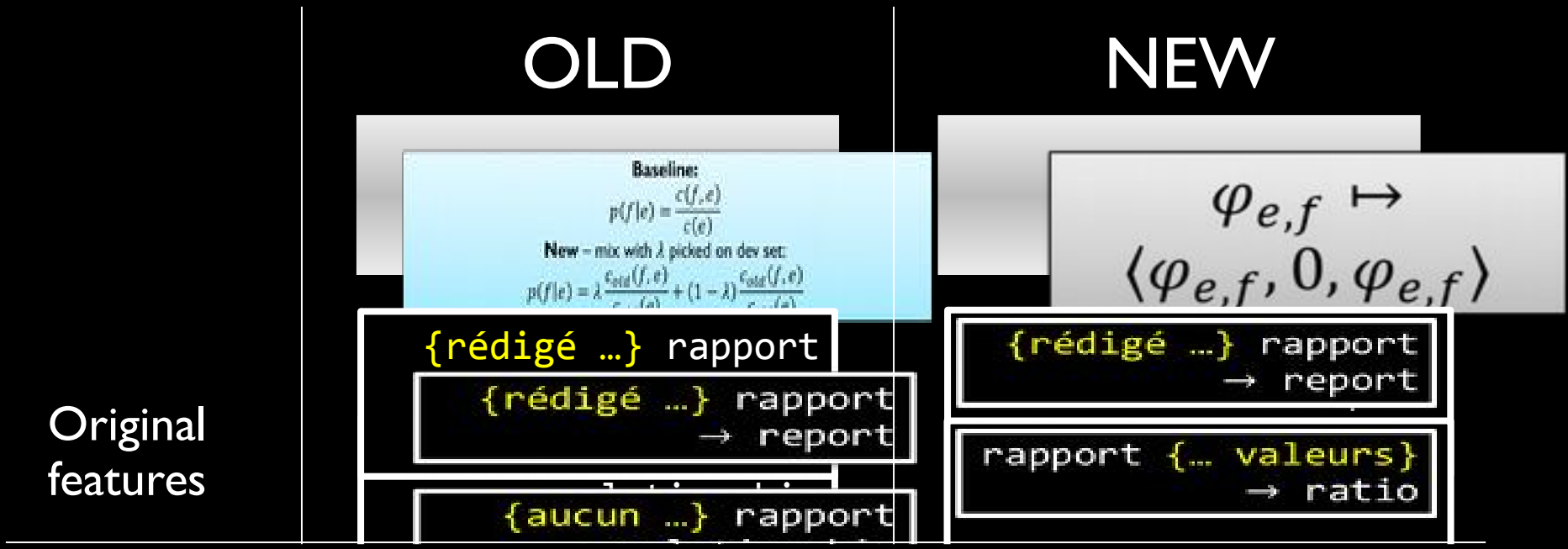
Baseline:

$$\alpha_1 \log p(f|e) + \alpha_2 \log p(e) + \ ...$$

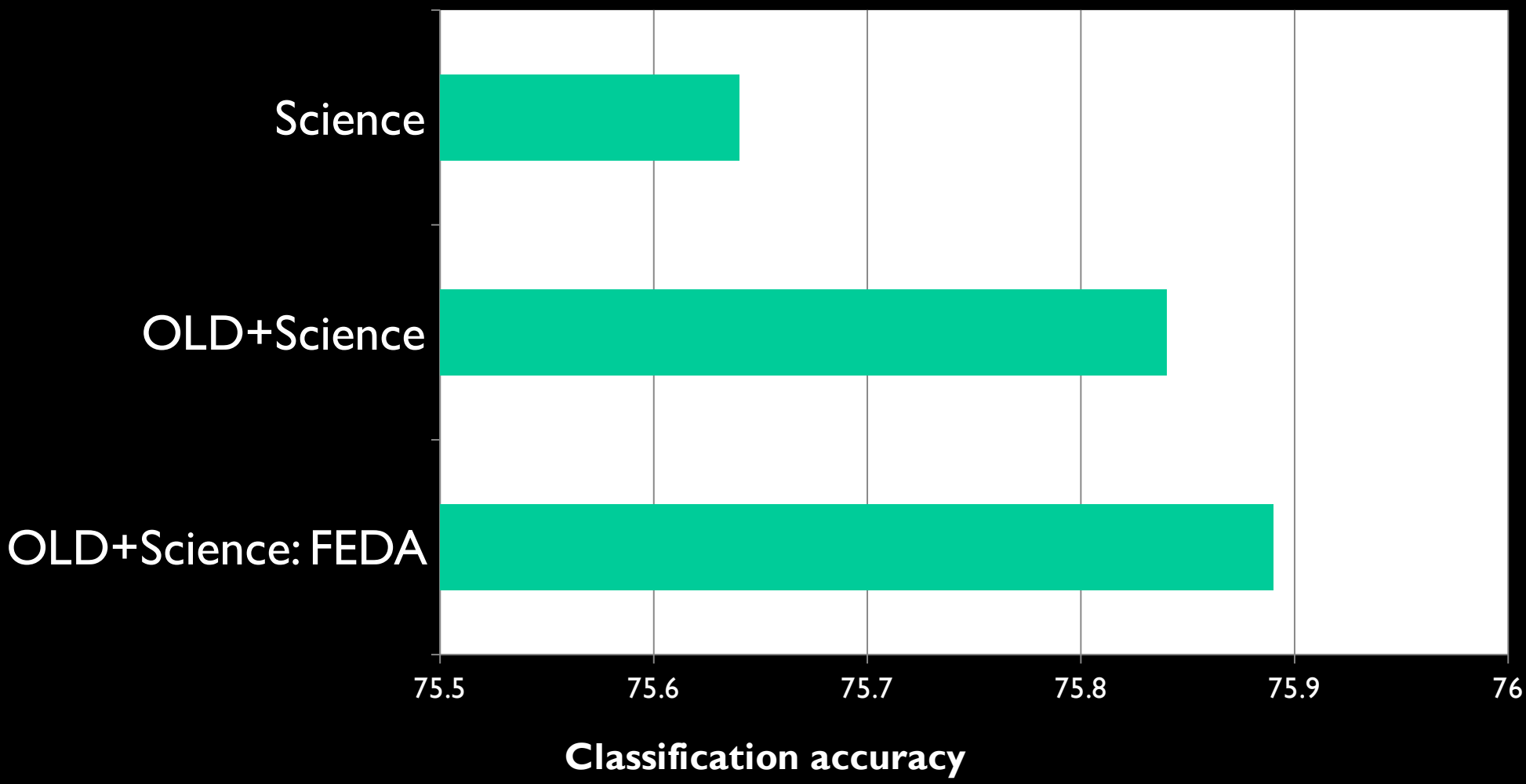
New:

$$\alpha_{1OLD} \log p_{OLD}(f|e) + \alpha_{1NEW} \log p_{NEW}(f|e) + \alpha_2 \log p(e) +$$

Feature augmentation



Domain adaptation results: Science



PSD in Moses: VW-Moses integration

- First general purpose classifier in Moses
- Tight integration
  - Can be built and run out-of-the-box, extended with new features, etc
  - Fast!
    - 180% run time of standard Moses, fully parallelized in training (multiple processes) and decoding (multithreading)

Other areas of investigation

PSD for Hierarchical phrase-based translation

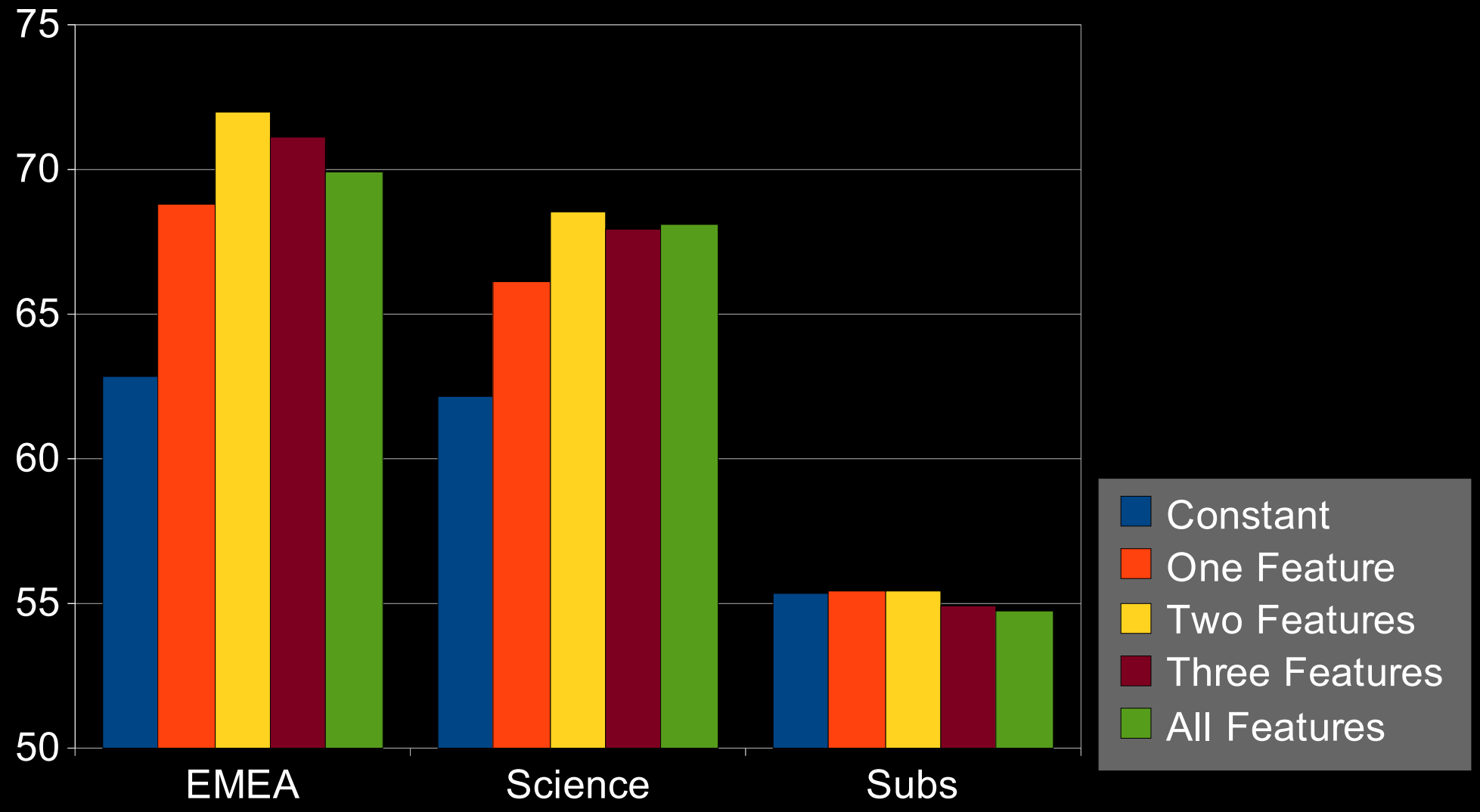
Discovering latent topics from parallel data

Spotting new senses: determining when a source word gains a new sense (needs a new translation)

Spotting New Senses

- Binary classification problem:
  - +ve: French token has previously unseen sense
  - -ve: French token is used in a known way
- Gold standard as byproduct of S4 analysis
- Many features considered
  - Frequency of words/translations in each domain
  - Language model perplexities across domains
  - Topic model "mismatches"
  - Marginal matching features
  - Translation "flow" impedence

Experimental Results



Selected features:

```
EMEA:    ppl   || matchm flow   || matchm topics flow
Science: ppl   || matchm ppl    || matchm topics ppl
Subs:    topcs || matchm topics || matchm topics flow
```

Discussion

- Introduced taxonomy and measurement tools for adaptation effects in MT
- "Score" errors – target of prior work – only a part of what goes wrong
- Marginal matching introduced as a model for addressing *all* S4 issues simultaneously: +2.4 BLEU
- Data and outputs released for you to use (both in MT and as a stand-alone lexical selection task)
- Feature-rich approaches integrated into Moses via VW library, applied to adaptation
- Range of other problems to work on: identifying new senses, cross-domain topic models, etc.)

Marine Carpuat
(NRC-CNRC)

Alex Fraser
(U. Stuttgart)

Thanks! Questions?

Hal Daumé III, me@hal3.name

Domain Shift Setting

Old domain: Hansard parliamentary proceedings

New Domain Datasets

Two methods for measuring adaptation effects

Macro Analysis: corpus-level analysis using BLEU

TETRA: table enhancement for translation analysis

Micro Analysis: word-level analysis using word alignments

WADE: word alignment driven evaluation

[Irvine et al. TACL 2013]

SenseSpotting

Why? MT performance across domains degrades due to lexical choice errors

What? New task to identify word occurrences (tokens) that gain a new sense in new domains

How? Automatic annotation from parallel text + supervised learning

SenseSpotting Task Definition

Old domain translation lexicon

```
rapport ||| report ||| 0.8
rapport ||| connection ||| 0.1
rapport ||| study ||| 0.05
rapport ||| relationship ||| 0.05
```

New domain sentences

ces données sont basées sur le **rapport** d' étude clinique

le **rapport** cholestérol total / hdlc est resté stable

Key aspects of SenseSpotting

Sense inventory is defined by the MT lexicon

[Chan et al. 2007, Carpuat & Wu, 2007, inter alia]

New Senses are detected at the token-level

SenseSpotting is related to...

novel sense detection, but SenseSpotting…

- operates at the token-level
- is specific to domain-adaptation

[Sagi et al. 2009, Cook & Stevenson, 2010, Gulordava & Baroni, 2011, Lau et al. 2012, inter alia]

most frequent sense shift detection, but SenseSpotting

- considers *all* previously seen senses

[McCarthy et al. 2004, 2007; Erk 2006; Chan & Ng, 2007]

SenseSpotting is related to…

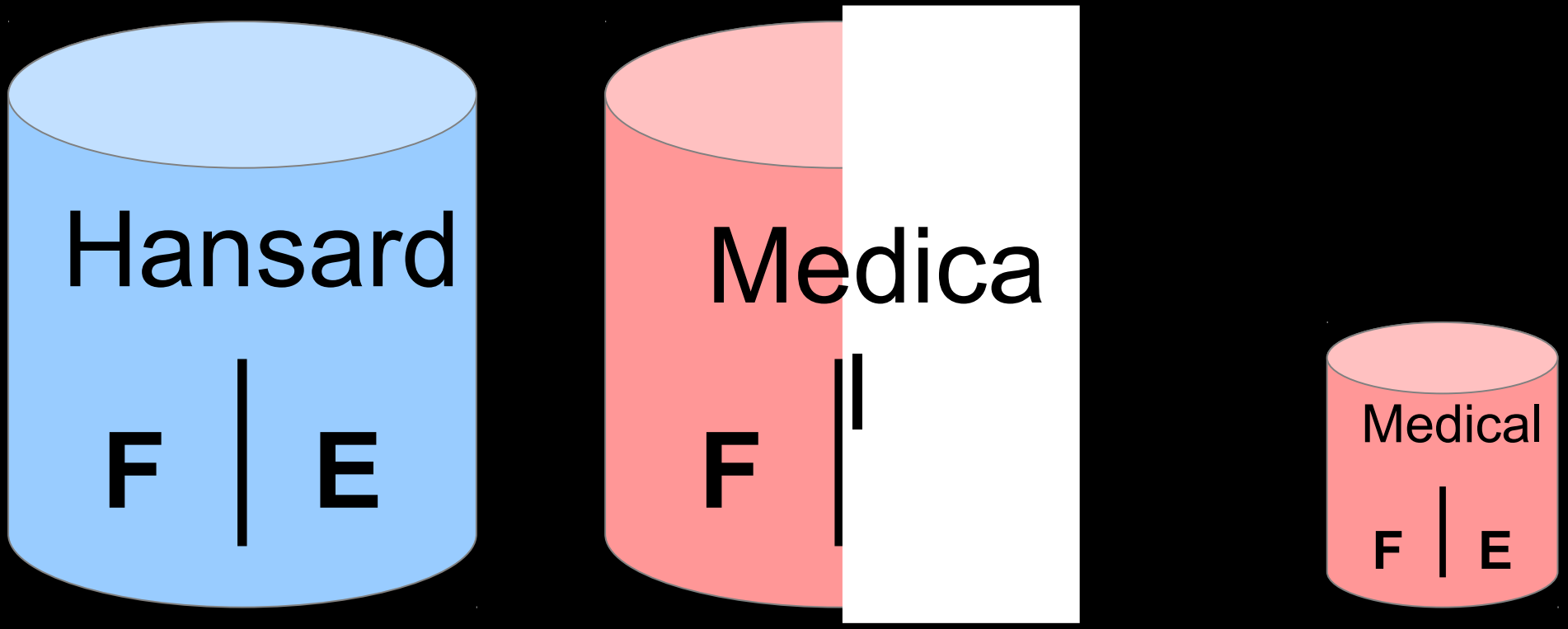
word sense disambiguation, but SenseSpotting expects sense inventory to grow…

[e.g., Sinha et al. 2010, Lefever & Hoste 2010]

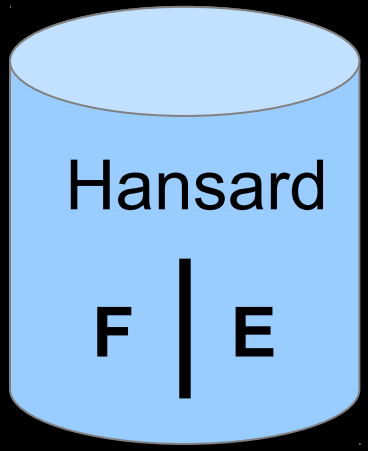
… similar to **word sense induction**, but we have old senses defined

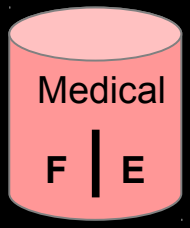
[e.g., Agirre and Soroa 2007]

Data Requirements

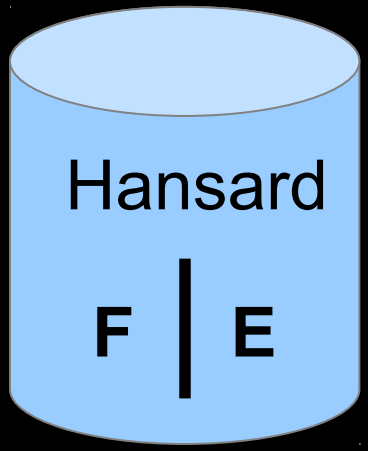


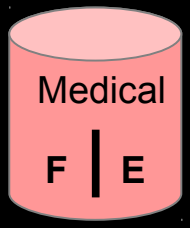
Generating Annotated Data

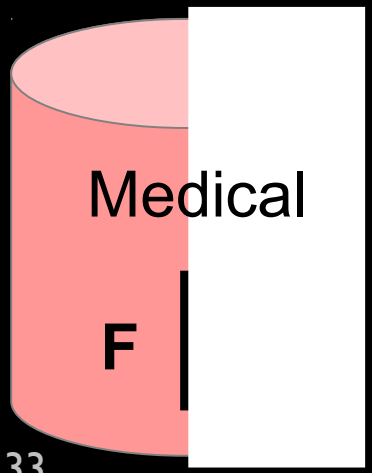




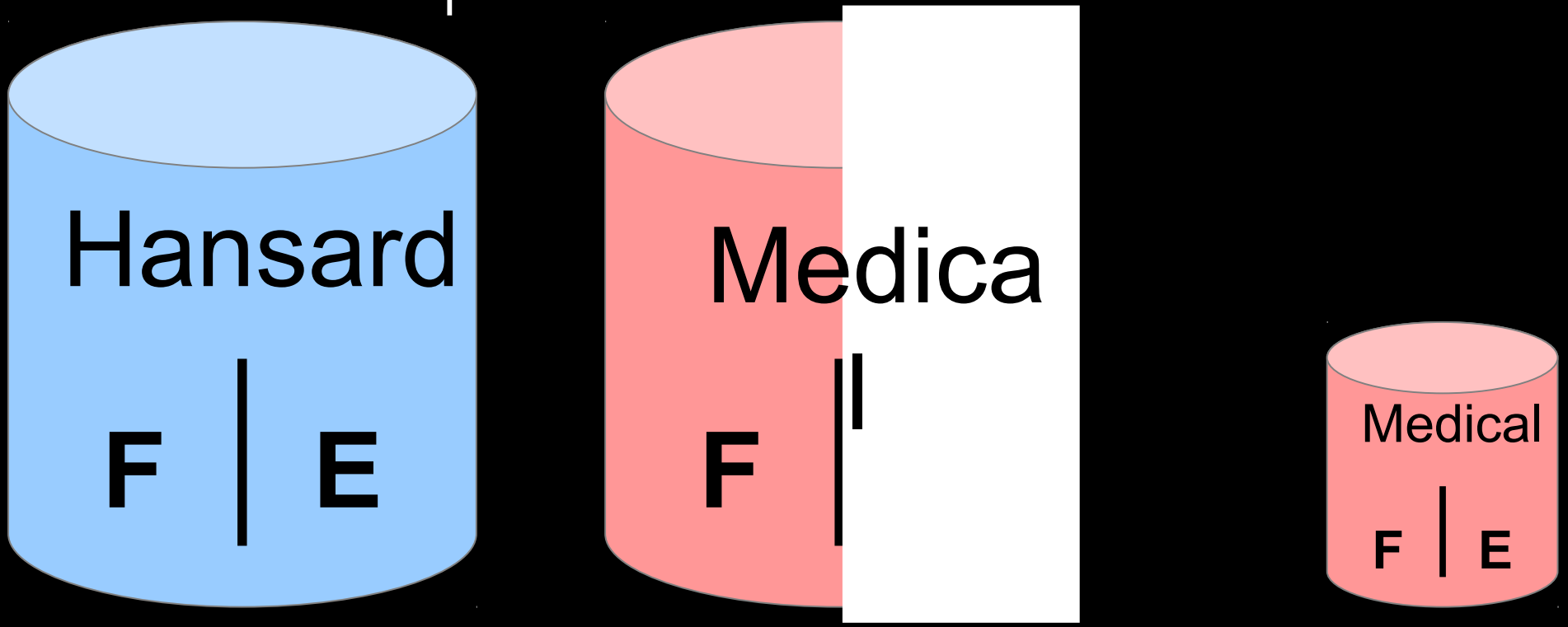
Generating Annotated Data



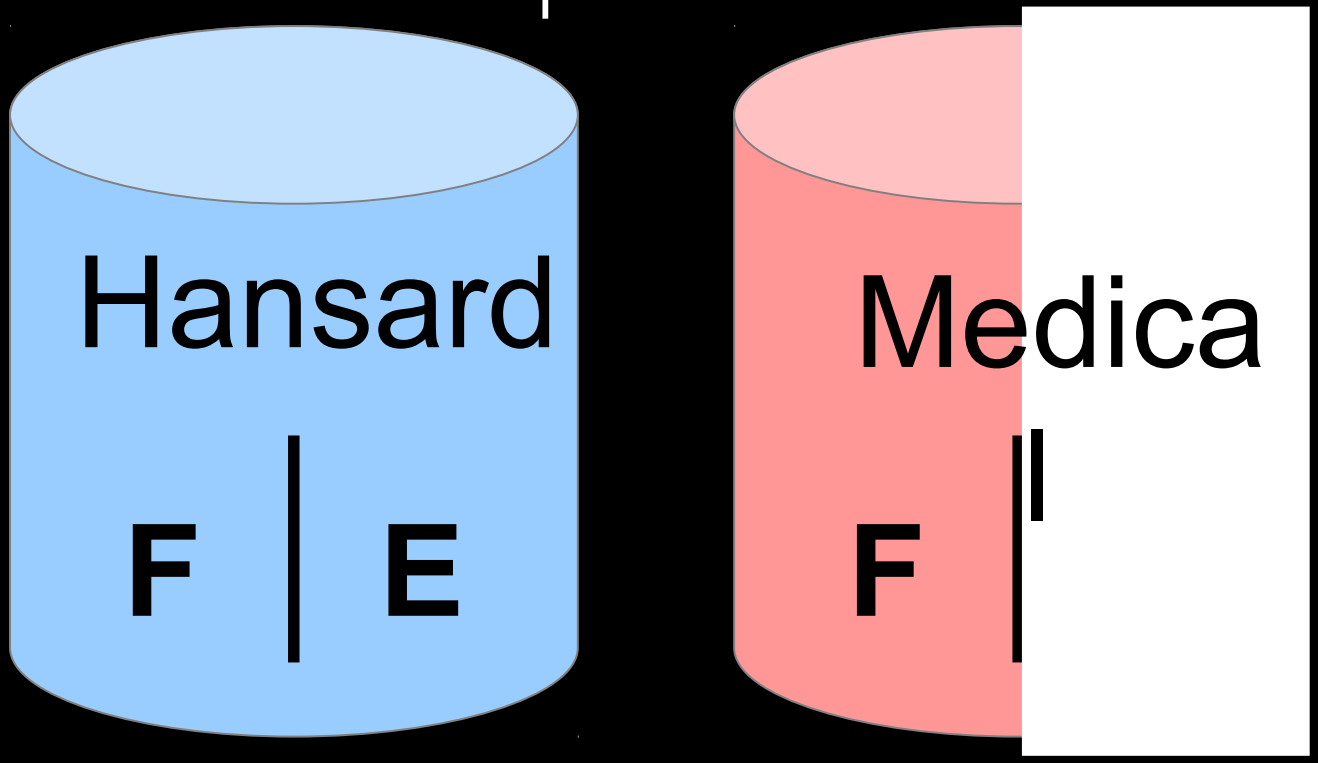


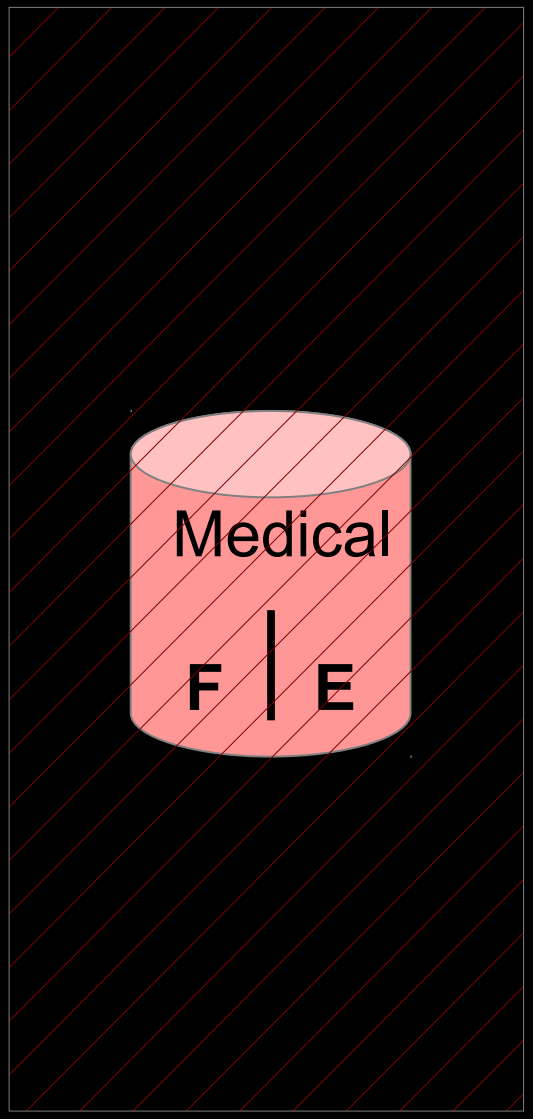


Data Requirements



Data Requirements





we can remove this requirement using a surrogate new domain

Classification set-up

Logistic regression model trained with vw

- L1 or L2 regularized based on tuning data

16-fold cross validation at the type level

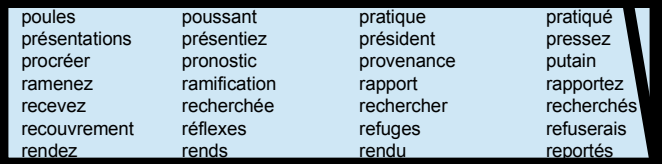
- Never test on type seen in training!
- E.g., train on "mode", "administration"; test on "rapport"

Evaluation metric: AUC

- area under the ROC curve

post vocabulary

ramenez	ramification	**rapport**	rapportez
recevez	recherchée	rechercher	recherchés
recouvrement	réflexes	refuges	refuserais
rendez	rends	rendu	reportés
rigidité	rigueur	rompre	rond
scalaire	sébastien	service	signalant
sorties	souches	souhaitée	soulèvement
stériles	subissant	substituant	suis



poules	poussant	pratique	pratiqué
présentations	présentiez	président	pressez
procréer	pronostic	provenance	putain
ramenez	ramification	rapport	rapportez
recevez	recherchée	rechercher	recherchés
recouvrement	réflexes	refuges	refuserais
rendez	rends	rendu	reportés

New Sense Indicators

New senses alter corpus-level word frequency

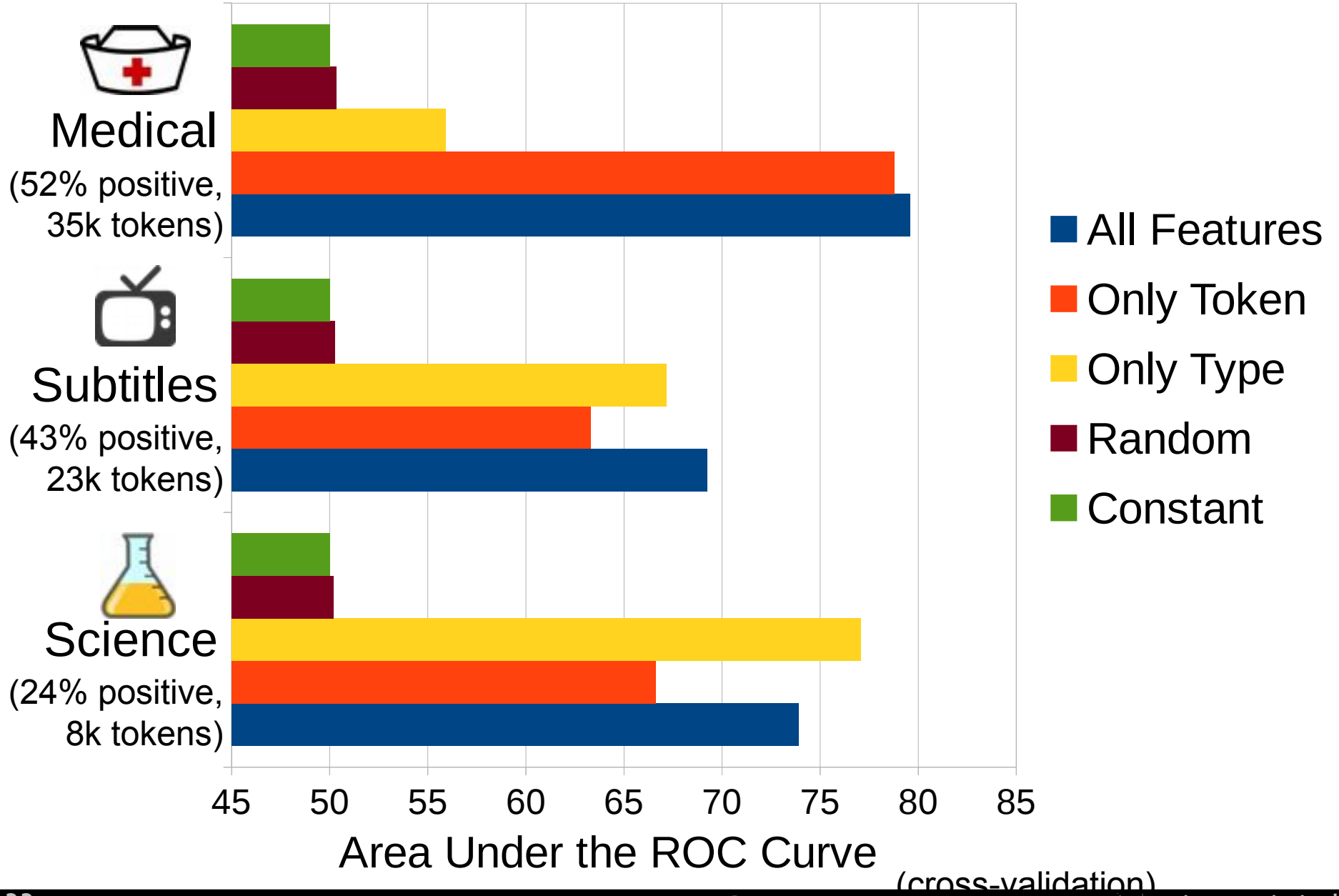
New senses alter document-level context

- topic distribution

New senses alter local context

- n-gram language model
- distributional similarity
- context-dependent translation model

SenseSpotting Results

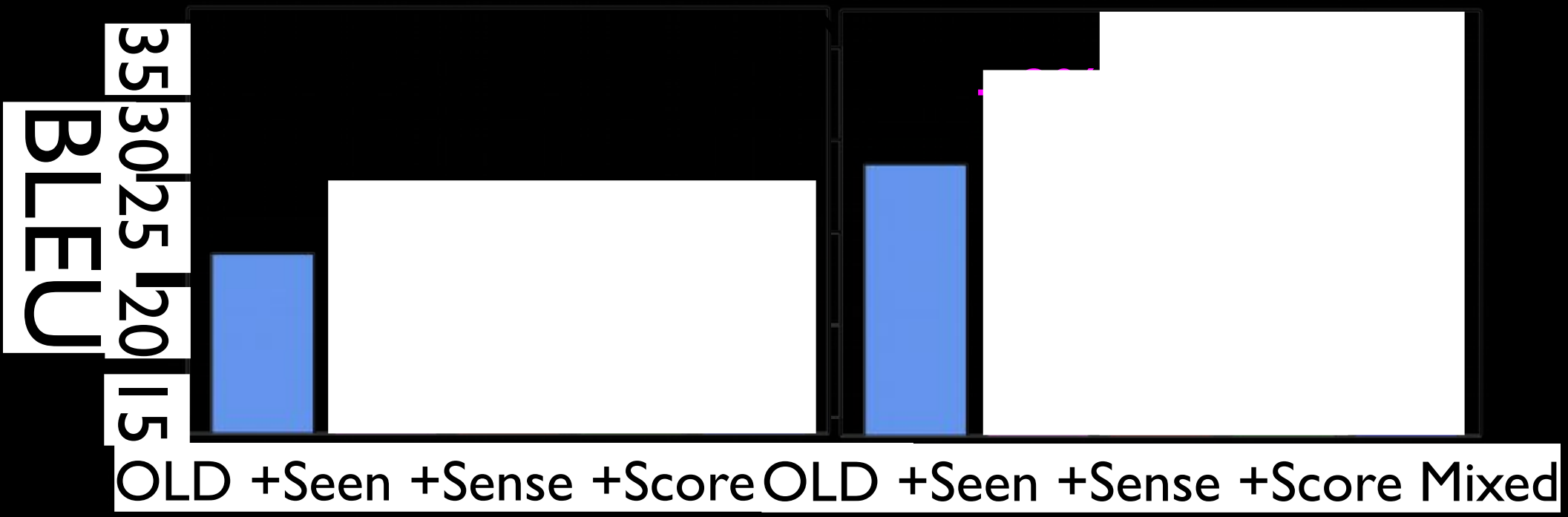




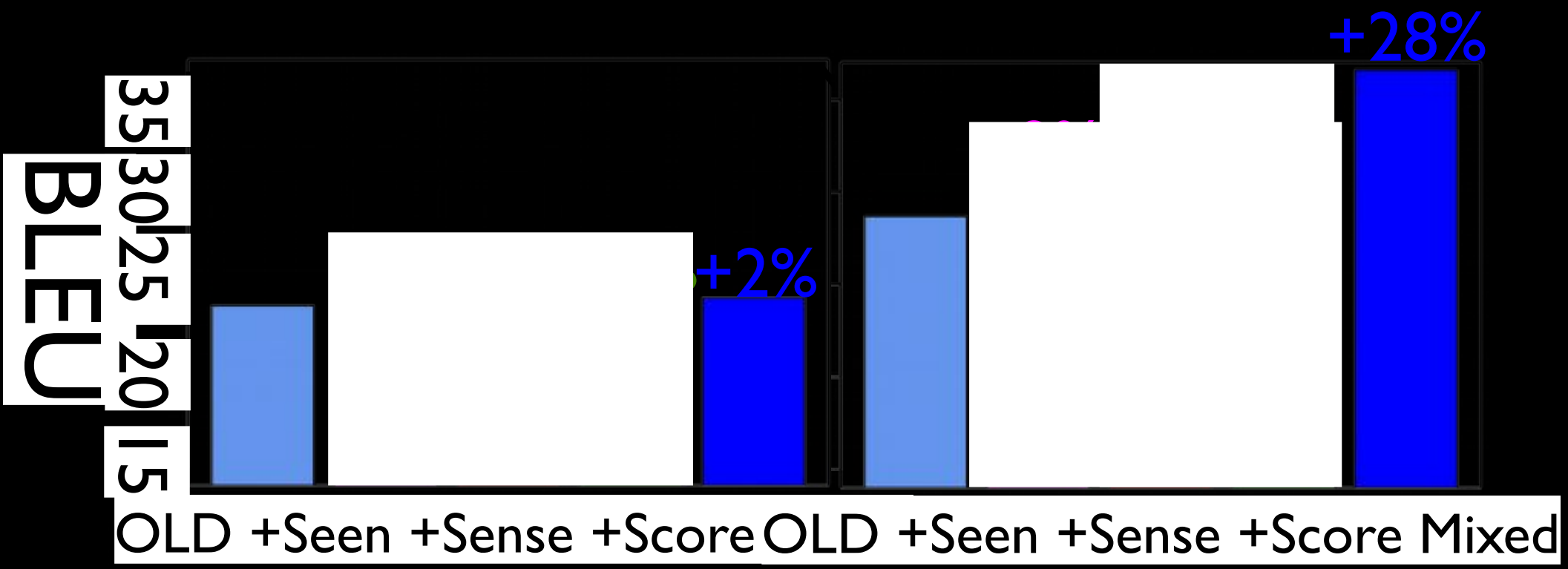


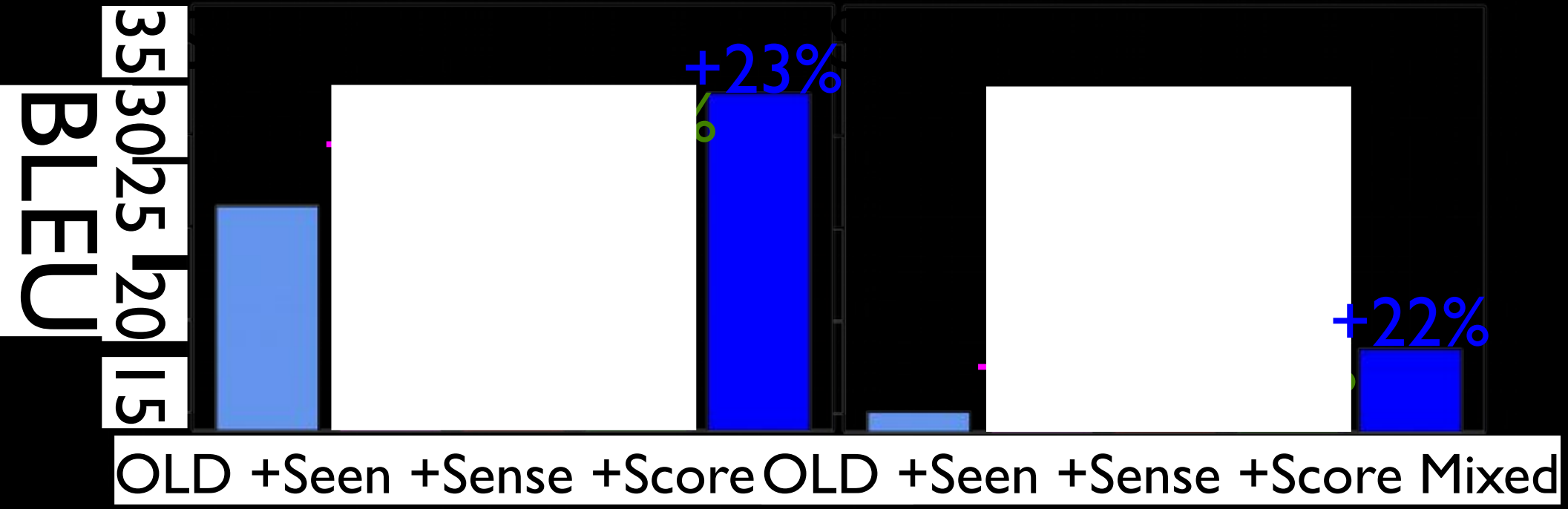
SenseSpotting summary

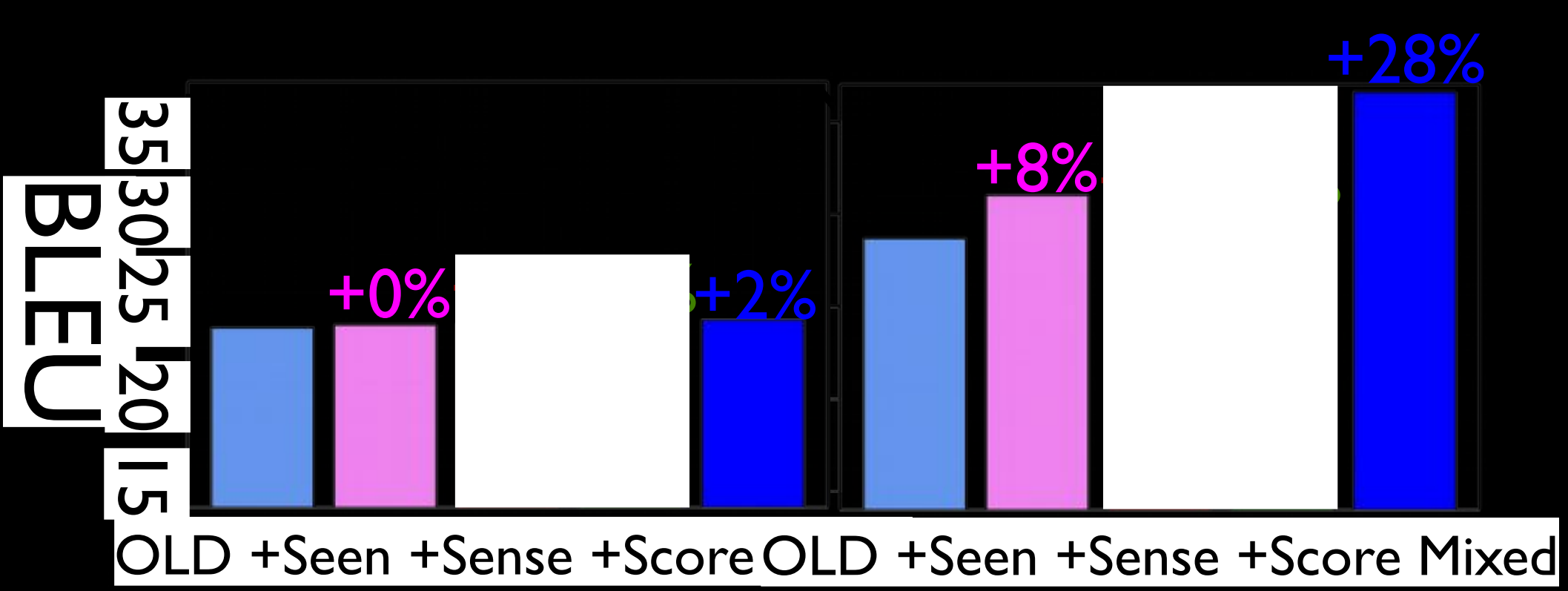
- new task motivated by cross-domain machine translation errors
- free **token**-level annotation from parallel text
  - minimal new domain parallel text required
- AUC as high as 80%
  - on word types **never seen** during training
  - requires both type and token level indicators

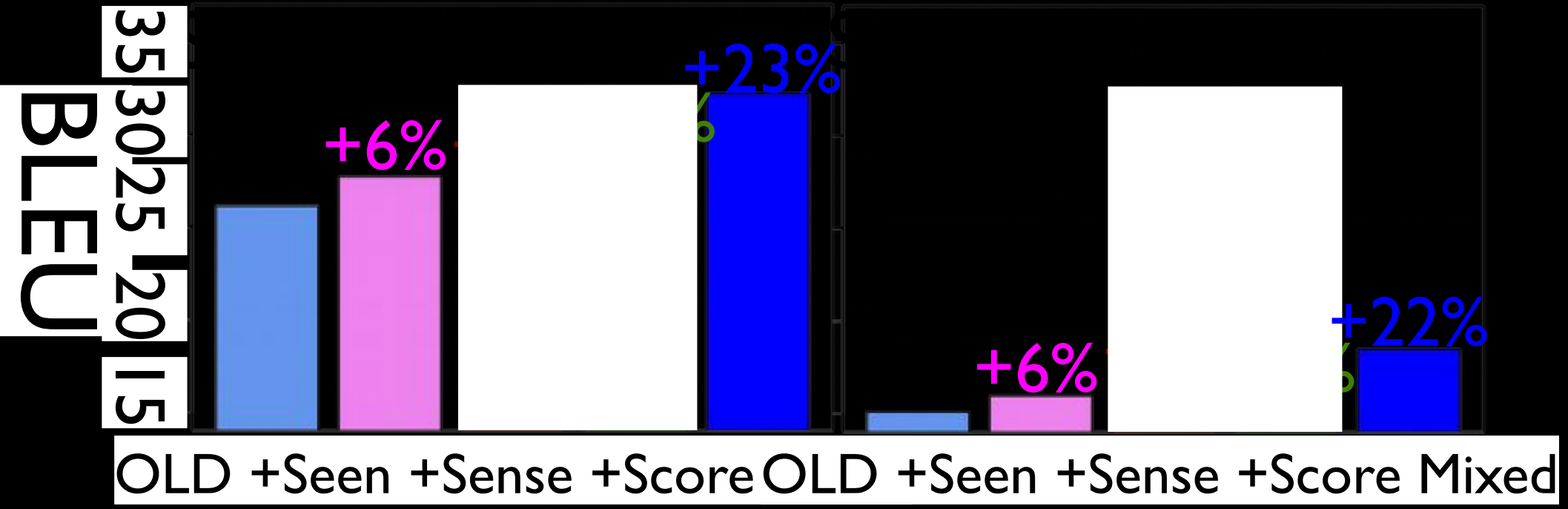

BLEU
35
30
25
20
15
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

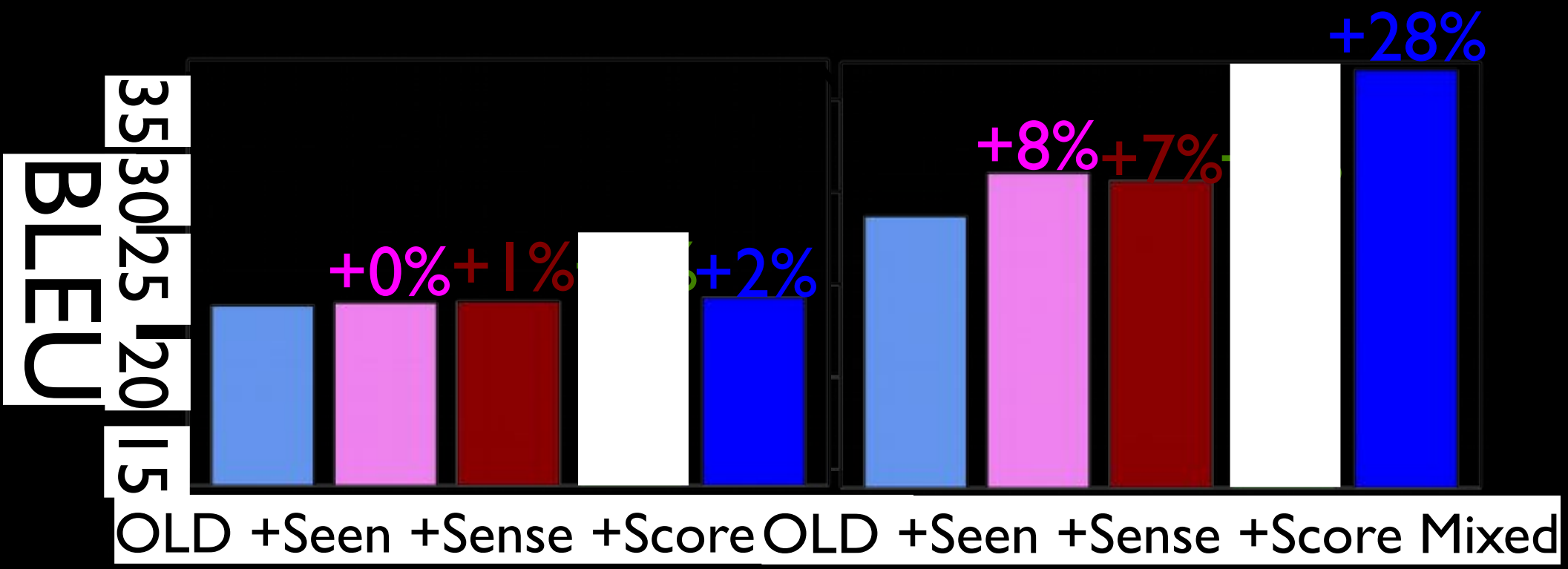

BLEU
35
30
25
20
15
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

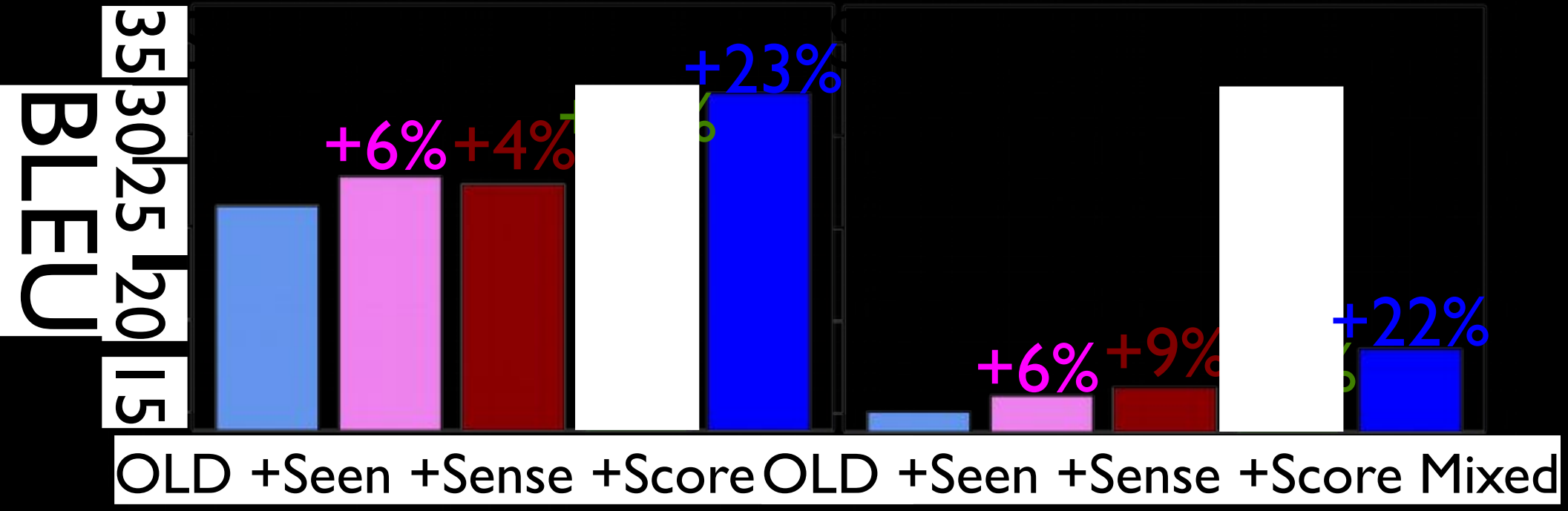

BLEU
35
30
25
20
15
+2%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

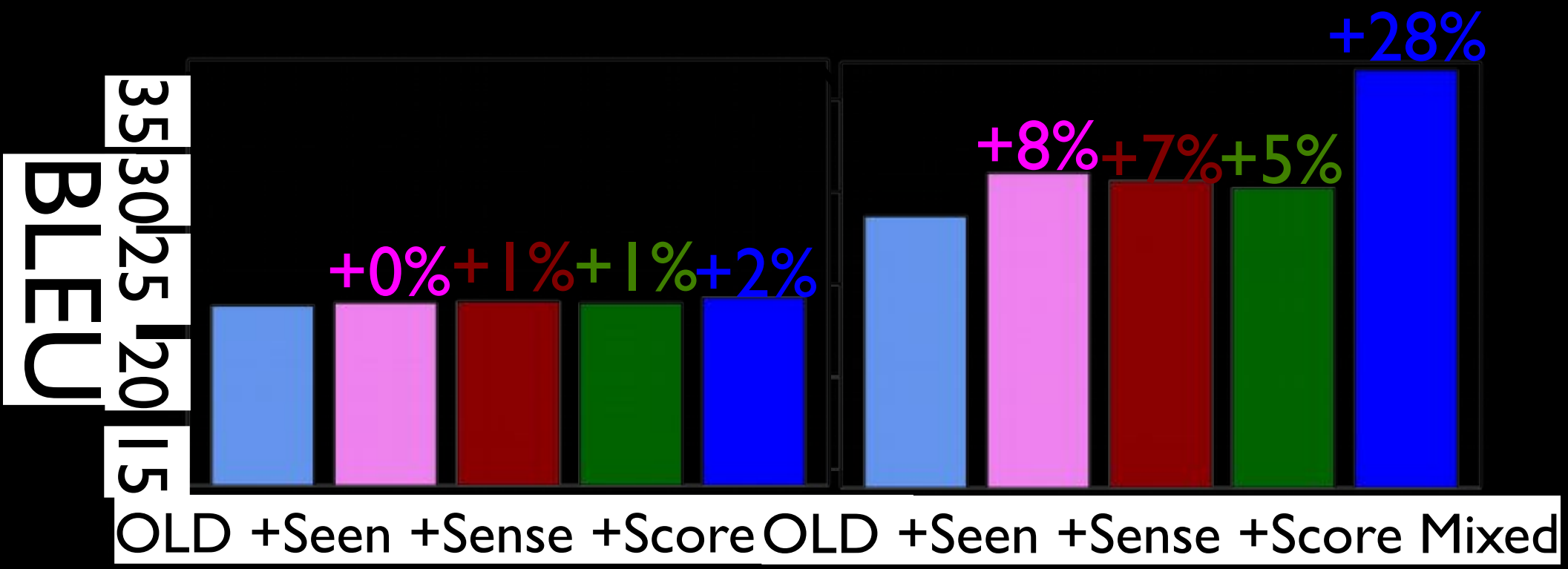

BLEU
35
30
25
20
15
+23%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

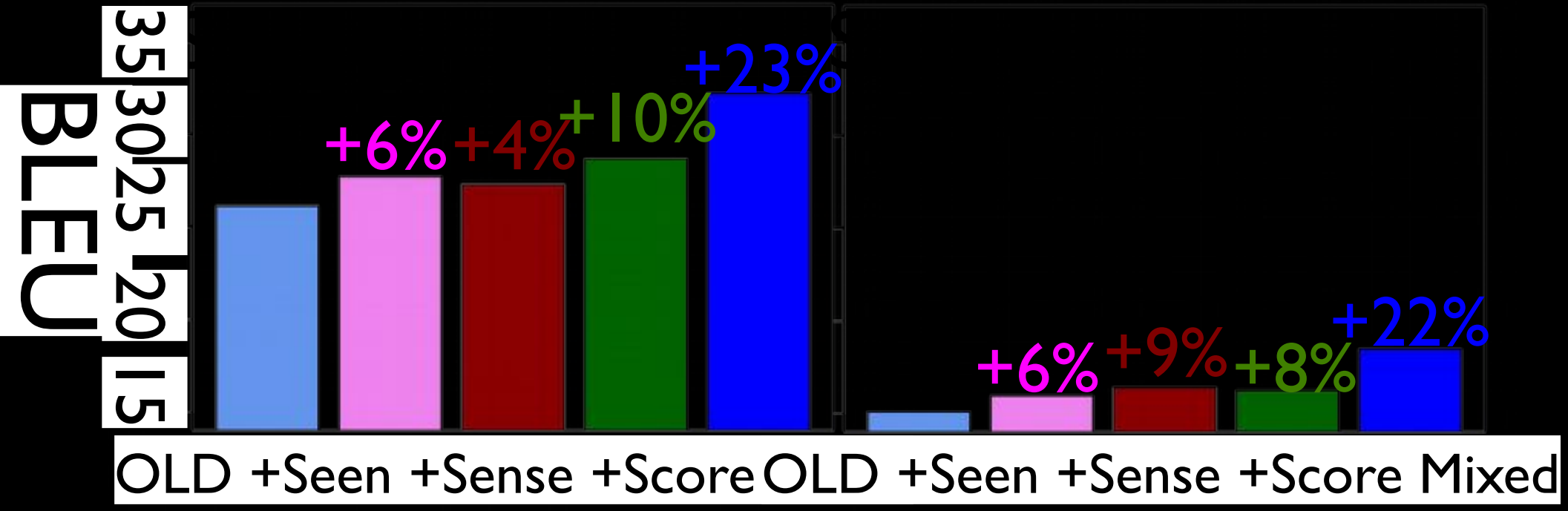

BLEU
35
30
25
20
15
+0%
+2%
+8%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

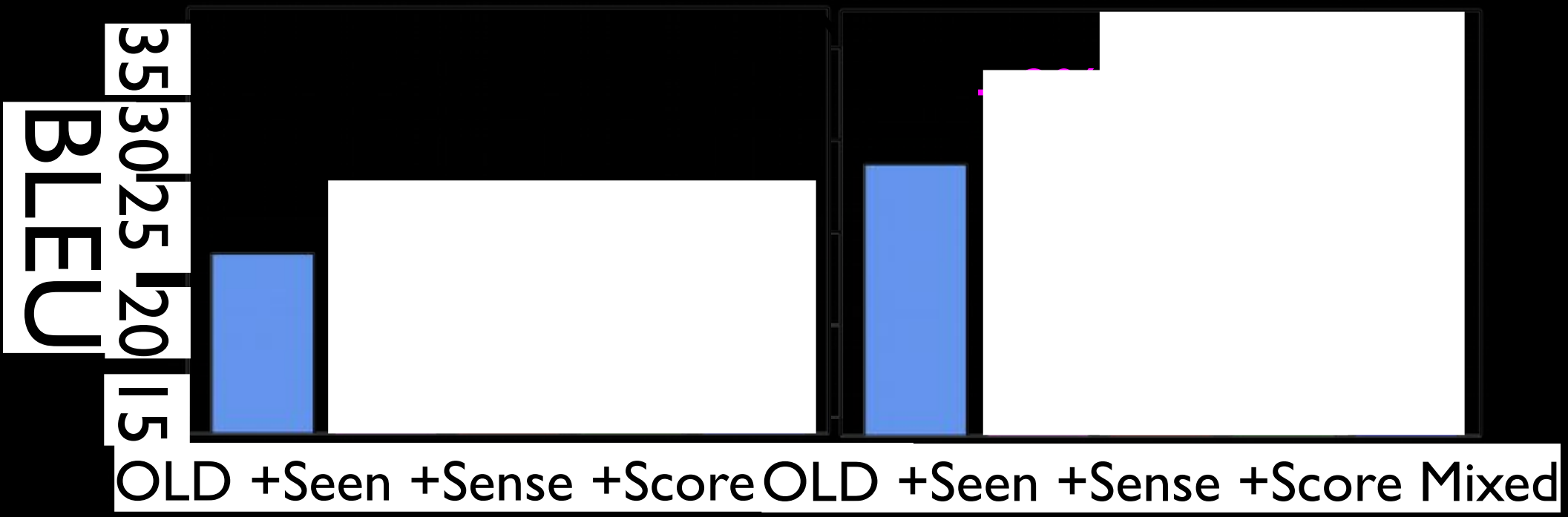

BLEU
35
30
25
20
15
+6%
+23%
+6%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed


BLEU
35
30
25
20
15
+0%
+1%
+2%
+8%
+7%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

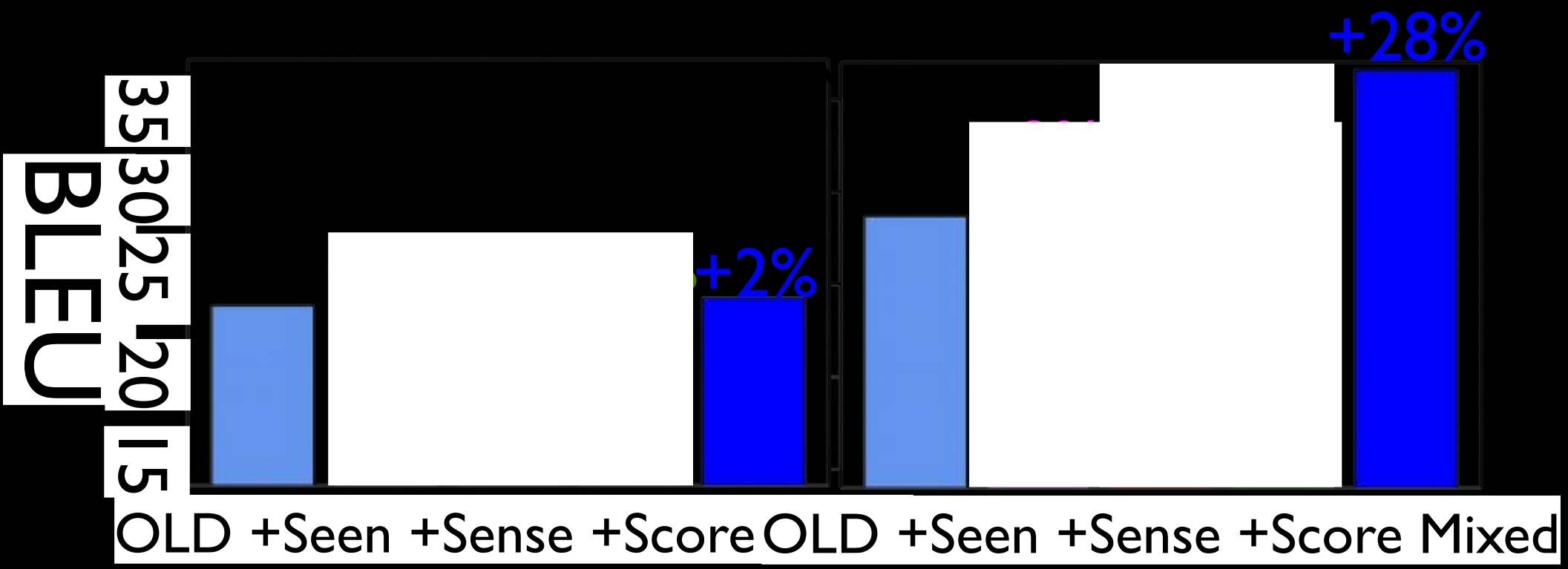

BLEU
35
30
25
20
15
+6%
+4%
+23%
+6%
+9%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

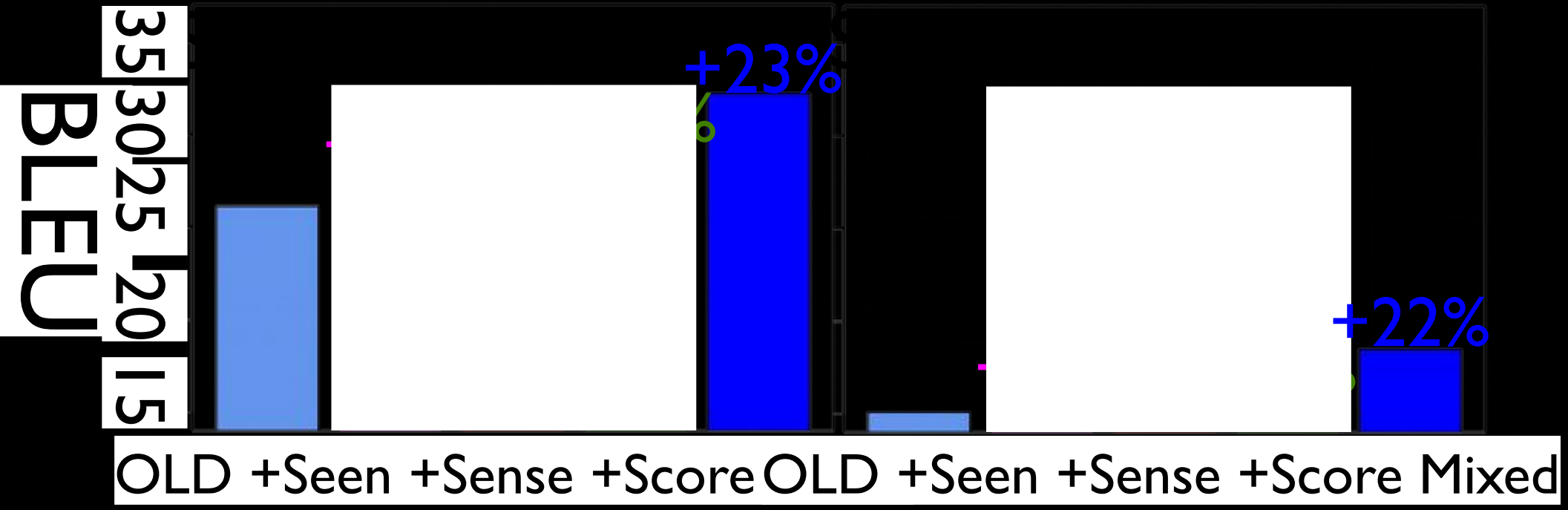

BLEU
35
30
25
20
15
+0%
+1%
+1%
+2%
+8%
+7%
+5%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

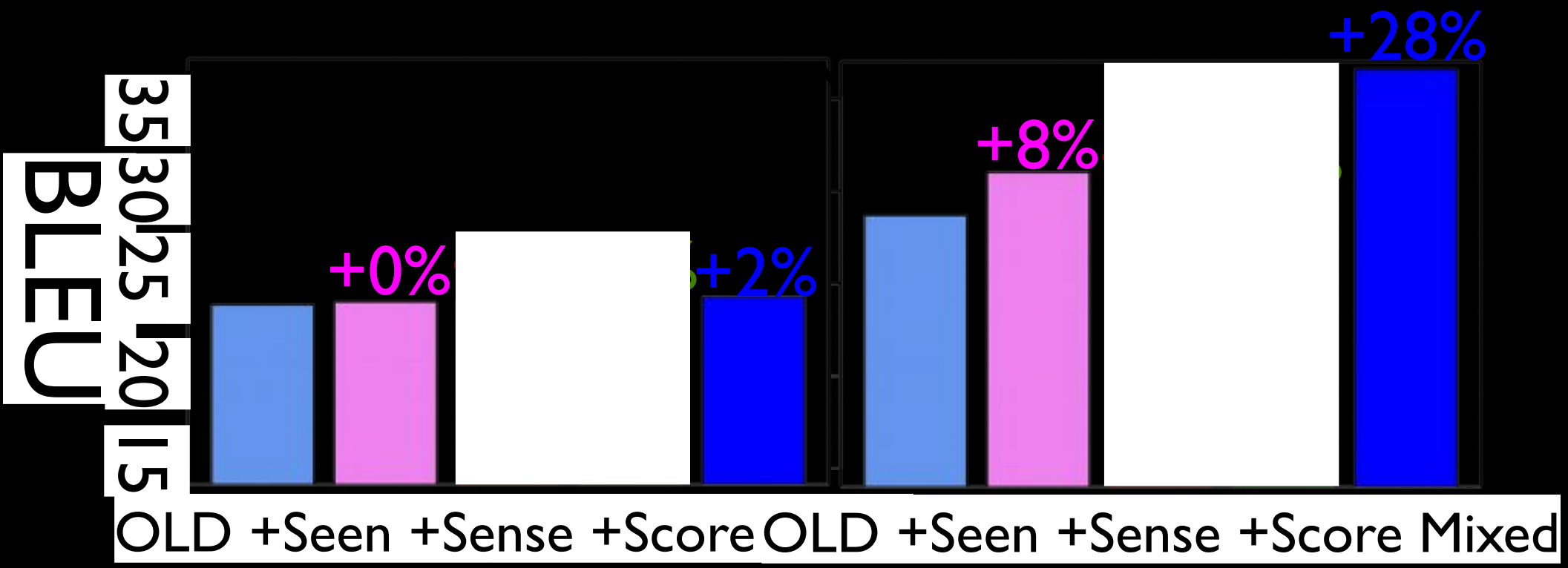

BLEU
35
30
25
20
15
+6%
+4%
+10%
+23%
+6%
+9%
+8%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

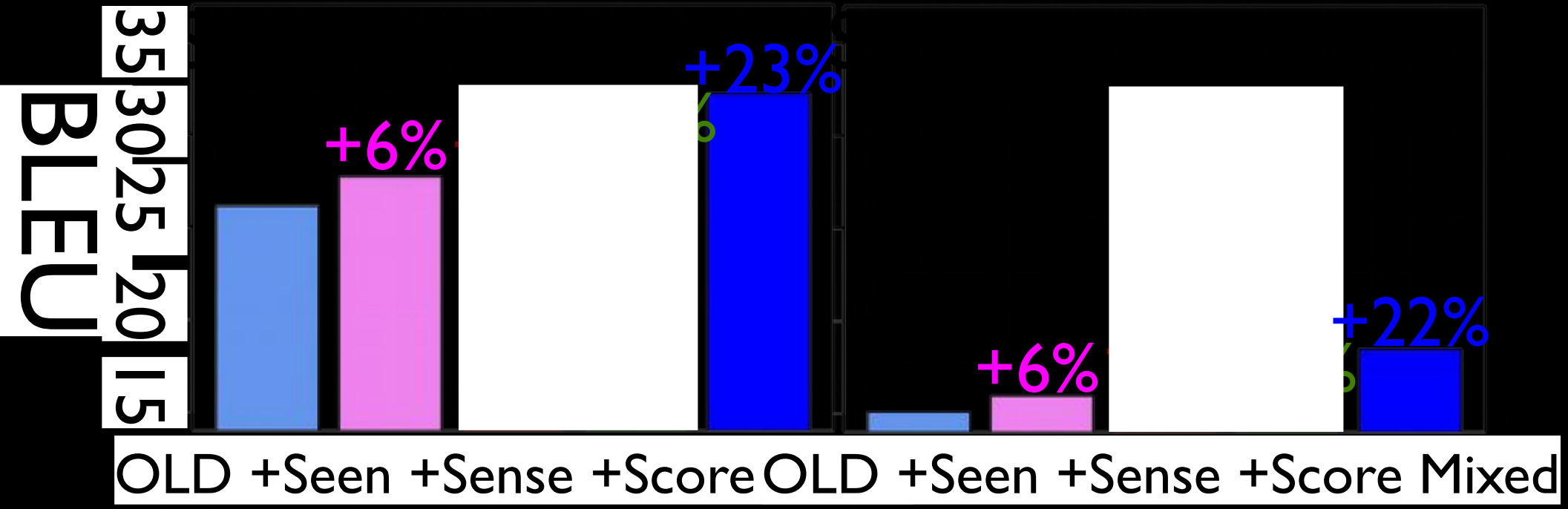

BLEU
35
30
25
20
15
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

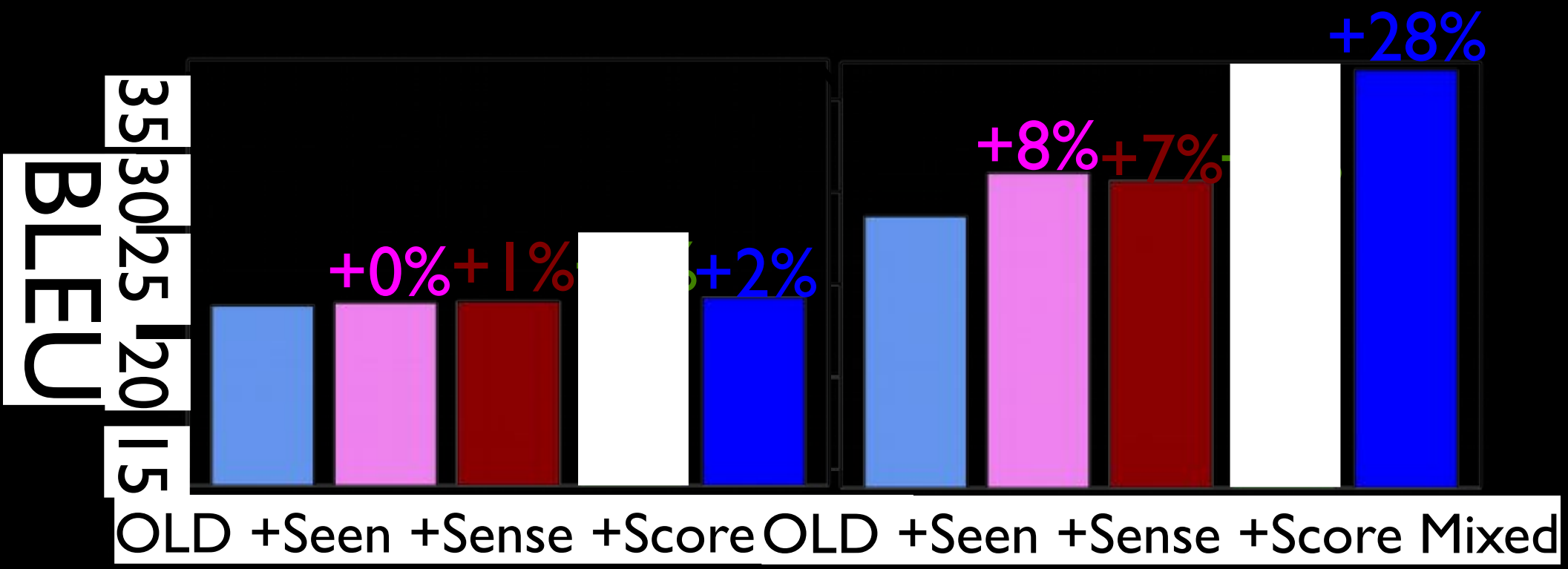

BLEU
35
30
25
20
15
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

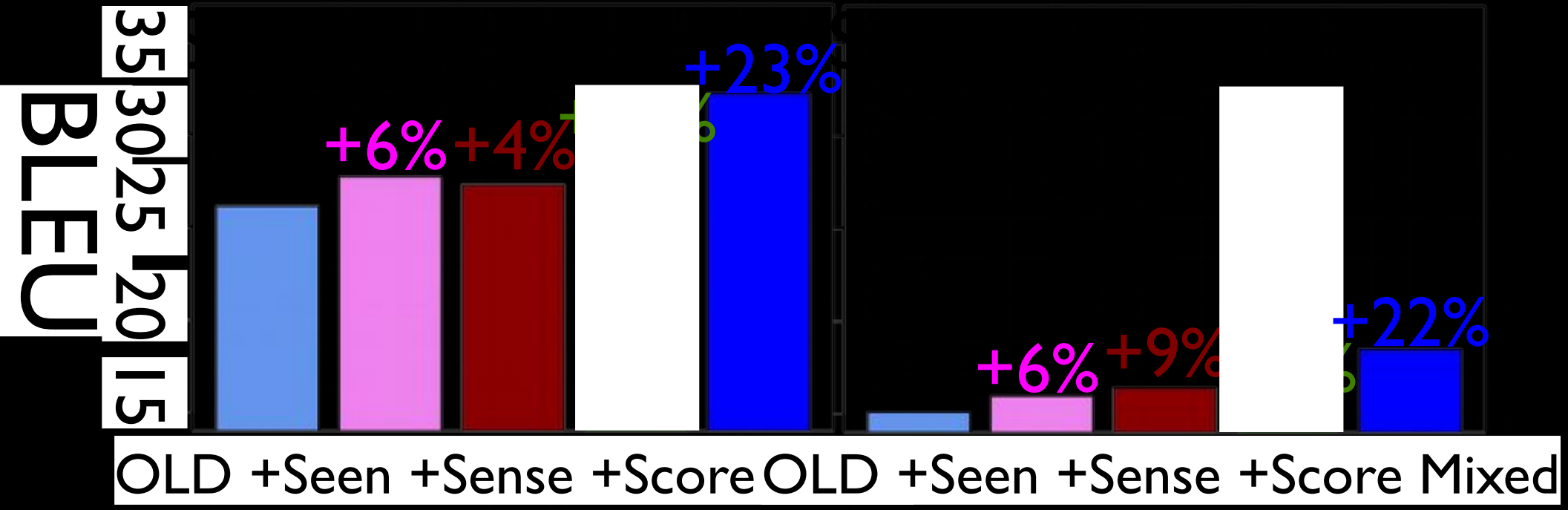

BLEU
35
30
25
20
15
+2%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

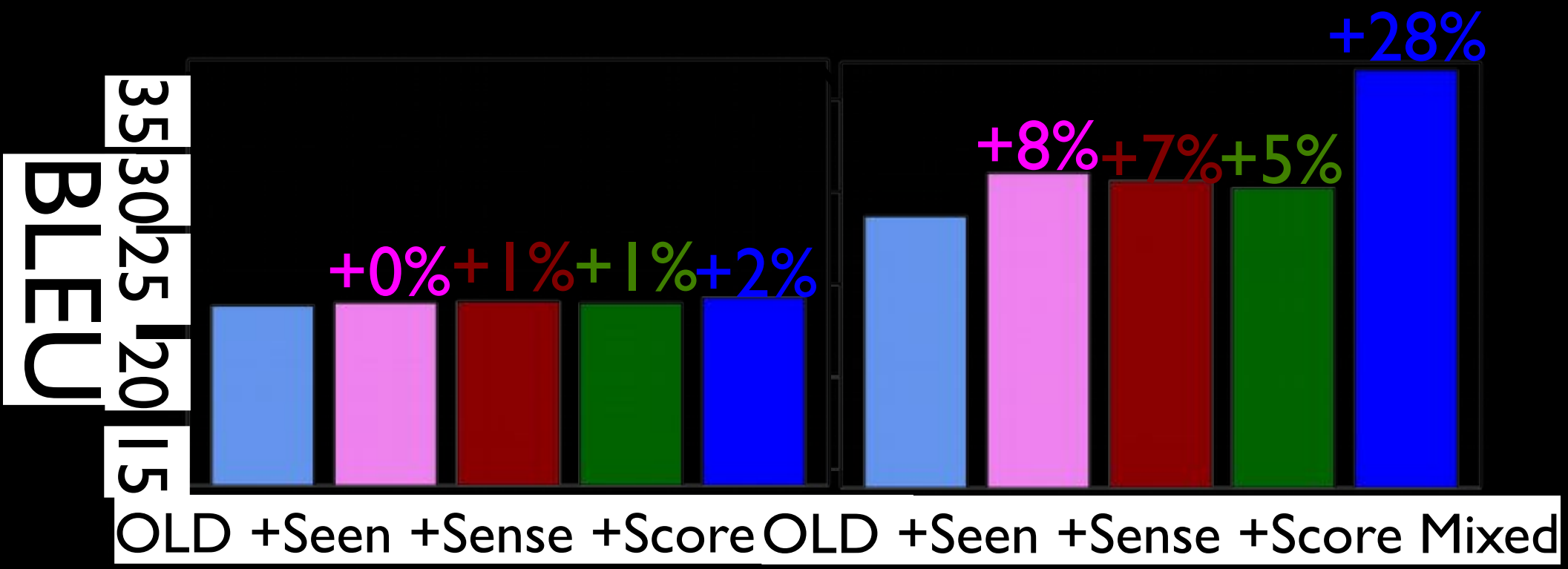

BLEU
35
30
25
20
15
+23%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

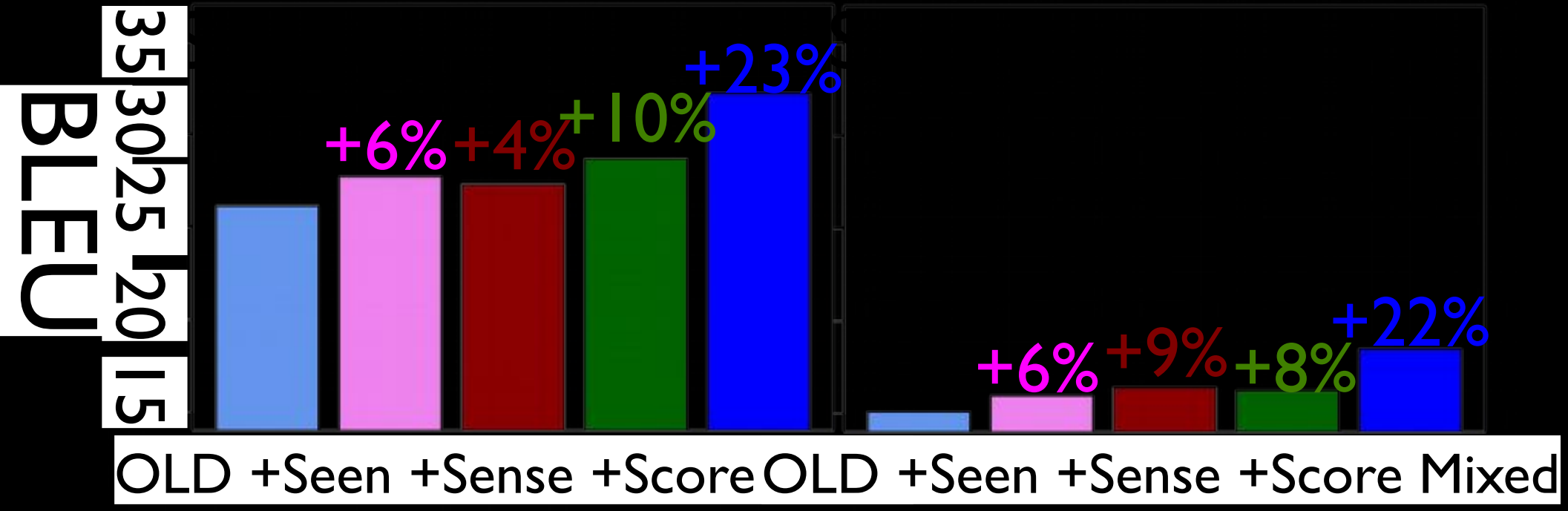

BLEU
35
30
25
20
15
+0%
+2%
+8%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed


BLEU
35
30
25
20
15
+6%
+23%
+6%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

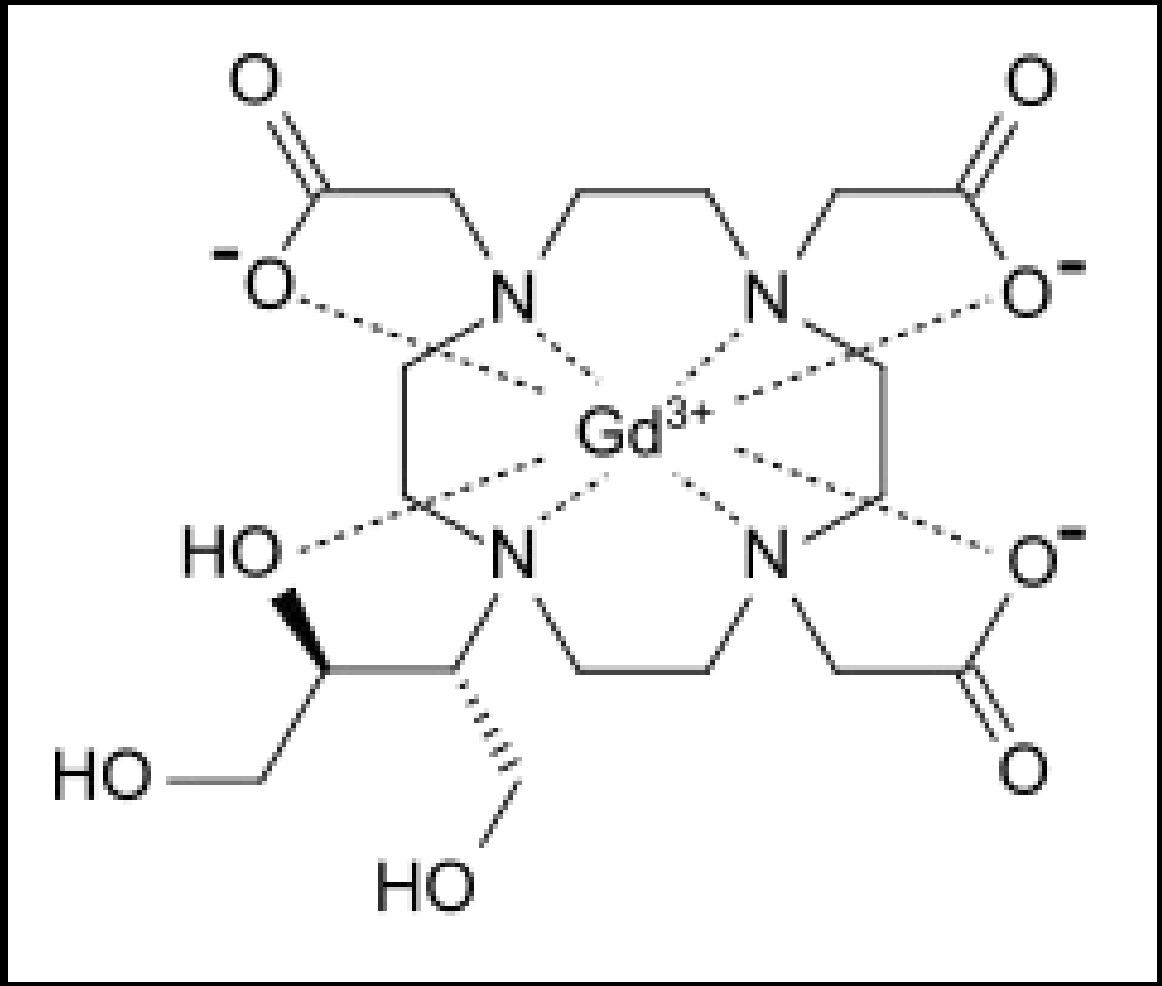

BLEU
35
30
25
20
15
+0%
+1%
+2%
+8%
+7%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed


BLEU
35
30
25
20
15
+6%
+4%
+23%
+6%
+9%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed


BLEU
35
30
25
20
15
+0%
+1%
+1%
+2%
+8%
+7%
+5%
+28%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

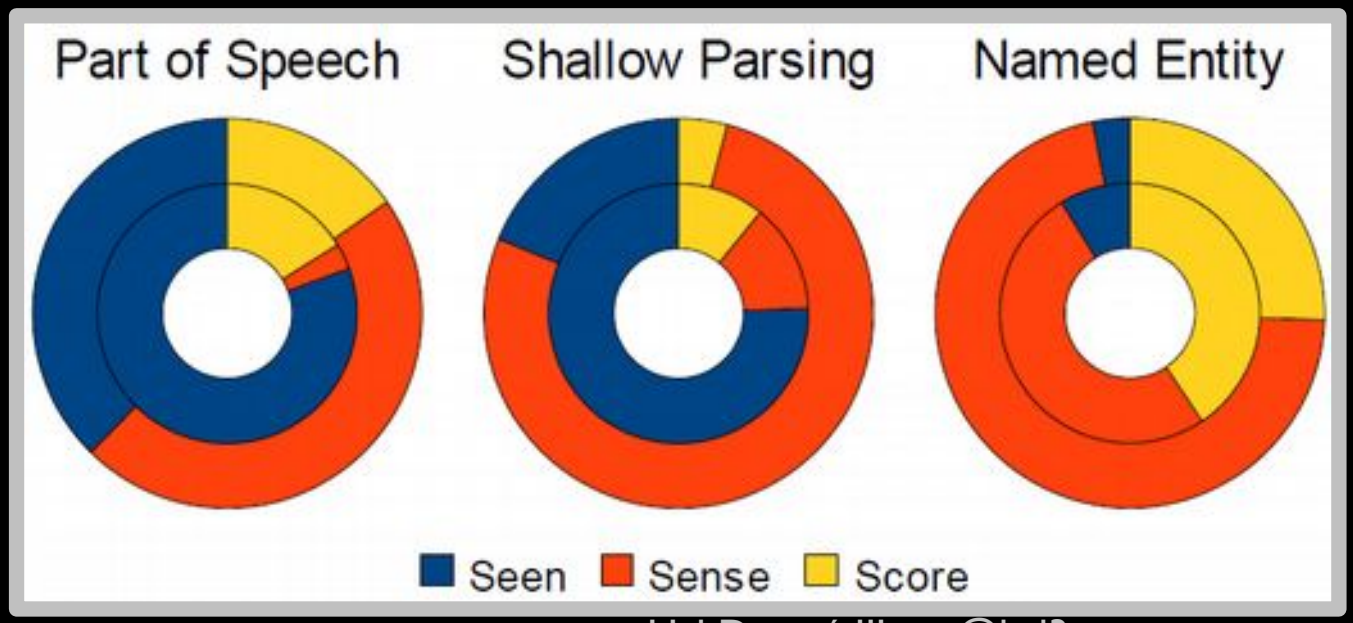

BLEU
35
30
25
20
15
+6%
+4%
+10%
+23%
+6%
+9%
+8%
+22%
OLD
+Seen
+Sense
+Score
OLD
+Seen
+Sense
+Score
Mixed

Gadovist



S^4 ontology of adaptation effects

- **Seen:** Never seen this word before
  - News to medical: "diabetes mellitus"
- **Sense:** Never seen this word used in this way
  - News to technical: "monitor"
- **Score:** The wrong output is scored higher
  - News to medical: "manifest"
- **Search:** Decoding/search erred *(ignored)*



(inside=old domain
outside=new domain)

Adaptation effects in MT

- Quick observations:
  - New D language model helps (10%-63% impro
  - Tuning on new D data helps (10%-90% improve
  - Weighting new D data helps (4%-150% improv

Consistent in:
* movie subtitles
* scientific pubs
* PHP tech docs

- Identifying errors in MT (w/o parallel ne data):
  - **Seen:** old-only model + unseen input word pairs
  - **Sense:** old-only model + seen input/unseen out t pairs
  - **Score:** intersect old and mixed model, score fr m old

	News	Medical
Seen	Little effect	~ 40% of error
Sense	Little effect	~ 40% of error
Score	~ 90% of error	~ 20% of error

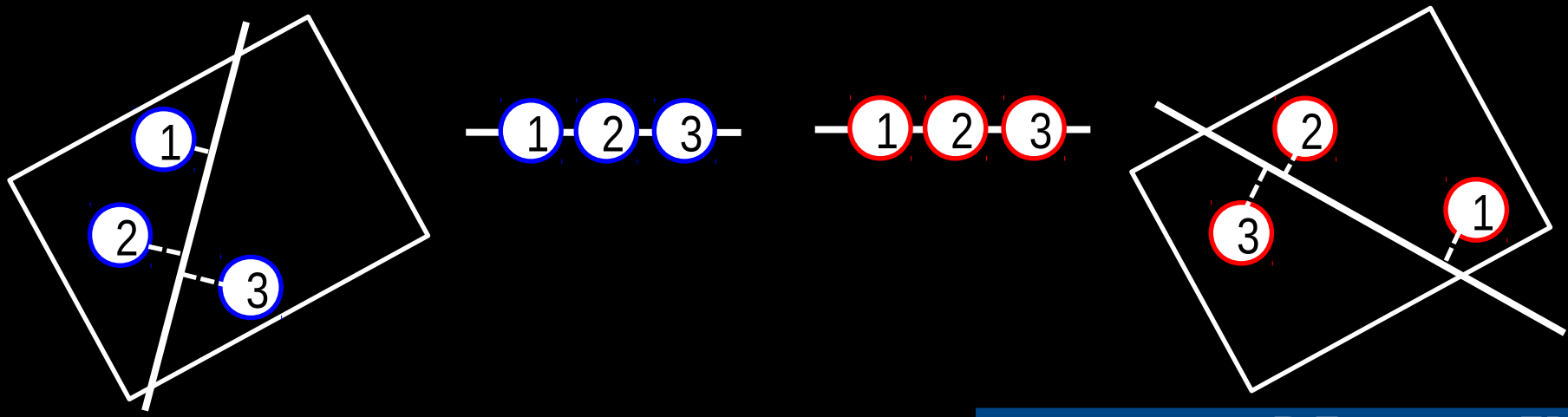
(as measured by Bleu score)

Translating across domains is hard

Dom	Most frequent OOV Words			
News (17%)	behavior	favor	neighbors	fueled
	neighboring	abe	wwii	favored
	favorable	zhao	ahmedinejad	bernanke
	favorite	phelps	ccp	skeptical
Medical (49%)	renal	hepatic	subcutaneous	irbesartan
	ribavirin	olanzapine	serum	patienten
	dl	eine	sie	pharmacokinetics
	ritonavir	hydrochlorothiazide	erythropoietin	efavirenz
Movies (44%)	gonna	yeah	mom	hi
	b****	daddy	s***	later
	f*****g	f***	gotta	wanna
	uh	namely	bye	dude

Dictionary mining for "seen" errors

- Find frequent terms in new domain
- Use those that exist in old domain as "training data"
- Extract context and orthographic features
- Find low-dimensional subspace on training data (CCA)



- Pair input words with <=5 output words
- Add four features to SMT model
- Rerun parameter tuning

	DE	FR
News	+0.80	+0.36
Emea	+1.44	+1.51
Subs	+0.13	+0.61
PHP	+0.28	+0.68

(Bleu score improvements)

Marginal matching for "sense" errors

Given:

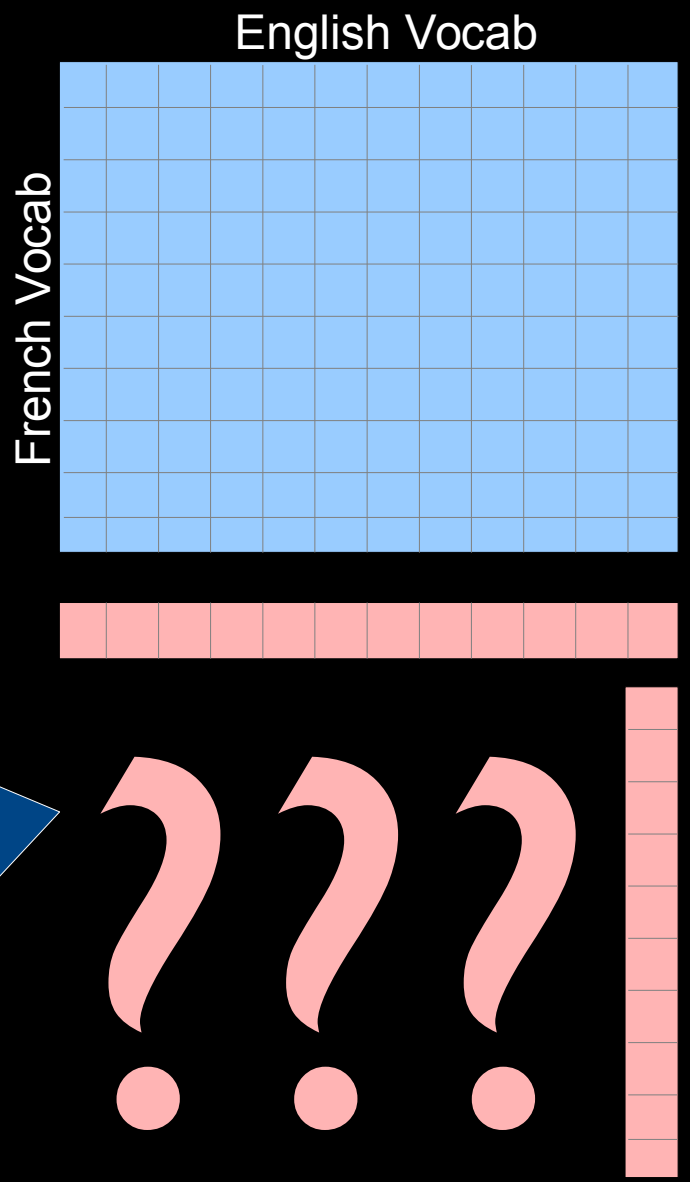
- A joint p(x,y) in the old domain
- Marginals q(x) and q(y) in the new domain

Recover:

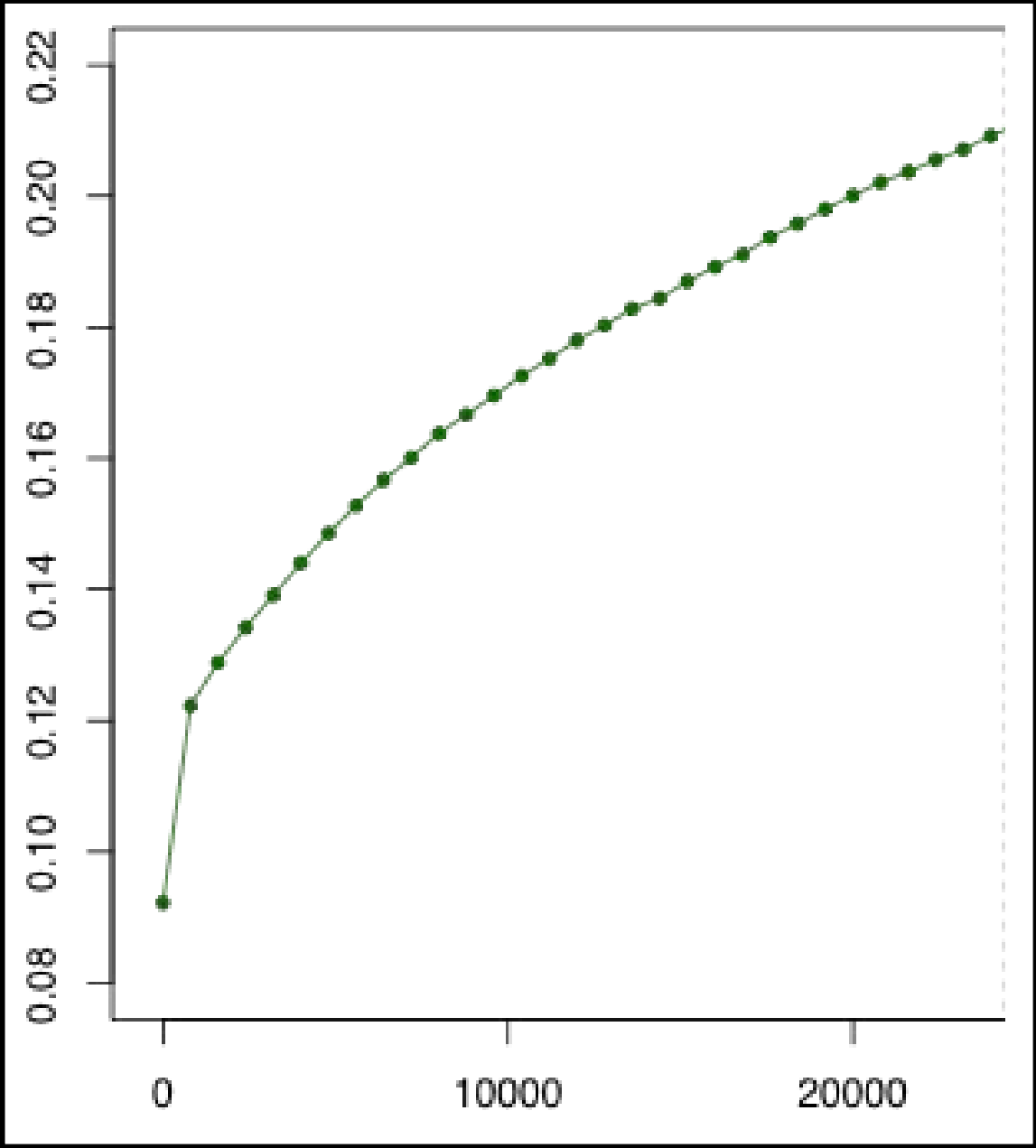
- Joint q(x,y) in the new domain

We formulate as a L1-regularized linear program

Easier: *many* q(x) and q(y)s



Intrinsic evaluation: MRR



- Ranked document pairs: learning from most science-like first
- Work in progress: increasing document pairs
- Relative gain expected to slow, documents less and less science-y

Example Learned Translations

French	Correct	Learned Translations
cisaillement	shear	viscous crack shear
chromosomes	chromosomes	chromosomes chromosome chromosomal
caractérisation	characterization	characterization characteristic π
araignées	spiders	spiders ant spider
tiges	stems	usda centimeters flowering

Bleu Scores

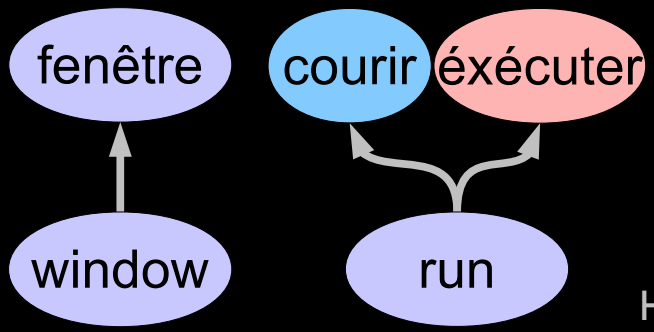
Baseline	21.91
Baseline + Strip Accents	22.20
Append Top-1 translation for OOVs	23.25
Append Top-1 translation for freq(fr)<11	23.86
Append Oracle OOV translations	26.38

Automatically identifying new senses

- **Context + existence of translations in comparable data**

is a **window** of opportunity	via une **fenêtre** insérée .
have a **window** of opportunity	vers ma **fenêtre** ou vers
in the **run** up to	voulons pas **courir** le risque
, we **run** the risk	, sans **courir** le risque

the browser **window** ' s	dans la **fenêtre** . cet
in the **window** to give	dans la **fenêtre** . </s>
time to **run** when applied	*courir not found*
or have **run** vcvars.bat ,	



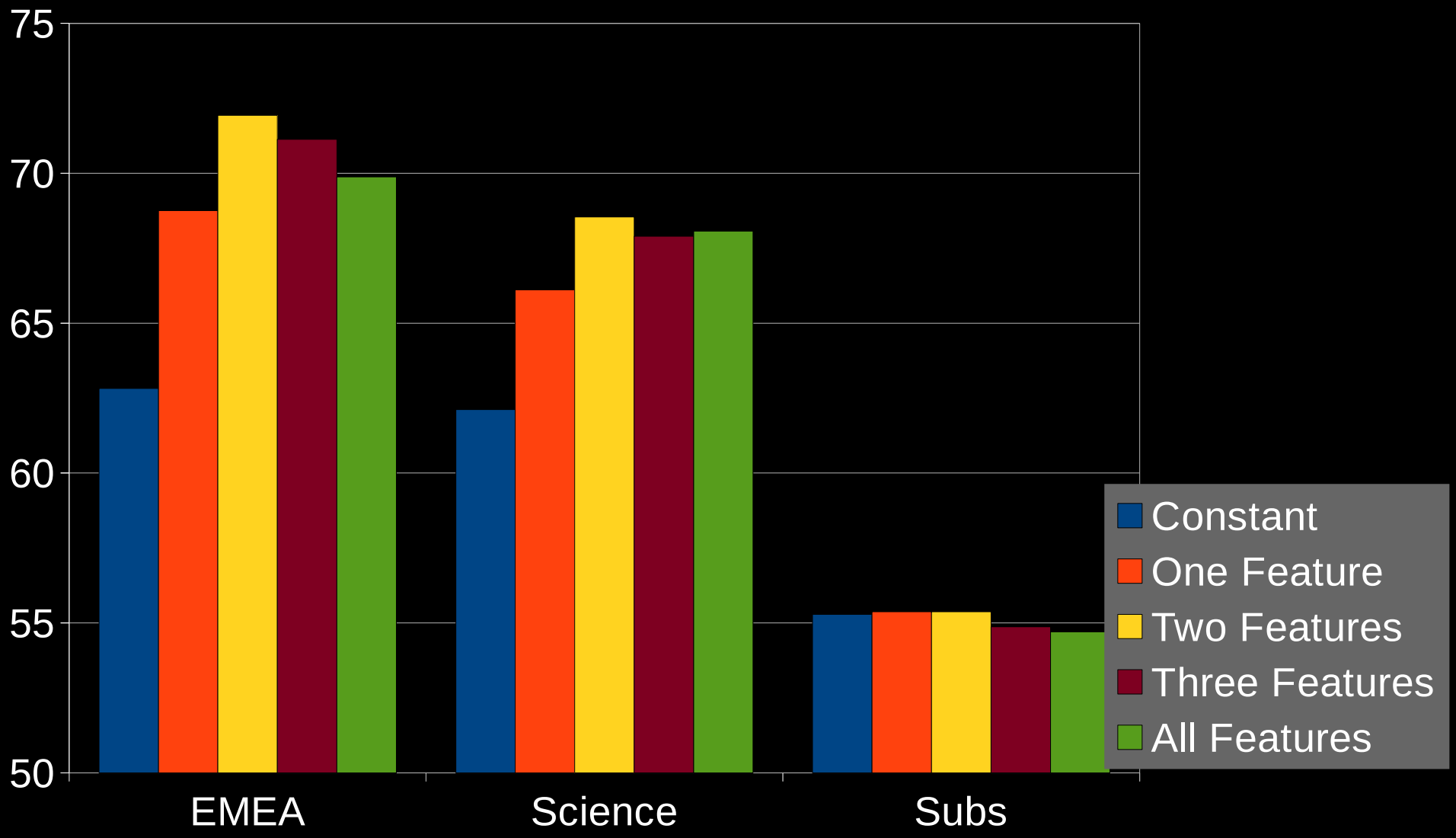
ne pouvez **éxécuter** que les
pour l' **éxécuter** elle va

Spotting New Senses

- Binary classification problem:
  - +ve: French token has previously unseen sense
  - -ve: French token is used in a known way

- Lots of features considered...
  - Frequency of words/translations in each domain
  - Language model perplexities across domains
  - Topic model "mismatches"
  - Marginal matching features
  - Translation "flow" impedence

Experimental Results



Selected features:

```
EMEA:    ppl   || matchm flow   || matchm topics flow
Science: ppl   || matchm ppl    || matchm topics ppl
Subs:    topcs || matchm topics || matchm topics flow
```

Discussion

- Introduced taxonomy and measurement tools for adaptation effects in MT
- "Score" errors – target of prior work – only a part of what goes wrong
- Marginal matching introduced as a model for addressing *all* S4 issues simultaneously: +2.4 BLEU
- Data and outputs released for you to use (both in MT and as a stand-alone lexical selection task)
- Feature-rich approaches integrated into Moses via VW library, applied to adaptation
- Range of other problems to work on: identifying new senses, cross-domain topic models, etc.)

Thanks! Questions?


Marine Carpuat
(NRC-CNRC)


Alex Fraser
(U. Stuttgart)

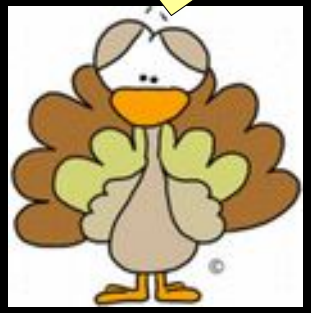

Chris Quirk
(Microsoft)

Discussion and Future Work

- With labels, or labeler, for target, simple methods go a long way
- Active sampling can use adaptation knowledge
- Adaptive co-regularization ties domains together
- Similar ideas can be applied to multitask learning

- What happens with tons of domains?
  - Or non-discrete domains?
  - Or only domain *features*?
- Can we make better use of unlabeled data (ala Blitzer)

Happy Thanksgiving!



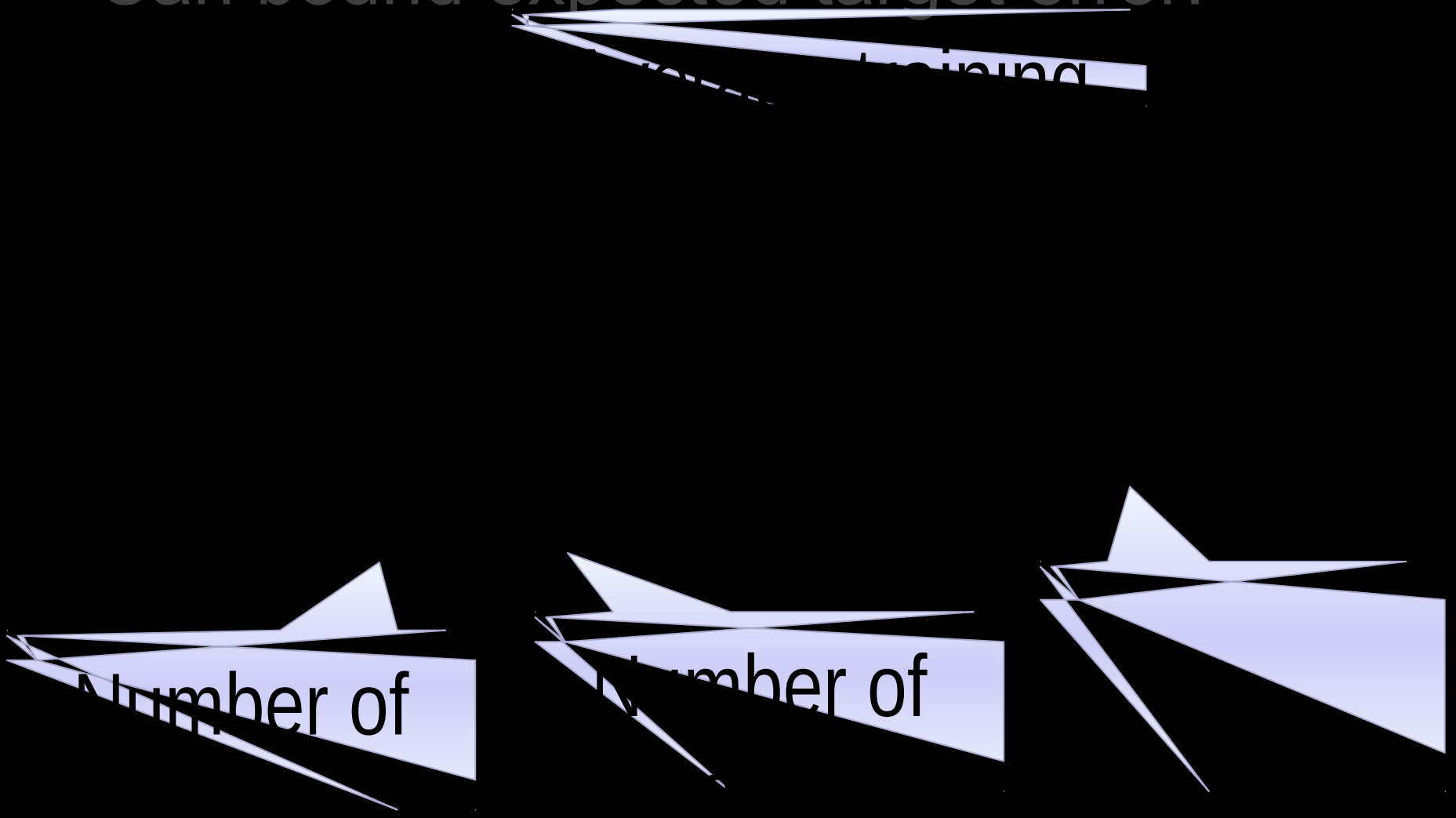
Thanks! Care to actively query?

Some experimental results

NER	wi	2.45	2.45	**2.12** (all)	2.41
2.12					
	un	3.67	2.46	2.16 (linint)	2.03
1.91					
	bn	2.83	0.46	0.40 (all)	**0.34**
0.32					
CoNLL	tgt	2.49	2.95	**1.75** (wgt/li)	1.89
1.76					
PubMed	tgt	12.02	4.15	3.95 (linint)	3.99
3.61					
CNN	tgt	10.29	3.82	3.44 (linint)	**3.35**
3.37					
	wsj	6.63	4.35	4.30 (weight)	4.27
4.11					
	swbd3	15.90	4.15	4.09 (linint)	3.60
3.51					
	br-cf	5.16	6.27	**4.72** (linint)	5.22
5.15					
Tree	br-cg	4.32	5.36	**4.15** (all)	4.25
4.90					
bank-	br-ck	5.05	6.32	**5.01** (prd/ll)	5.27
5.41					

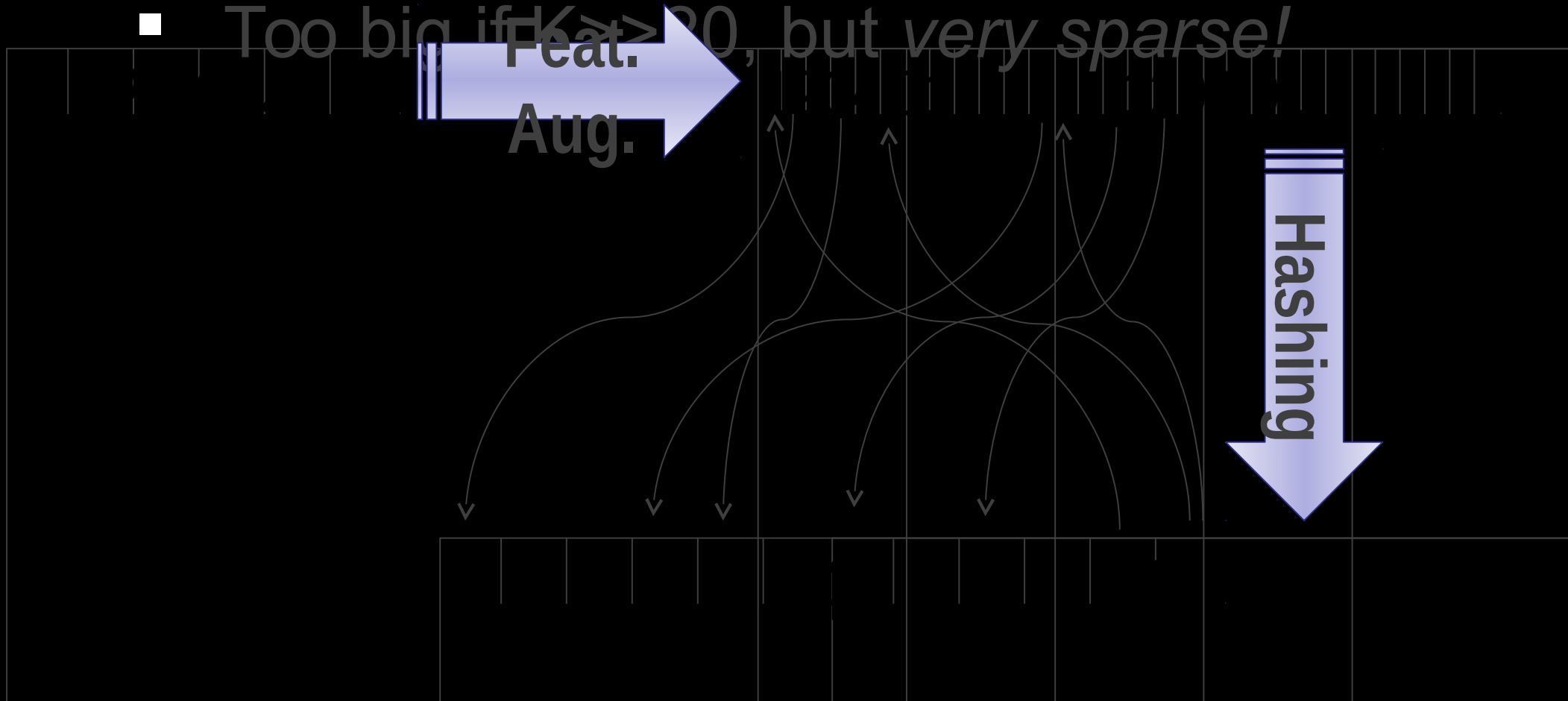
Some Theory

- Can bound expected target error:

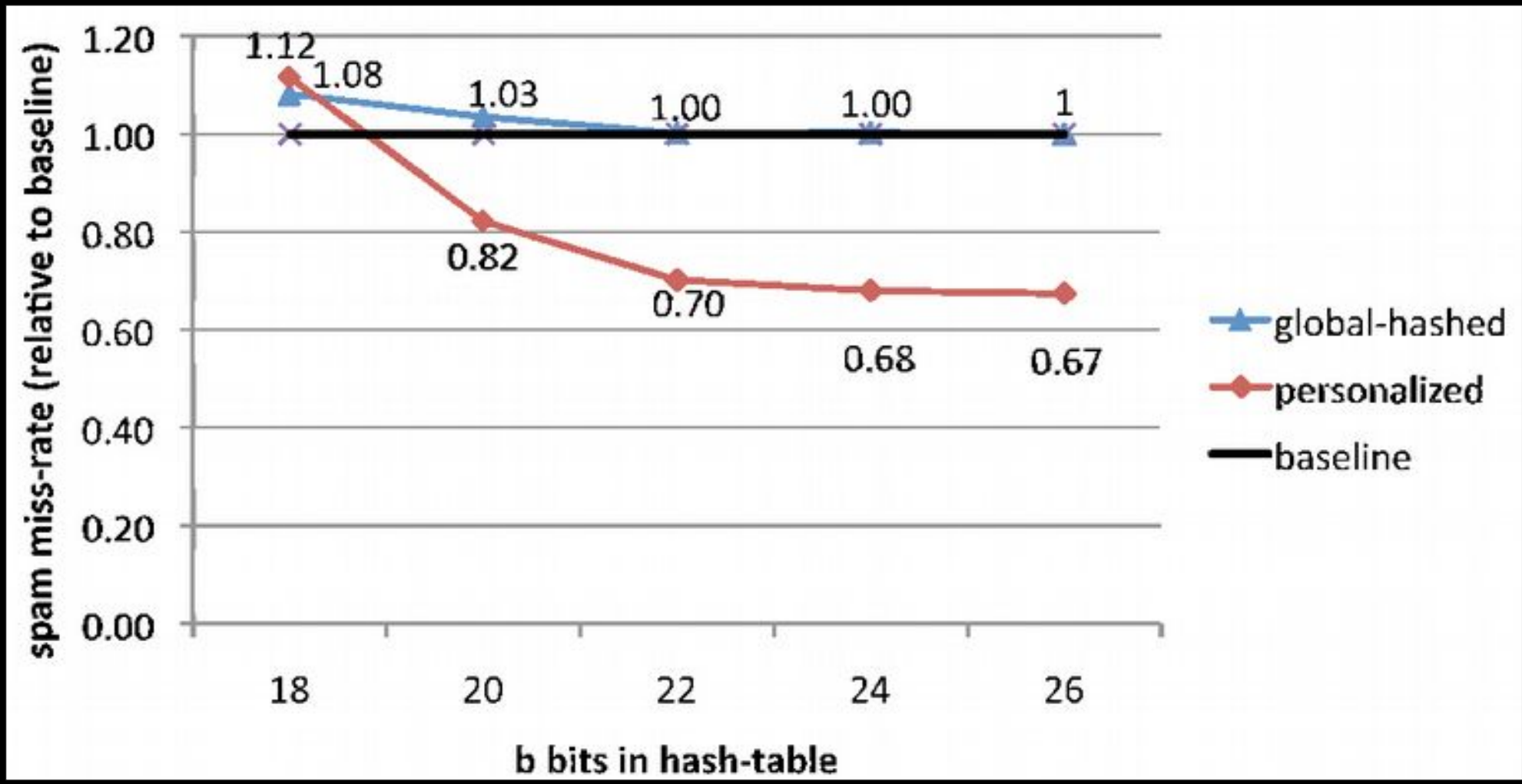


Feature Hashing

- Feature augmentation creates (K+1)D parameters
- Too big if K>>20, but *very sparse!*



Hash Kernels



What is a domain anyway?

- Time?
  - News the day I was born *vs* news today?
  - News yesterday *vs* news today?
- Space?
  - News back home *vs* news in Haifa?
  - News in Tel Aviv *vs* news in Haifa?
- Do my [illegible] main specifie[illegible]

Stream of <*x*,*y*,*d*> data with *y* and *d* sometimes hidden?

We're *all* domains: ~~personalization~~

- adapt learn across millions of "domains"?
- share enough information to be useful?
- share little enough information to be safe?
- avoid negative transfer?
- avoid DAAM (domain adaptation spam)?

Discussion