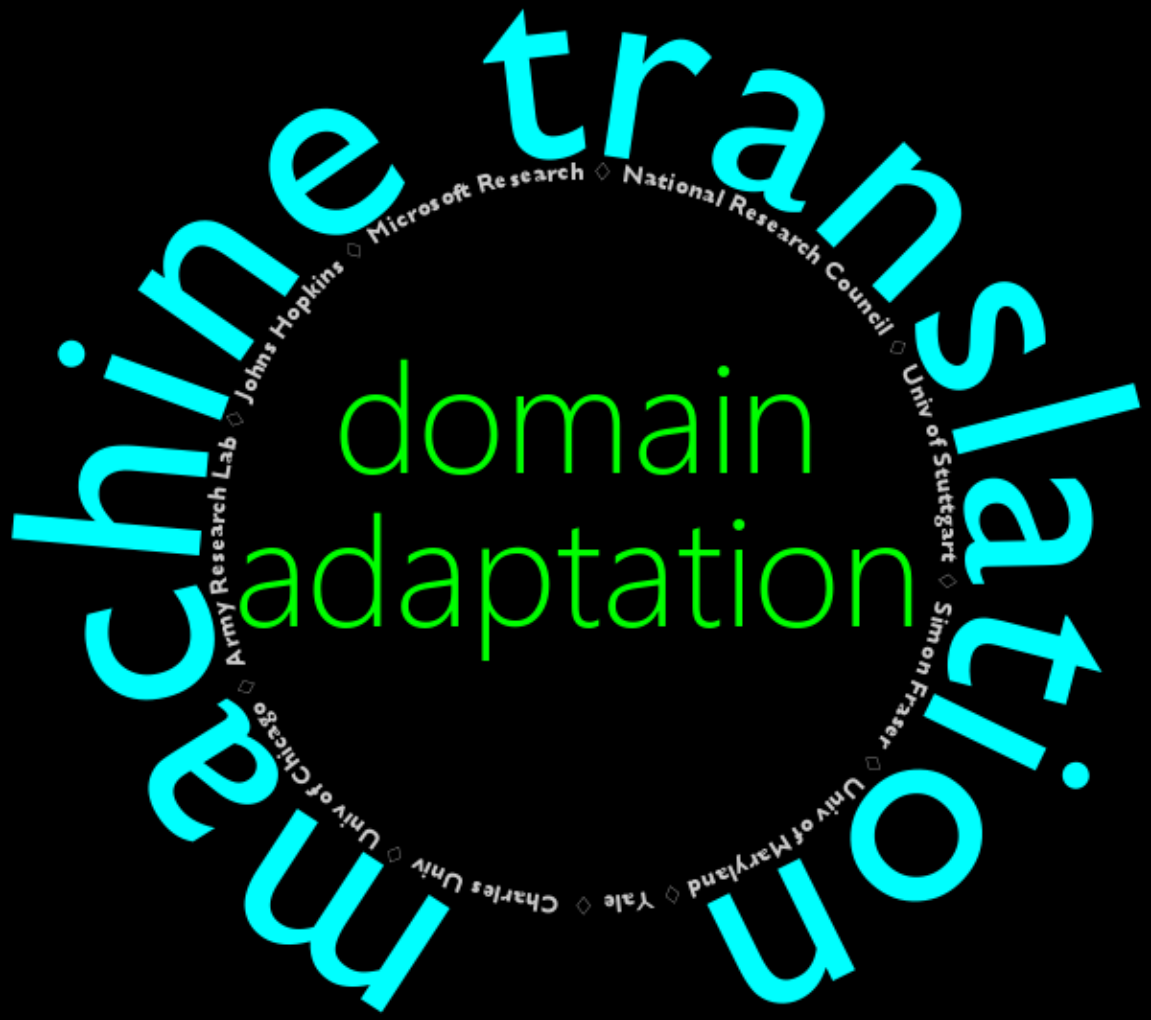


Fabienne Braune
Marine Carpuat
Ann Clifton
Hal Daumé III
Alex Fraser
Katie Henry
Anni Irvine
Jagadeesh Jagarlamudi
John Morgan
Chris Quirk
Majid Razmara
Rachel Rudinger
Ales Tamchyna

Special thanks:
George Foster
Dragos Munteanu
Everyone at CLSP

Eighth Machine Translation Marathon
Charles University, Prague
September 13th, 2013

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
(Science?) Joel Tetreault sings Greg Crowther	Jenny, what is this number? Tell me how it's defined. Jenny, plug in this number: Three point one four one five nine. (Three point one four one five nine).

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**

Outline

- Quick introduction to domain adaptation for SMT
- What is the problem really?
  - a new taxonomy for domain-related SMT errors
- Towards solving the errors
  - with comparable corpora
  - with parallel corpora

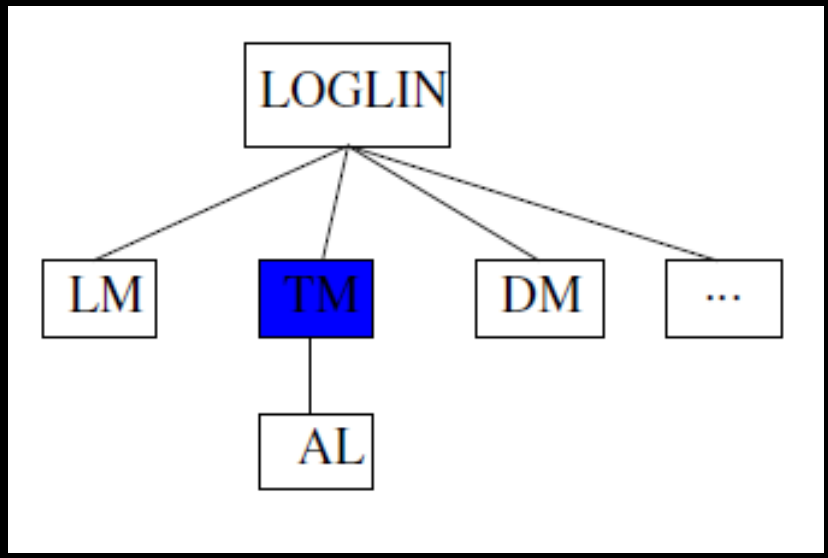
Domain Adaptation for SMT

- Problem: **domain mismatch** between test and training data can cause severe degradation in translation quality
- General solution: adjust SMT parameters to optimize performance for a test set, based on some knowledge of its domain
- Various settings:
  - amount of in-domain training data: small, dev-sized, none (just source text)
  - nature of out-of-domain data: size, diversity, labeling
  - monolingual resources: source and target, in-domain or not, comparable or not
  - latency: offline, tuning, dynamic, online, (interactive)

Slide adapted from Foster 2012

What to adapt?

- Log-linear model
  - limited scope if in-domain tuning set (dev) is available
- Language model (LM)
  - effective and simple
  - previous work from ASR
  - perplexity-based interpolation popular
- Translation model (TM):
  - most popular target, gains can be elusive
- Other features: little work so far
- Alignment: little work, possibly limited scope due to “one sense per discourse"



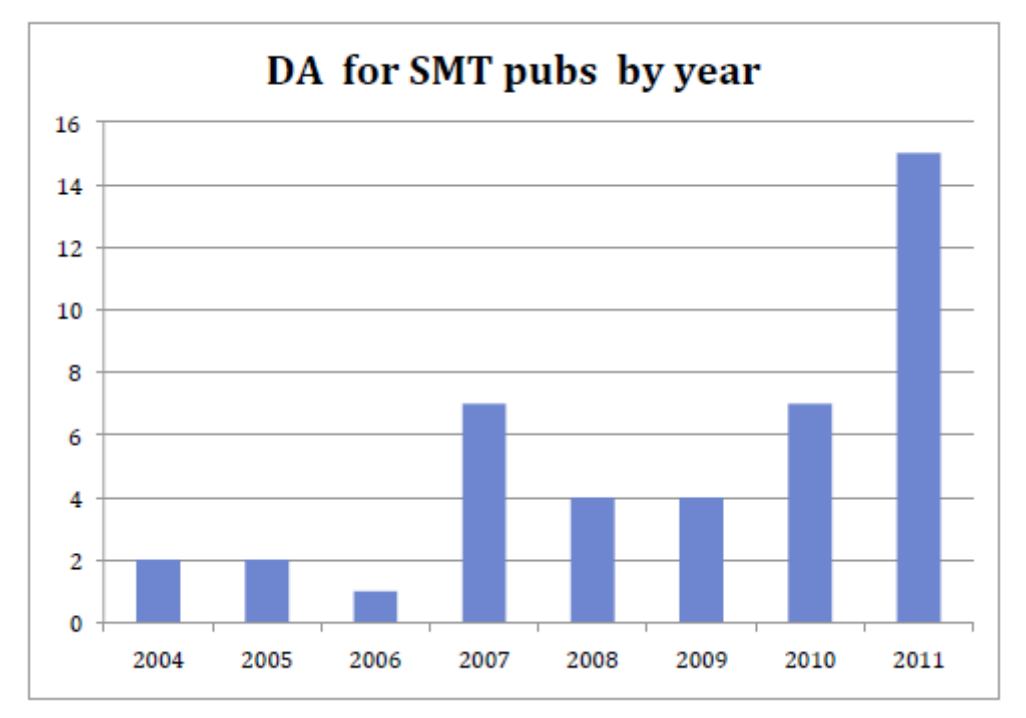
Slide adapted from Foster 2012

How to adapt to a new domain?

- Filtering training data
  - select from out-of-domain data based on similarity to our domain
- Corpus weighting (generalization of filtering)
  - Done at sub-corpora, sentence, or phrase-pair levels
- Model combination
  - train sub-models on different sub-corpora
- Self-training
  - generate new parallel data with SMT
- Latent semantics
  - exploit latent topic structure
- Mining comparable corpora
  - extend existing parallel resources

Slide adapted from Foster 2012

Domain adaptation is a hot topic…



Slide adapted from Foster 2012

The state of affairs (early 2012)

- Lots of interest in domain adaptation!
- Many settings and approaches
  - Most domain adaptation research is extremely tied to the available resources and SMT setup, highly heterogeneous
- Many key problems are not well understood
- Research reports somewhat conflicting findings
  - Lack of consensus as to what has worked so far
- No clearly-defined baselines
  - Some research does not even have a comparison to baselines
- No standard data sets or tasks

Slide adapted and completely mangled from Foster 2012

What happened next?

- Johns Hopkins CLSP 2012 summer workshop on domain adaptation for machine translation (DAMT)
  - 13(+) researchers working for 6 weeks on-site
  - Perhaps most important contribution: standard data and tasks for domain adaptation – see our website!
  - I won't tell you about that though, instead I'll tell you about some cool work we did
    - We did mostly use these resources, so you will get an idea...
  - (Tangent: the JHU summer workshop era is over – time for a European university to step in?)

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**

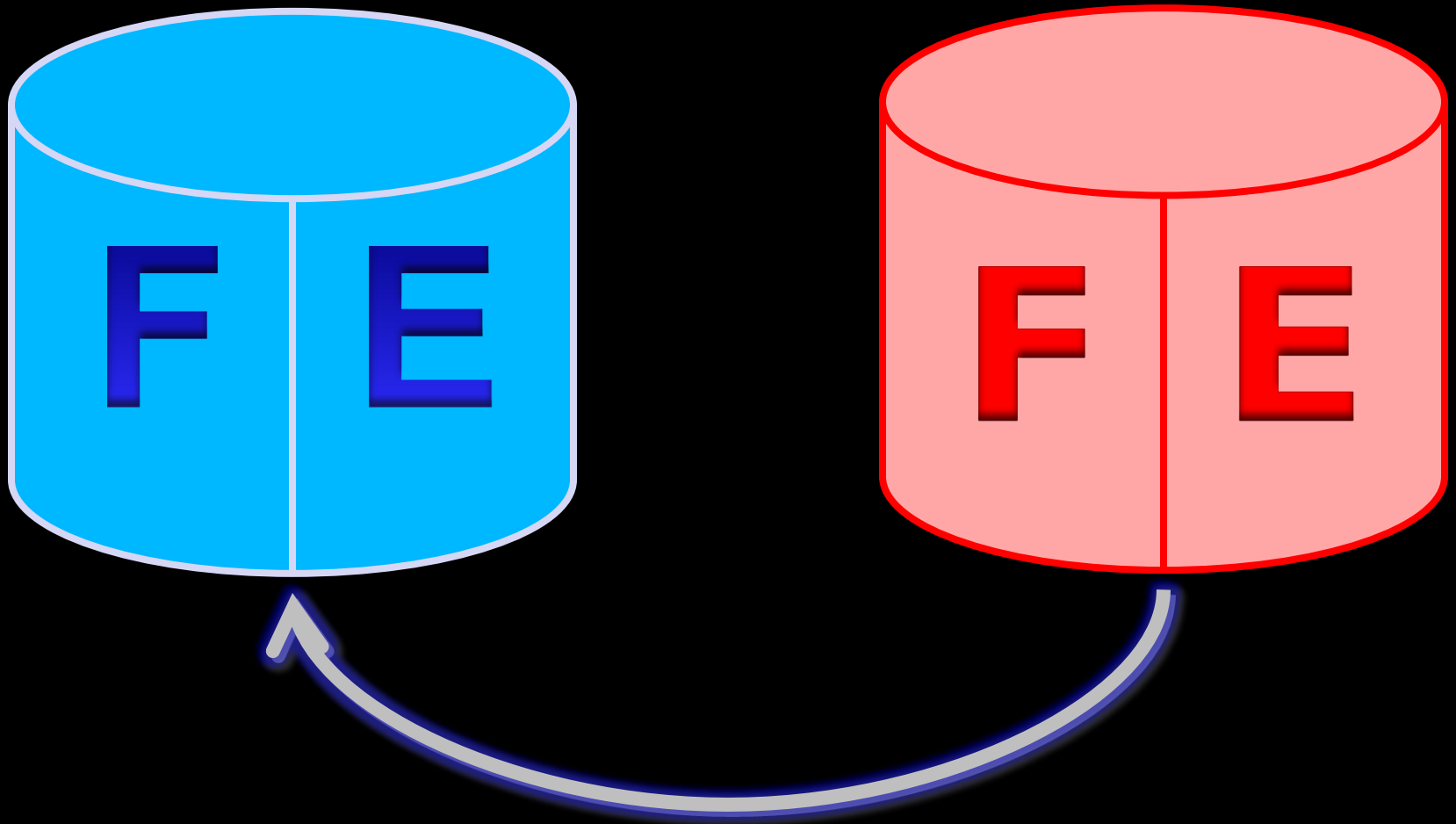
Key Question: What went wrong?

S^4 taxonomy of adaptation effects

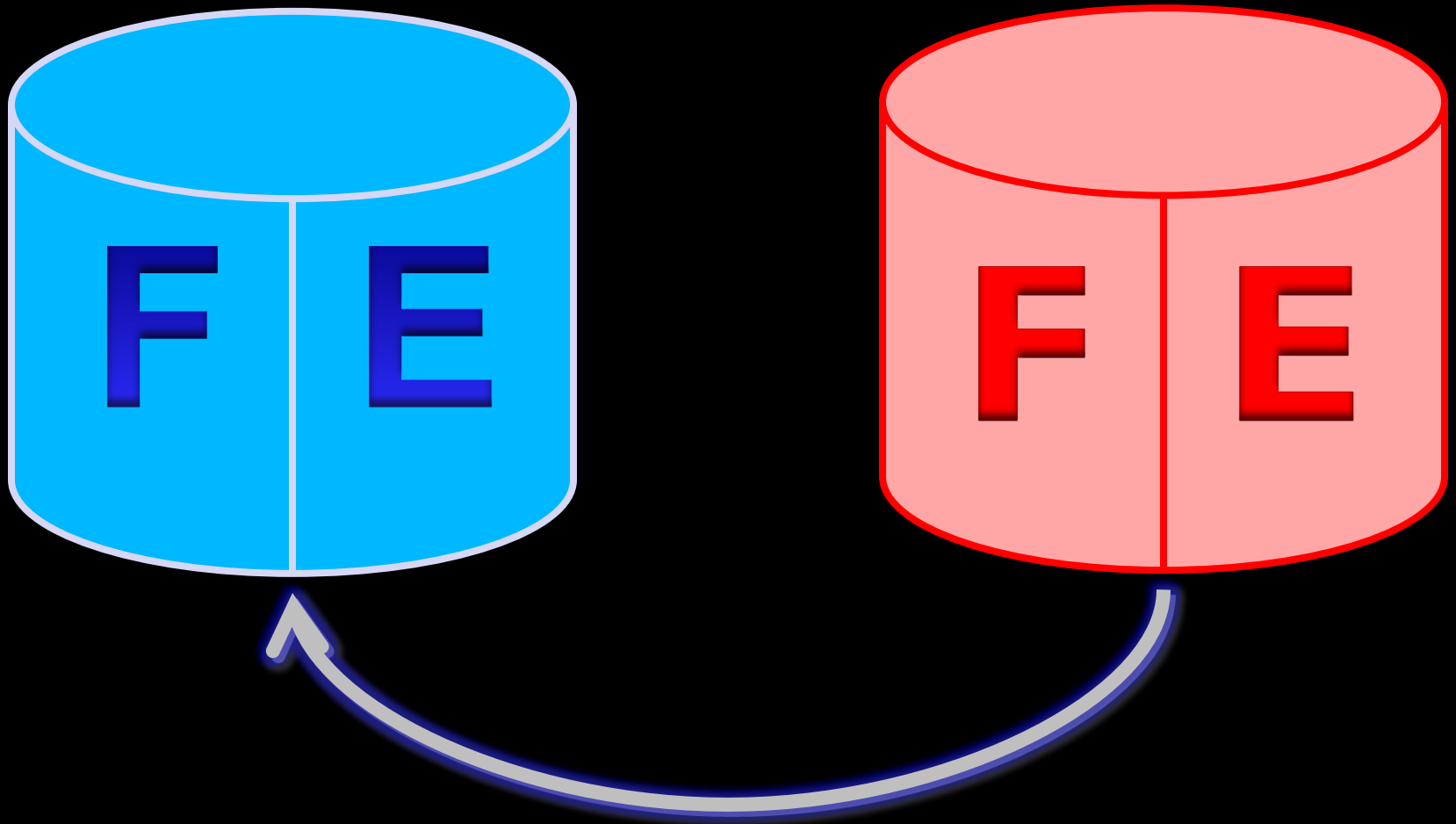
- **Seen:** Never seen this word before
  - News to medical: "diabetes mellitus"
- **Sense:** Never seen this word used in this way
  - News to technical: "monitor"
- **Score:** The wrong output is scored higher
  - News to medical: "manifest"
- **Search:** Decoding/search erred

Working with *no* new domain parallel data!

Measuring SEEN effects

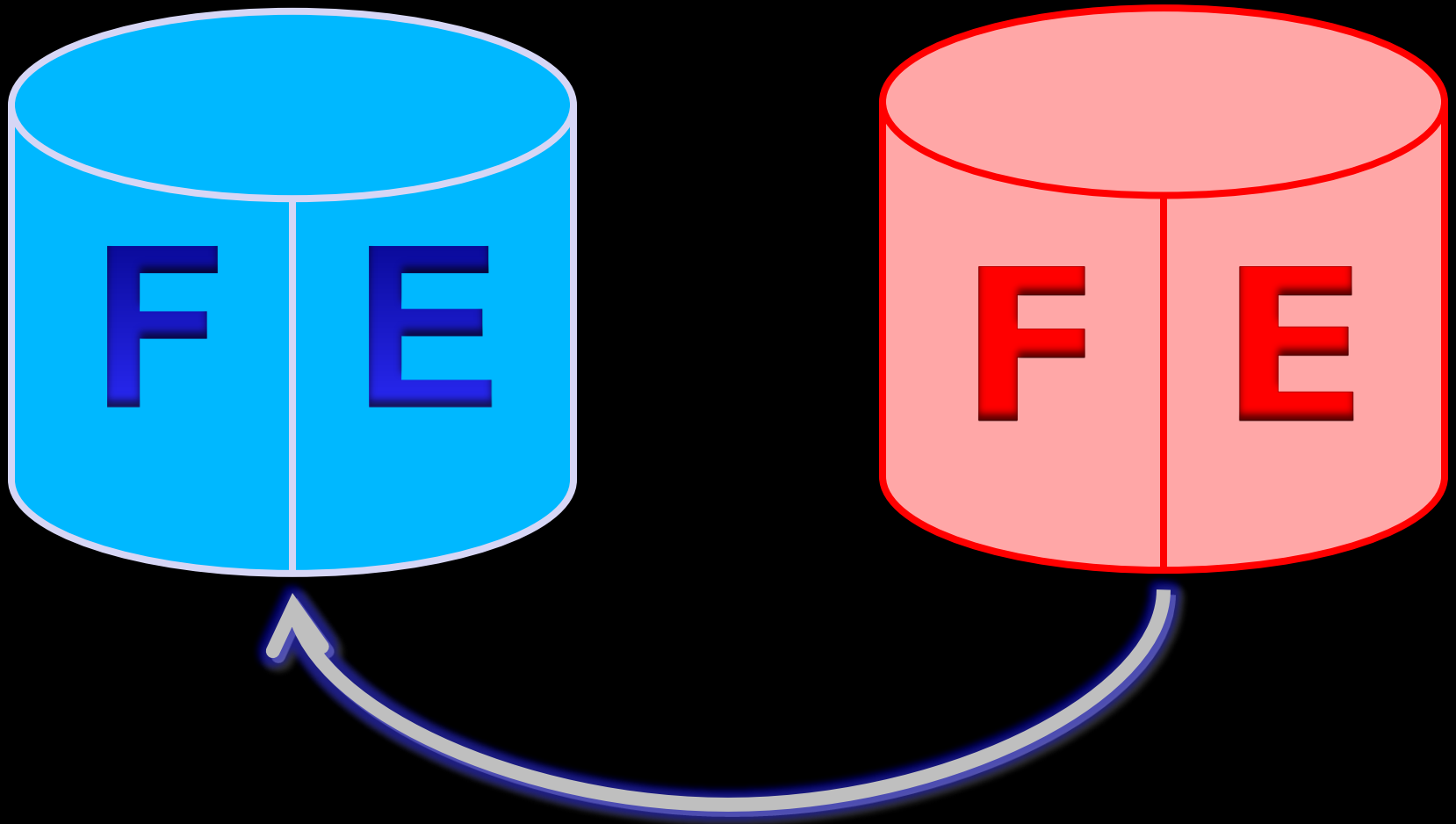


Measuring SENSE effects



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)

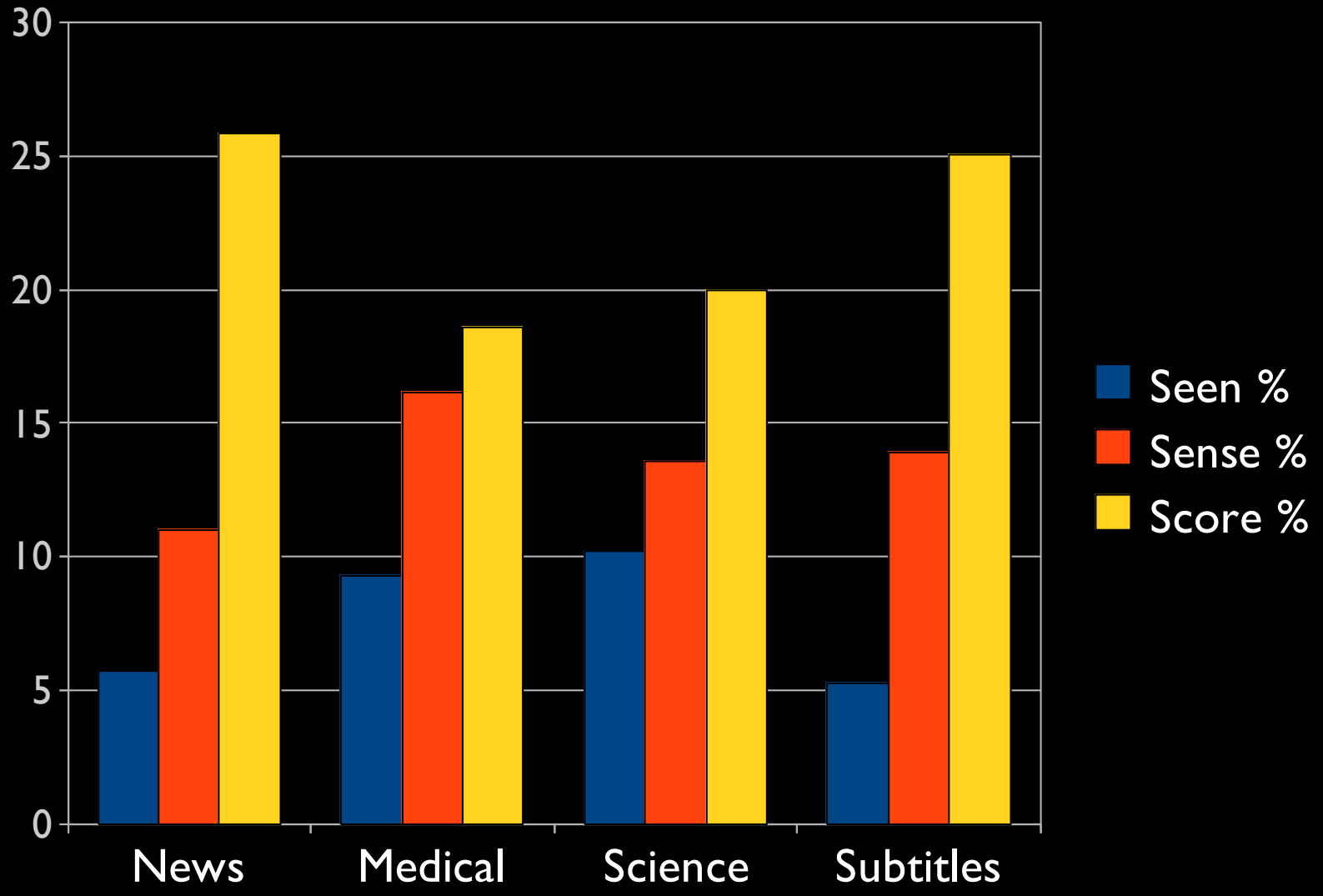
Macro-analysis of S^4 effects

- Evaluation using BLEU

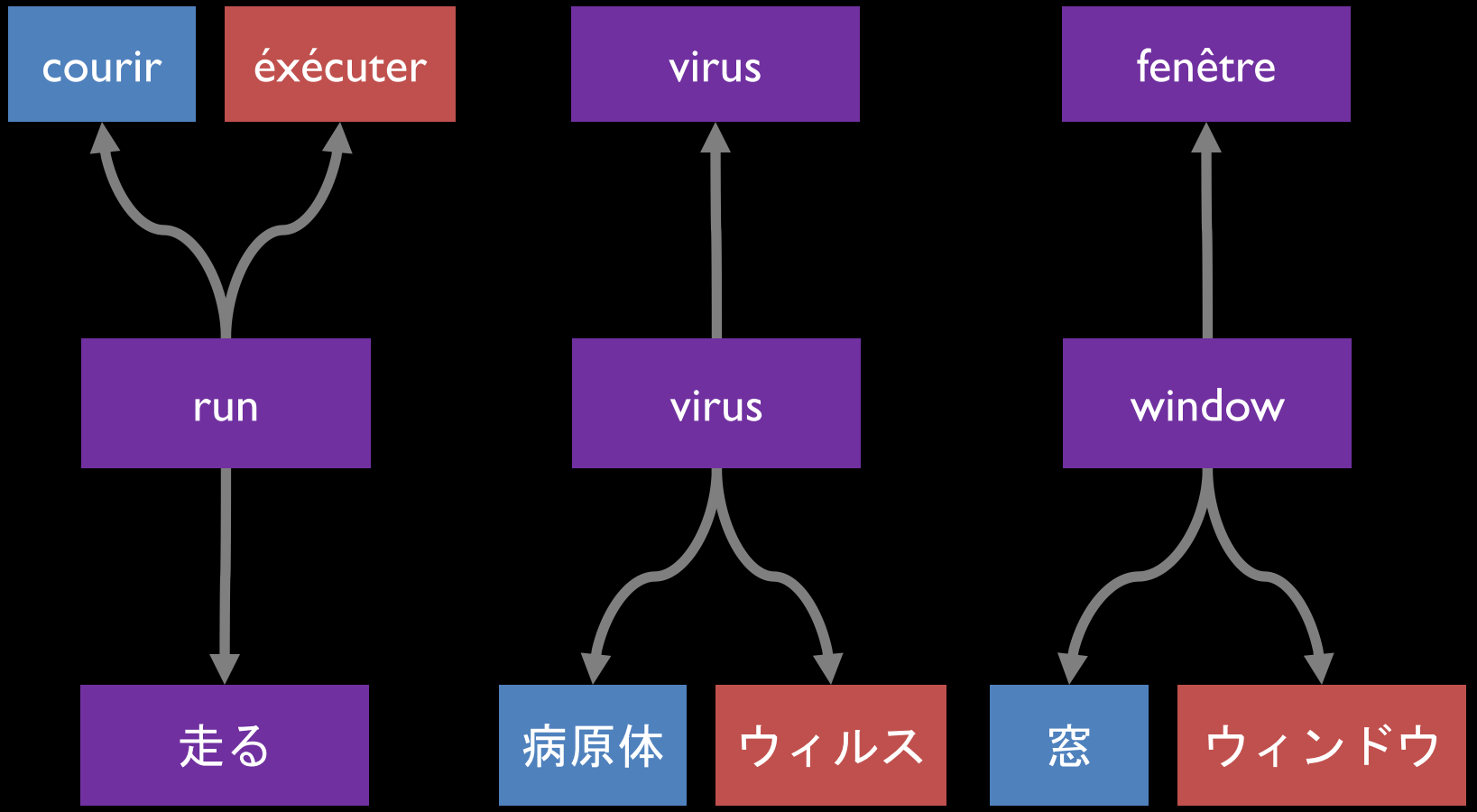
	News	Medical	Science	Subtitles
Seen	+0.3%	+8.1%	+6.1%	+5.7%
Sense	+0.6%	+6.6%	+4.4%	+8.7%
Score	+0.6%	+4.5%	+9.9%	+8.4%

- 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

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-regularized 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

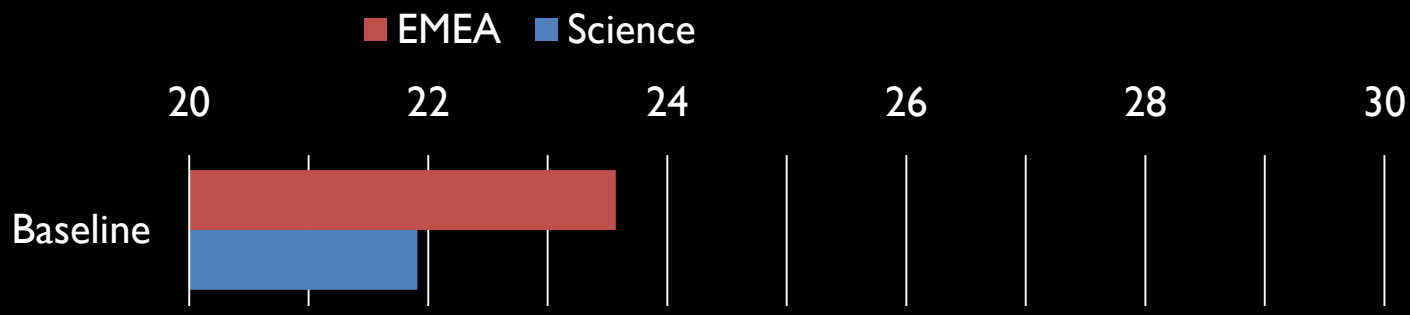
E
the
…
spiders
…

F
le
…
araignées
…

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

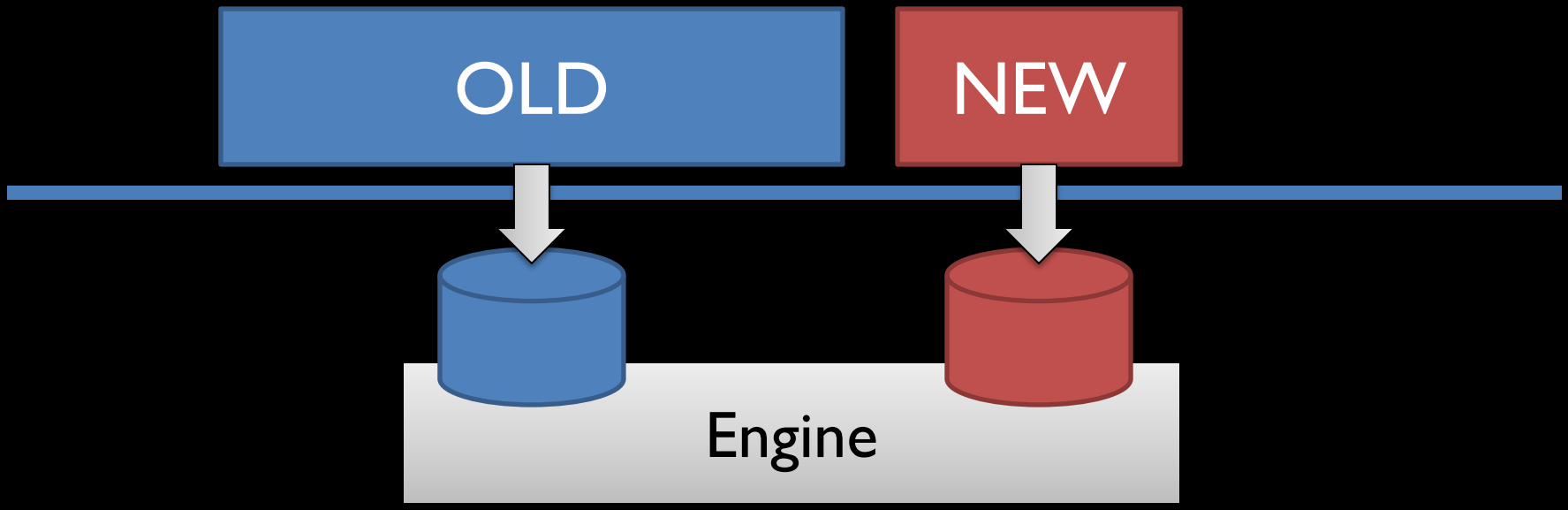


Case 2: Add NEW domain parallel data

- 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)



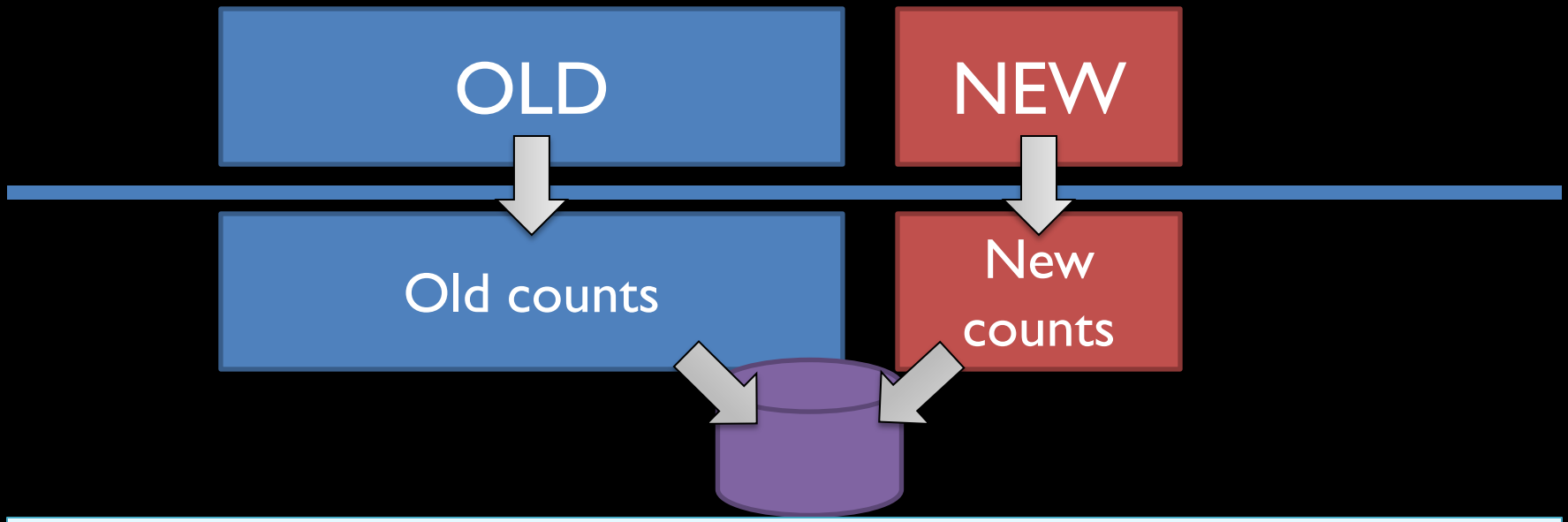
Baseline:

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

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:

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

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)

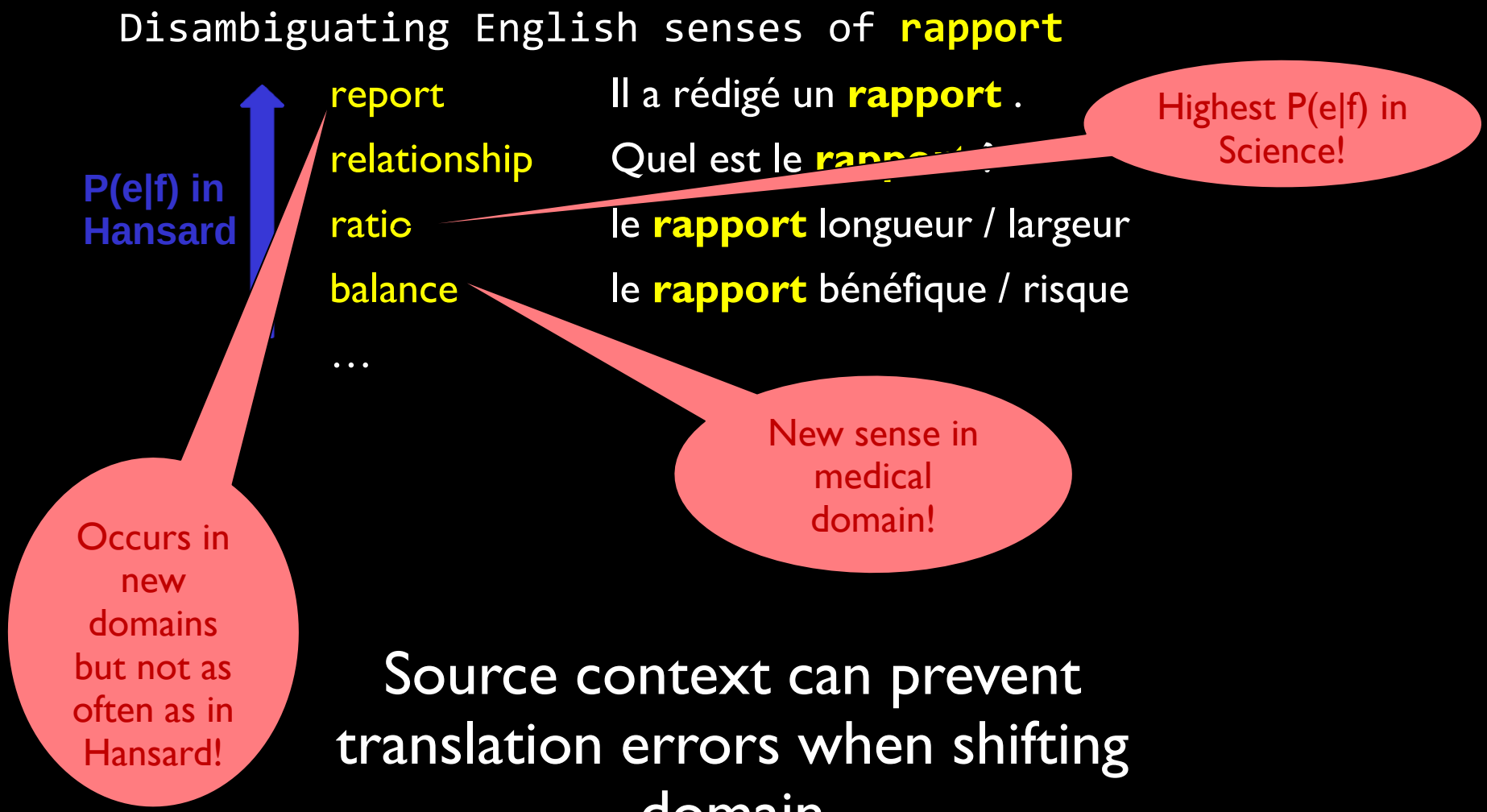
Proposed solution: Phrase Sense Disambiguation

[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

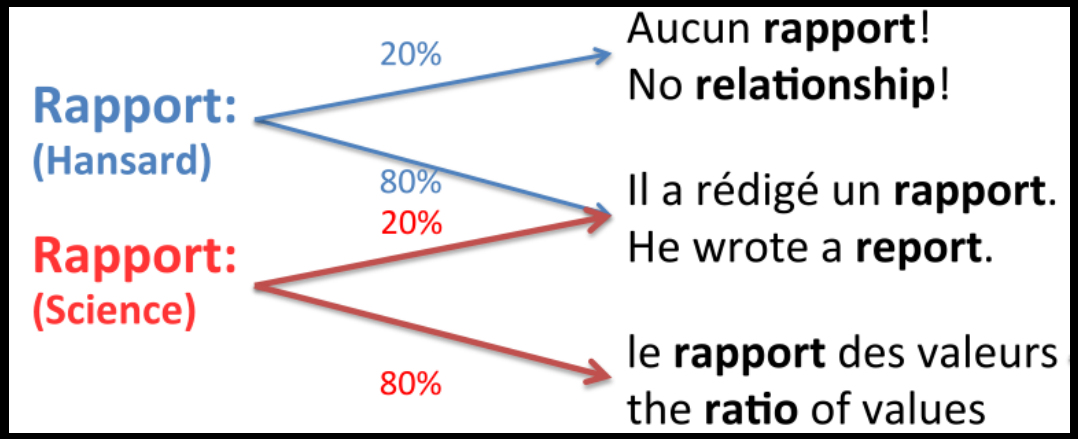
[Chan et al. 2007, Stroppa et al. 2007, Gimenez & Màrquez 2008, Jeong et al. 2010, Patry & Langlais 2011, ...]

Why PSD for domain adaptation?



Source context can prevent translation errors when shifting domain

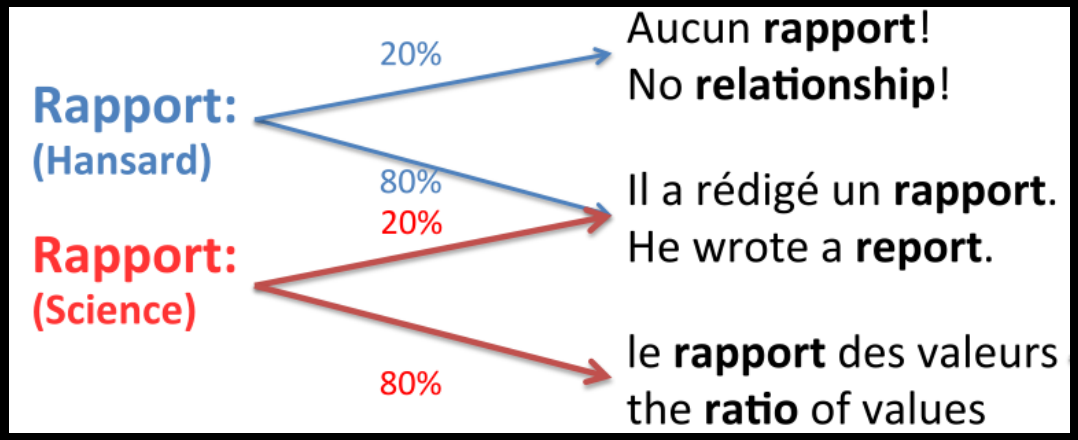
Phrase Sense Disambiguation



- 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

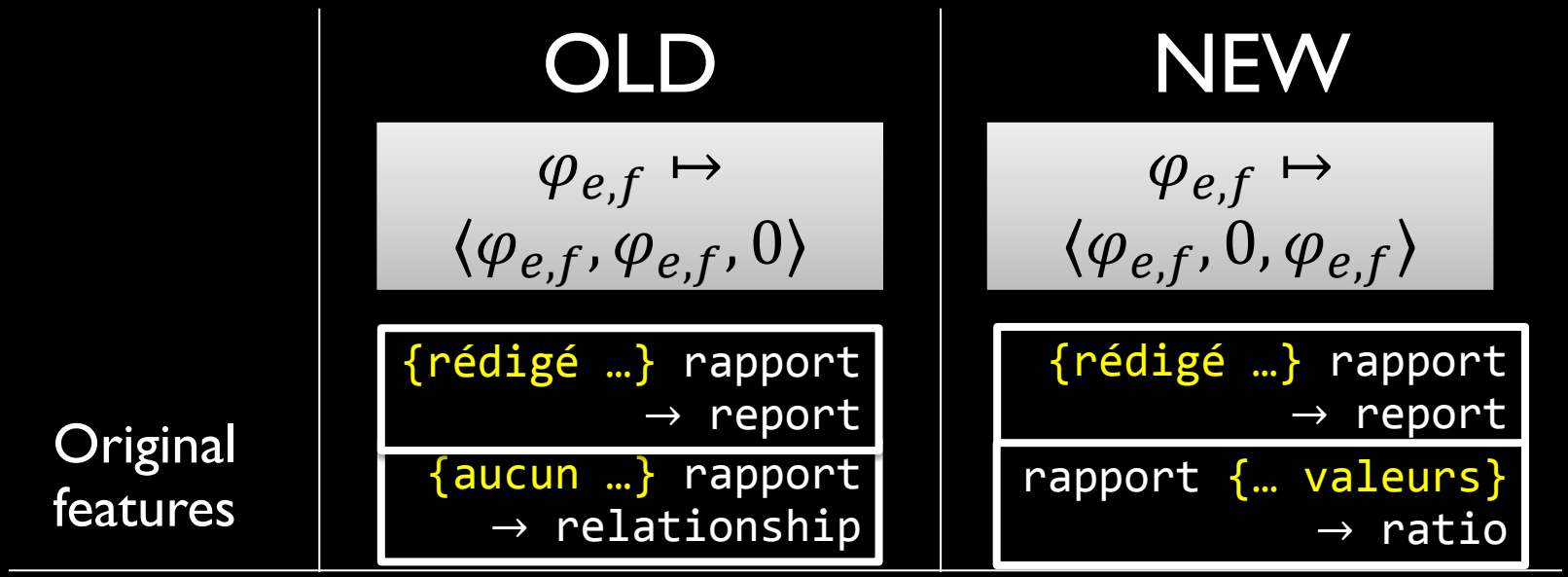
[Blitzer & Daumé 2010]

Domain adaptation in PSD

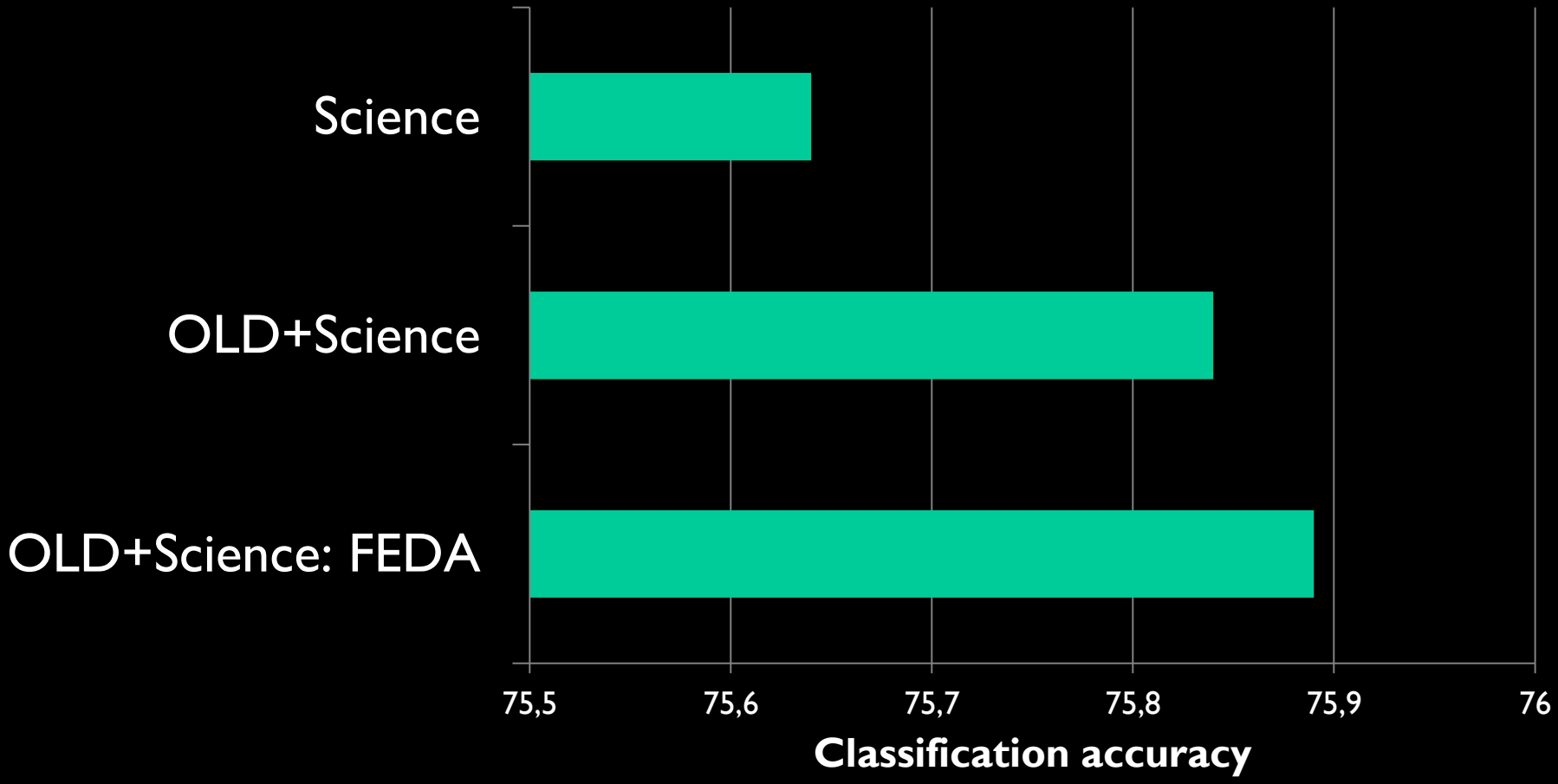


- Train a classifier over OLD and NEW data
- Allow classifier to:
  - share some features
    {rédigé …} rapport → report
  - keep others domain specific
    rapport {… valeurs} → ratio

Feature augmentation



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

Discussion

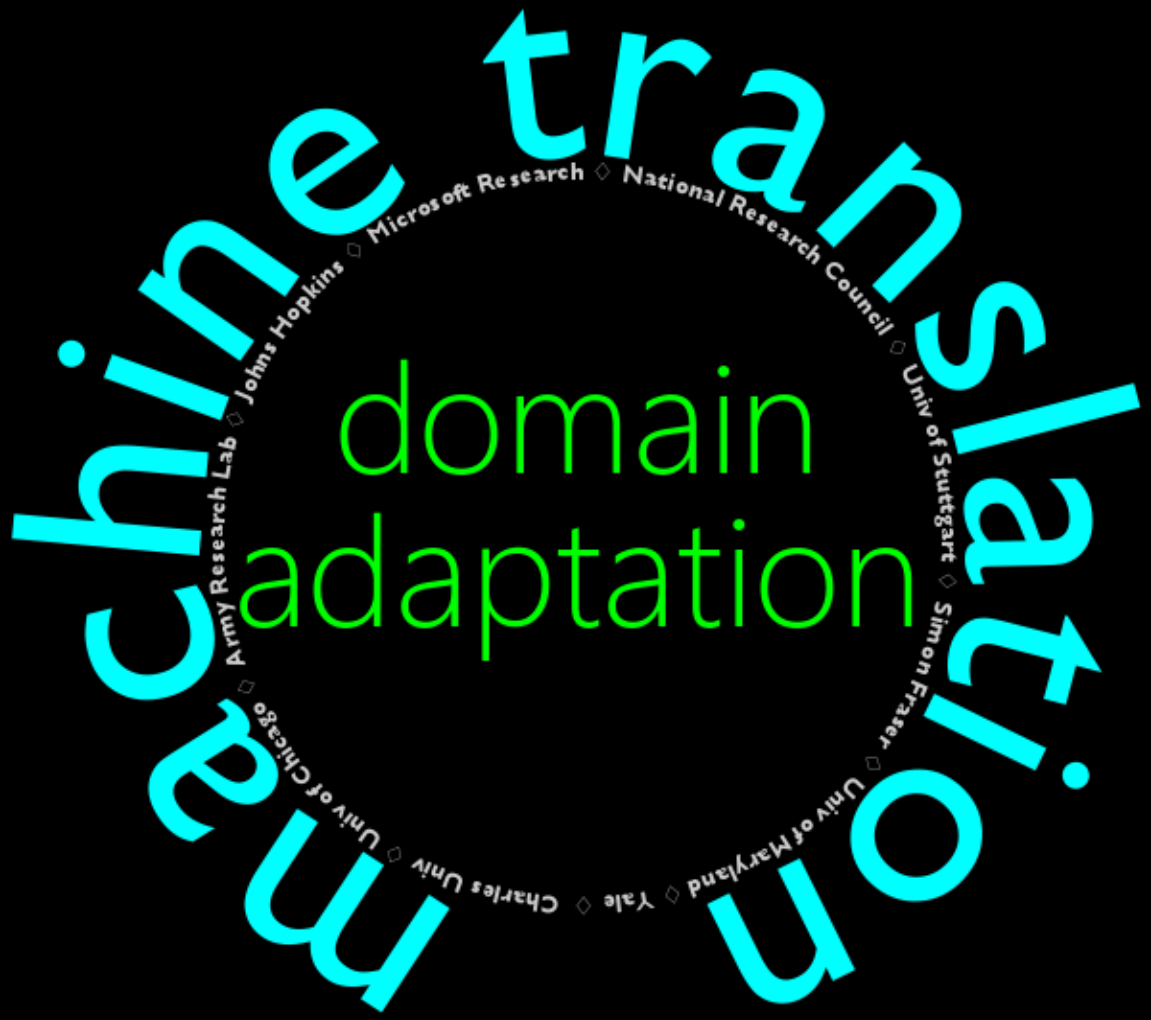
- Introduced taxonomy and measurement tools for adaptation effects in MT
- "Score" errors – target of prior work – only a part of what goes wrong
- Marginal matching introduced as a model for addressing *all* S^4 issues simultaneously: +2.4 BLEU
- Data and outputs released for you to use (both in MT 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

Hal Daume
U Maryland

Chris Quirk
MSR

Thank you!



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
Everyone at CLSP

Eighth Machine Translation Marathon
Charles University, Prague
September 13th, 2013