machine translation

domain adaptation

Microsoft Research ◇ National Research Council ◇ Univ of Stuttgart ◇ Simon Fraser ◇ Univ of Maryland ◇ Yale ◇ Charles Univ ◇ Univ of Chicago ◇ Army Research Lab ◇ Johns Hopkins

Fabienne Braune
Marine Carpuat
Ann Clifton
Hal Daumé III
Alex Fraser
Katie Henry
Anni Irvine
Jagadeesh Jagarlamudi
John Morgan
Chris Quirk
Majid Razmara
Rachel Rudinger
Ales Tamchyna

Special thanks:
George Foster
Dragos Munteanu

Conference of the Association for Machine Translation in the Americas
Wed 31 Oct 2012

Domains really are different

- 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.

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**

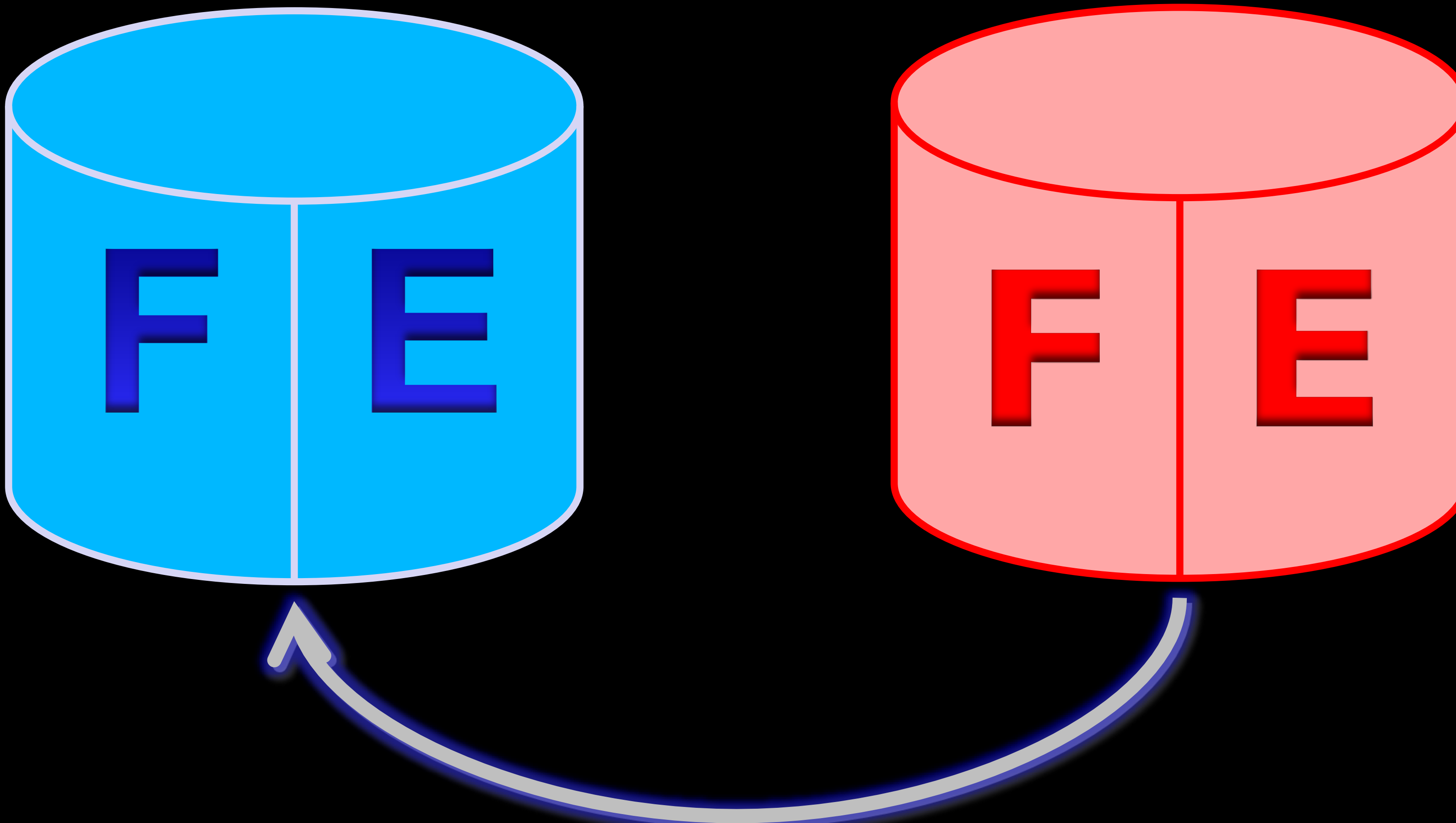
Key Question: What went wrong?

S^4 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

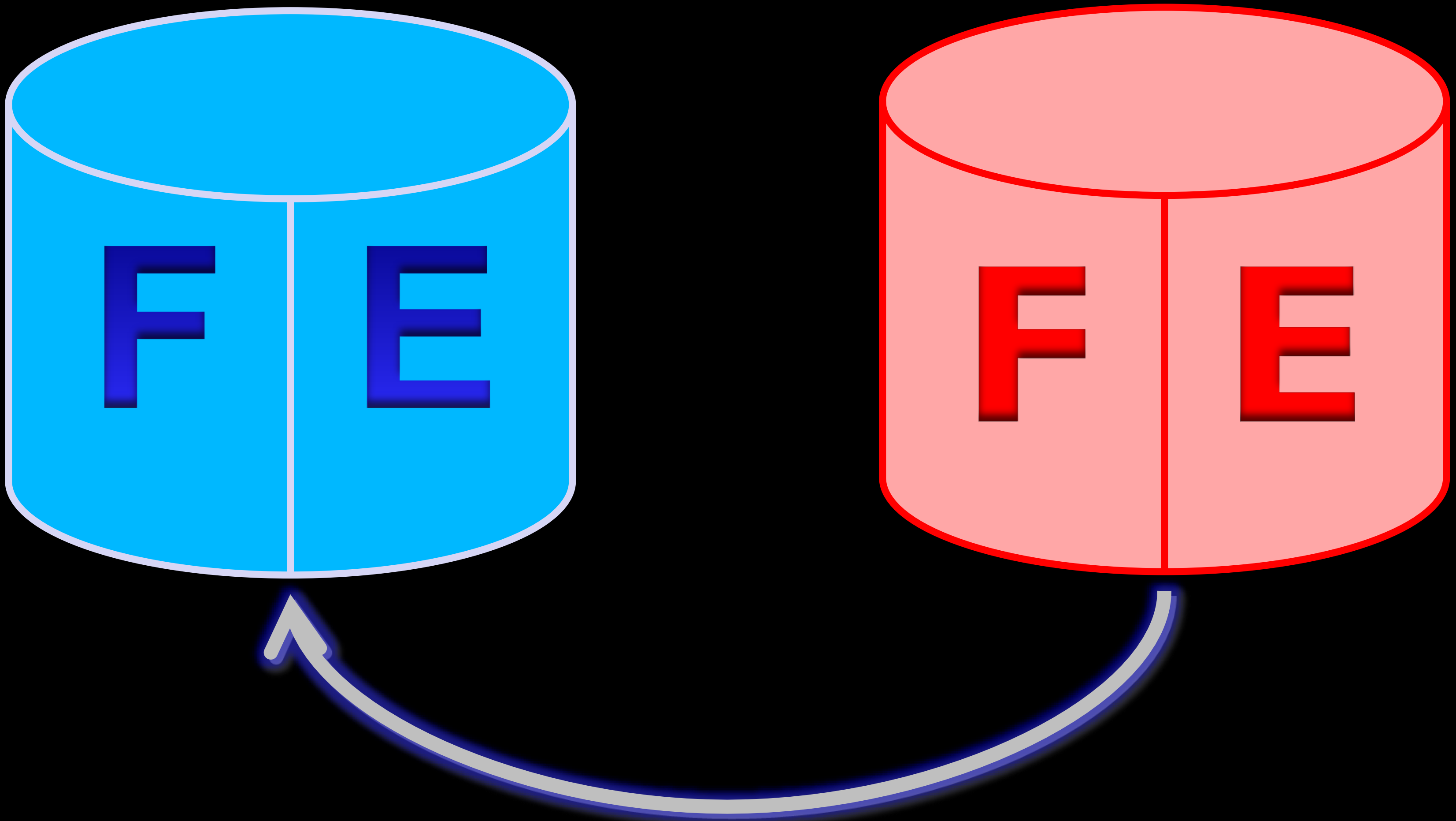
Working with *no* new domain parallel data!

Measuring SEEN effects



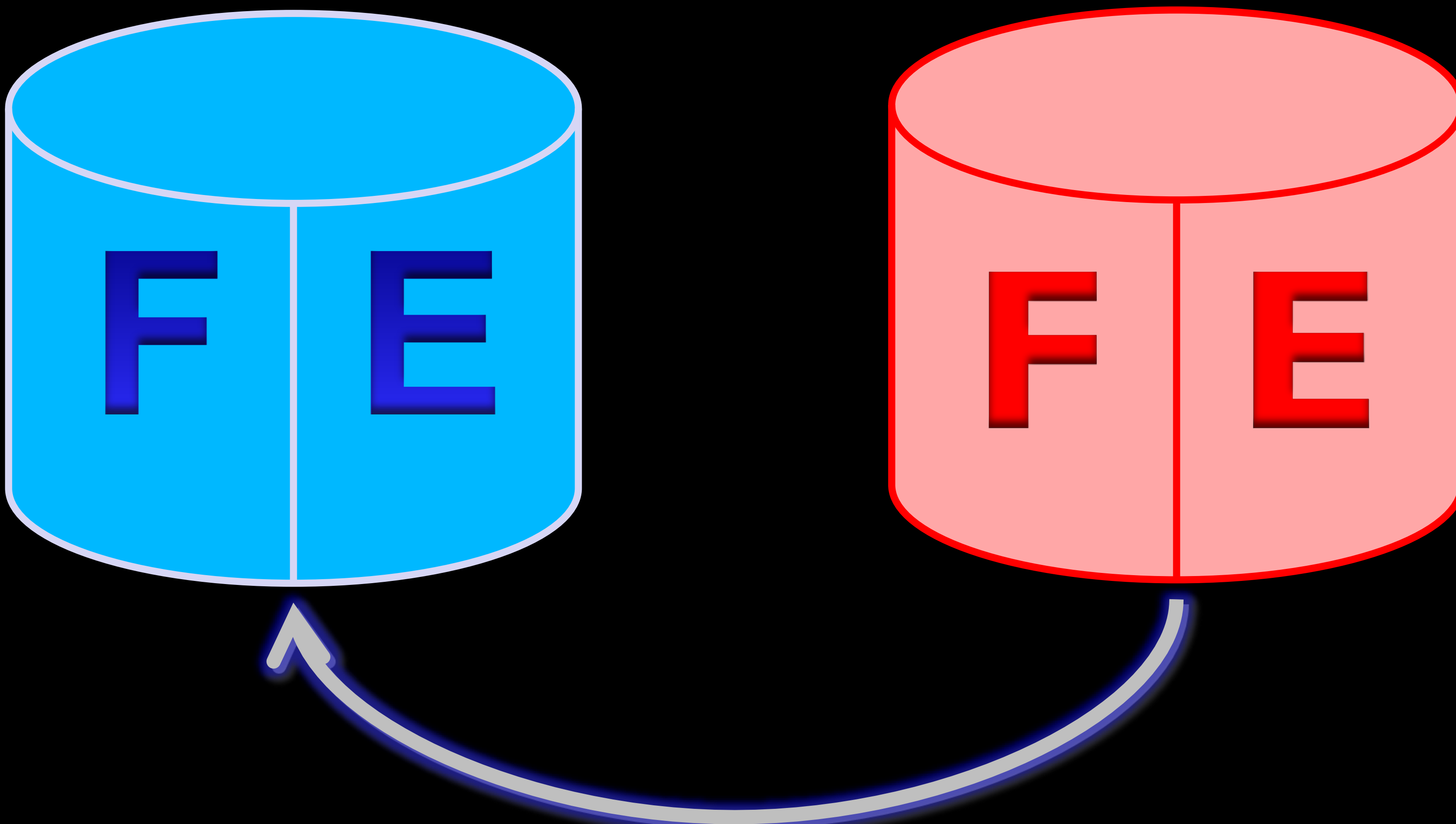
Add all phrase pairs with previously unseen F side

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 S^4 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 S^4 effects

Output (en):	initial	dose	:
Input (fr):	dose	initiale	:
Ref (en):	starting	dose	:

Correct

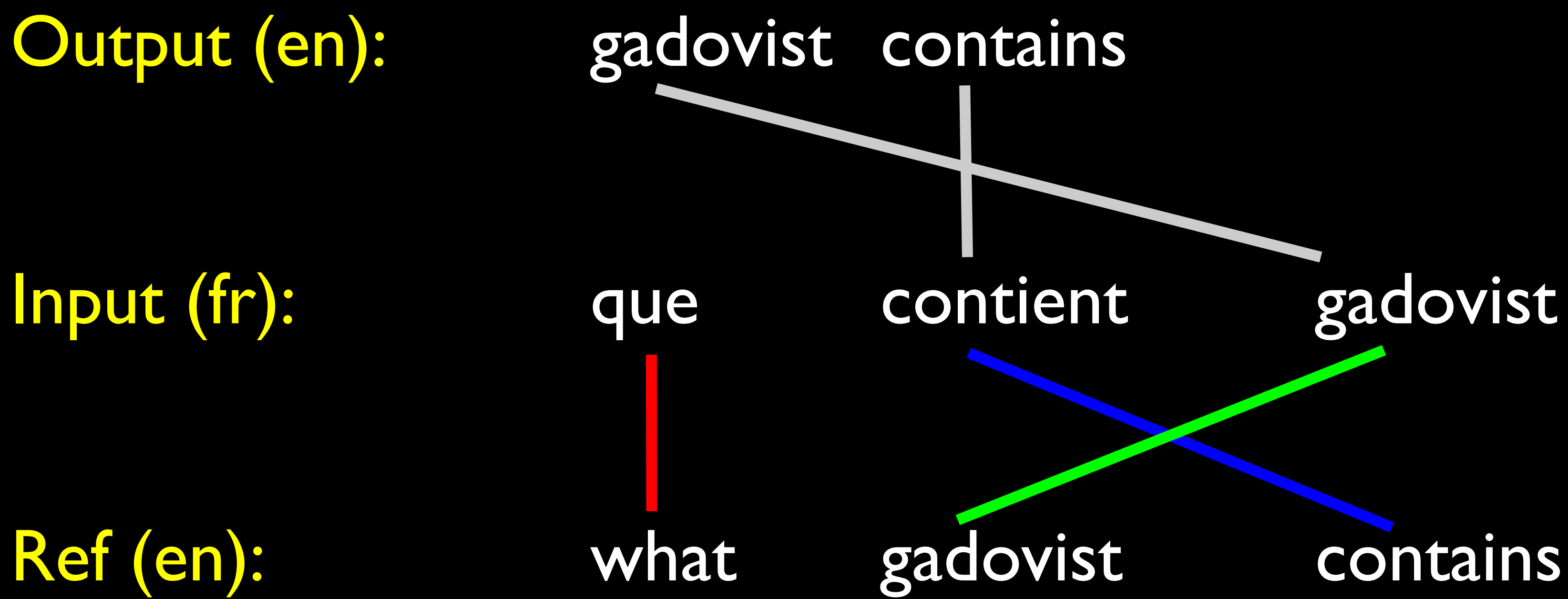
Seen-freebie

Seen-error

Sense-error

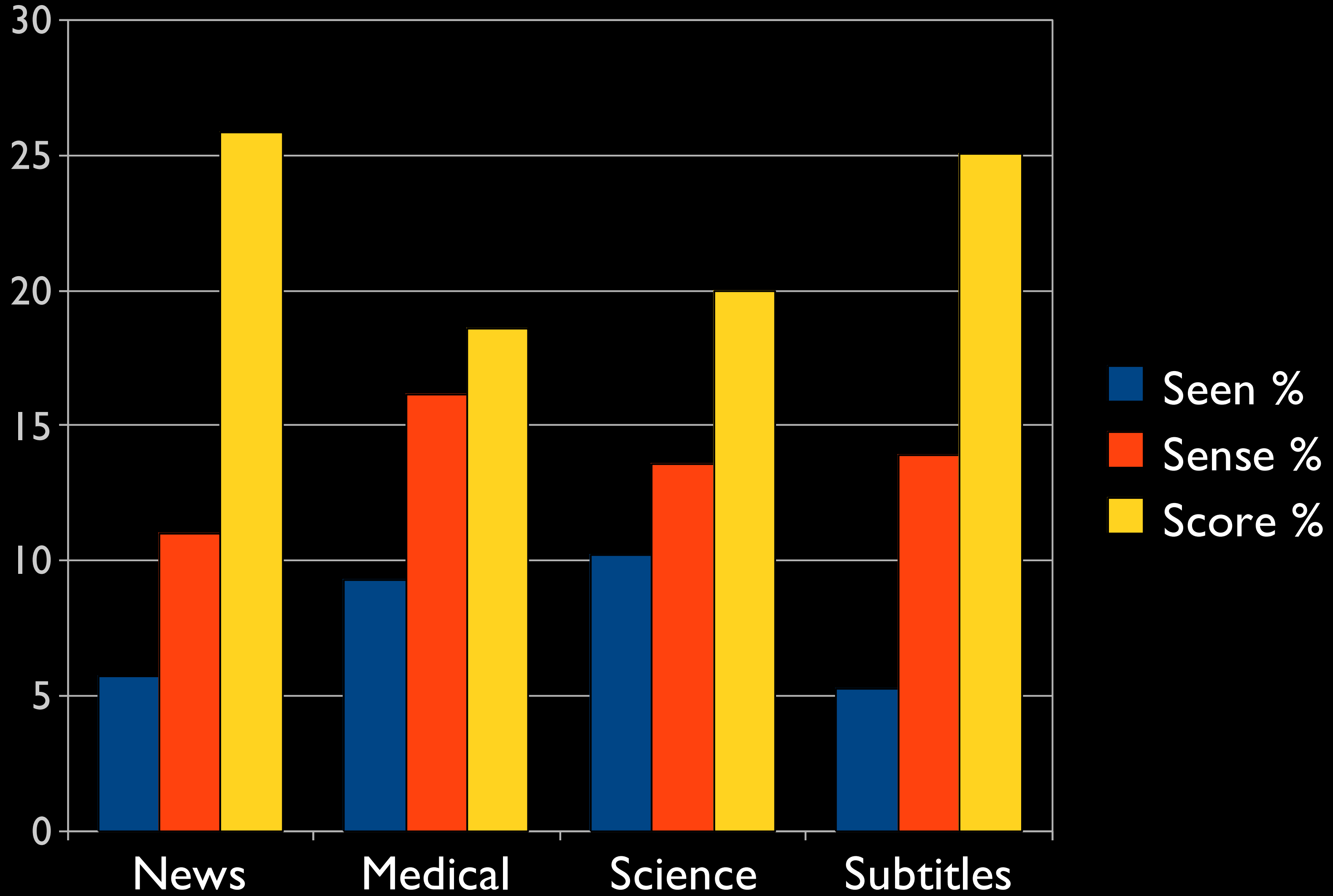
Score-error

Micro-analysis of S^4 effects

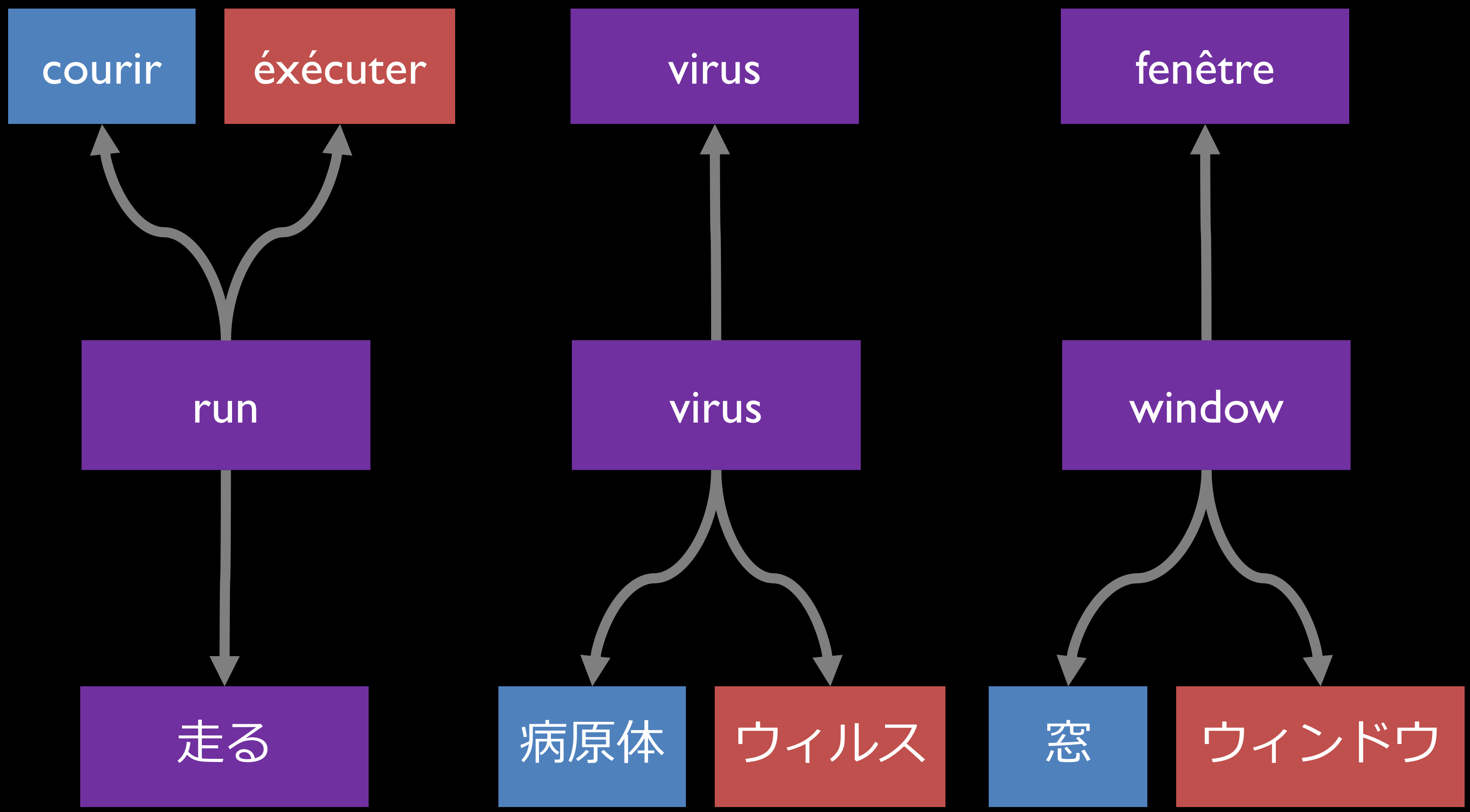


Correct
Seen-freebie
Seen-error
Sense-error
Score-error

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	…	Σ
accorder	9	1	…	10+…
…	…	…	…	…
Σ	9+…	1+…	…	

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

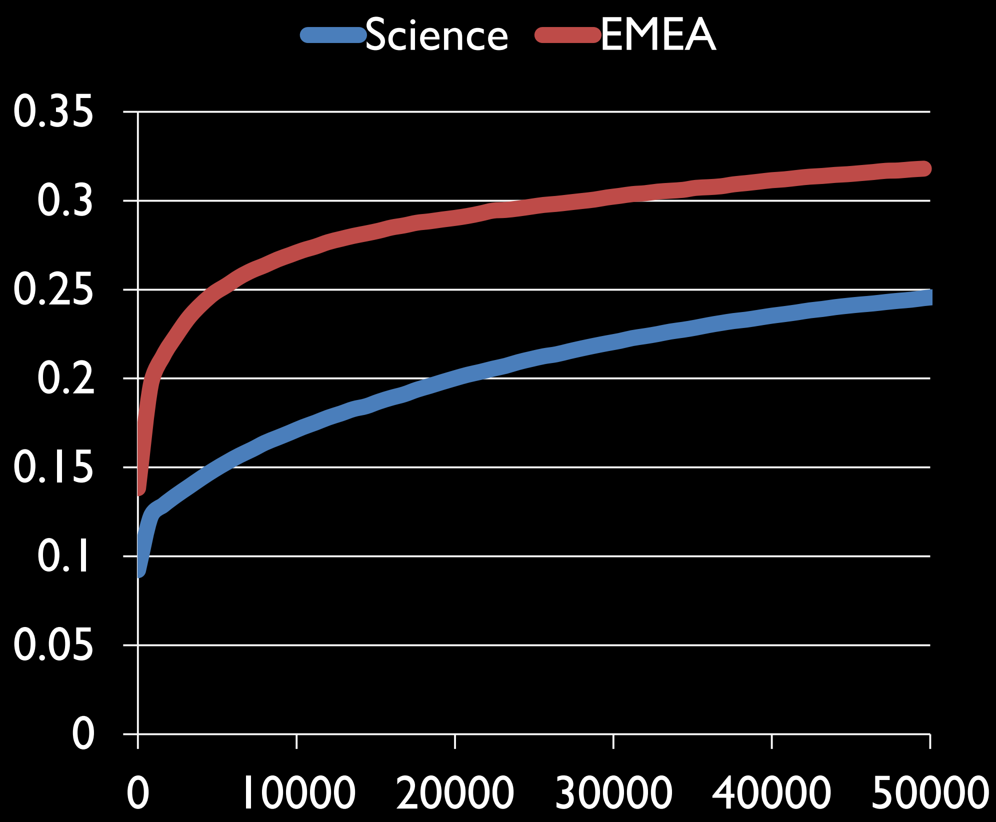
Additional features

- 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)

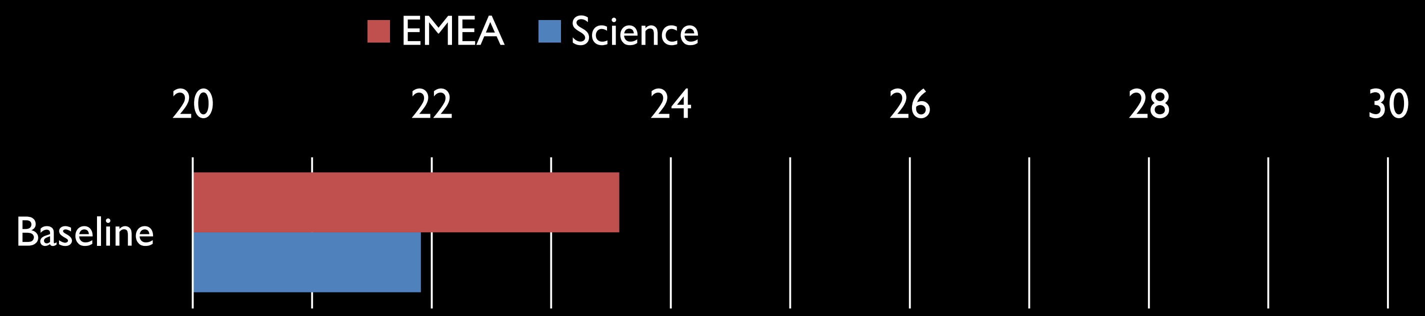


- Ranked Wikipedia document pairs: learning from most science-like first
- Decreasing benefits after ~50,000 document pairs
- Relative gain expected to slow, as documents are less and less science-y

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

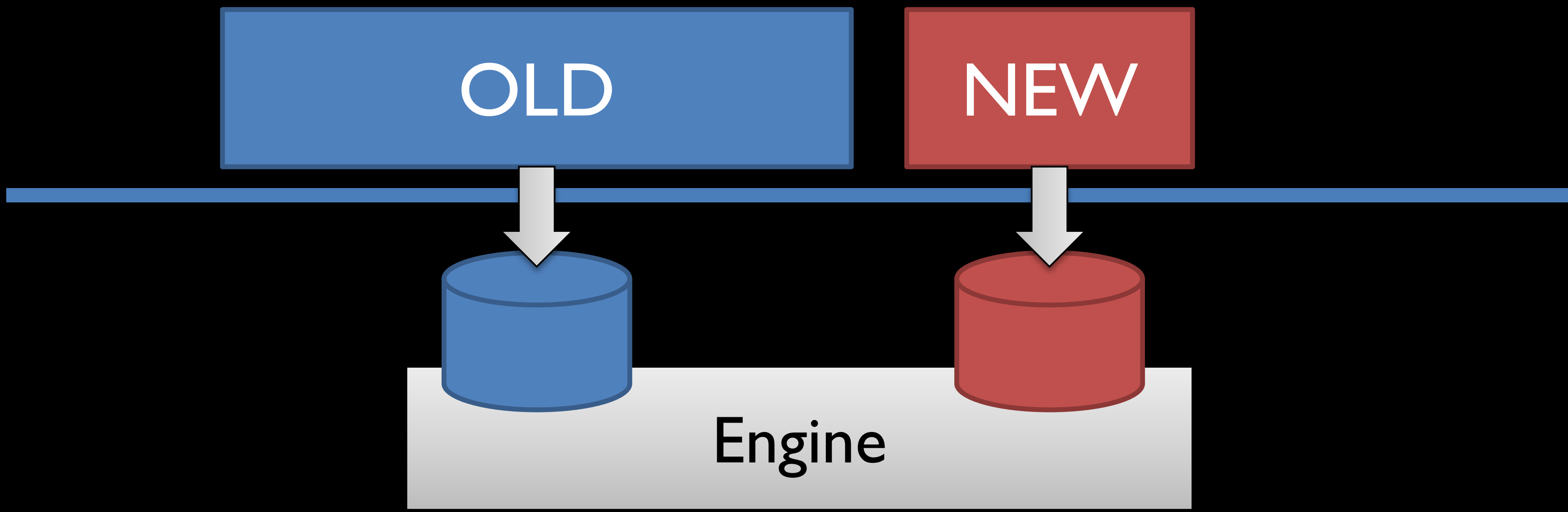


Case 2: Add NEW domain parallel data

- 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)



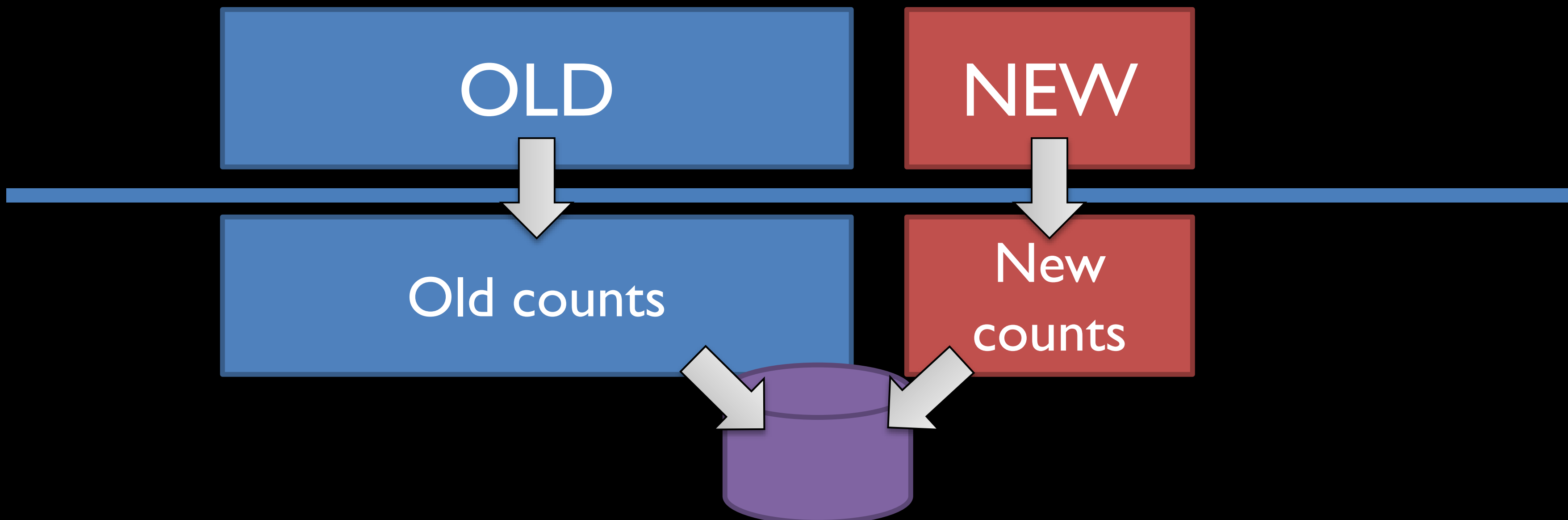
Baseline:

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

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?

Disambiguating English senses of **rapport**

P(e|f) in Hansard

report	Il a rédigé un **rapport** .
relationship	Quel est le **rapport** ?
ratio	le **rapport** longueur / largeur
balance	le **rapport** bénéfique / risque
…	

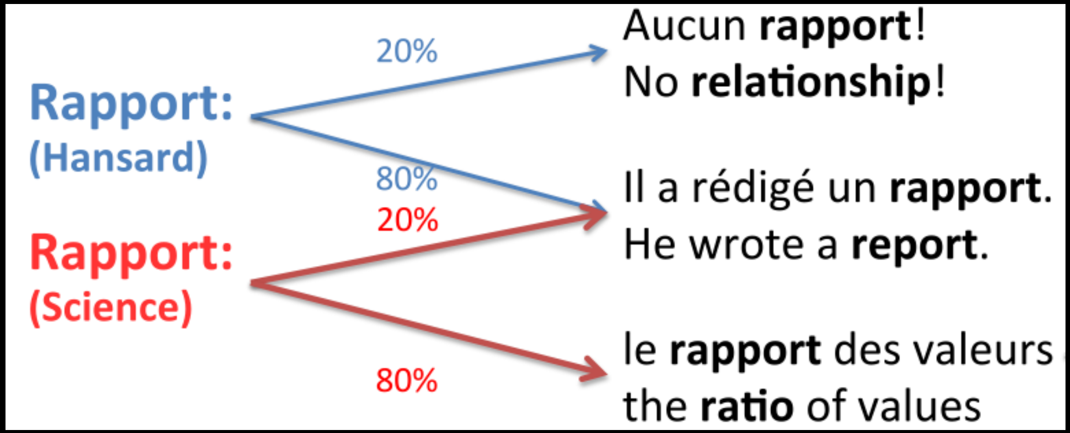
Highest P(e|f) in Science!

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

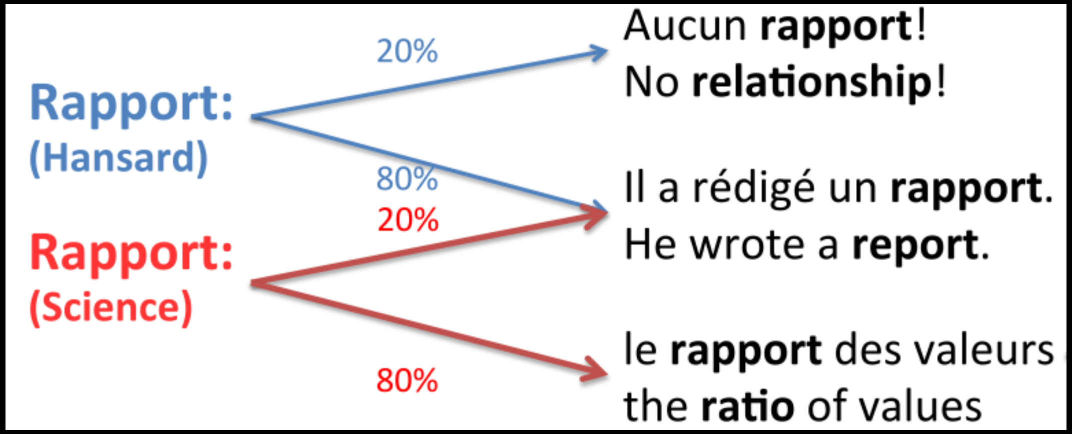
Phrase Sense Disambiguation



- 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

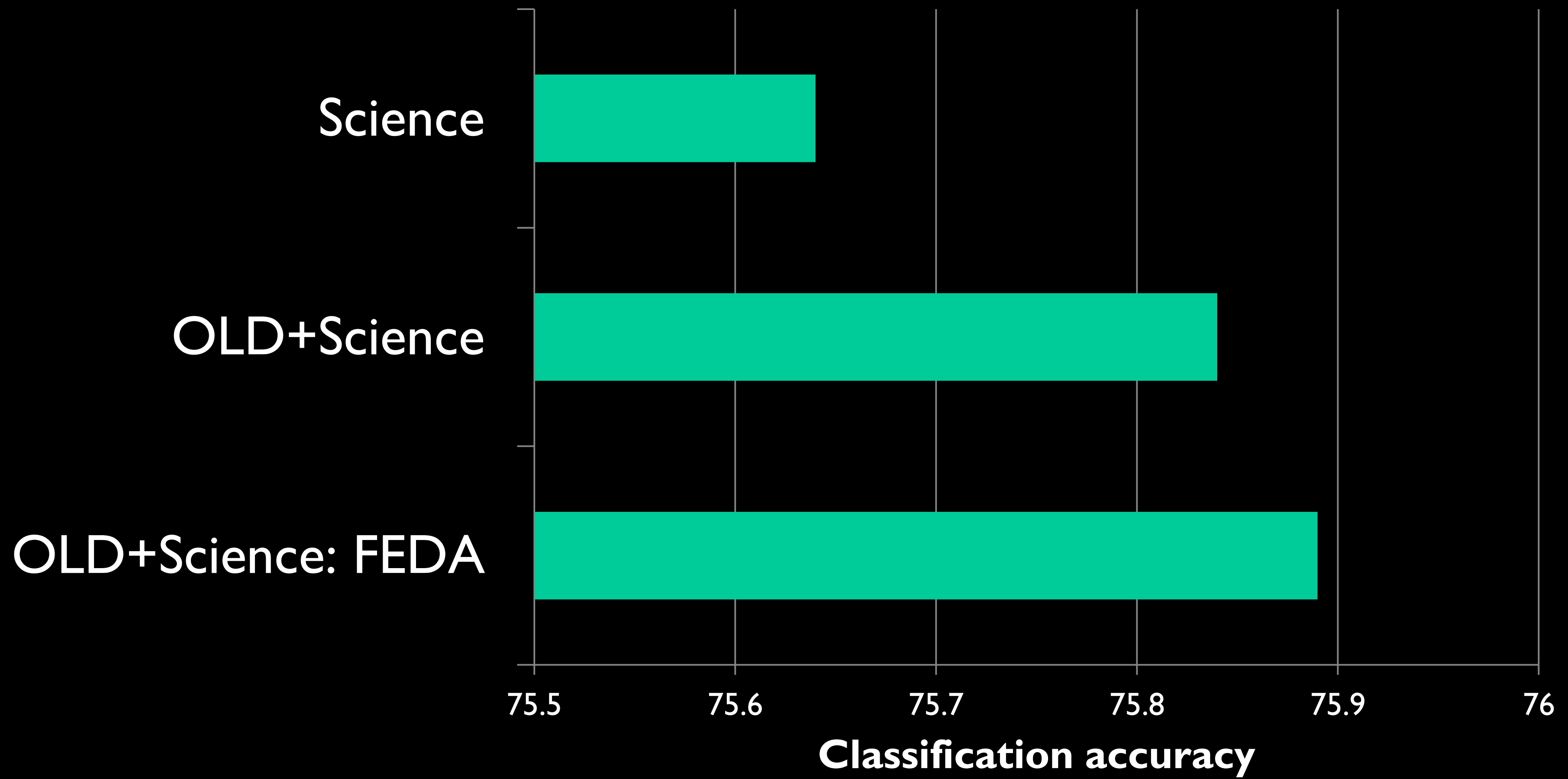


- Train a classifier over OLD and NEW data
- Allow classifier to:
  - share some features
    `{rédigé …} rapport → report`
  - keep others domain specific
    `rapport {… valeurs} → ratio`

Feature augmentation

	OLD	NEW
	$\varphi_{e,f} \mapsto \langle \varphi_{e,f}, \varphi_{e,f}, 0 \rangle$	$\varphi_{e,f} \mapsto \langle \varphi_{e,f}, 0, \varphi_{e,f} \rangle$
Original features	{rédigé …} rapport → report {aucun …} rapport → relationship	{rédigé …} rapport → report rapport {… valeurs} → ratio

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

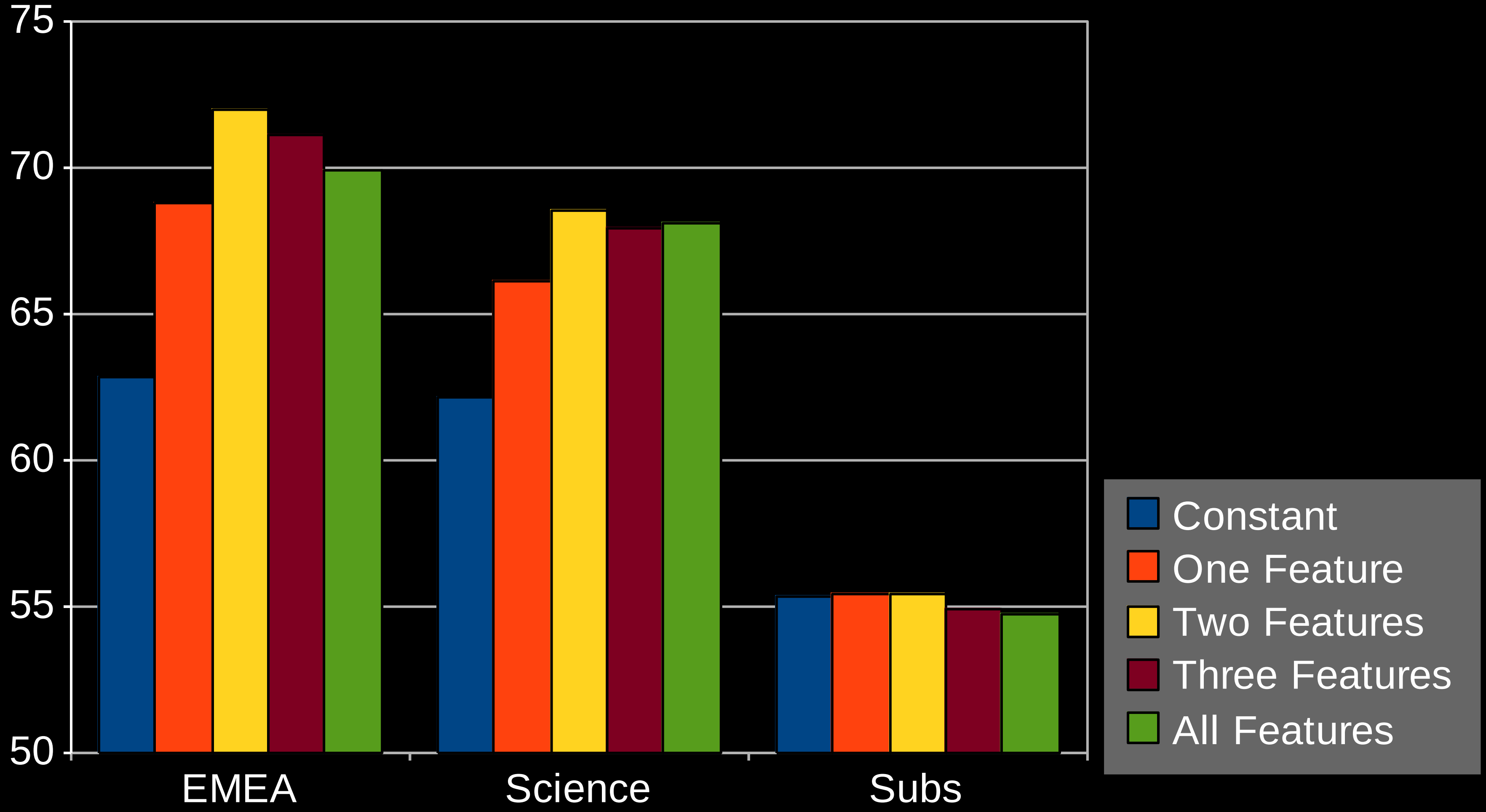
Discovering latent topics from parallel data

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 S^4 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* S^4 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?