

# machine translation

domain adaptation

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

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

George Foster

# Translating across domains is hard

## Old Domain (Parliament)

<b>Original</b>	monsieur le président, les pêcheurs de homard de la région de l'atlantique sont dans une situation catastrophique.
<b>Reference</b>	mr. speaker, lobster fishers in atlantic canada are facing a disaster.
<b>System</b>	mr. speaker, the lobster fishers in atlantic canada are in a mess.

## New Domain

<b>Original</b>	comprimés pelliculés blancs pour voie orale.
<b>Reference</b>	white film-coated tablets for oral use.
<b>System</b>	white <b>pelliculés</b> tablets to oral.

## New Domain

<b>Original</b>	mode et voie(s) d'administration
<b>Reference</b>	method and route(s) of administration
<b>System</b>	<b>fashion</b> and <b>voie(s)</b> of <b>directors</b>

**Key Question: What went wrong?**

# Goals of workshop

- **Understand domain divergence in parallel data and build models to improve cross-domain translation quality**
- Analyze data
  - Identify lexical divergences across domains
- Domain adaptation for phrase sense disambiguation
  - Build adaptable phrase- and Hiero-based systems to new domains
  - Find useful context features (beyond sentence level)
  - Discover domains from large heterogeneous corpora
- Translation/sense discovery
  - Design algorithms for spotting new senses
  - Learn new translations for them

# Background: DA in SMT

- Optimistic assumptions about domain
  - **new** parallel data available for training
  - not too divergent from **old** (Europarl to News)
- Past Approaches [FGK10]
  - Concatenate **old** + **new** data
    - Doesn't usually help
    - Can hurt if **old** is large and very different from **new**
  - Mix **old** + **new** model
    - Doesn't hurt
    - But crude: entire **old** corpus is uniformly down-weighted
  - Sentence weighting
    - Find sentences in **old** that are more similar to **new**
    - Still too coarse-grained

# Limitations of past research

- Understanding the translation adaptation problem:
  - Universally focuses on lexical choice
  - Sense divergence is ignored
  - Focuses on non-representative data
- Building adaptable translation models:
  - Can (*mostly*) only reweight existing translation candidates
  - Cannot extend to new word senses
  - Ignores (large) document context
- Methodology for statistical domain adaptation:
  - Assumes all possible “labels” are observed old domain data
  - Works on labeled (“parallel”) or unlabeled (“monolingual”) data, does not extend to “comparable” data

# Senses are domain/language specific

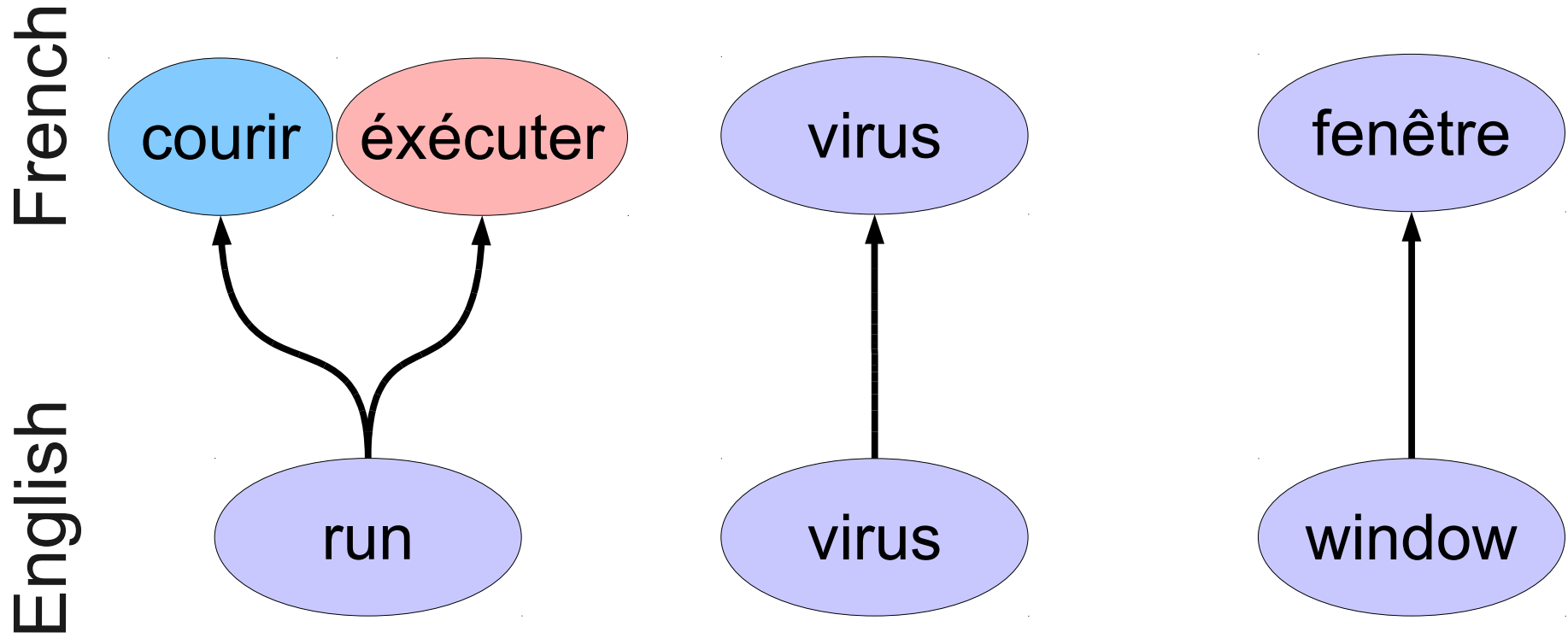
English

run

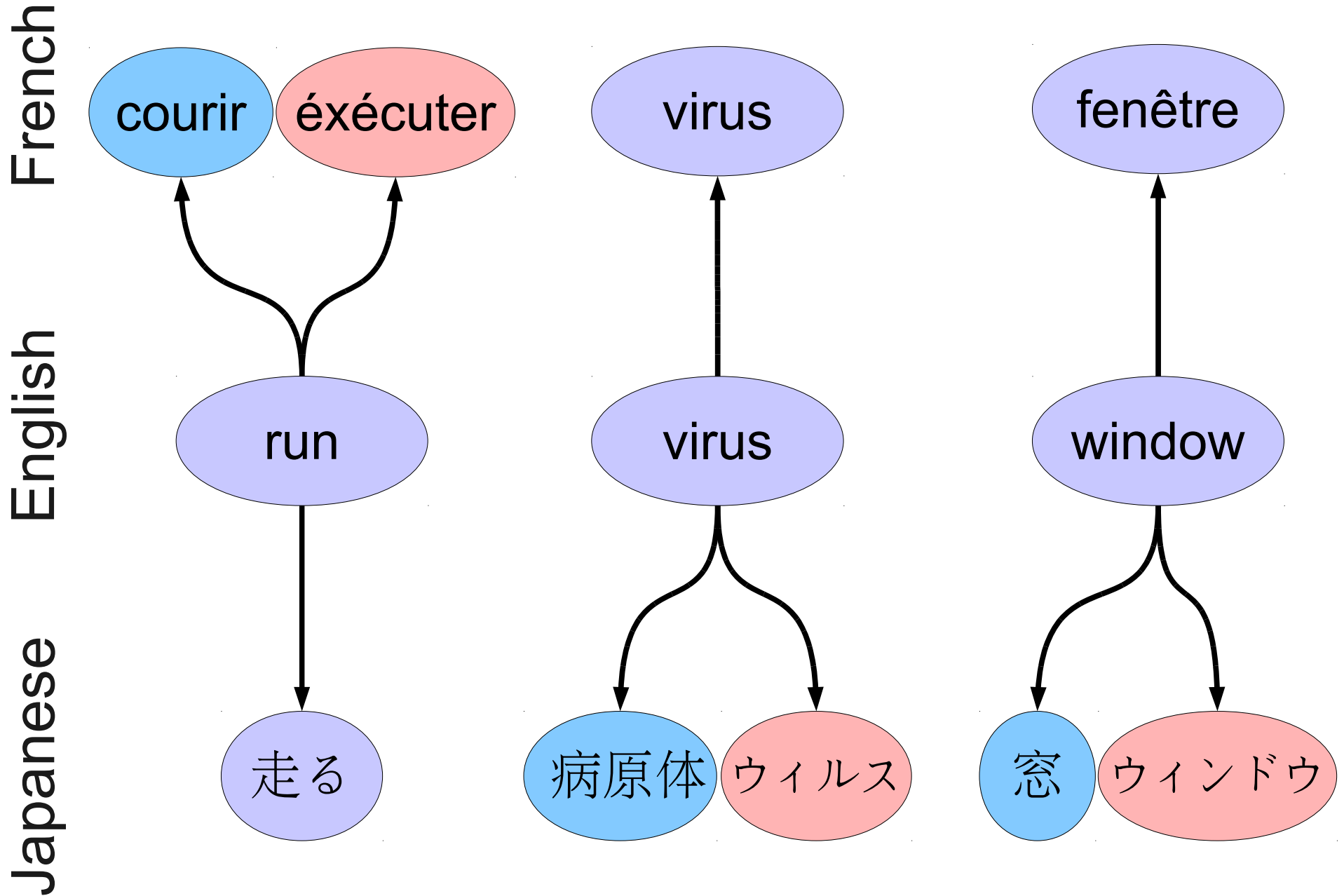
virus

window

# Senses are domain/language specific

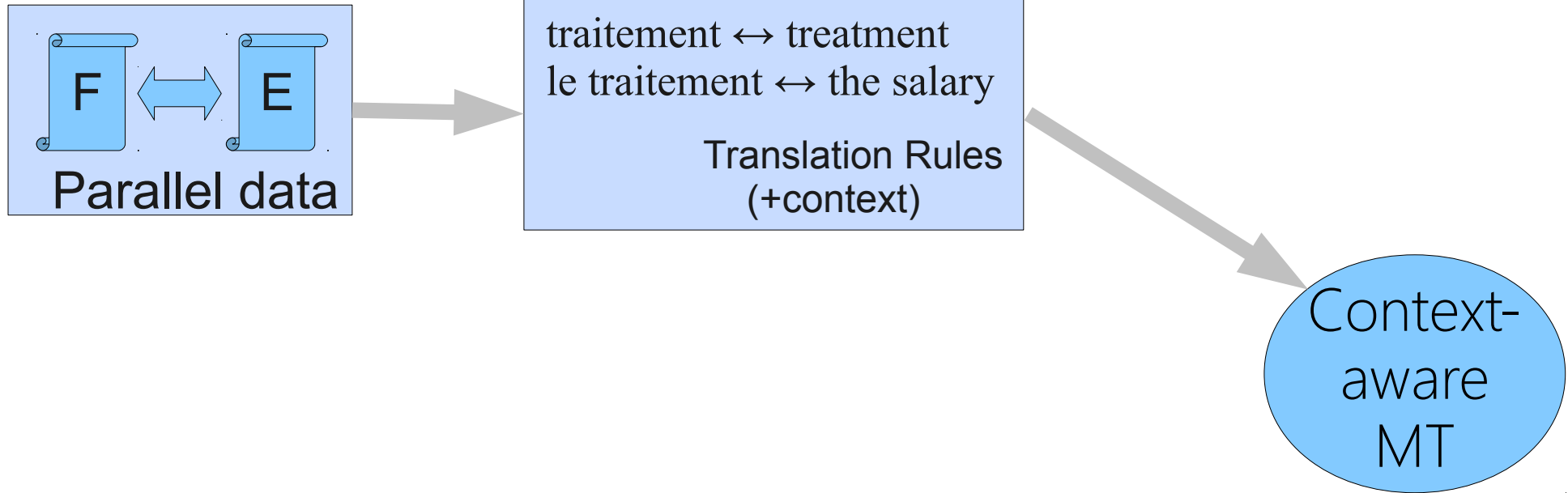


# Senses are domain/language specific

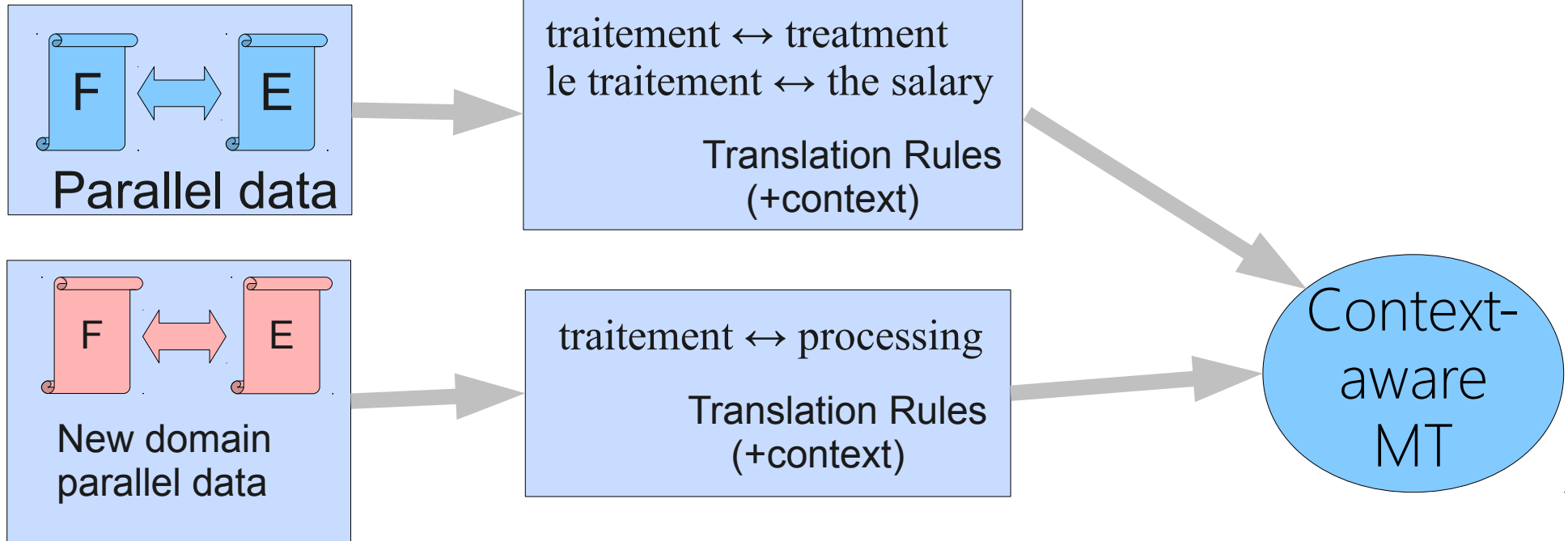




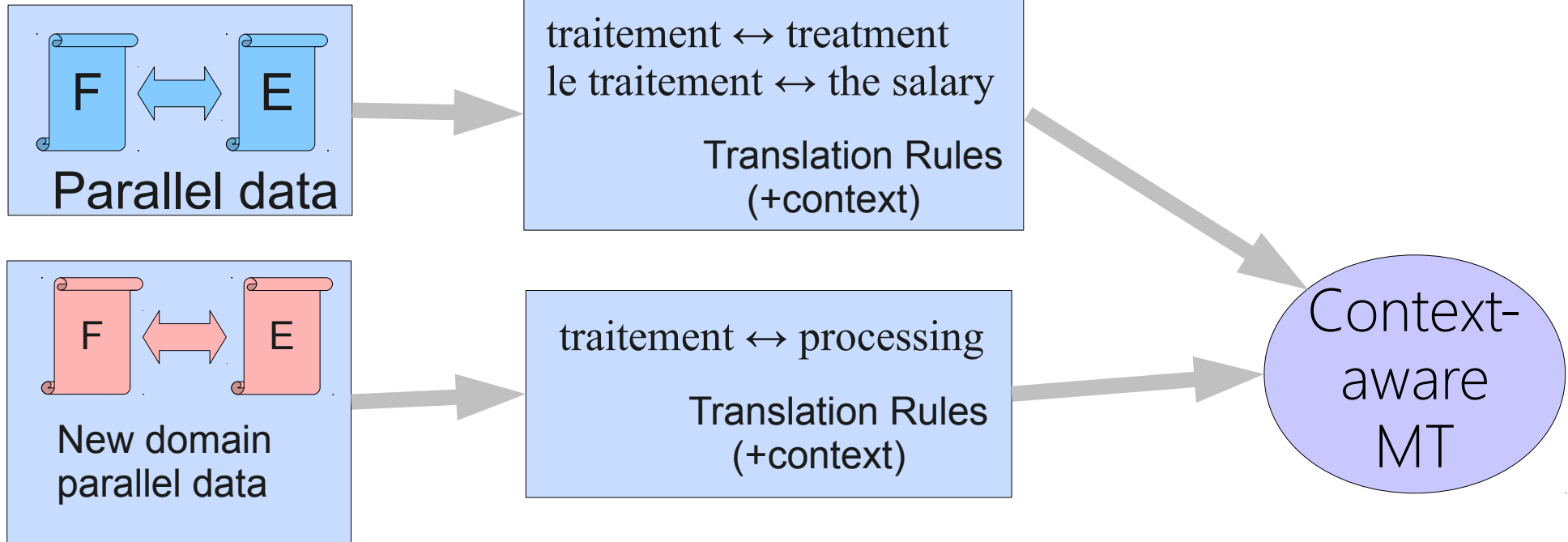
# Approach



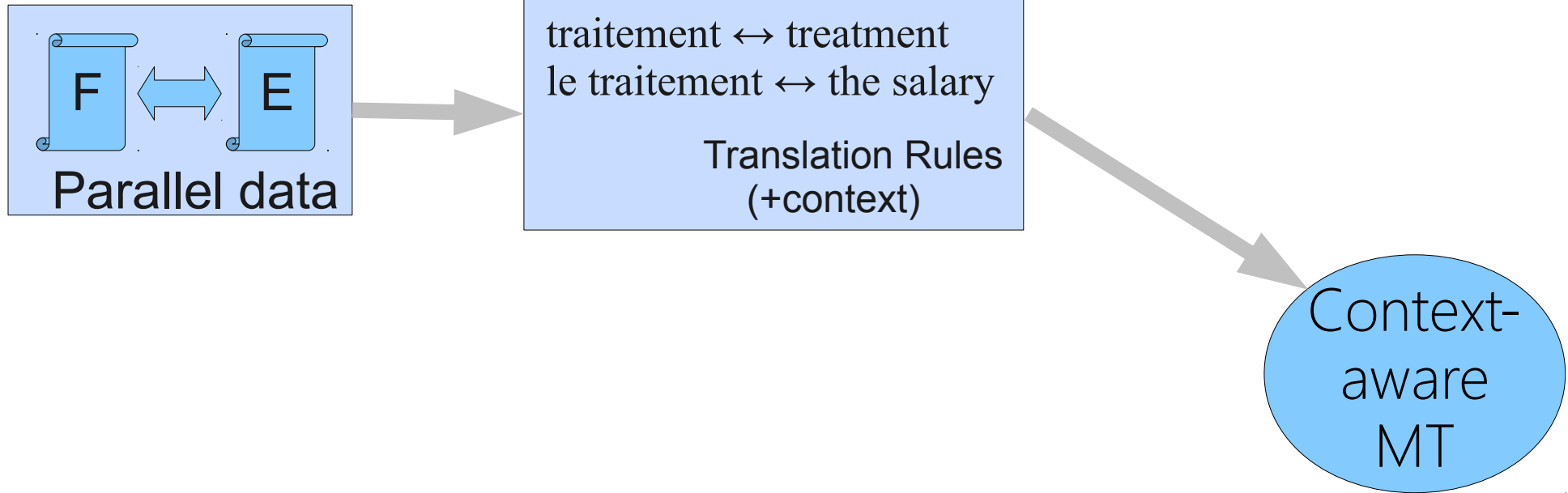
# Approach



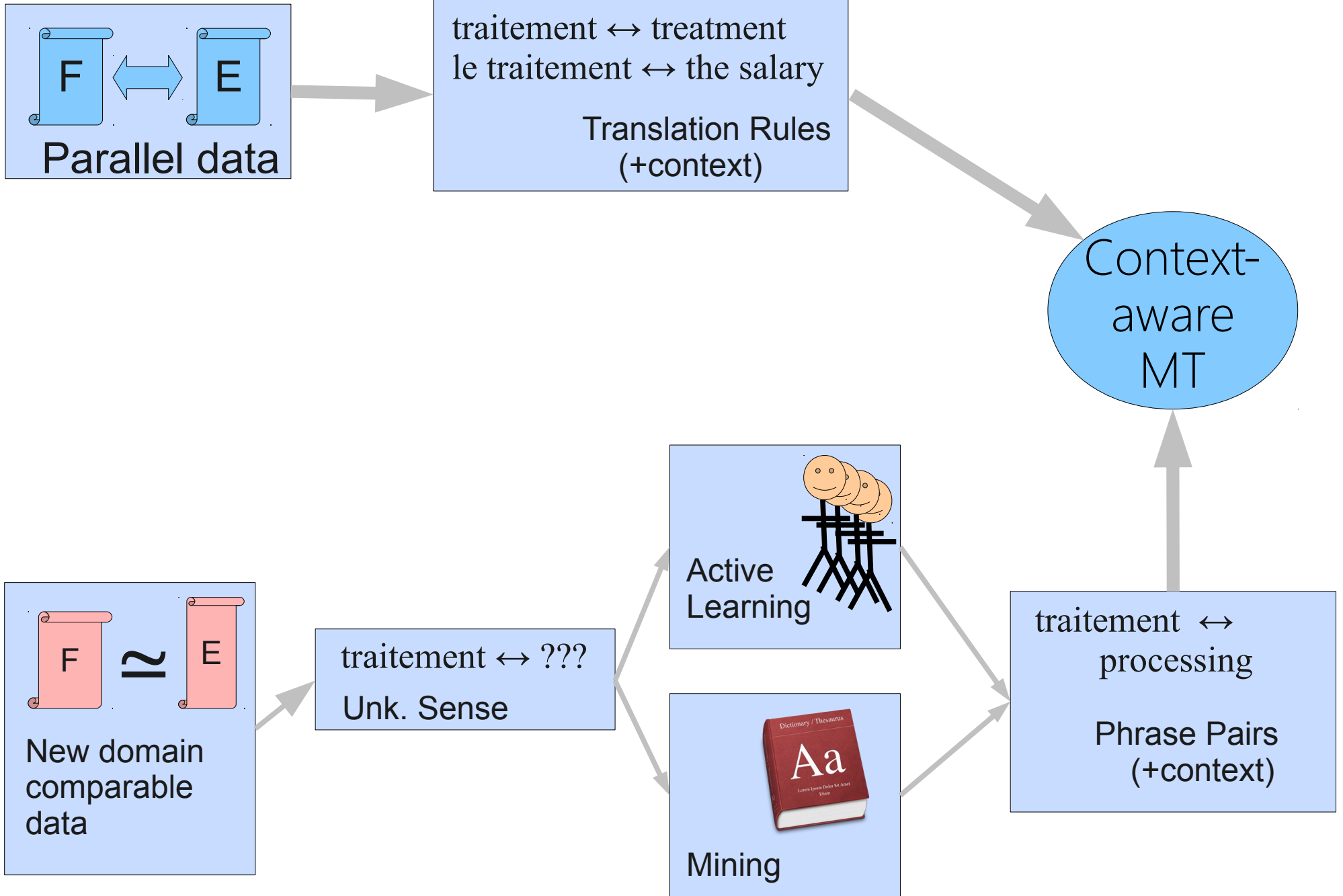
# Approach



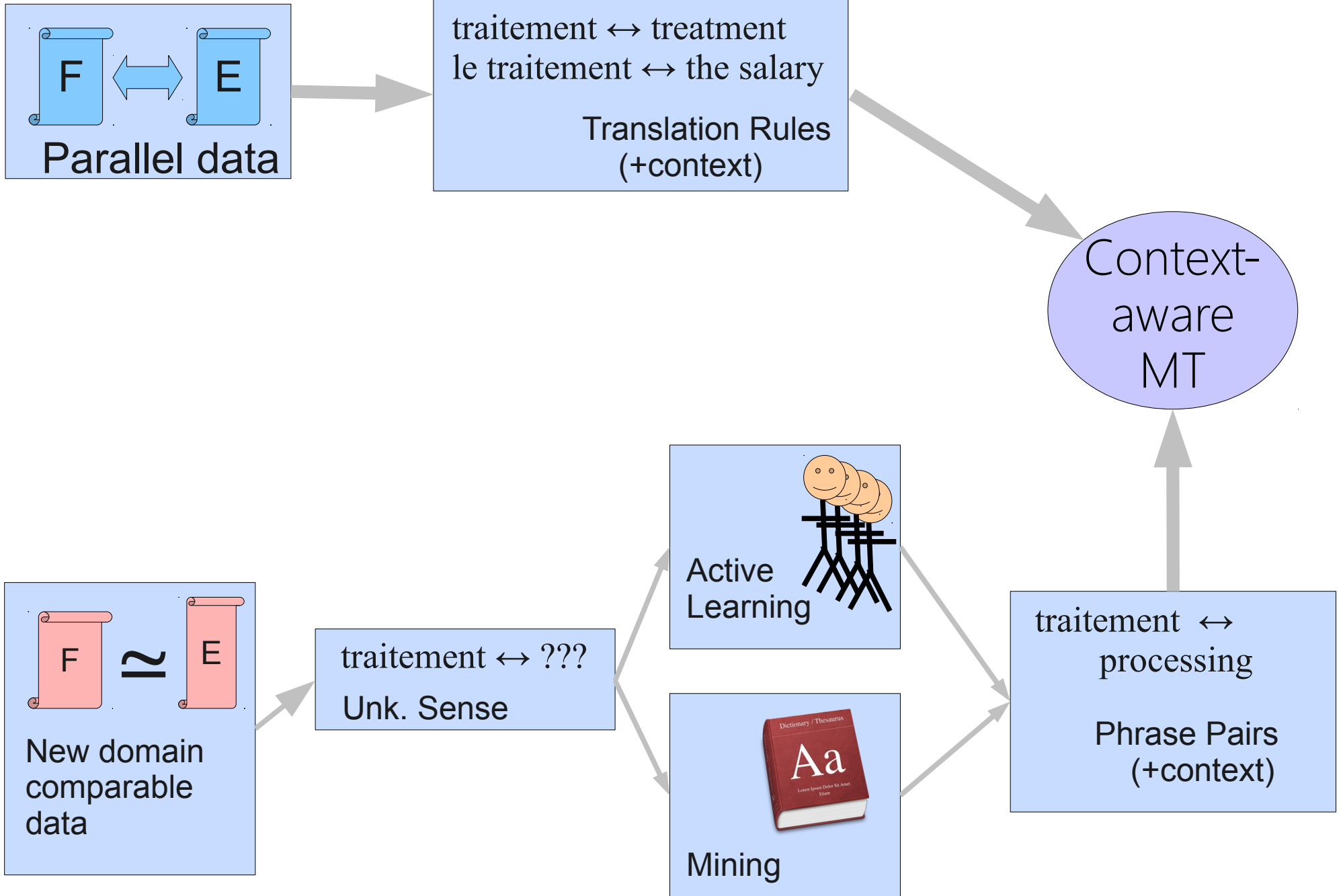
# Approach



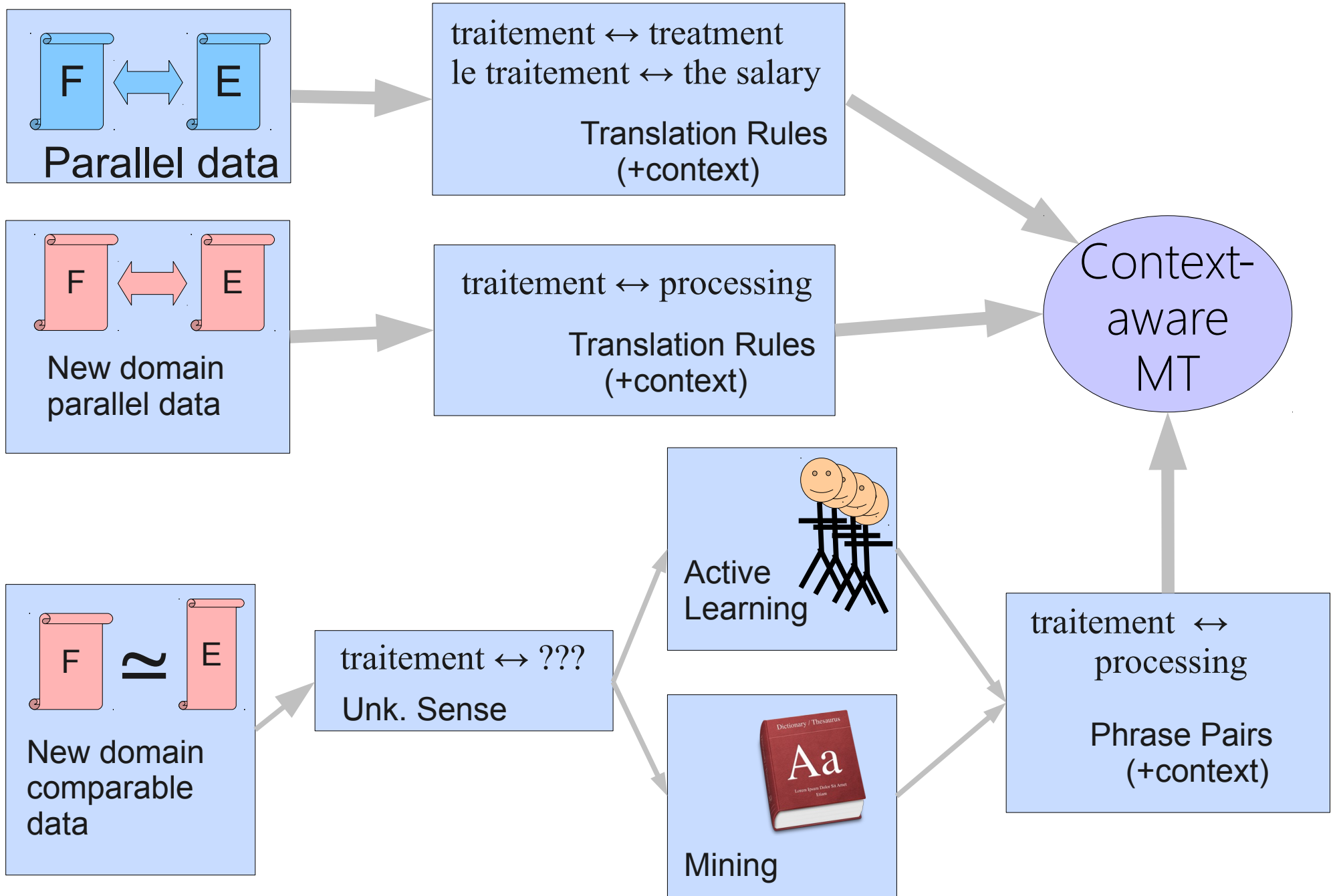
# Approach



# Approach



# Approach



# Goals I: Framework for adaptation

- **Create standardized conditions for MT adaptation**
  - Resources available to other researchers
  - Understanding of intricacies of domains
  - Methodology for and analysis of adaptation effects
- **Develop intrinsic lexical choice accuracy task**
  - Given a source phrase in context, predict correct translation
  - Annotated data released in old domain and all new domains
  - Variety of conditions and experimental setups
- **Automatic translation quality evaluation**
  - Using standard metrics (Bleu, Meteor)
  - Compare performance before and after adaptation
  - New domain parallel data vs. only new domain comparable data





# Goals II: Algorithms

- **Context-sensitive discriminative translation**
  - Fully integrated in open-source MT system Moses
  - Algorithms to adapt discriminative translation to new domains
  - Adapted models for phrase- and Hiero-based systems
  - Find useful features for these systems
- **Discover new senses and their translations**
  - Algorithms for spotting new senses (applies beyond MT)
  - Algorithms for discovering subdomains (applies beyond MT)
  - Discover new translations for these senses
    - Human-based active learning
    - Fully automatic dictionary mining






# How you will spend your afternoon...

- Analysis of data

- About the data 
- Errors of MT systems 





Chris Quirk  
John Morgan, Anni Irvine

- Discriminative models for lexical selection

- Overview of translation via classification 
- Lexical selection as a stand-alone task  
- Lexical selection in MT  
- Adaptation experiments

Alex Fraser  
Katie Henry  
Ales Tamchyna, Fabienne Braune  
Majid Razmara

- Spotting new senses and their translations

- Overview and new techniques 
- Spotting new senses  
- Topic models and parallel data 

Anni Irvine  
Rachel Rudinger  
Ann Clifton, Jagadeesh Jagarlamudi

- Wrap-up

- Conclusions and future work
- Questions and answers

Marine Carpuat  
all of you and all of us

Chris Quirk

# Outline

- Introduction
- **Analysis**
  - Domains: examples, sizes, and overlap
  - Baseline and simple adaptation results
    - BLEU, lexical choice
  - Error analysis with S4 (before adaptation)
  - New diagnostic metric, Sanjeeval
- PSD for domain adaptation
- Mining new terminology
- Conclusion

# Language pair

- French to English
  - SMT systems work well on this language pair...  
...which can be a liability
  - Lots of OLD domain data
  - Many NEW domains possible
  - Several speakers on the team
- Techniques should not be language specific

# Stereotypical domain examples

**Hansards:** Parallel English-French documents from the Canadian government.

Voulez-vous que l'on vote au sujet de la motion modifiée?  
Do we want to vote on the amended motion?

Avalez le comprimé en entier.  
Swallow the tablet whole.

Z0 bosons obtiennent leurs masses de la brisure de la symétrie du vide  
Z0 bosons obtain masses from vacuum spontaneous symmetry breaking

Rocky, l'aliment pour tortues, ça se paye.  
You gotta pay for that turtle food, rock head.

# Stereotypical domain examples

**EMA:** European Medicines Agency. Mostly information about pharmaceuticals.

Voulez-vous que l'on vote au sujet de la motion modifiée?  
Do we want to vote on the amended motion?

Avalez le comprimé en entier.  
Swallow the tablet whole.

Z0 bosons obtiennent leurs masses de la brisure de la symétrie du vide  
Z0 bosons obtain masses from vacuum spontaneous symmetry breaking

Rocky, l'aliment pour tortues, ça se paye.  
You gotta pay for that turtle food, rock head.

# Stereotypical domain examples

**Science:** Abstracts from scientific articles across many domains (computer science, biology, etc.)

Voulez-vous que l'on vote au sujet de la motion modifiée?  
Do we want to vote on the amended motion?

Avalez le comprimé en entier.  
Swallow the tablet whole.

Z0 bosons obtiennent leurs masses de la brisure de la symétrie du vide  
Z0 bosons obtain masses from vacuum spontaneous symmetry breaking

Rocky, l'aliment pour tortues, ça se paye.  
You gotta pay for that turtle food, rock head.



# Stereotypical domain examples

## **Subs:** Parallel movie subtitles.

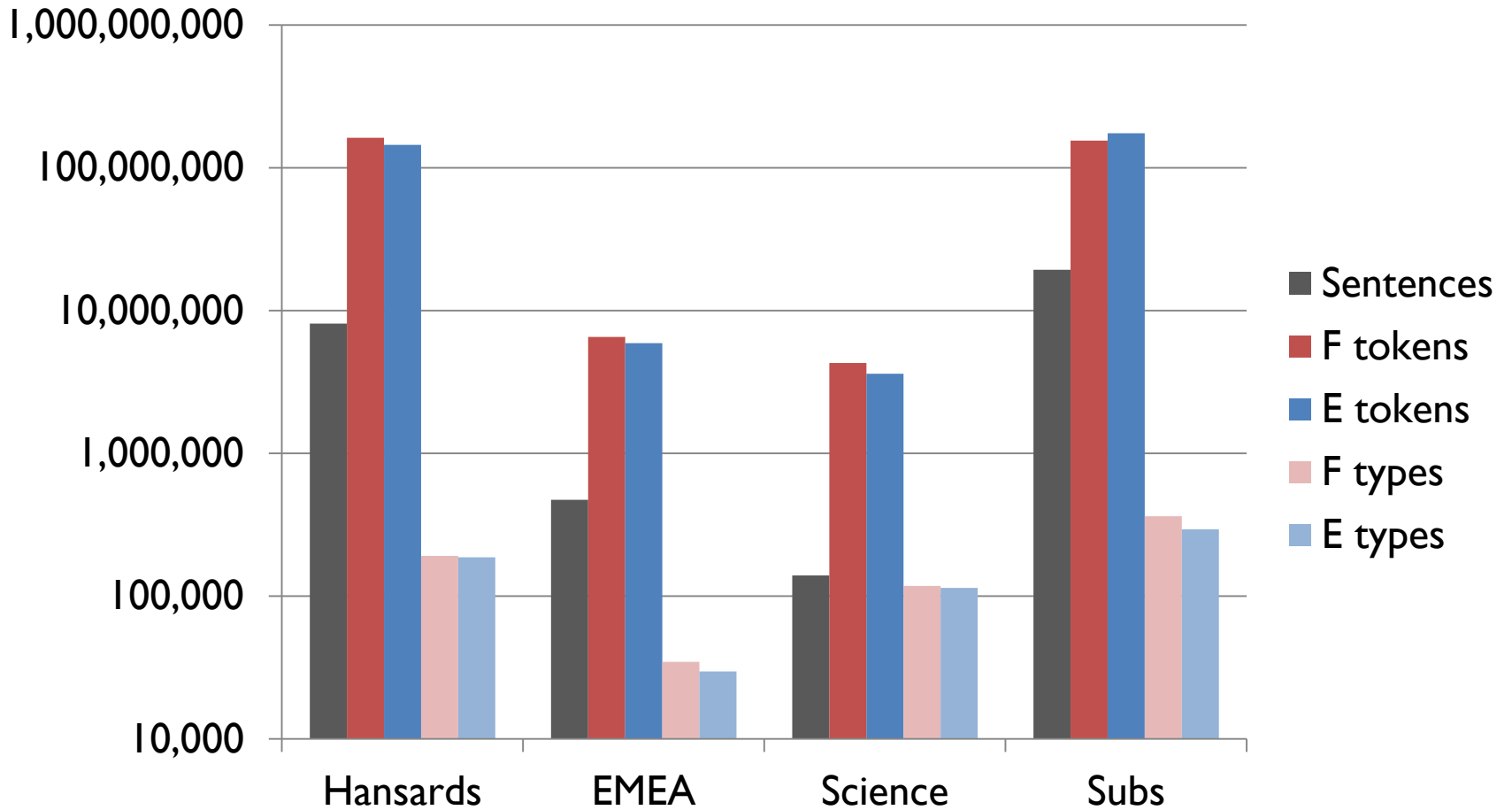
Voulez-vous que l'on vote au sujet de la motion modifiée?  
Do we want to vote on the amended motion?

Avalez le comprimé en entier.  
Swallow the tablet whole.

Z0 bosons obtiennent leurs masses de la brisure de la symétrie du vide  
Z0 bosons obtain masses from vacuum spontaneous symmetry breaking

Rocky, l'aliment pour tortues, ça se paye.  
You gotta pay for that turtle food, rock head.

# Domain sizes



# Measuring domain overlap

- Gauge difficulty of the domain adaptation task
  - Gather information from training data for OLD domain and training data for NEW domain
- What do we measure?
  - Focus here is on unigrams: certainly not sufficient to have unigram coverage, but necessary

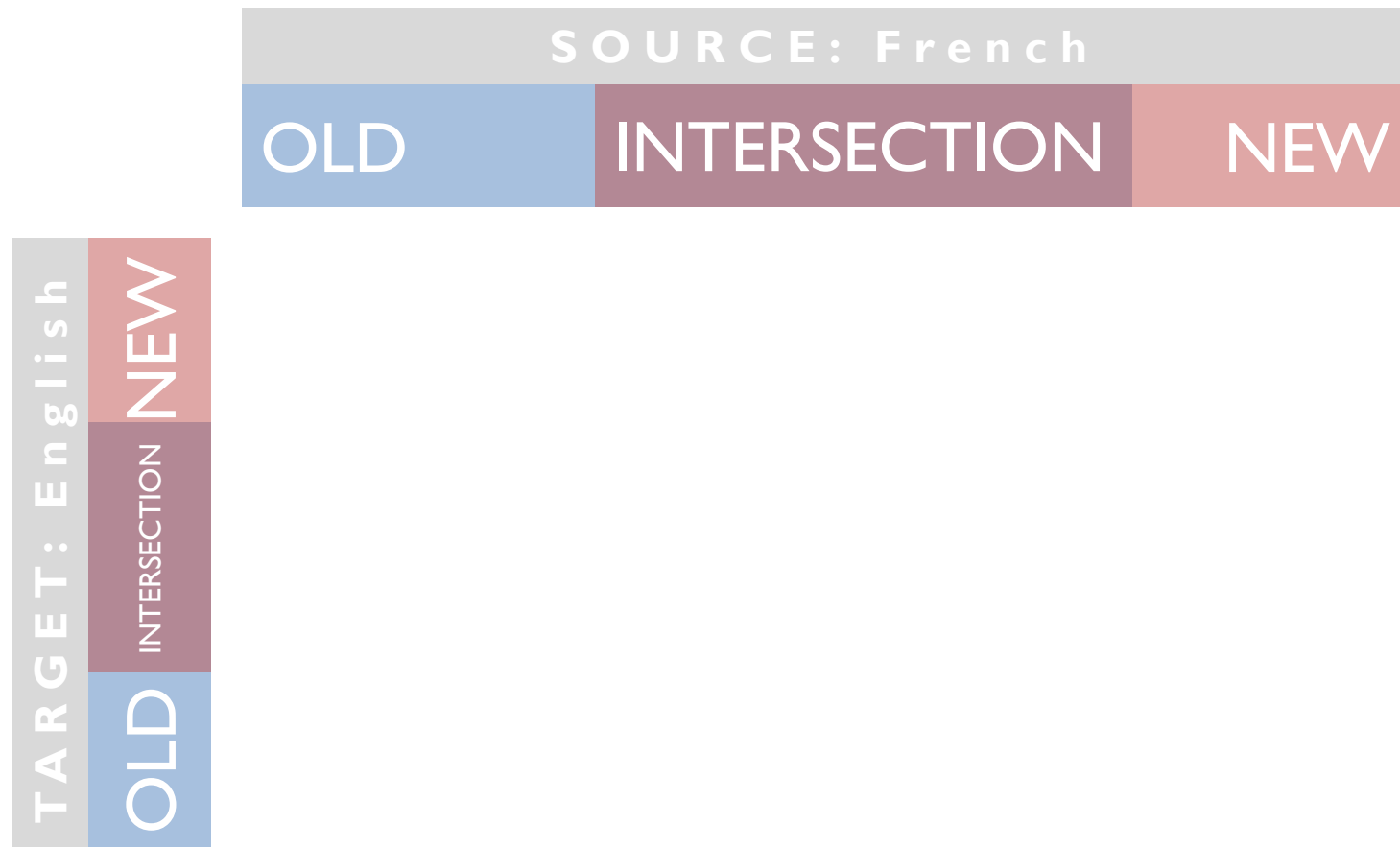
# Measuring domain overlap (cont'd)

- **Multiple possible items to count:**



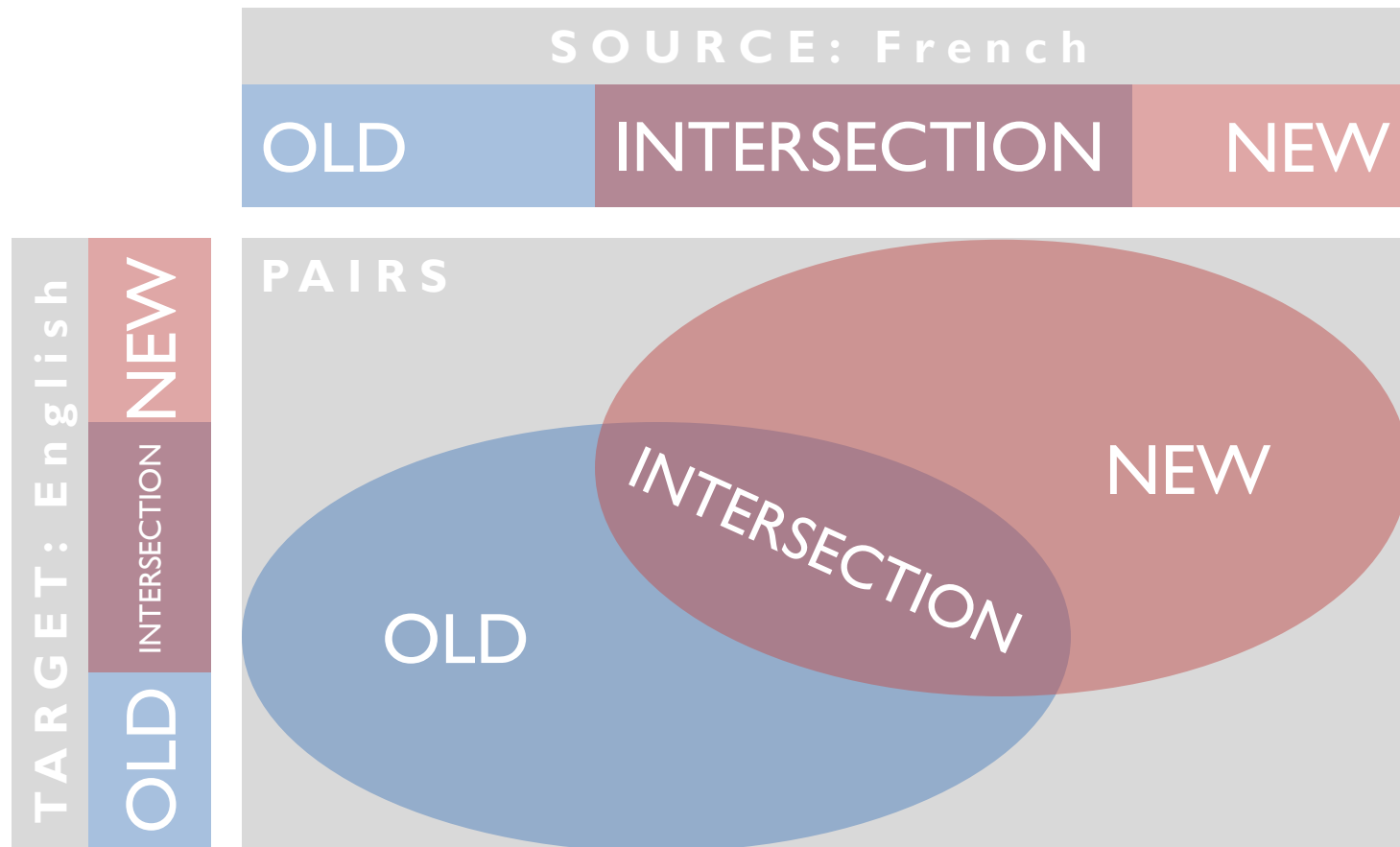
# Measuring domain overlap (cont'd)

- **Multiple possible items to count:**



# Measuring domain overlap (cont'd)

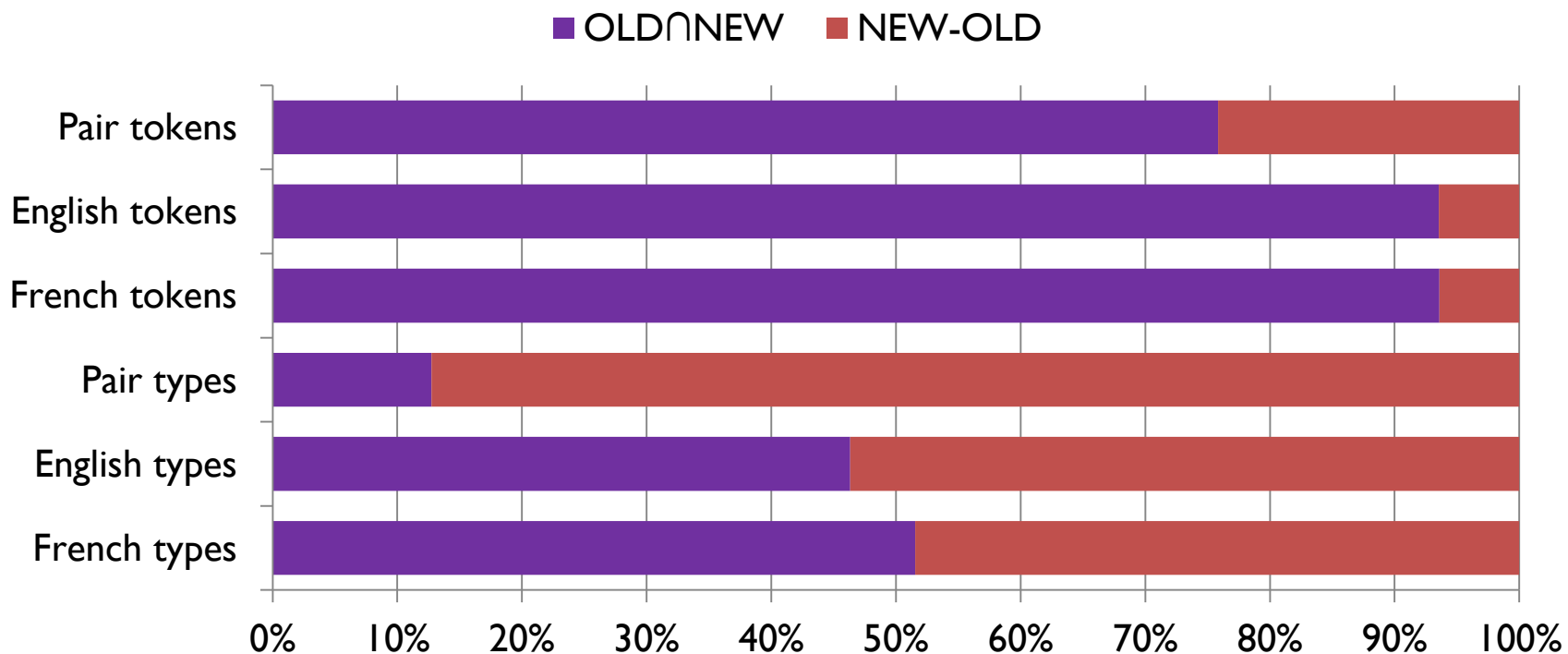
- **Multiple possible items to count:**



# Measuring domain overlap (cont'd)

- Three ways to count
  - **Tokens**: count the number of space-delimited items
  - **Types**: count the number of distinct words
  - **Singletons**: number of items that occur exactly once
- Three combinations to consider
  - OLD = Hansards (Canadian parliamentary discussions)
  - NEW = { EMEA (medical data), Science, Subtitles }

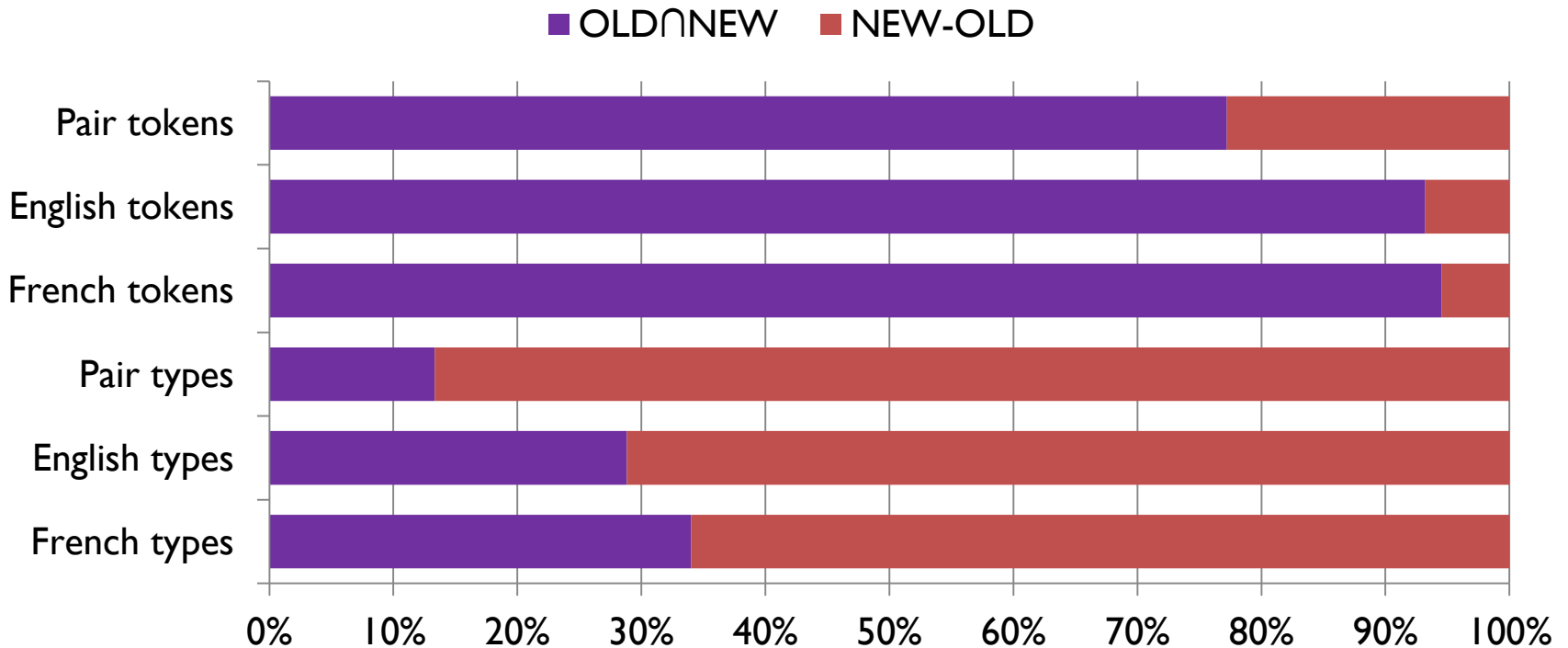
# Hansards $\mapsto$ EMEA



	French types	English types	Pair types	French tokens	English tokens	Pair tokens
OLD $\cap$ NEW	17845	13743	63087	6124518	5522972	6290162
NEW-OLD	16779	15920	431877	419575	381324	2002943

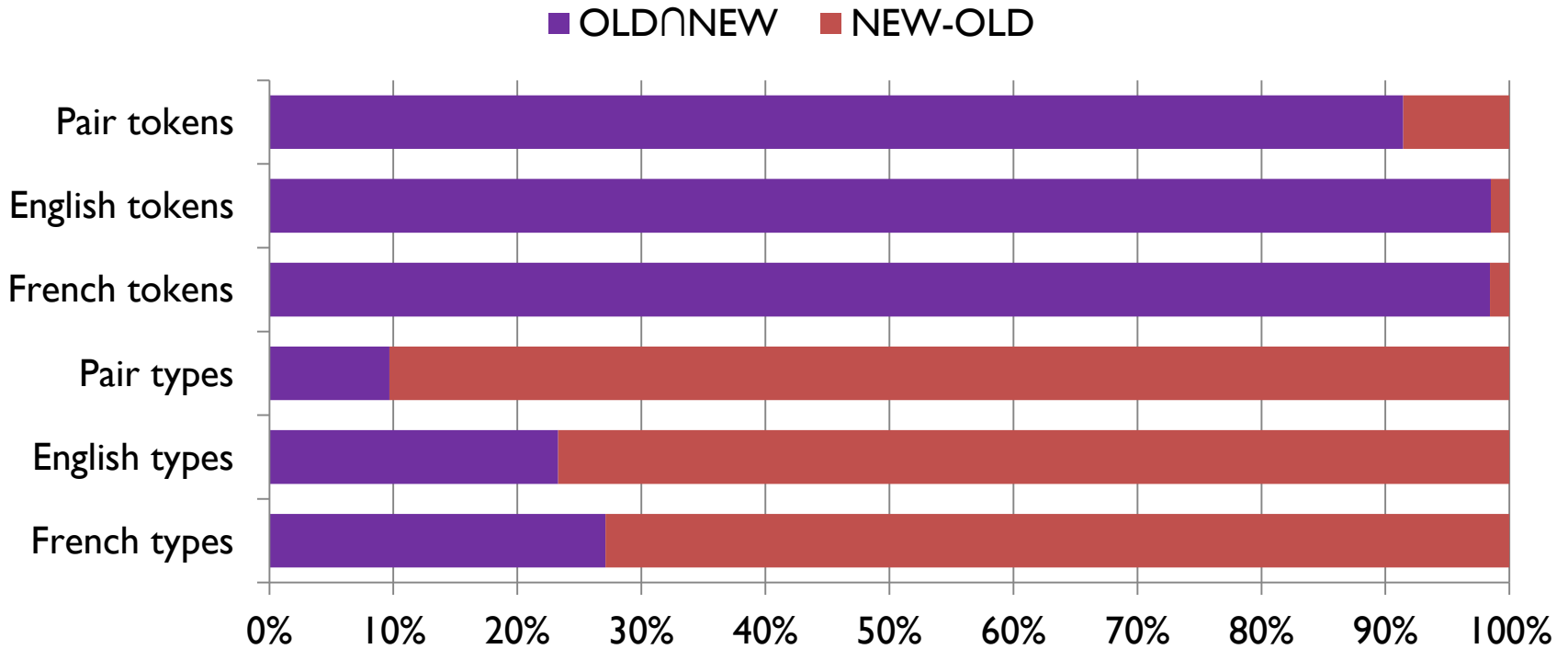


# Hansards $\leftrightarrow$ Science



	French types	English types	Pair types	French tokens	English tokens	Pair tokens
OLD $\cap$ NEW	40016	32947	135247	4057191	3358471	3995699
NEW-OLD	77653	81270	879423	235429	244328	1179428

# Hansards $\leftrightarrow$ Subs



	French types	English types	Pair types	French tokens	English tokens	Pair tokens
OLD ∩ NEW	98048	68274	694212	152519138	171806360	199375051
NEW-OLD	263536	224975	6471868	2433294	2624046	18649558

# SMT quality across domains: Coarse mixture models can help BLEU, sometimes

Simply concatenating old and new domain is not always a good idea!

Domain	News	EMEA	Science	Subs
Old	22.61	22.72	21.22	13.64
New	20.33	<b>34.83</b>	32.49	20.57
Old+New	<b>23.82</b>	34.76	??	??



Learning mixing weights for old and new domain is slightly better  
[Foster & Kuhn 2007]

Domain	News	EMEA	Science	Subs
New	20.27	40.84	32.48	<b>25.50</b>
Mix LM	21.57	40.95	32.60	<b>25.51</b>
Mix LM + Mix TM	<b>23.50</b>	<b>41.47</b>	<b>32.78</b>	25.38

Warning:  
scores not  
comparable  
across 2  
tables!  
(different MT  
systems, larger  
test sets!)

# Analysis: how difficult is lexical choice across domains?

EMEA	Micro Accuracy	Macro Accuracy
Old domain phrase-table (hansard)	43.98	49.50
New domain phrase-table	59.19	76.86
Old domain Moses	77.77	55.22
New domain Moses	<b>92.58</b>	<b>77.28</b>
Old+New domain Moses	92.02	74.88

- quite difficult with old domain only!
- much easier with lots of new domain data
- yet, concatenating old+new is too crude to help

# Analysis: accuracy patterns differ across French types

EMEA	Enceinte	Régime	Formation	Rapport	Etat
Old domain phrase-table (hansard)	0	0	12.50	42.42	67.24
New domain phrase-table	<b>100</b>	<b>92.30</b>	<b>87.50</b>	09.09	24.14
Old domain Moses	<b>100</b>	0	37.50	36.36	56.89
New domain Moses	<b>100</b>	84.61	81.25	03.03	74.13
Old+New domain Moses	<b>100</b>	53.84	81.25	<b>45.54</b>	<b>77.58</b>

New domain data might be sufficient, but  
- we need better local context models  
- old domain shouldn't hurt

New domain not sufficient!  
Better context models are needed

John Morgan



# Analysis

# Taxonomy of Errors

Categorize errors in translation according to cause:

- Seen: NEW domain source words or phrases not in OLD
  - Out of Vocabulary Words and Phrases.

Science anisotropie

Subs zut

Medical pelliculé

- Sense: source NEW domain phrase is in OLD, translation is not

Medical membres

Science régime

Subs campagne

- Score: phrase pair is in both OLD and NEW, but correct translation has lower score
- Search: correct translation has higher score, search fails to find it



# Seen and Sense

How can we measure the impact of SEEN and SENSE errors?

- Approach – selectively fix errors in OLD, measure improvement.
- To identify where SEEN and SENSE errors occur:
  - Train concatenated OLD and NEW new system (CAT)
  - Identify phrase pairs in CAT where

**UNSEEN** Source side of phrase pair in NEW phrase-table only.

**NEW SENSE** Source side of phrase pair in OLD, but phrase pair in NEW only.

**SEEN ADD:** Add just UNSEEN phrase pairs to OLD phrase table.

**SENSE ADD:** Add just NEW SENSE phrase pairs to OLD phrase table.

- Tune and Test SEEN ADD and SENSE ADD on NEW.
- Measure improvements against OLD tuned on NEW.

## Seen and Sense Analysis Results

domain	OLD	SEEN ADD	SENSE ADD
News	23.82	23.68	23.93
Medical	22.62	32.77 (45%)	32.58 (44%)
Science	21.22	26.36 (24%)	25.58 (21%)
Subtitles	13.64	16.60 (22%)	17.75 (30%)

**Table:** BLEU scores before and after adding OOVs and new senses to OLD phrase table.

## Comments on SEEN and SENSE Errors

- OOVs and new senses are not the sources of errors in the News domain.
- The impact of OOVs and new senses is similar in the other 3 domains.
- The largest impact came from OOVs in the medical domain (44%).

## How can we measure the impact of SCORE errors?

- Intersect OLD and CAT phrase tables.

SCORE NEW: Use phrase pair scores from NEW.

- Tune and Test SCORE NEW on NEW.
- Compare results with OLD, SCORE NEW, and CAT tuned on NEW.

domain	OLD	SCORE NEW	CAT
News	23.82	23.93	23.82
Medical	22.62	30.69 (36%)	40.53
Science	21.22	26.20 (23%)	30.17
Subtitles	13.64	17.65 (29%)	20.41

**Table:** BLEU scores before and after adding scores from either OLD or CAT to intersection of OLD and NEW phrase tables.

## Comments on SCORE Errors

- Scores are again not the source of errors in the News domain.
- All other domains benefit from better scores, especially medical.
- There is potential for substantial benefit with better scores.

Anni Irvine

# Extrinsic Word-Level Evaluation

(Sanjeeval)

# Extrinsic Word-Level Evaluation

## (Sanjeeval)

- S4 is a *macro*-level analysis of end-to-end MT



# Extrinsic Word-Level Evaluation

## (Sanjeeval)

- S4 is a *macro*-level analysis of end-to-end MT
- Sanjeeval is a *micro*-level analysis of end-to-end MT
  - Unit of analysis: alignments between source language test (English) data and target language reference (French) data

Correct: Blue

OOV-Freebie: Green

New-Sense-Freebie: Purple

Score/Search Errors: Red

OOV-Wrong: Orange

New-Sense-Wrong: Pink

Phrase-Span: Gray Dashed

# Extrinsic Word-Level Evaluation

## (Sanjeeval)

- S4 is a *macro*-level analysis of end-to-end MT
- Sanjeeval is a *micro*-level analysis of end-to-end MT
  - Unit of analysis: alignments between source language test (English) data and target language reference (French) data

- Tools:
  - Sentence-level visualizer
  - Aggregate statistics

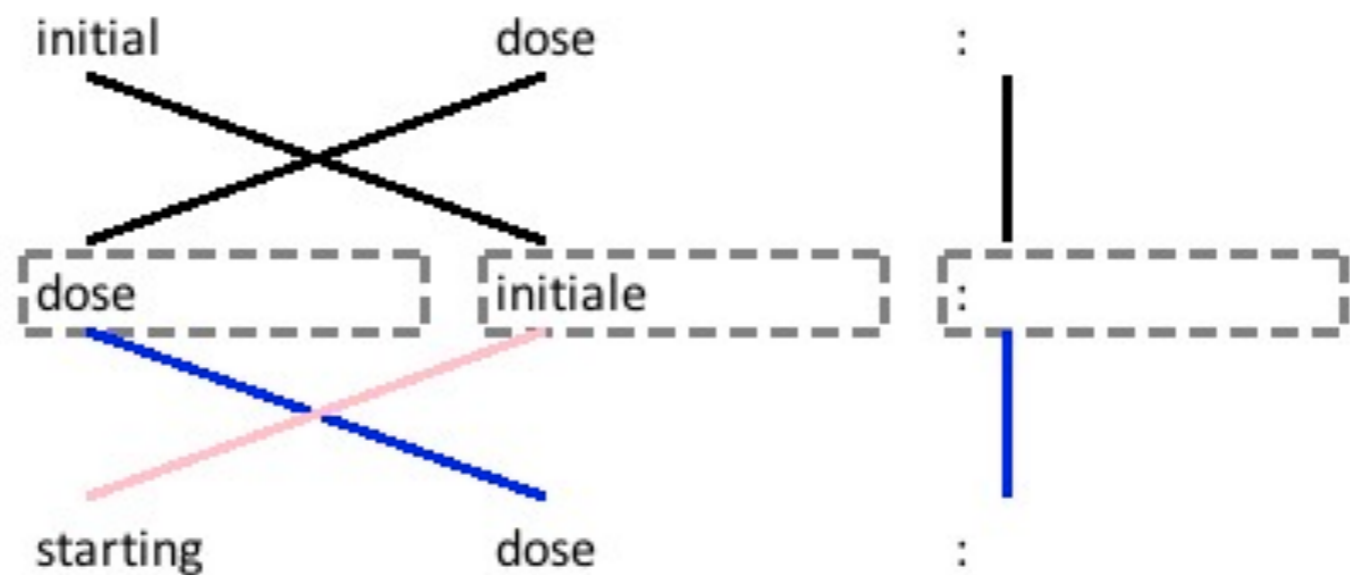
Correct: Blue  
OOV-Freebie: Green  
New-Sense-Freebie: Purple  
Score/Search Errors: Red  
OOV-Wrong: Orange  
New-Sense-Wrong: Pink  
Phrase-Span: Gray Dashed

# Extrinsic Word-Level Evaluation (Sanjeeval)

Output: English

Input: French

Reference: English



Correct: Blue

OOV-Freebie: Green

New-Sense-Freebie: Purple

Score/Search Errors: Red

OOV-Wrong: Orange

New-Sense-Wrong: Pink

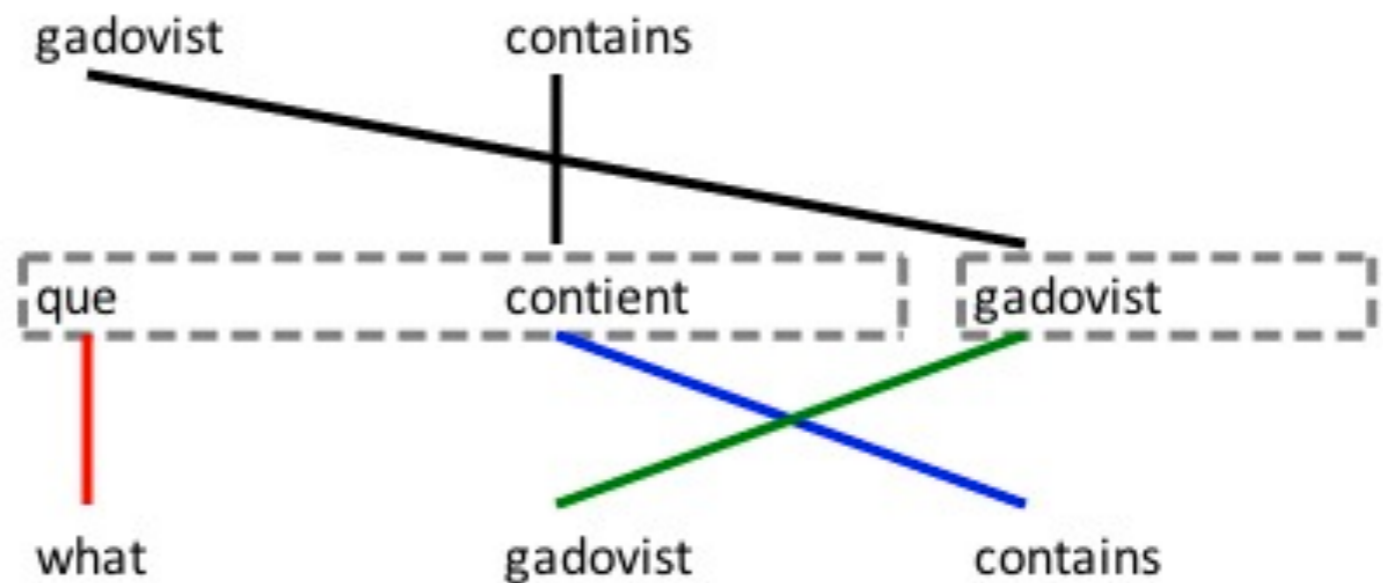
Phrase-Span: Gray Dashed

# Extrinsic Word-Level Evaluation (Sanjeeval)

Output: English

Input: French

Reference: English



Correct: Blue

OOV-Freebie: Green

New-Sense-Freebie: Purple

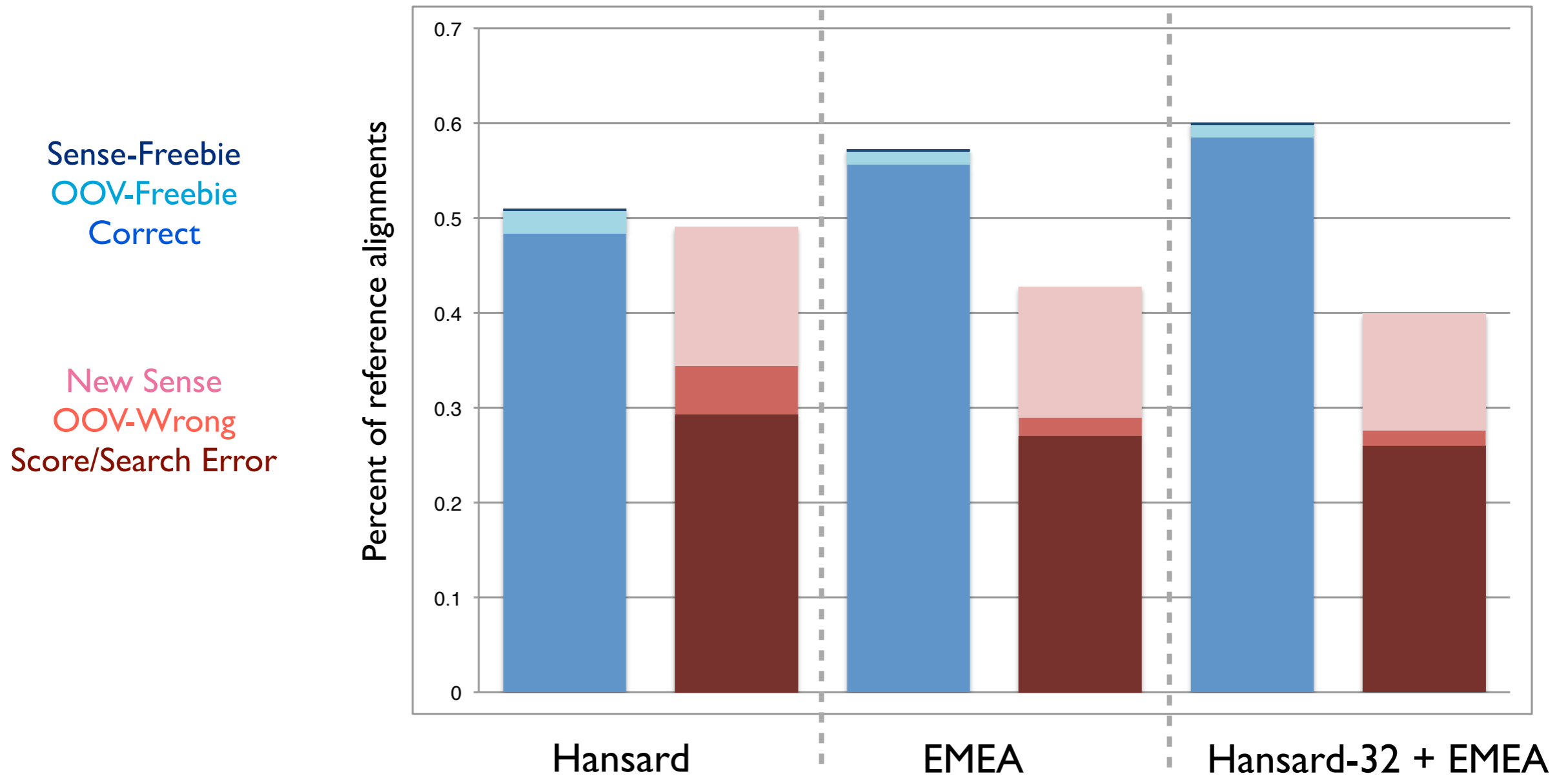
Score/Search Errors: Red

OOV-Wrong: Orange

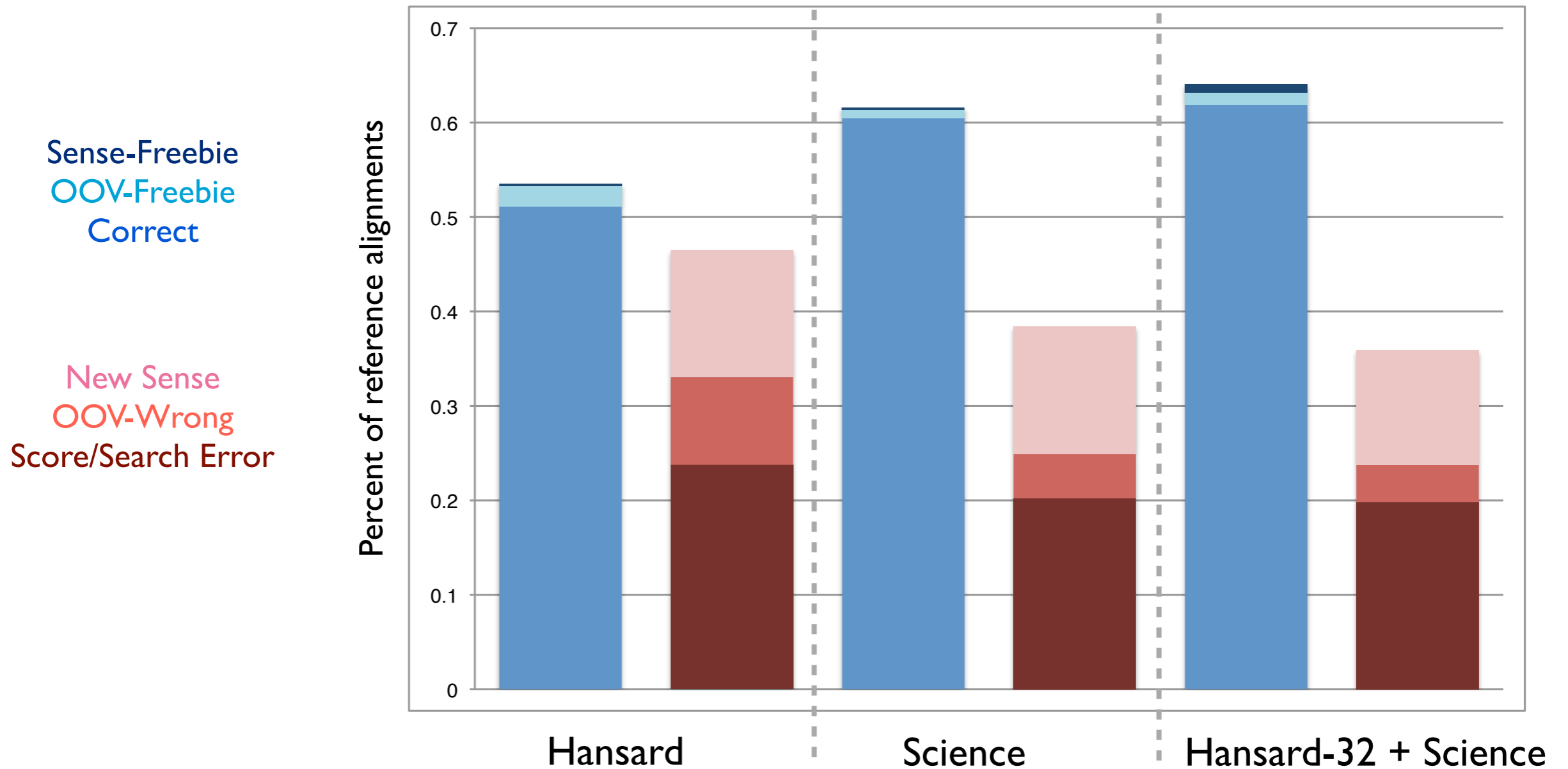
New-Sense-Wrong: Pink

Phrase-Span: Gray Dashed

# Extrinsic Word-Level Evaluation (Sanjeeval)



# Extrinsic Word-Level Evaluation (Sanjeeval)



5 MIN BREAK

Alex Fraser



# Phrase Sense Disambiguation for Domain Adapted SMT

- Introduction
  - Phrase Sense Disambiguation (PSD)
  - Vowpal Wabbit (VW) Classifier
- Phrase Sense Disambiguation (PSD) - Evaluation
  - Decoder and classifier focused
  - Lexical selection
- Integrating VW into the Moses decoder for PSD
- Source sentence context features for PSD
- Hiero and soft-syntactic features for PSD
- Domain adaptation using PSD

# Phrase Sense Disambiguation for Domain Adaptation in SMT

- Lessons from analysis:
  - domain shift yields different types of lexical choice errors
  - coarse uniform adaptation at the domain level doesn't work
- Proposed solution: **Phrase Sense Disambiguation**
  - Discriminative, context-dependent translation lexicon
  - Unlike phrase-table translation probabilities

[Carpuat & Wu 2007]

# Phrase Sense Disambiguation

Disambiguating English senses of **rapport**

$P(e|f)$  ↑  
[report] Il a rédigé un **rapport** .  
[relationship] Quel est le **rapport** ?  
...

- PSD = phrase translation as classification
- PSD at test time
  - use context to predict correct English translation of French phrase
- PSD at train time
  - use word alignment to extract training instances
  - occurrences of French phrases **in context** are annotated with their English translations

# Why PSD for DAMT?

- **Source context** can prevent some translation errors when shifting domain
  - Even without DA
- PSD can flexibly incorporate **rich domain-relevant features** in SMT
  - without adding tuning/decoding complexity
- PSD can capture different behavior of **general vs domain-specific French phrases**
  - unlike more standard coarse mixtures for DA
- PSD can directly **leverage existing ML algorithms**, for classification and adaptation
  - unlike standard SMT

# Vowpal Wabbit



- Fast implementation of stochastic gradient descent and L-BFGS for many losses
- Recently built into a library (for this workshop)
- Very widely used for ML tasks (>6 companies)
- Built-in support for:
  - Feature hashing (scaling to billions of features)
  - Caching (no need to re-parse text)
  - Different losses and regularizers
  - Reductions framework to binary classification
  - Multithreaded/multicore support

# Vowpal Wabbit

- Our “weird” setting: label-dependent features
  - Normal for NLPers, impossible for MLers to grasp
  - Think of it like ranking:

x = le croissant rouge  
y1 = the red croissant  
y2 = the croissant red  
y3 = the croissant  
y4 = the red

x = mange  
y1 = eat  
y2 = eats  
y3 = ate

- Different inputs have different #s and definitions of possible “labels,” each with it's own features
- Define feature space as  $X*Y$  cross-product and:
- Regress on loss (“csoaa\_ldf”)
- Classifier all-versus-all (“wap\_ldf”)

# Evaluating PSD

- Ways to evaluate PSD
  - Best way: use in decoder
  - $P(e\_phrase|f\_phrase,f\_context)$  added as a feature function to the decoder
    - Tuned along with standard phrase table features such as  $p(e\_phrase|f\_phrase)$  estimated using relative frequency
  - Easily extended for Domain Adaptation
  - Measure test set BLEU
  - More on this in a few minutes...
- Problem: slow to run experiments, difficult to analyze/assign blame for problems

# Classifier Accuracy on All Phrases

- Another way to evaluate:
  - Classifier accuracy on held-out VW training data
  - This is easy to do, just run feature extraction, build a classifier, and measure accuracy on the held-out set
  - Very useful for testing domain adaptation algorithms!
- However, there is one classifier training example per phrase pair token (worse in Hiero)
  - Includes overlapping phrases - problem: assigns equal weight to all phrases!
- Depends on the word alignment to the English reference translation of the held-out set
  - So the so-called gold standard can contain errors
  - No idea of importance of (possibly overlapping) phrase pairs



Katie Henry

# Evaluate Lexical Choice in Isolation

## Target Domain Specific/Ambiguous Words

812 Representative Phrases

accessoire  
actualisé  
additif  
.  
.  
.  
virus  
visage  
vue  
zut

How do we translate these words in different contexts?

# Examples of Representative Phrases

Rep  
Phrase

Hansard

EMEA

Science

Subs

état



enceinte



formation



rapport



régime



# Lexical Selection on Representative Phrases

## Goal:

Evaluate performance on translating representative phrases

## Task:

Compute translation accuracy for each representative phrase

## Advantages:

Allows comparison of output from a PSD classifier and a full MT system

Cheaper way of evaluating features

# What could we gain with Multiple References?

## Percent of Alignments Made by X

Train	Exact	Stem	Synonym	Paraphrase
Hansard	78.02%	0.85%	1.86%	9.17%
EMEA	93.16%	0.68%	0.75%	0.86%
Hansard + EMEA	92.52%	0.45%	0.56%	1.92%

Meteor alignments  
between representative  
phrases from Moses  
output and reference set

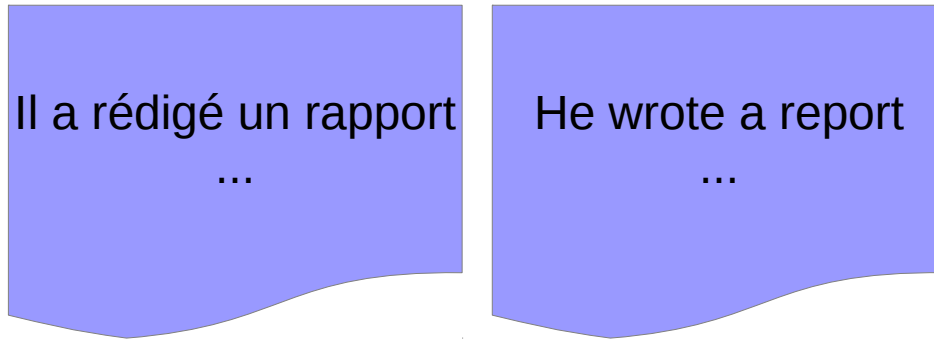
## Precision of Alignments

Train	Synonym	Paraphrase	Either
Hansard	0.98	0.47	0.50
EMEA	0.98	0.95	0.95
Hansard + EMEA	0.97	0.68	0.73

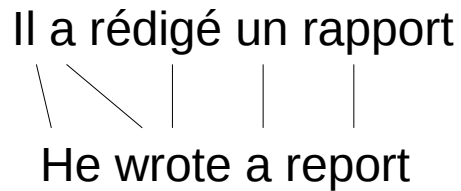
Aleš Tamchyna

# PSD Pipeline

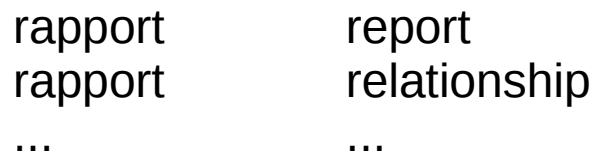
Standard Pipeline:



*Align words*



*Extract phrases*

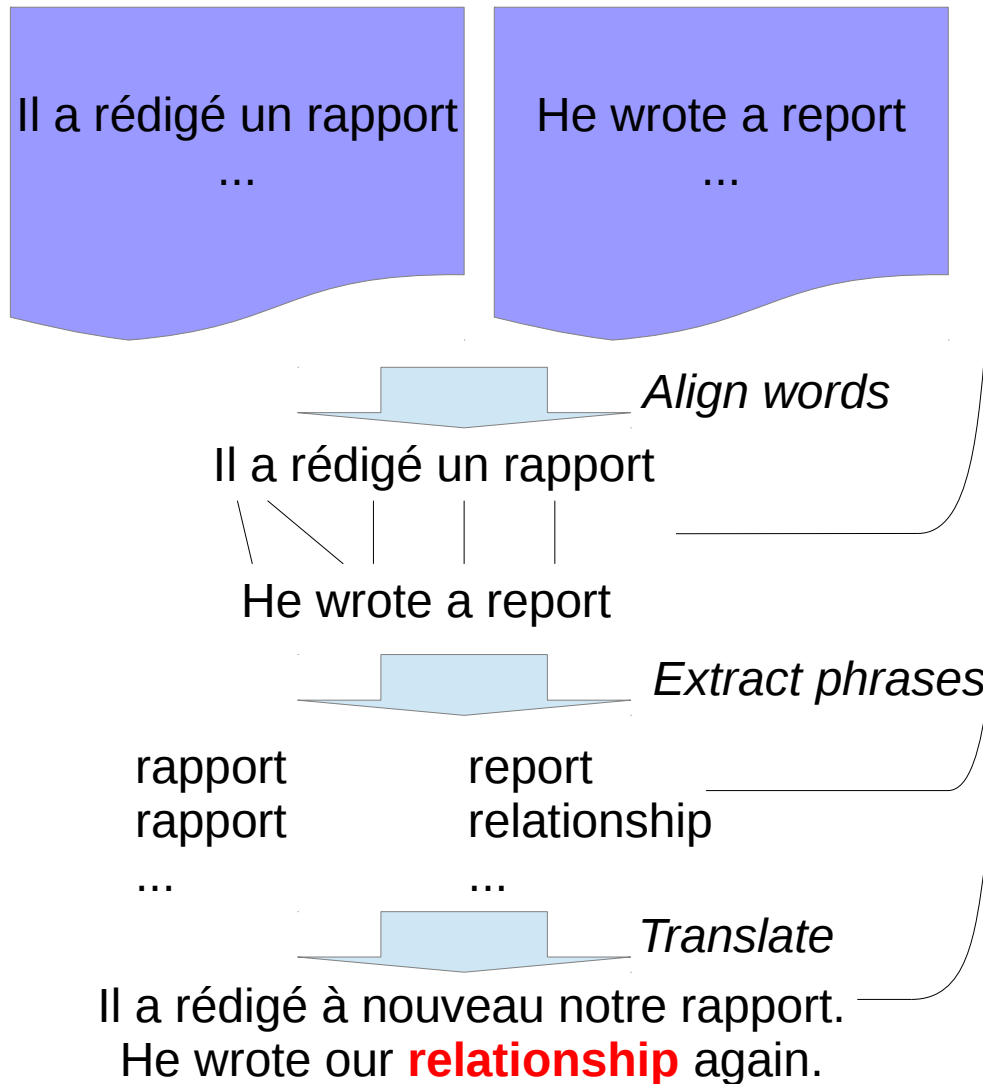


*Translate*

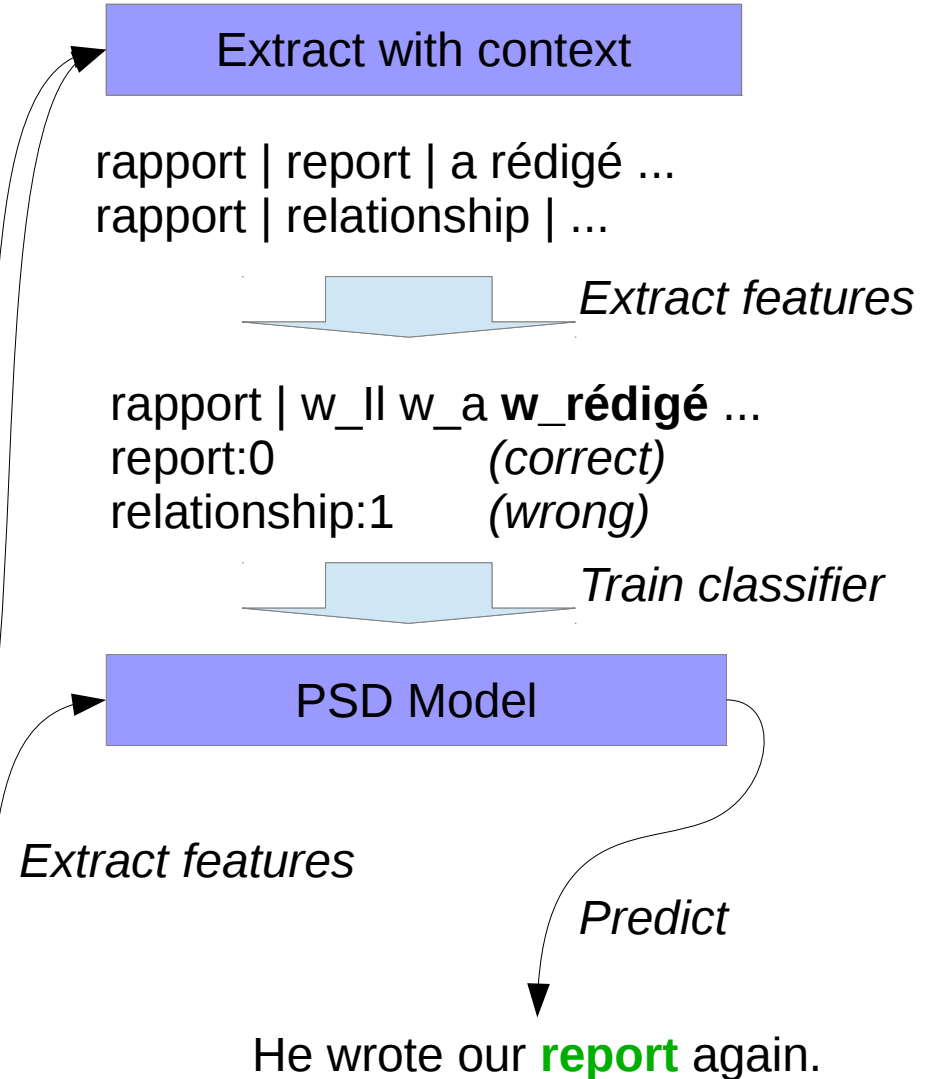
Il a rédigé à nouveau notre rapport.  
He wrote our **relationship** again.

# PSD Pipeline

Standard Pipeline:



With PSD:





# Phrase-Sense Disambiguation in Moses

- ▶ In branches `damt_phrase` and `damt_hiero`.
- ▶ Vowpal Wabbit linked with Moses.
- ▶ PSD integrated in training and decoding.
- ▶ Extensible interface for creating new features.
- ▶ Fully integrated in EMS, a system for managing experiments bundled with Moses.
- ▶ Support parallelism in training (multiple processes) and decoding (multithreading).
- ▶ Classifier predictions can now be used as features in Moses.

# Phrase-Sense Disambiguation in Decoding

- ▶ PSD is a feature function in Moses.
  - ▶ Scores depend on source context.
- ▶ Integrated into the log-linear model, its weight is tuned.
- ▶ Scores calculated before decoding.
  - ▶ Saves repeated computation.
  - ▶ Initial pruning can include PSD scores.
- ▶ VW is queried for each possible translation of a source span.
- ▶ Scores (inverse losses) are exponentiated and locally normalized.

# Basic Features for Phrase-Sense Disambiguation (1/2)

## Context

Form:	nous	ne	le	savons	pas	encore	.
Lemma:	il	ne	le	savon	pas	encore	.
Tag:	CLS	ADV	DET	NC	ADV	ADV	PONCT

## Phrase Pair

Source: ne le savons  
Target: do not know

## Features

Source indicator: p<sup>^</sup>ne\_le\_savons  
Target indicator: p<sup>^</sup>do\_not\_know  
Source internal: w<sup>^</sup>ne w<sup>^</sup>le w<sup>^</sup>savons  
Target internal: w<sup>^</sup>do w<sup>^</sup>not w<sup>^</sup>know  
Context: c<sup>^</sup>0\_-1\_nous c<sup>^</sup>1\_-1\_il c<sup>^</sup>2\_-1\_CLS c<sup>^</sup>0\_1\_pas ...

# Basic Features for Phrase-Sense Disambiguation (2/2)

## Phrase Pair

Source: ne le savons  
Target: do not know  
Alignment: 0-1 1-2 2-2  
Scores: -7.5 -9.2 -1.6 -7.5

## Features

Paired: p<sup>^</sup>ne\_not p<sup>^</sup>le\_know p<sup>^</sup>savons\_know  
Scores: sc<sup>^</sup>0\_-10 sc<sup>^</sup>0\_-9 sc<sup>^</sup>0\_-8 sc<sup>^</sup>1\_-10 sc<sup>^</sup>2\_-10 ...

# Phrase-Based MT: Evaluation

## Lexical Selection

- ▶ Evaluated on representative phrases in Science domain.

Training Data	Accuracy	
	Baseline	PSD
Hansard	—	73.6
Hansard + Science	69.0	73.7

## Machine Translation Experiments

- ▶ Can run full Moses pipelines.
- ▶ No improvements in BLEU so far, still looking for bugs.

Fabienne Braune



# PSD and Syntactic Features in Hierarchical PBSmt

Fabienne Braune

# Hierarchical Rules for Word Sense Disambiguation

F personne diabétique **enceinte**

E **pregnant** diabetic person

F confiné dans une **enceinte**

E hidden in a **building**

Unseen patiente diabétique **enceinte** → **pregnant** diabetic patient

Unseen diabétique **enceinte** → **pregnant** diabetic

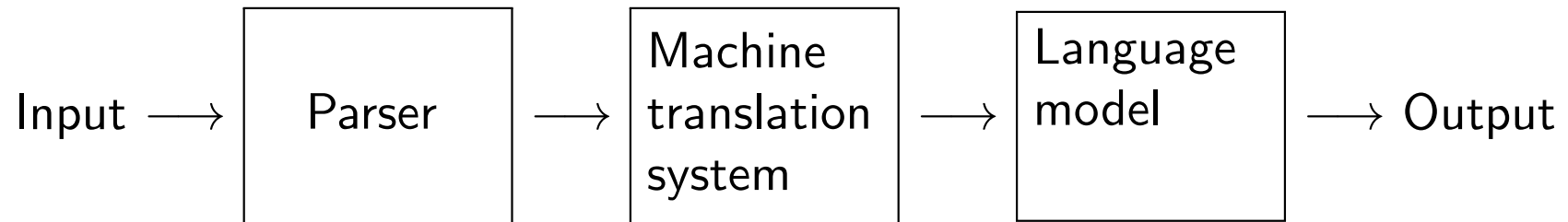
Seen X **enceinte** → **pregnant** X

Seen X **enceinte** → X **building**

- |                     |          |   |          |  |                  |
|---------------------|----------|---|----------|--|------------------|
| personne diabétique | enceinte | ⇒ | pregnant | <table border="1"><tr><td>diabetic patient</td></tr></table> | diabetic patient |
| diabetic patient    |          |   |          |  |                  |
- |                  |          |   |   |             |          |
|------------------|----------|---|---|-------------|----------|
| confiné dans une | enceinte | ⇒ | <table border="1"><tr><td>hidden in a</td></tr></table> | hidden in a | building |
| hidden in a      |          |   |   |             |          |



# Syntax Based SMT



- Parser (SCFG rules):
  - SENT/SENT → <NP enceinte , pregnant NP>
  - NP/NP → <NN ADJ , ADJ NN>
  - NN → <personne , person>
  - ADJ → <diabétique , diabetic>

## Why syntactic features instead of hard constraints

- confiné dans une enceinte  $\Rightarrow$  hidden in a building
- X **enceinte**  $\rightarrow$  X **building**
- X does not match a syntactic constituent
- $\Rightarrow$  Use hierarchical (unlabeled) rules with syntactic **features**

# More ambiguity in Hiero than Phrase-Based

- Source segment : patiente diabétique **enceinte**
  - $N$  rule source sides :
    - $X/X \rightarrow \langle X \text{ enceinte}, \dots \rangle$
    - $X/X \rightarrow \langle \text{patiente } X, \dots \rangle$
    - $X/X \rightarrow \langle \text{patiente } X \text{ enceinte}, \dots \rangle$
  - **For each source side :**
    - $M$  target sides :
      - $X/X \rightarrow \langle X \text{ enceinte}, \text{pregnant } X \rangle$
      - $X/X \rightarrow \langle X \text{ enceinte}, X \text{ building} \rangle$
- For each source side of a rule, get score of all targets

# PSD Features in Hiero

- PSD features (source) trigger choice of right rules:  
⇒ Chose rule with right terminal items

- **personne diabétique** enceinte ⇒ pregnant **diabetic patient**
- **confiné dans une** enceinte ⇒ **hidden in a** building

- Integrated PSD features in hiero :
  - French (source) context of rule
  - Source and Target of rule
  - Bag of words inside of rule
  - Bag of words outside of rule
  - Aligned terminals
  - Rule scores (e.g.  $p(e|f)$ )

# Syntax Features in Hiero

- Syntax features (source) guide right rule application :  
⇒ Apply non-terminals in the right place
  - ( (personne<sub>NN</sub> diabétique<sub>ADJ</sub>)<sub>NP</sub> enceinte)<sub>NP</sub>  
⇒ pregnant diabetic patient
  - ( confiné<sub>VPART</sub> dans<sub>PREP</sub> une<sub>DET</sub> enceinte)<sub>SENT</sub>  
⇒ hidden in a building
- Integrated syntactic features in hiero :
  - Constituent and Parent of applied rule
  - Span width of applied rule
  - Type of reordering (multiple non-terminals)

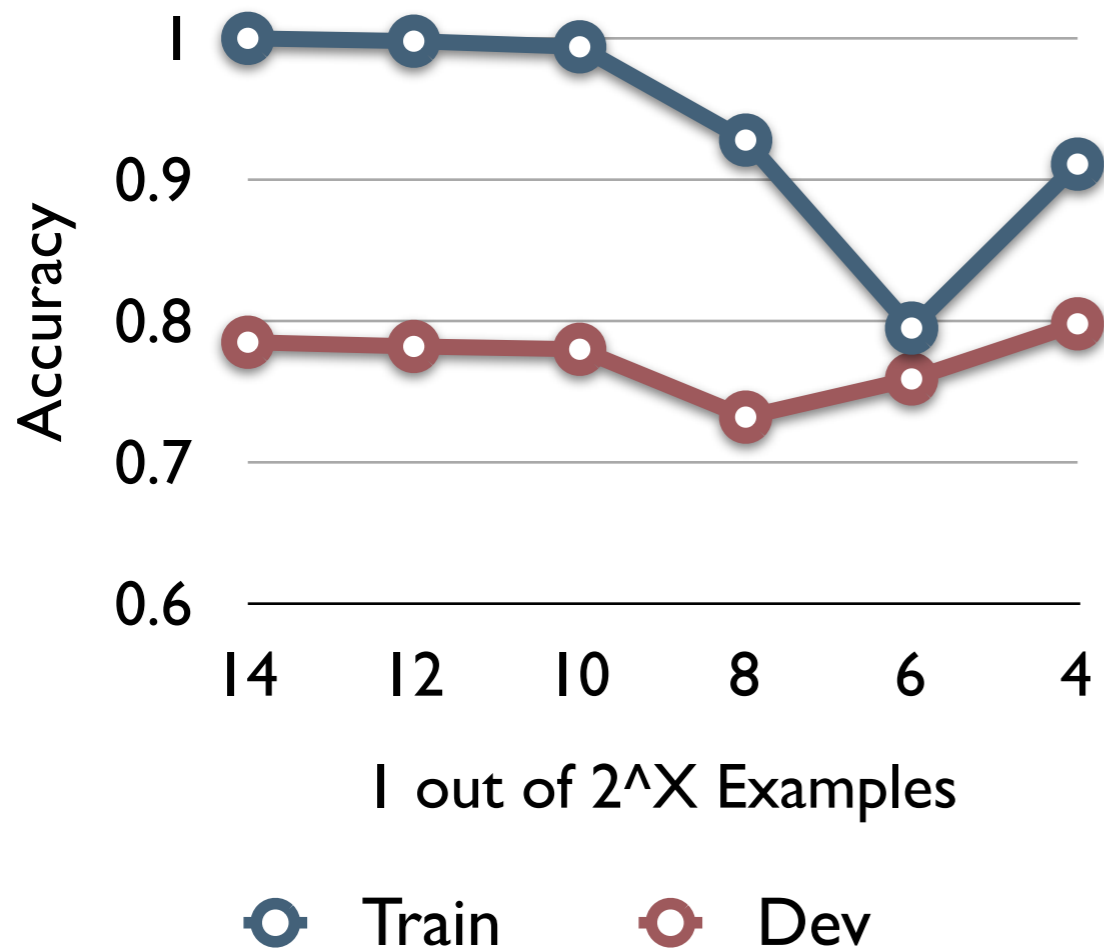
## Contributions and Future Work

- Integration of a classifier into a Hierarchical Phrase-Based SMT system
- PSD and basic soft syntactic features integrated and working
- Running end-to-end experiments but no results yet
- Room for more features :
  - Near term : CCG style incomplete constituents
  - Long term : More complex rules on subtrees

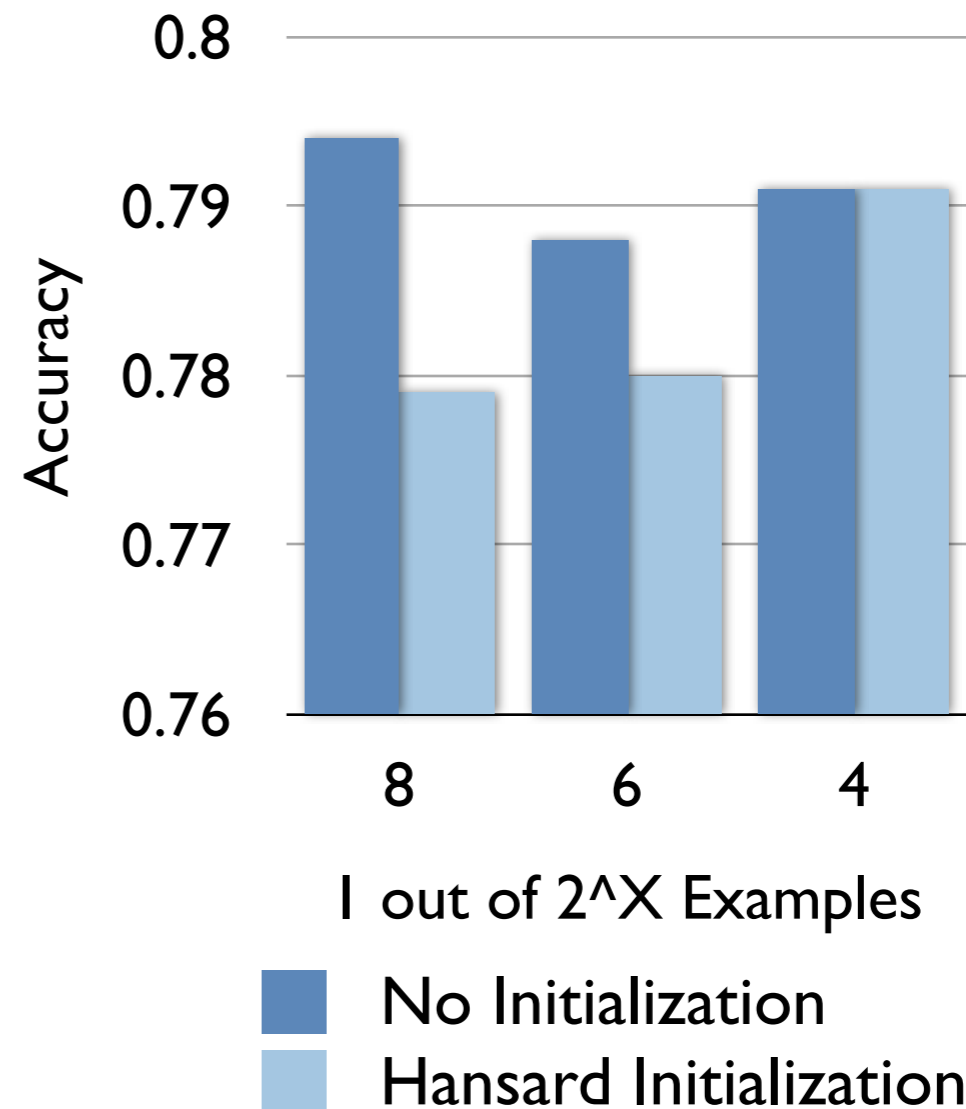
Majid Razmara

# Classifier Accuracy on All Phrases

## Accuracy vs Training Data Size 4 Iterations of VW



## Different Initializations of VW Classifier





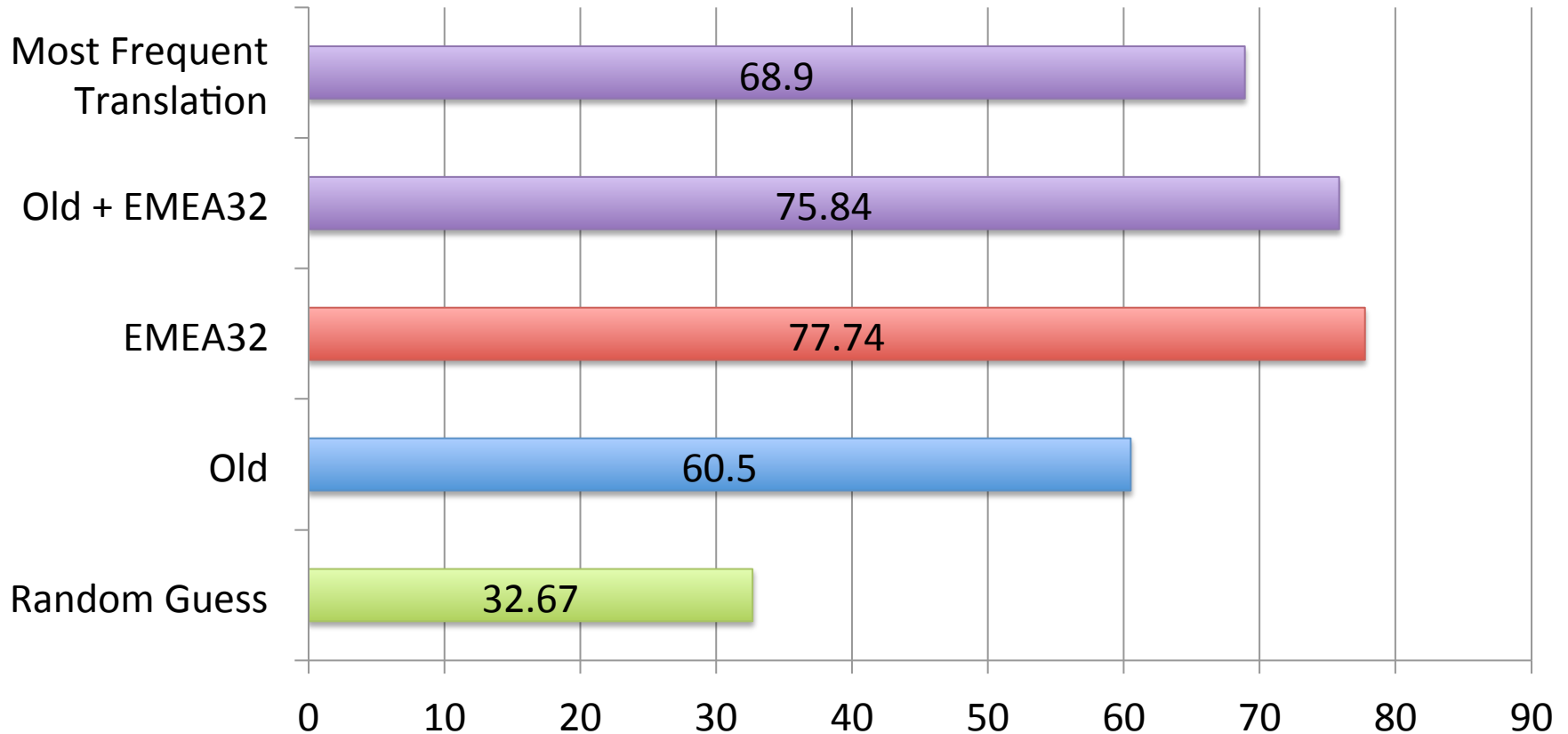
# PSD Domain Adaptation

Damt

# EMEA32 Baselines

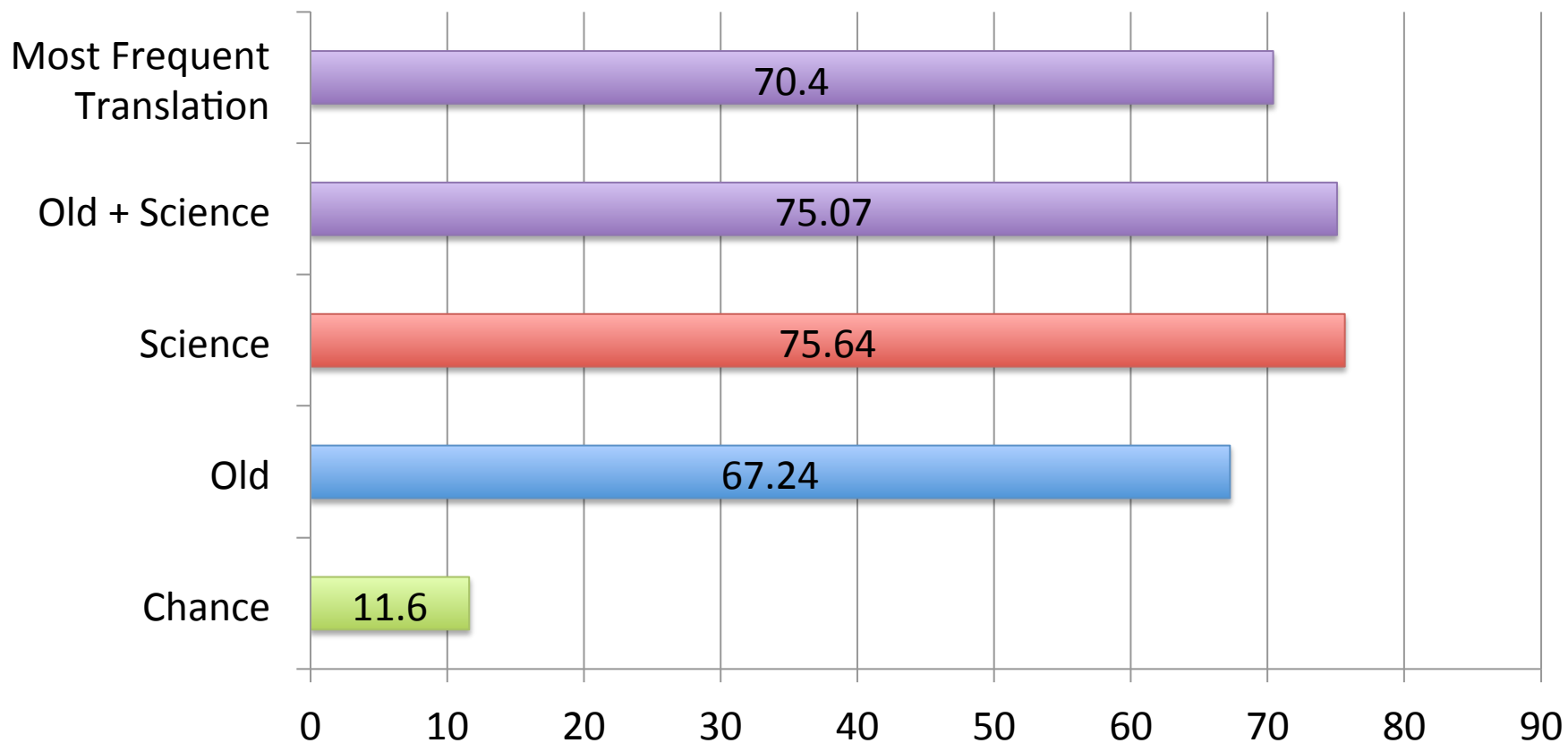
21%  
Unambiguous  
Cases

## Accuracy



# Science Baselines

## Accuracy



# Old and New Agreement

<b>Old</b> \ <b>EMEA</b>	<b>Correct</b>	<b>Incorrect</b>
<b>Correct</b>	57%	4%
<b>Incorrect</b>	21%	18%

# Old and New Agreement

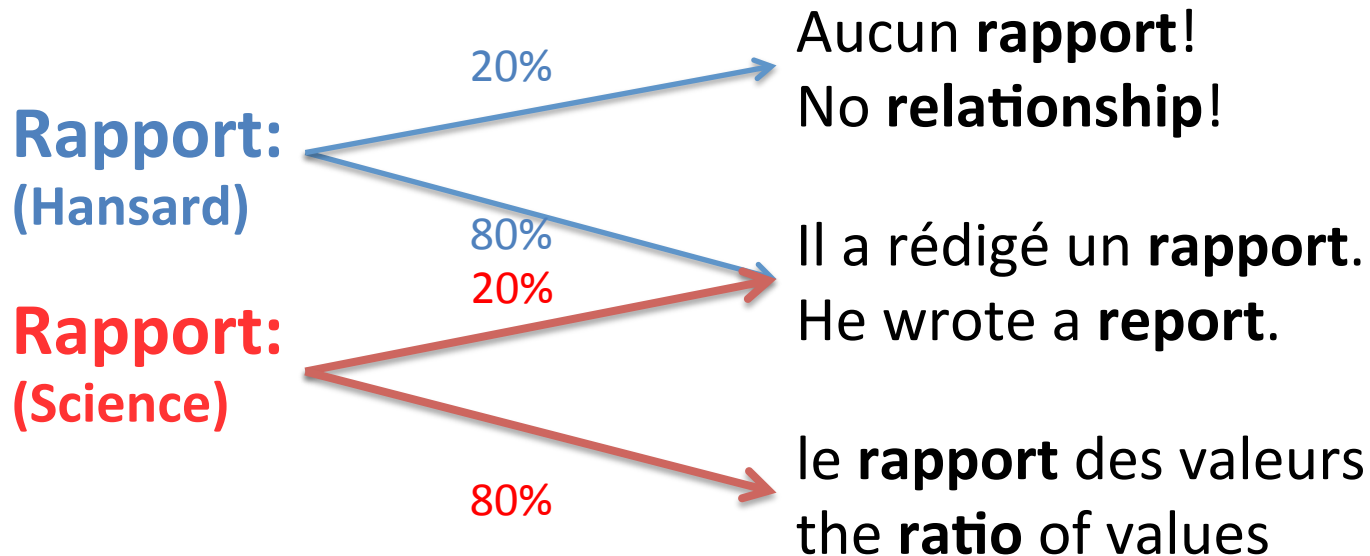
<b>Old</b> \ <b>EMEA</b>	<b>Correct</b>	<b>Incorrect</b>
<b>Correct</b>	57%	4%
<b>Incorrect</b>	21%	18%

<b>Old</b> \ <b>Science</b>	<b>Correct</b>	<b>Incorrect</b>
<b>Correct</b>	62.7%	4.5%
<b>Incorrect</b>	12.9%	19.9%

# Domain Adaptation Techniques

- Frustratingly Easy Domain Adaptation
  - [Blitzer and Hal, 2010]
- Instance Weighting
- Using Old Prediction in New
- Model Interpolation

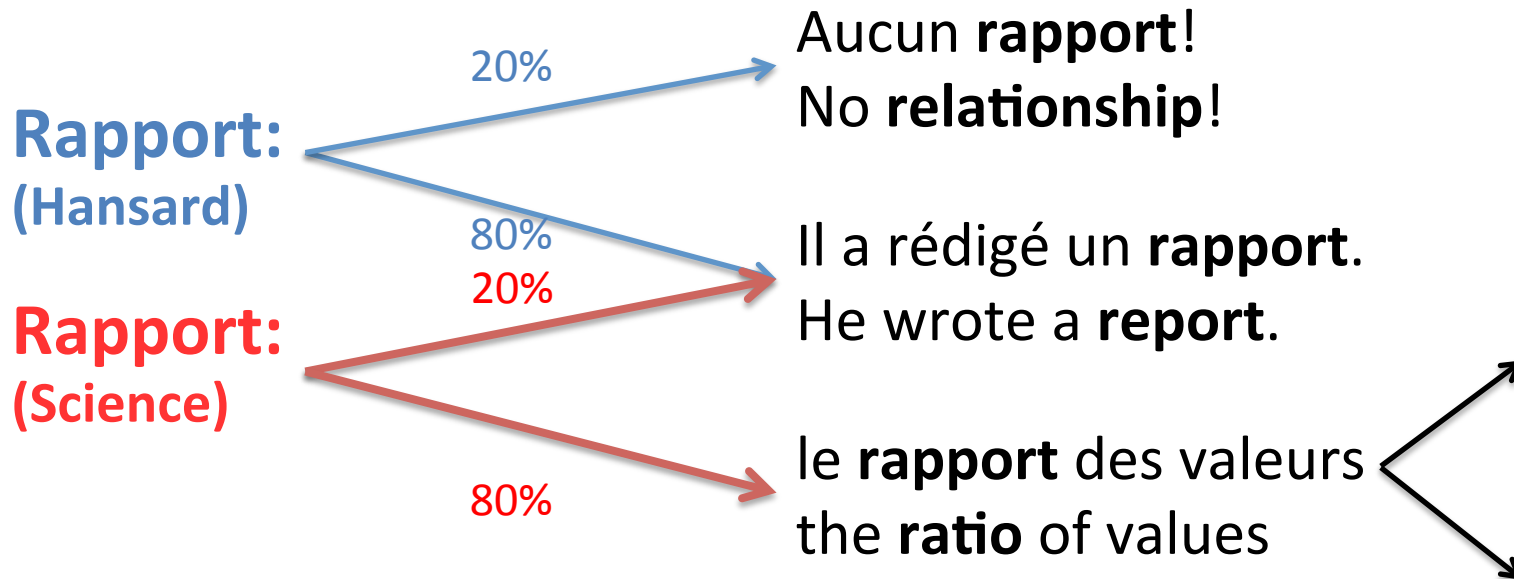
# Frustratingly Easy DA



## Key Idea:

Share some features (e.g. rédigé )  
Don't share others (e.g. rapport)

# Frustratingly Easy DA



## Key Idea:

Share some features (e.g. rédigé )  
Don't share others (e.g. rapport)



# Feature Augmentation

Old:  $x \rightarrow \langle x, x, 0 \rangle$

New:  $x \rightarrow \langle x, 0, x \rangle$

Hansard

Science

Original  
Features

Augmented  
Features

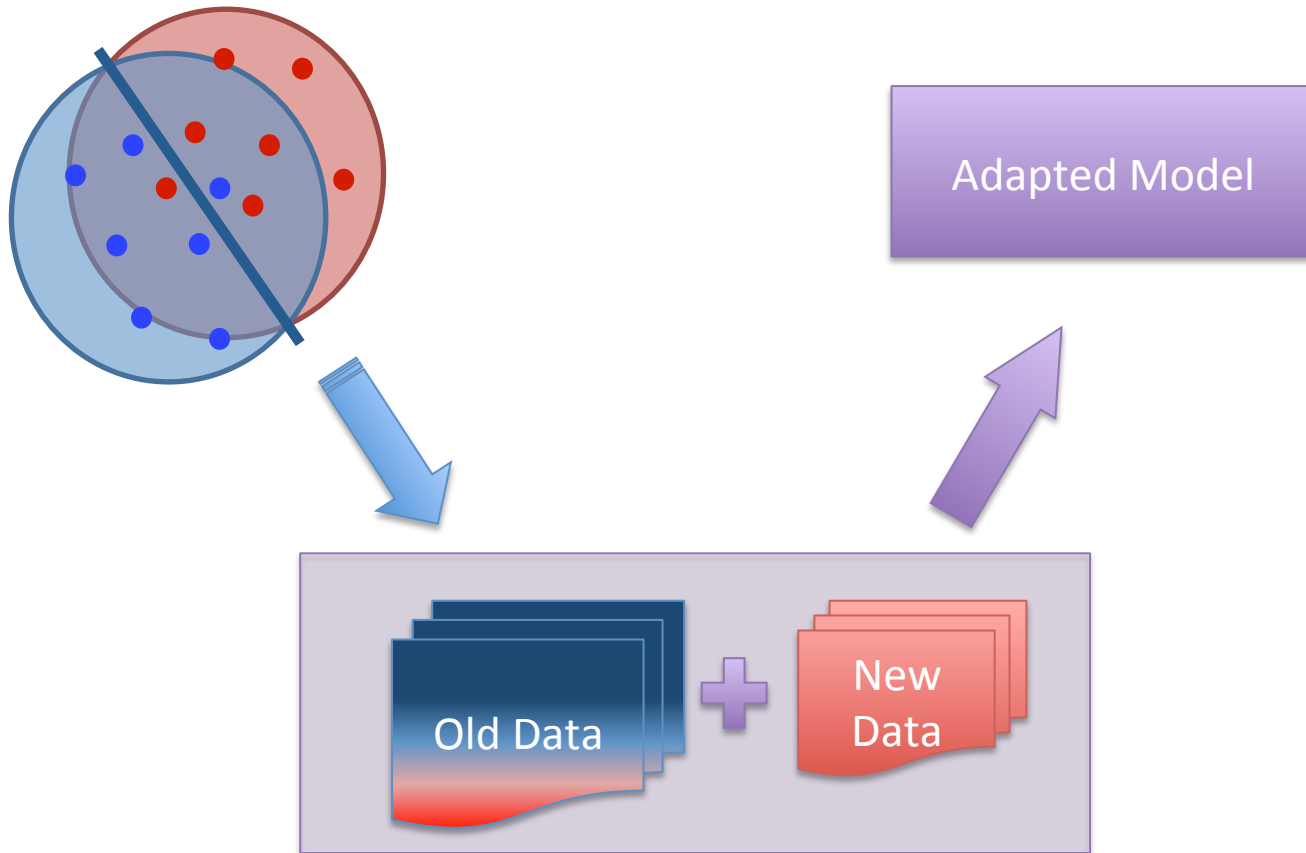
AUCUN  
RÉDIGÉ

VALEURS  
RÉDIGÉ

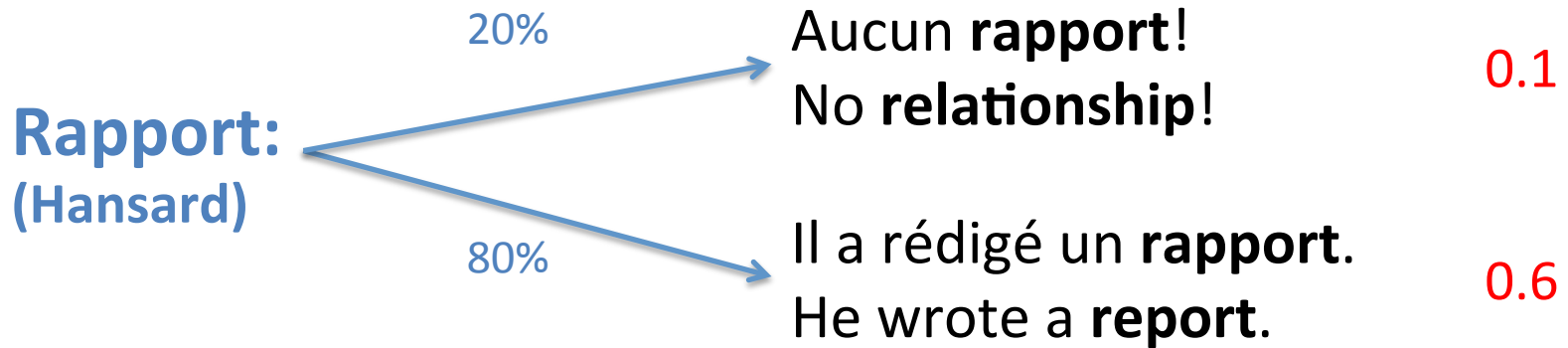
O\_AUCUN  
O\_RÉDIGÉ

N\_VALEURS  
N\_RÉDIGÉ

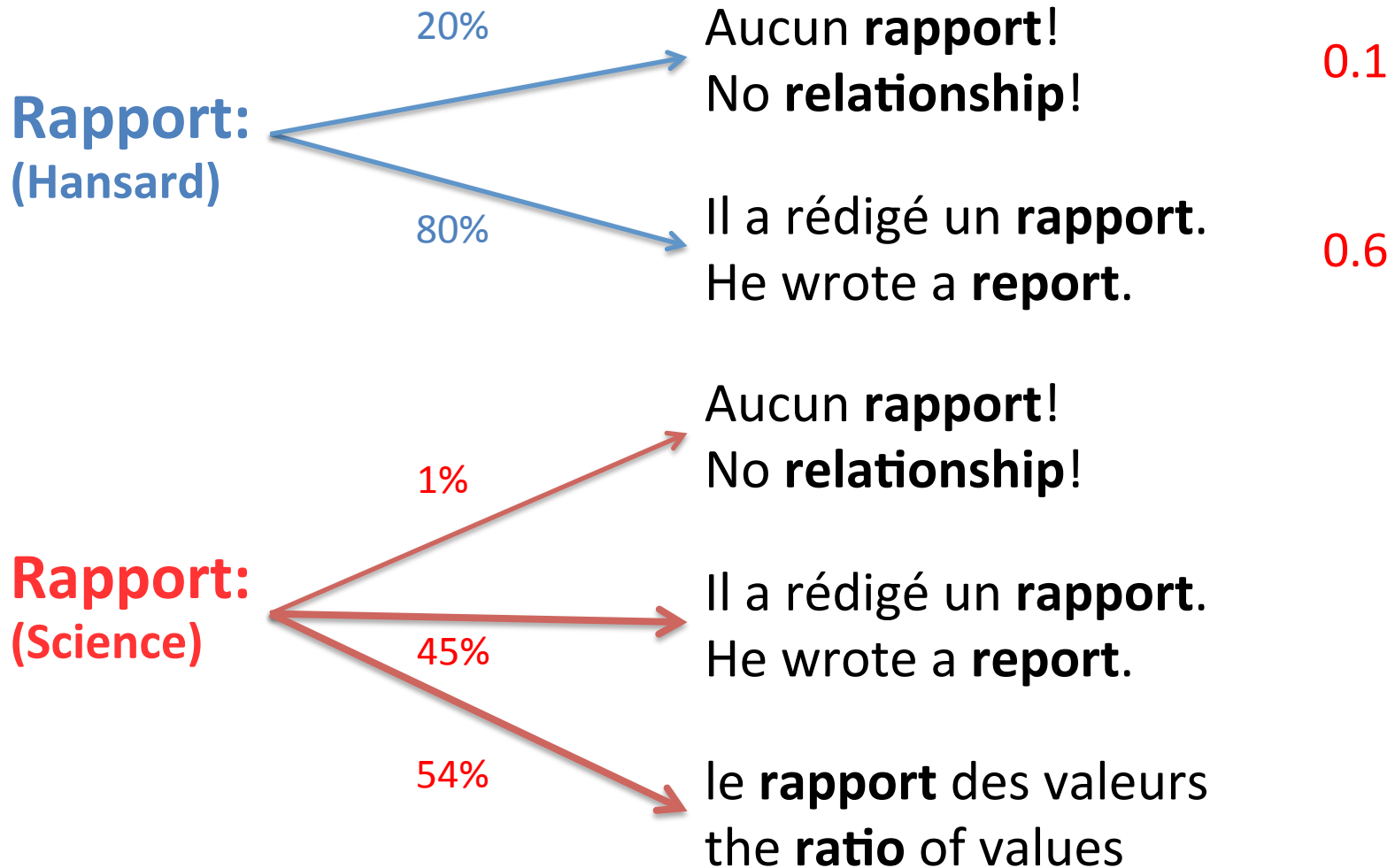
# Instance Weighting



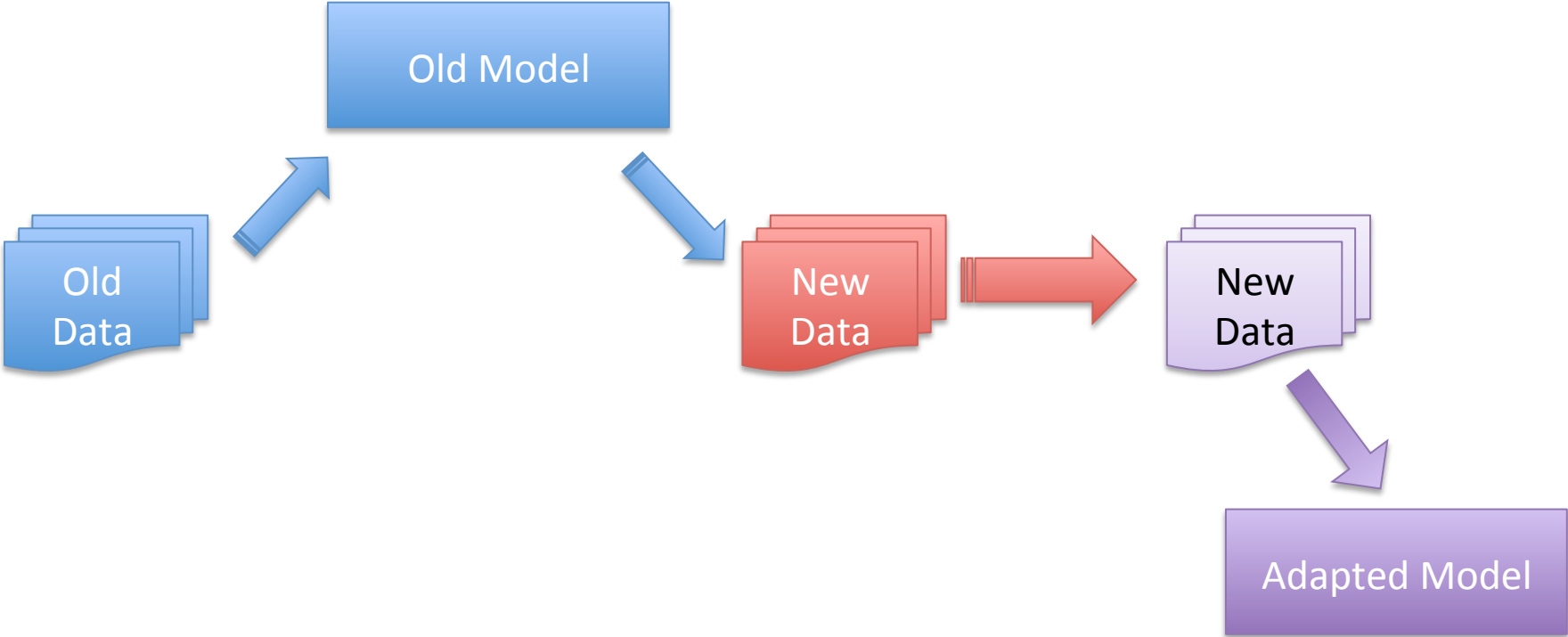
# Instance Weighting



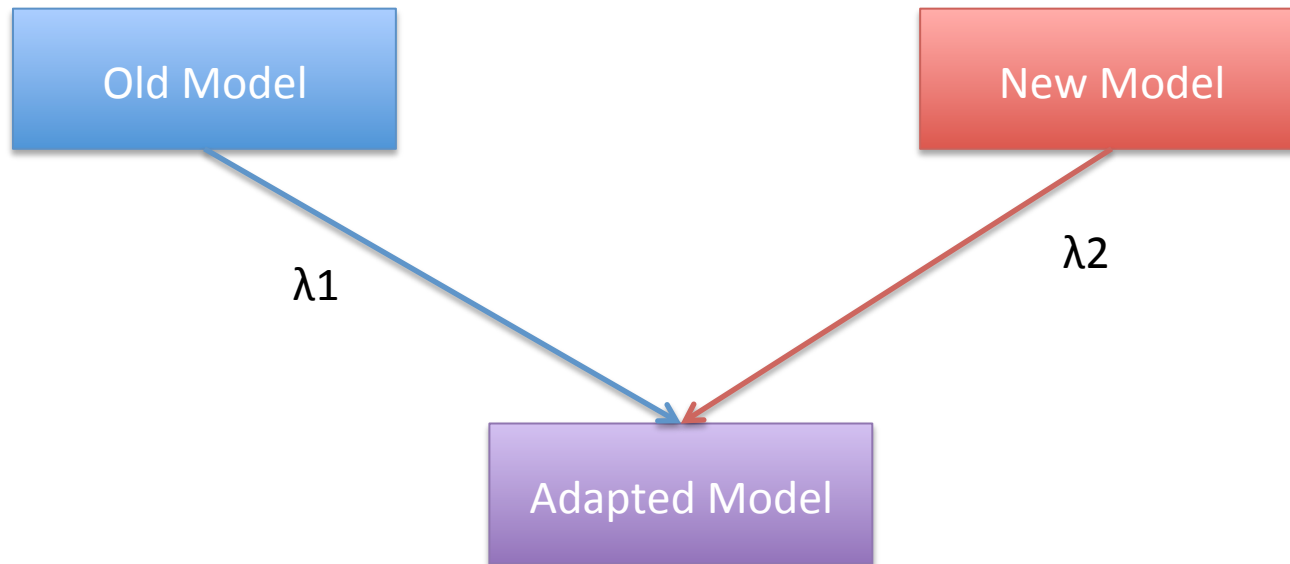
# Instance Weighting



# Old Predictions Feature in New

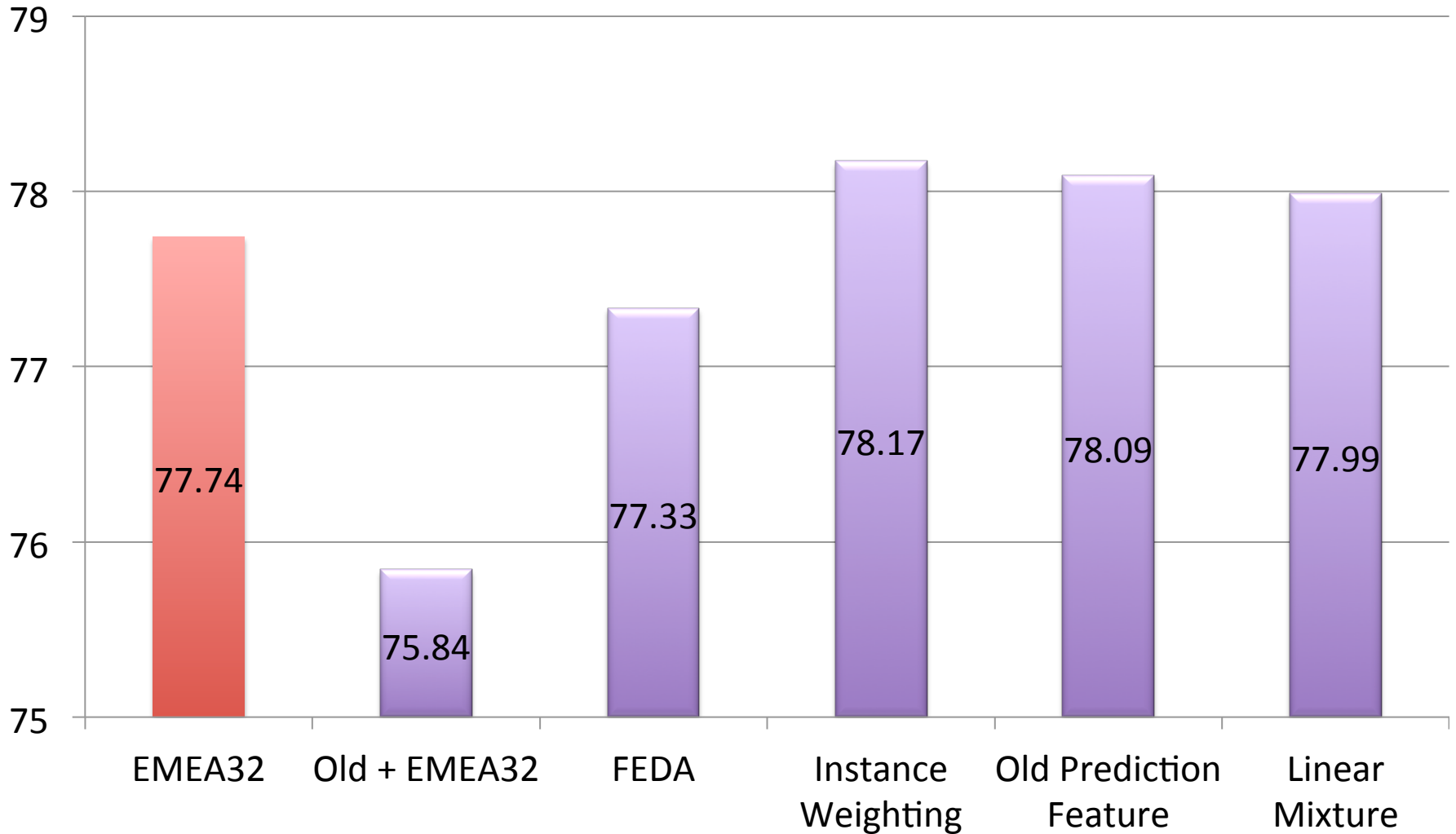


# Model Interpolation

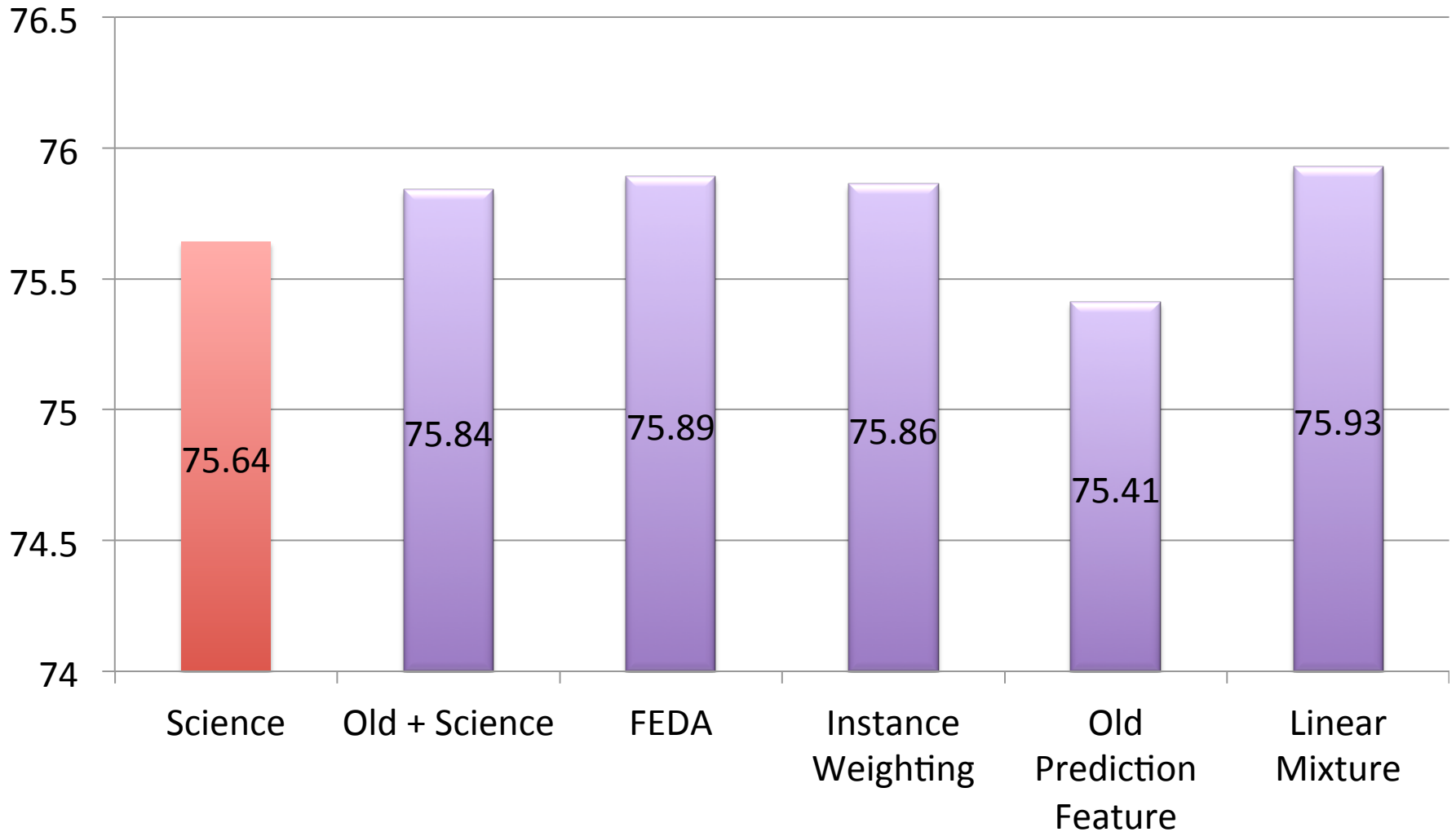


- Linear
- Log-linear
- Cross Validation

# Domain Adaptation Results on EMEA32



# Domain Adaptation Results on Science

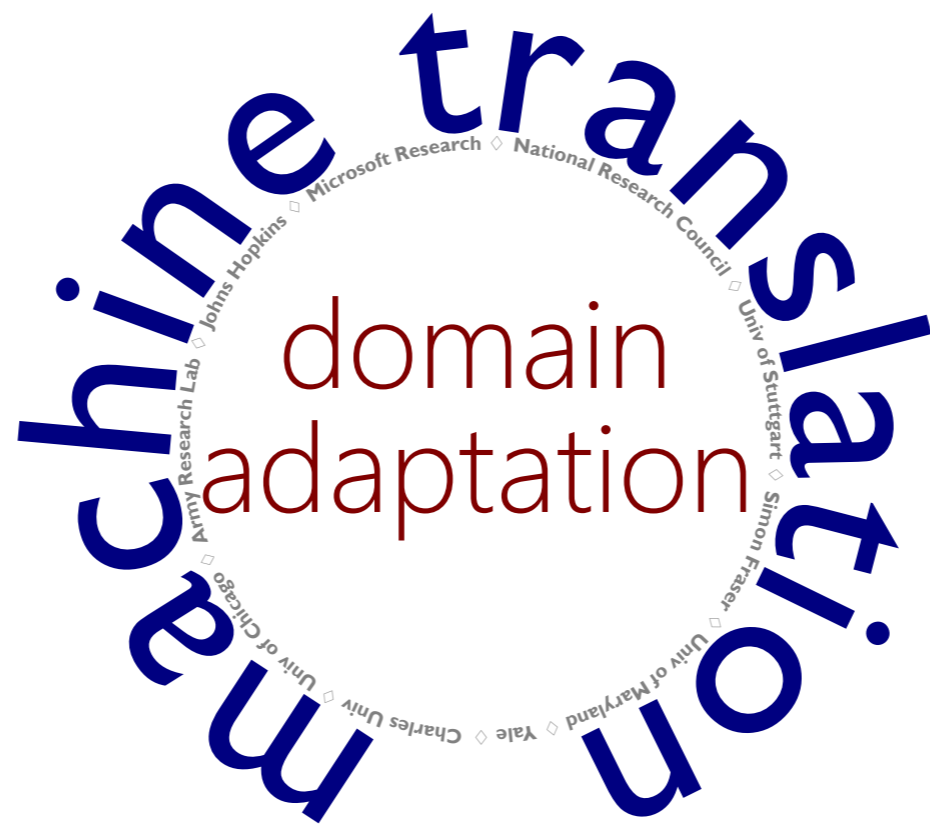




5 MIN BREAK

Anni Irvine

# Translation Mining



# Translation Mining

- Both OOV and sense errors account for a large fraction of translation problems (S4, Sanjeeval)
- Two basic tasks:
  - Find French words that are:
    - OOV (easy)
    - Likely to have a *new* translation (new sense)
  - Get translations for them
- Useful to separate two tasks because different techniques might be useful to solve each

# Spotting New Senses

- Given a stream of *monolingual text* in the new domain, discover word tokens (in context) that appear to have new senses
- General approach:
  - Design features that are indicative of new senses
  - Train a classifier to predict new senses  
(trained on small amounts of parallel data)
  - Apply it to large monolingual corpora

# Translation Mining

# Translation Mining

- Learn translations for:

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*



# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - *OOV* word (types)

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:
  - Dictionary mining approaches using:

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:
  - Dictionary mining approaches using:
    - Old domain parallel data, *comparable* new domain data

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - *OOV* word (types)
- Two ways to translate:
  - Dictionary mining approaches using:
    - Old domain parallel data, *comparable* new domain data
    - Old domain parallel data, *parallel* new domain data

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:
  - Dictionary mining approaches using:
    - Old domain parallel data, *comparable* new domain data
    - Old domain parallel data, *parallel* new domain data
  - Ask bilingual speakers  
(hypothesis: people are better at translating in context than hallucinating words that might be used in a new way)

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:
  - Dictionary mining approaches using:
    - Old domain parallel data, *comparable* new domain data
    - Old domain parallel data, *parallel* new domain data
  - Ask bilingual speakers  
(hypothesis: people are better at translating in context than hallucinating words that might be used in a new way)
- *Spotting* words with new senses

# Translation Mining

- Learn translations for:
  - Words (types/tokens) with *new senses*
  - OOV word (types)
- Two ways to translate:
  - Dictionary mining approaches using:
    - Old domain parallel data, *comparable* new domain data
    - Old domain parallel data, *parallel* new domain data
  - Ask bilingual speakers  
(hypothesis: people are better at translating in context than hallucinating words that might be used in a new way)
- *Spotting* words with new senses
  - Features from above techniques



# Translation Mining:

Learning from document pair marginal distributions

# Old Domain

## French-English Parallel Data

Fr

En

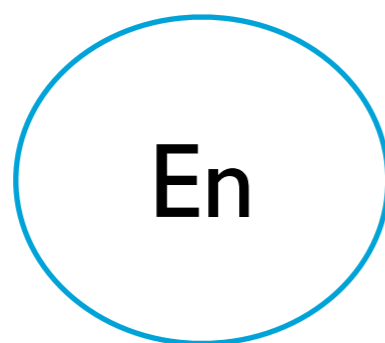
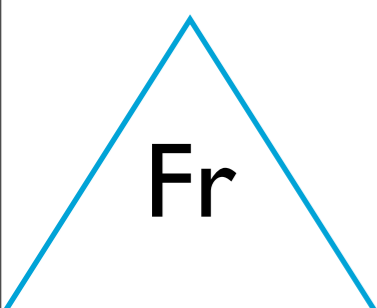


	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

## Old Domain French-English Parallel Data



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				



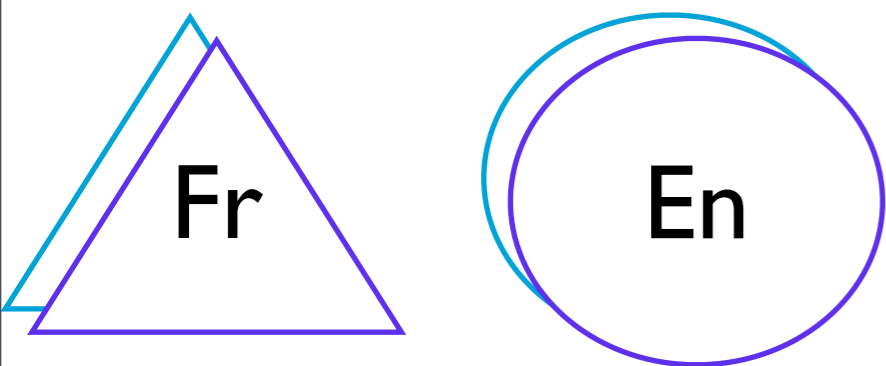
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

## *K* New Domain French-English Comparable Document Pairs

# Old Domain French-English Parallel Data



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				



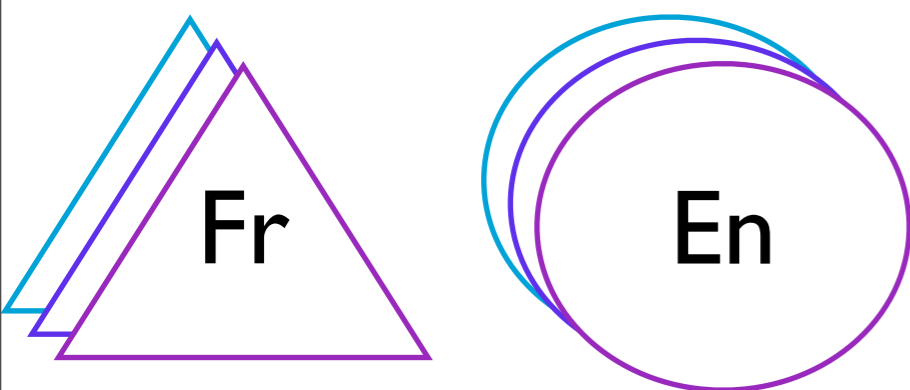
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

# *K* New Domain French-English Comparable Document Pairs

# Old Domain French-English Parallel Data



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				



## *K* New Domain

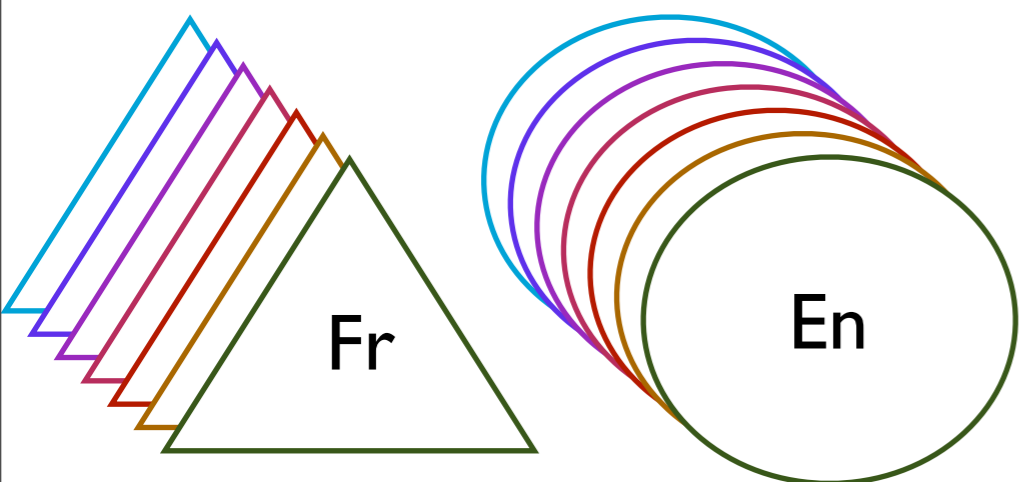
# French-English Comparable Document Pairs

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

# Old Domain French-English Parallel Data



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				



# *K* New Domain French-English Comparable Document Pairs

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	<b>?</b>					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

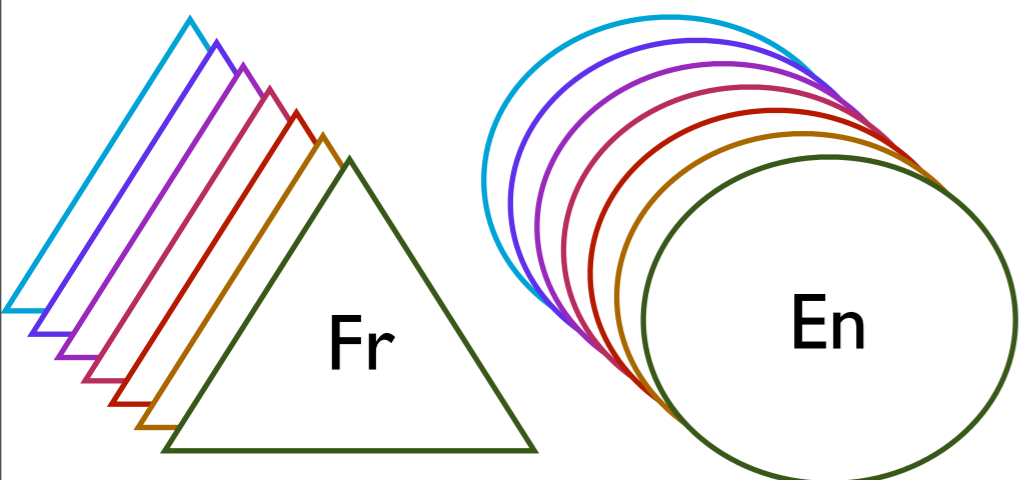
# Old Domain French-English Parallel Data



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

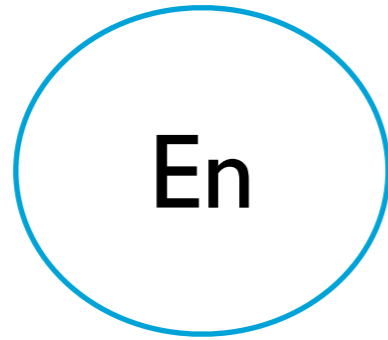
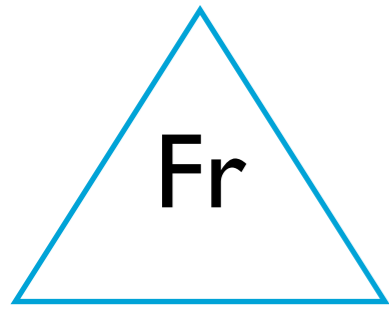
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

New, improved, domain-adapted  $p_k(e, f)$ , updated w.r.t  $k$  comparable documents



# **K New Domain French-English Comparable Document Pairs**

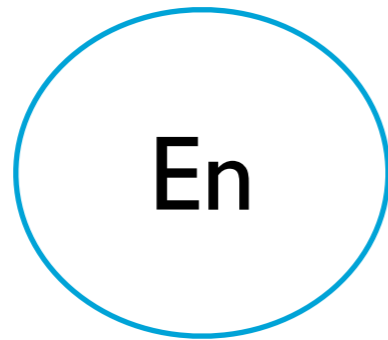
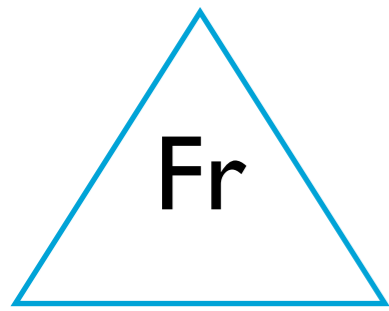
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
						$q(e_1)$



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	<b>?</b>					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...



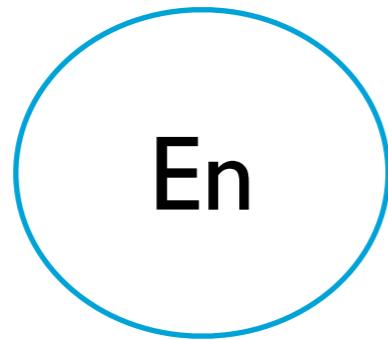
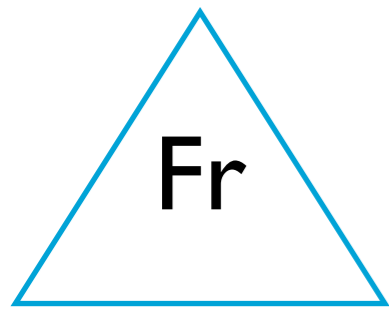


	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$



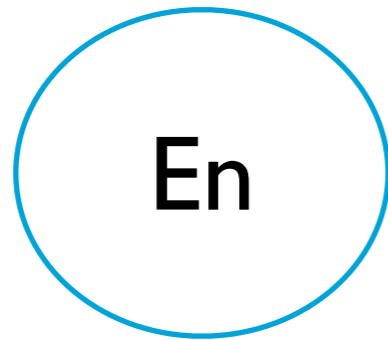
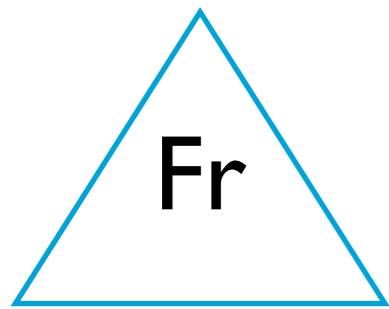
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from  
original joint



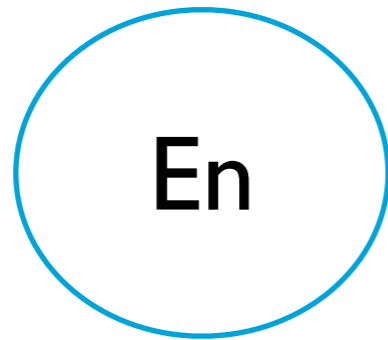
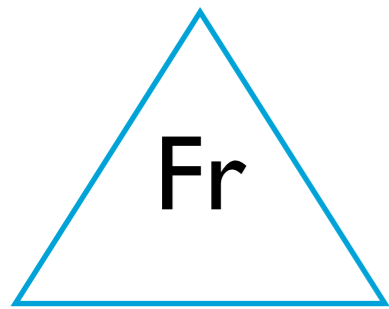
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ : monolingual relative frequency difference

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint



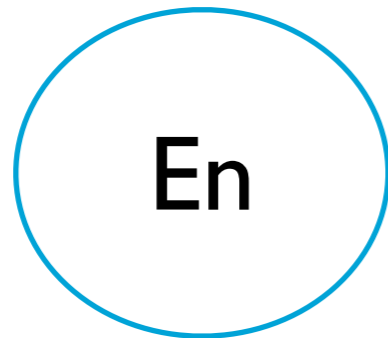
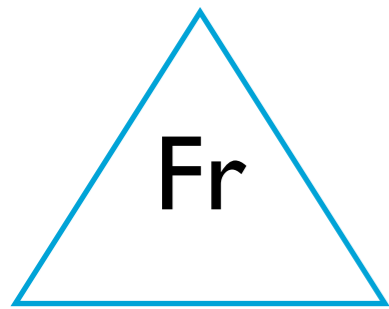
	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ : monolingual relative frequency difference

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint
string edit distance between e and f



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

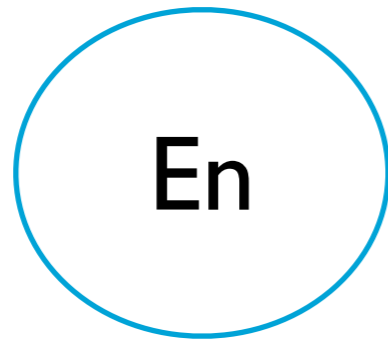
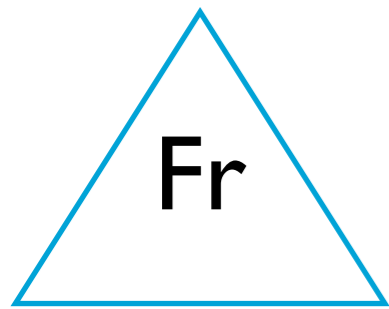
monolingual relative frequency difference

difference between wikipedia page distributions

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint

string edit distance between e and f



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

monolingual relative frequency difference

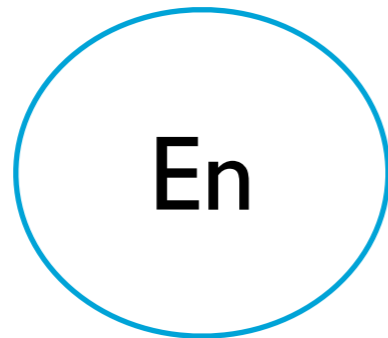
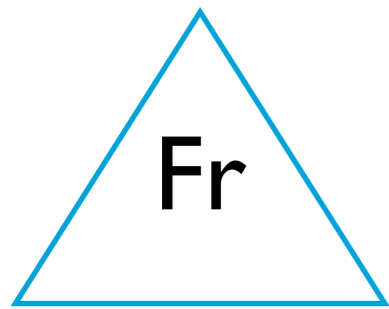
difference between wikipedia page distributions

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint

string edit distance between e and f

Sparsity Penalty



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

monolingual relative frequency difference

difference between wikipedia page distributions

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint

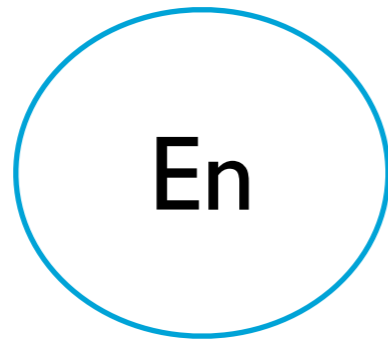
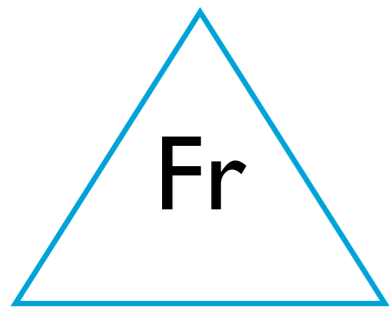
string edit distance between e and f

Sparsity Penalty

Subject to constraints:

$$\sum_{f \in F} \hat{p}(e, f) - q(e) < \epsilon$$

$$\sum_{e \in E} \hat{p}(e, f) - q(f) < \epsilon$$



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$	
$f_1$	?					$q(f_1)$
$f_2$						$q(f_2)$
...						...
$f_{m-1}$						$q(f_{m-1})$
$f_m$						$q(f_m)$
	$q(e_1)$	$q(e_2)$	...	$q(e_{n-1})$	$q(e_n)$	

For each comparable document pair...

Minimize over  $\hat{p}(e,f)$ :

monolingual relative frequency difference

difference between wikipedia page distributions

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) * (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$

distance from original joint

string edit distance between e and f

Sparsity Penalty

Subject to constraints:

Update  $p(e,f)$  in the direction of learned joint:  

$$p_k(e, f) = p_{k-1}(e, f) + \lambda(\hat{p}(e, f) - p_{k-1}(e, f))$$

$$\sum_{f \in F} \hat{p}(e, f) - q(e) < \epsilon$$

$$\sum_{e \in E} \hat{p}(e, f) - q(f) < \epsilon$$



# Translation Mining: Evaluation

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

# Translation Mining: Evaluation

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

**Gold Standard:  
New Domain  
French-English Parallel Data**

Fr

En



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

# Translation Mining: Evaluation

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

- Mean Reciprocal Rank

where is  $\max p_{\text{new}}(e|f)$  in  $p_{\text{learned}}(e|f)$  ranked list over  $f$

**Gold Standard:  
New Domain  
French-English Parallel Data**

Fr

En



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

# Translation Mining: Evaluation

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

- Mean Reciprocal Rank

- Mean Average Precision

AUC under precision-recall curve,  
averaged over  $f$  words;  
recall only up to  $p_{\text{new}}(e|f) > 0.1$

**Gold Standard:  
New Domain  
French-English Parallel Data**

Fr

En



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

# Translation Mining: Evaluation

	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p_k(e_1, f_1)$	$p_k(e_2, f_1)$	...	$p_k(e_{n-1}, f_1)$	$p_k(e_n, f_1)$
$f_2$	$p_k(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p_k(e_1, f_{m-1})$				
$f_m$	$p_k(e_1, f_m)$				

- Mean Reciprocal Rank
- Mean Average Precision
- Conditional Prob. Overlap, Accuracy in Top-k, Divergence between  $p_{\text{new}}$  and  $p_{\text{learned}}$ , Number of OOVs learned about...

**Gold Standard:  
New Domain  
French-English Parallel Data**

Fr

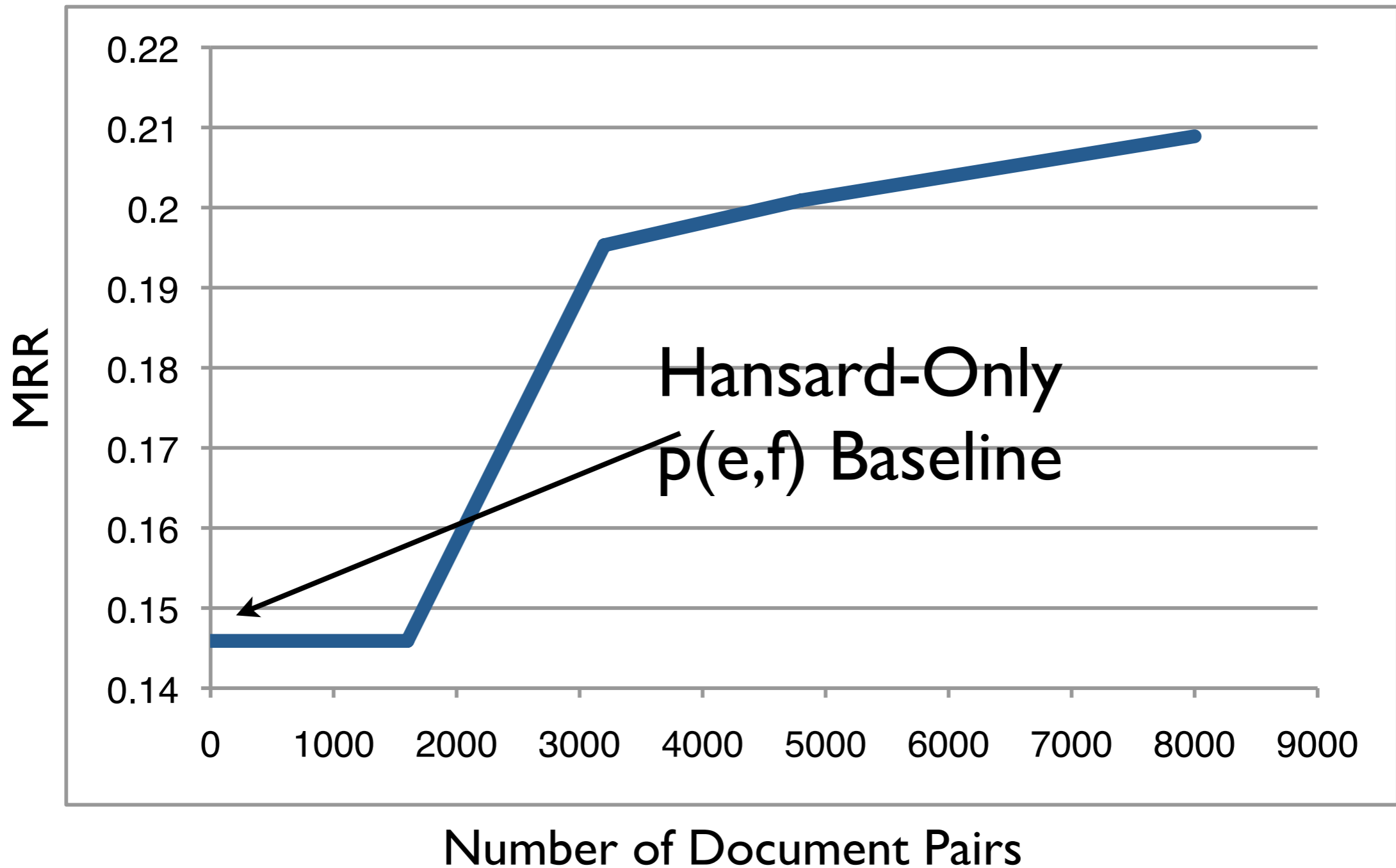
En



	$e_1$	$e_2$	...	$e_{n-1}$	$e_n$
$f_1$	$p(e_1, f_1)$	$p(e_2, f_1)$	...	$p(e_{n-1}, f_1)$	$p(e_n, f_1)$
$f_2$	$p(e_1, f_2)$	...			
...	...				
$f_{m-1}$	$p(e_1, f_{m-1})$				
$f_m$	$p(e_1, f_m)$				

# Translation Mining: Evaluation

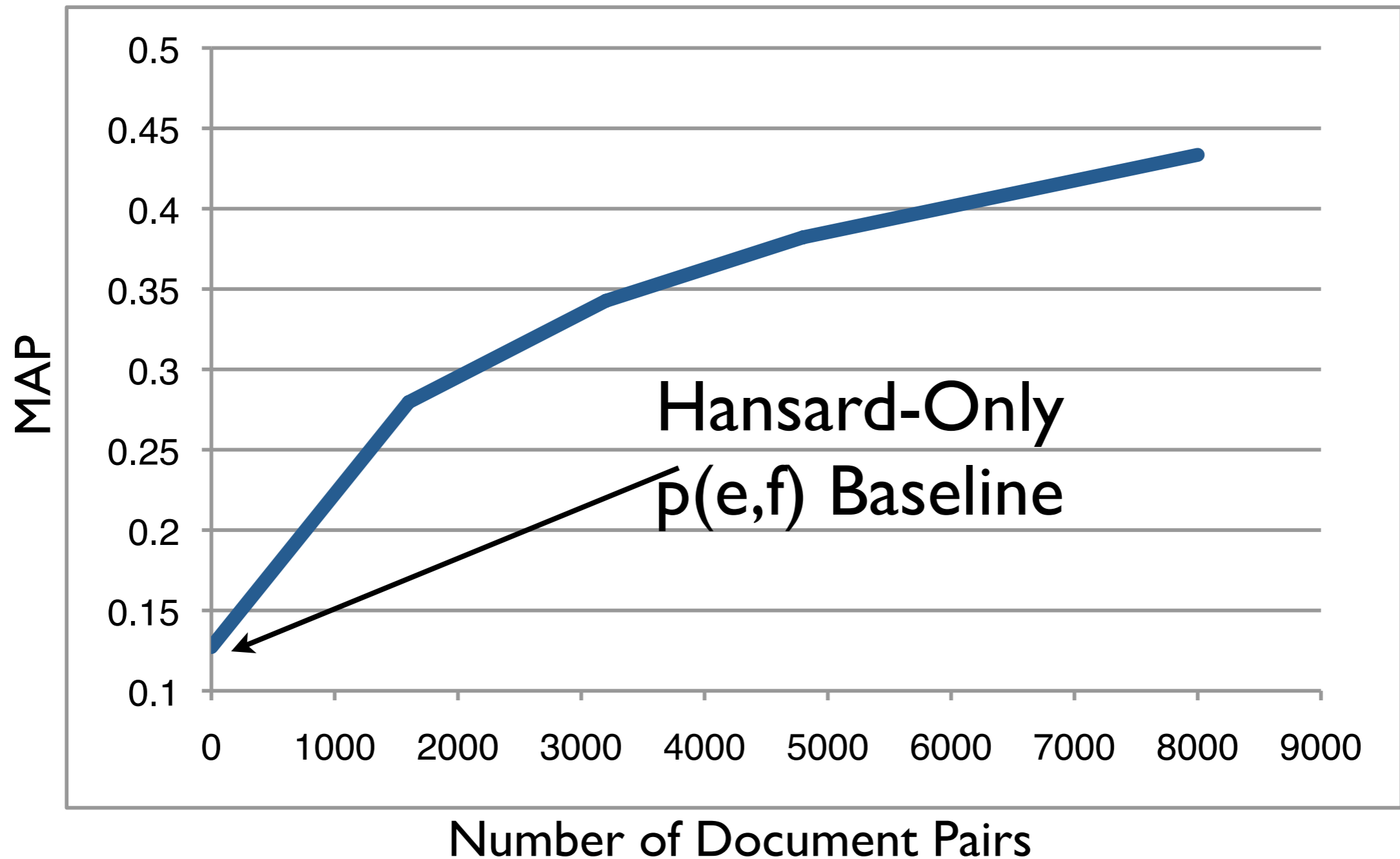
Domain: EMEA



results are similar for Science domain

# Translation Mining: Evaluation

Domain: EMEA



results are similar for Science domain

# Preliminary MT Results

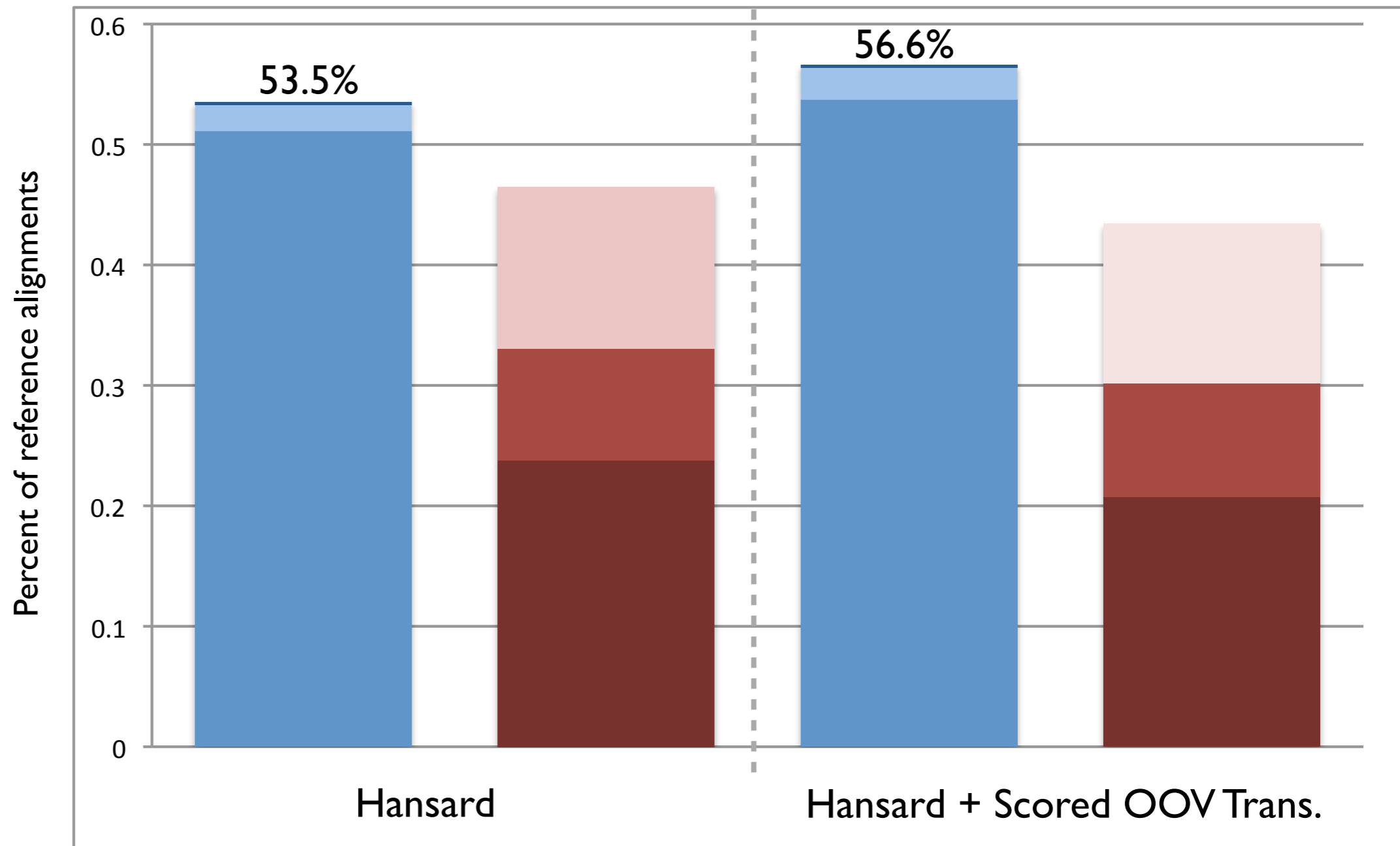
Experimental Setup:

Augment phrase table trained on Hansard-only data  
with OOV Translations  
and learned  $p(e|f)$ ,  $p(f|e)$  scores (as separate features)



# Preliminary MT Results

Sanjeeval



Domain: Science

# Preliminary MT Results

BLEU

Hansard-Trained	26.08
Hansard-Trained + Scored OOV Trans.	26.12

Domain: Science

Rachel Rudinger

# Spotting New Senses

- Binary classification problem:
  - +ve: French token has previously unseen sense
  - -ve: French token is used in a known way
- Experimental framework for feature exploration
  - Supports different classifiers
  - Features at a type or token level
  - Cross validation
  - Feature bucketing
- Results presented as area under the (ROC) curve

# Spotting: Baseline features

- Freq of French word in each domain
- Freq of its translations in the each domains
- Language model perplexities for this word type:
  - Averaged across occurances
  - With variance, max, min and other statistics

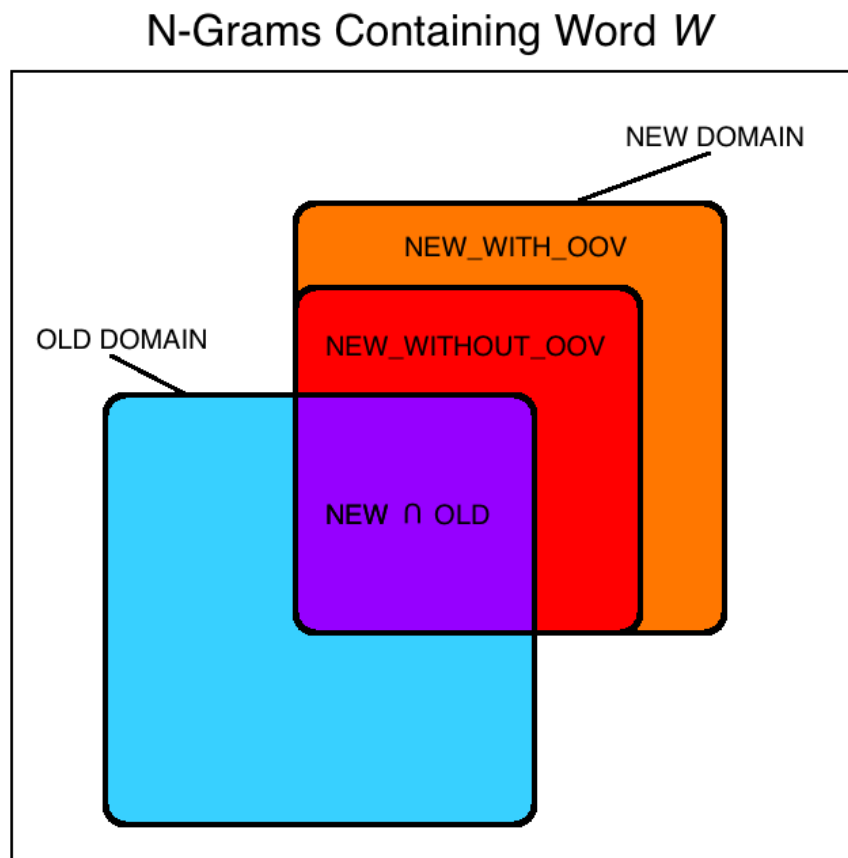
# Detecting Sense Change, Topic Model Approach

- For each word in source language vocabulary (intersection of Old and New domain), compute a score indicating likelihood of gaining new sense in new domain

$$Score(w) = \sum_{k \in topics_{new}} P_{new}(k|w) \times \max_{k' \in topics_{old}} (P_{old}(k'|w) \times cossim(k, k'))$$

- Potential limitations:
  - Noisy topic models
  - Topics may change even if sense does not change
- Preliminary results indicate topic model feature may improve sense classification performance.

# Detecting Sense Change, N-Gram Approach



$$ngram\_score(w) = \frac{|NEW\_WITHOUT\_OOV \setminus (NEW \cap OLD)|}{|NEW\_WITHOUT\_OOV|} \left( = \frac{|RED|}{|RED \cup PURPLE|} \right)$$

# Detecting Sense Change, N-Gram Approach

- Reasons to find word  $w$  in a new n-gram in new domain:
  1. Argument change, e.g. “**run** from bears” ; “**run** from lepidoptera”
  2. Sense change, e.g. “**run** for office” ; “**run** a program”
  3. Noise, e.g. n-gram overlaps with other phrase,  
“**run** and he” ; “done , run”
- Want to find words with many instances of reason 2.
- Ignoring phrases with OOVs may help reduce noise from reasons 1 and 3.
- If high scores correlate with words with new senses, score may be used as a feature in new sense detection.



# Document Pair Marginal Matching Features

- Word **type** features:

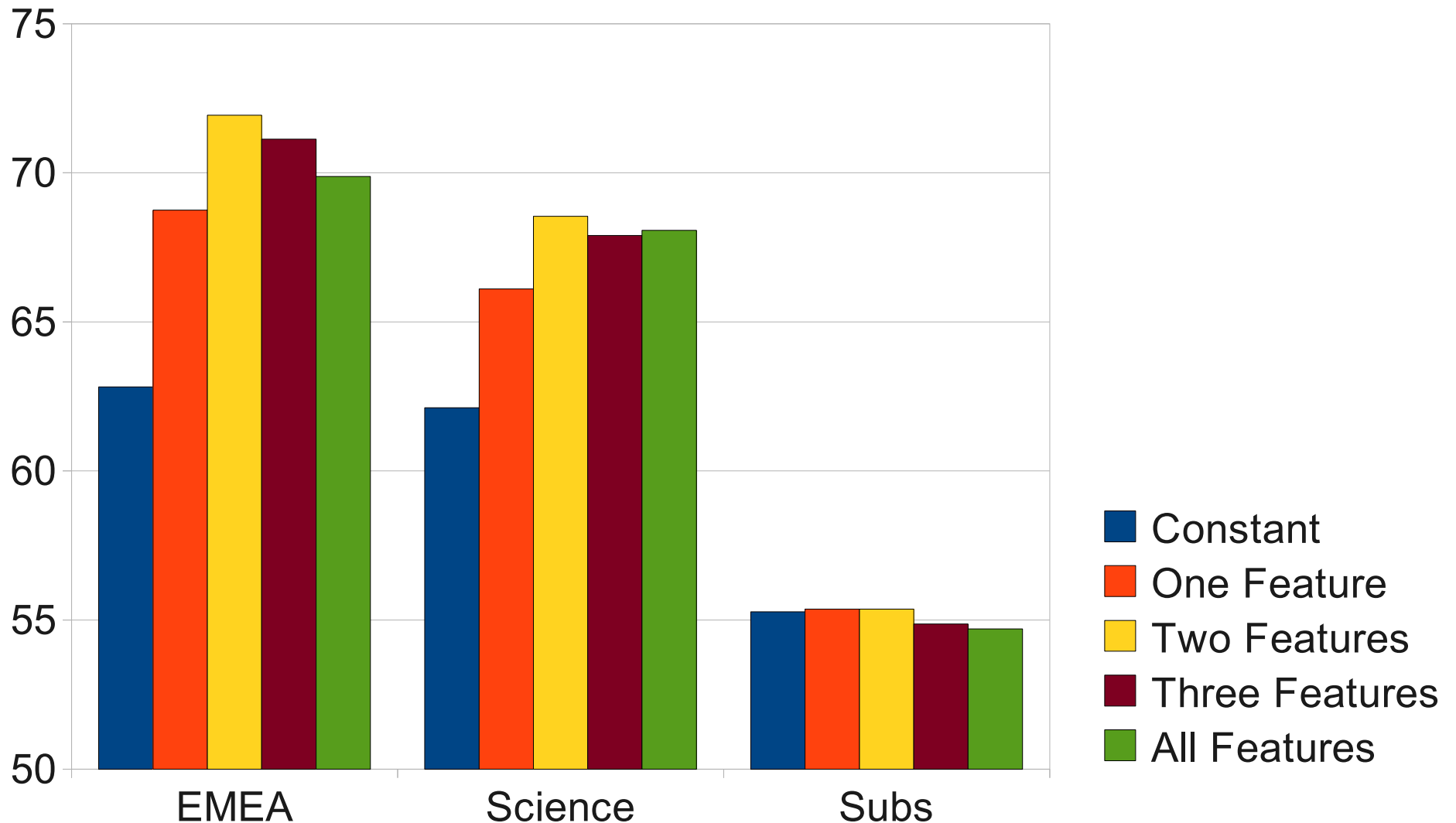
- $p_{\text{learned}}(f) > 0$  ?

- $\max_e p_{\text{old}}(e|f) = p_{\text{learned}}(e|f)$  ?

- $\text{overlap} [\text{top-5}_e p_{\text{old}}(e|f), \text{top-5}_e p_{\text{learned}}(e|f)] / 5$

- $\text{overlap} [\text{top-2}_e p_{\text{old}}(e|f), \text{top-2}_e p_{\text{learned}}(e|f)] / 2$

# 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

Ann Clifton

# Document-Level Info in MT

Topic Models for Machine Translation

# Topic Models for Machine Translation

Intuition: knowing the document-level topic of data can help resolve ambiguity

Example: 'he couldn't find a **match**.'

- ▶ '...they held Nigeria's first bone marrow drive. He couldn't find a match there..'
- ▶ 'The company allowed smoking in a designated indoor smoking room. However, he couldn't find a match.'

# Topic Models for Machine Translation

Intuition: knowing the document-level topic of data can help resolve ambiguity

Example: 'he couldn't find a **match**.'

- ▶ '...they held Nigeria's first bone marrow drive. He couldn't find a match there..'
- ▶ 'The company allowed smoking in a designated indoor smoking room. However, he couldn't find a match.'

# Topic Models for Machine Translation

Intuition: knowing the document-level topic of data can help resolve ambiguity

Example: 'he couldn't find a **match**.'

- ▶ '...they held Nigeria's first bone marrow drive. He couldn't find a match there..'
- ▶ 'The company allowed smoking in a designated indoor smoking room. However, he couldn't find a match.'

# Topic Models for Machine Translation

Intuition: knowing the document-level topic of data can help resolve ambiguity

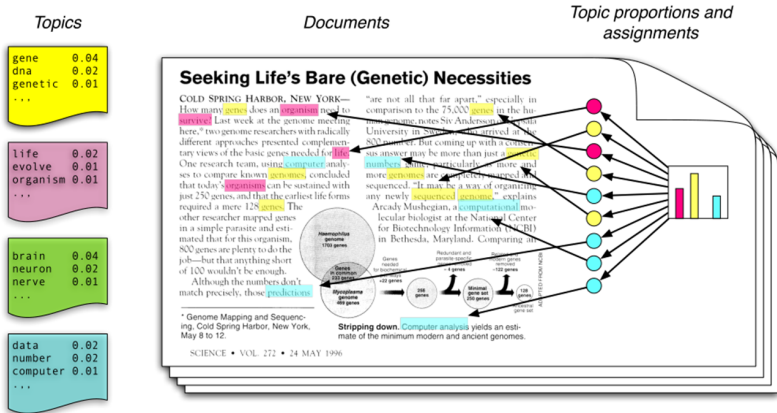
Example: 'he couldn't find a **match**.'

- ▶ '...they held Nigeria's first bone marrow drive. He couldn't find a match there..'
- ▶ 'The company allowed smoking in a designated indoor smoking room. However, he couldn't find a match.'



# Topic Models for Domain Adaptation in MT

Scenario: little new-domain parallel data, but plenty new-domain monolingual data



(Blei, 2011)

# Lexical Weighting with Topic Models: Using Comparable Data

Compute expected count  $e_{z_n}(e, f)$  under topic  $z_n$ :

$$e_{z_n}(e, f) = \sum_{d_i \in T} p(z_n | d_i) \sum_{x_j \in d_i} c_j(e, f)$$

Compute lexical probability conditioned on topic distribution:

$$p_{z_n}(e, |f) = \frac{e_{z_n}(e, f)}{\sum_e e_{z_n}(e, f)}$$

# Intrinsic Evaluation

	no-topic	doc-topic	word-topic
old-alignment, old topic	-1.78	-0.47	-0.48
new-domain, new-topic	-1.12	-0.26	-0.26
old-domain, new-topic	-1.78	-0.27	-0.27

Table: Average per-word log likelihood of EMEA data

# Lexical Weighting in Phrase-Based MT

As feature in phrase-based MT:

$$f_{z_n}(\bar{e}|\bar{f}) = -\log\{p_{z_n}(\bar{e}, |\bar{f})p(z_n|d)\}$$

$$\sum_p \lambda_p h_p(\bar{e}, \bar{f}) + \sum_{z_n} \lambda_{z_n} f_{z_n}(\bar{e}|\bar{f})$$

# Lexical Weighting with Topic Models: Using Parallel Data

## Issues with generative topic models:

- ▶ each word affects topic selection equally, regardless of how informative it is ('the' versus 'hexachordal')
- ▶ each topic learns an independent distribution, though some words' meaning change with topic ('the' versus 'play')

# Lexical Weighting with Topic Models: Using Parallel Data

Issues with generative topic models:

- ▶ each word affects topic selection equally, regardless of how informative it is ('the' versus 'hexachordal')
- ▶ each topic learns an independent distribution, though some words' meaning change with topic ('the' versus 'play')

# Lexical Weighting with Topic Models: Using Parallel Data

Issues with generative topic models:

- ▶ each word affects topic selection equally, regardless of how informative it is ('the' versus 'hexachordal')
- ▶ each topic learns an independent distribution, though some words' meaning change with topic ('the' versus 'play')

## A Discriminative Topic Model

conditional likelihood of a target document given a source document, using a mixture of latent topics:

$$P(T|S) = \sum_{z \in Z} \left( P(z|S) \prod_{(s,t) \in (S,T)} P(t|s,z) \right)$$

The topic distribution is predicted based on features of the whole source document:

$$P(z|S) \propto \exp(\theta \cdot F(S, z))$$

Each translation is predicted based only on the source word and a given topic likelihood:

$$P(t|s,z) \propto \exp(\phi \cdot G(s, z, t))$$



## A Discriminative Topic Model

conditional likelihood of a target document given a source document, using a mixture of latent topics:

$$P(T|S) = \sum_{z \in Z} \left( P(z|S) \prod_{(s,t) \in (S,T)} P(t|s,z) \right)$$

The topic distribution is predicted based on features of the whole source document:

$$P(z|S) \propto \exp(\theta \cdot F(S, z))$$

Each translation is predicted based only on the source word and a given topic likelihood:

$$P(t|s,z) \propto \exp(\phi \cdot G(s, z, t))$$

## A Discriminative Topic Model

conditional likelihood of a target document given a source document, using a mixture of latent topics:

$$P(T|S) = \sum_{z \in Z} \left( P(z|S) \prod_{(s,t) \in (S,T)} P(t|s,z) \right)$$

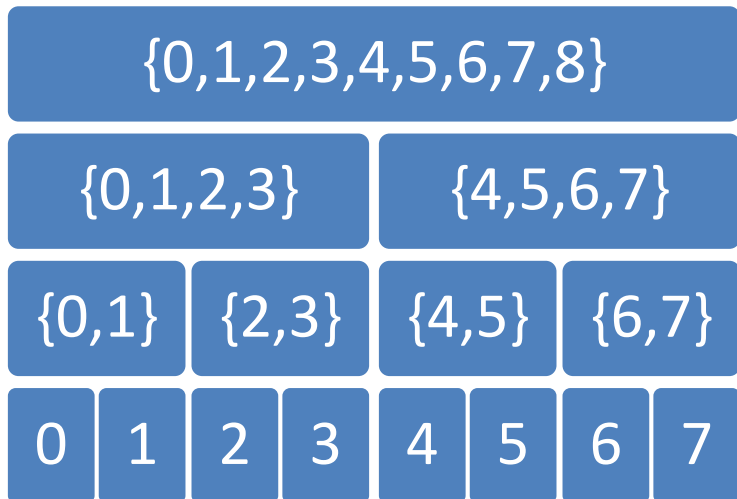
The topic distribution is predicted based on features of the whole source document:

$$P(z|S) \propto \exp(\theta \cdot F(S, z))$$

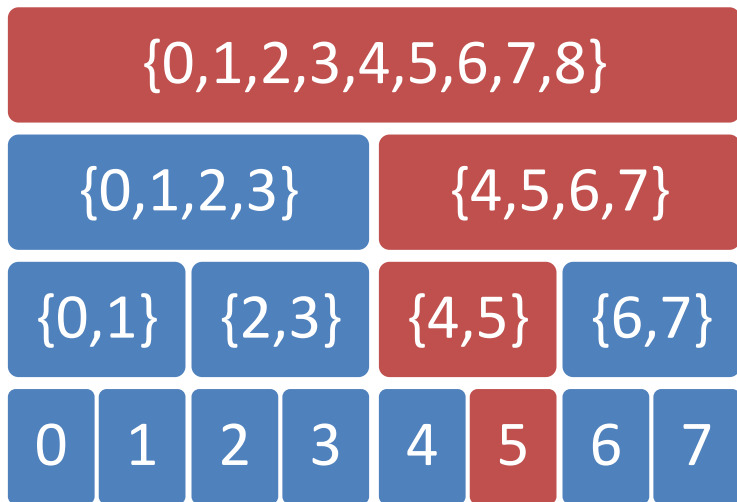
Each translation is predicted based only on the source word and a given topic likelihood:

$$P(t|s,z) \propto \exp(\phi \cdot G(s, z, t))$$

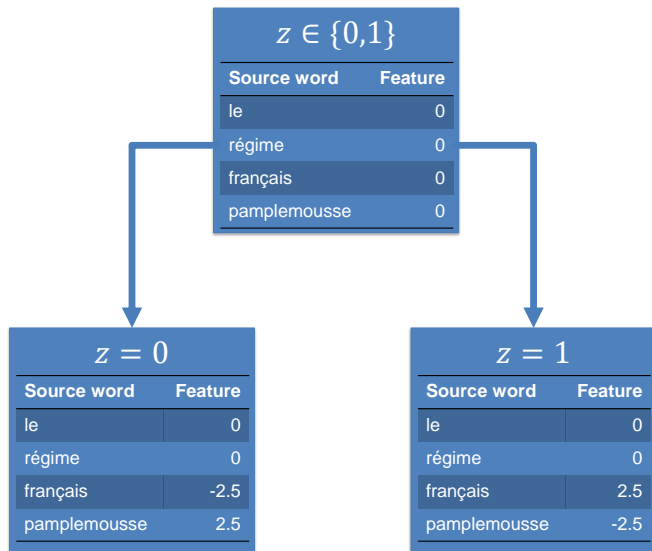
## A Discriminative Topic Model: Hierarchical Topic Features



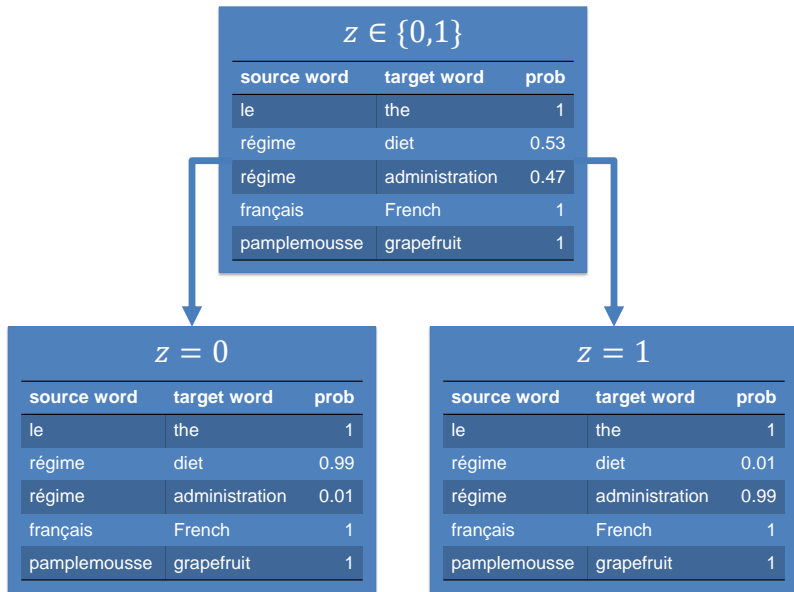
## A Discriminative Topic Model: Hierarchical Topic Features



# A Discriminative Topic Model: Hierarchical Topic Features



# A Discriminative Topic Model: Hierarchical Topic Distributions



# A Discriminative Topic Model: Example

(1a) le<sub>1</sub> régime<sub>2</sub> français<sub>3</sub>

(1b) the<sub>1</sub> French<sub>3</sub> administration<sub>2</sub>

(2a) le<sub>1</sub> régime<sub>2</sub> pamplemousse<sub>3</sub>

(2b) the<sub>1</sub> grapefruit<sub>3</sub> diet<sub>2</sub>

	topic 0	topic 1
sentence 1	0.01	0.99
sentence 2	0.99	0.01

# Discriminative Topic Model: Current Implementation Status

Improvements shown in log likelihoods on held-out data;  
further considerations:

- ▶ initialization
- ▶ regularization
- ▶ feature engineering



Jagadeesh  
Jagarlamudi

# **Mining Token Level Translations**

# From Type to Token

Adapt type level translations to token level

rapport	report	0.4
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**

# From Type to Token

Adapt type level translations to token level

rapport	report	0.4
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**

rapport	report	0.5
rapport	reporting	0.3



# From Type to Token

Adapt type level translations to token level

rapport	report	0.4
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**

rapport	report	0.5	<input checked="" type="checkbox"/>
rapport	reporting	0.3	

le **rapport** des valeurs

rapport	values	0.7	<input checked="" type="checkbox"/>
rapport	report	0.2	

**Take home !!!**

Intentionally left blank

# How can it help MT ?

1. Token level translations to be fed into MT  
Provide sentence specific translations
- 2.
- 3.

# How can it help MT ?

1. Token level translations to be fed into MT  
Provide sentence specific translations
2. Mine translations for the new Sense
- 3.



# How can it help MT ?

1. Token level translations to be fed into MT  
Provide sentence specific translations
2. Mine translations for the new Sense
3. Gather more training instances for PSD  
Add new sense/OOV words and their translations

# Main Idea

Step 1  
Word aligned parallel data

le rapport des valeurs



The ratio of values

# Main Idea

## Step 2

Learn vector representations

Vectors capture the word meaning

le

0.8
-0.2
0.1
...

rapport

-0.3
0.4
0.2
...

des

0.7
0.3
0.4
...

valeurs

0.5
-0.2
0.8
...

The

0.6
-0.1
0.2
...

ratio

-0.2
0.3
0.4
...

of

0.6
0.2
-0.1
...

values

0.4
-0.1
0.6
...

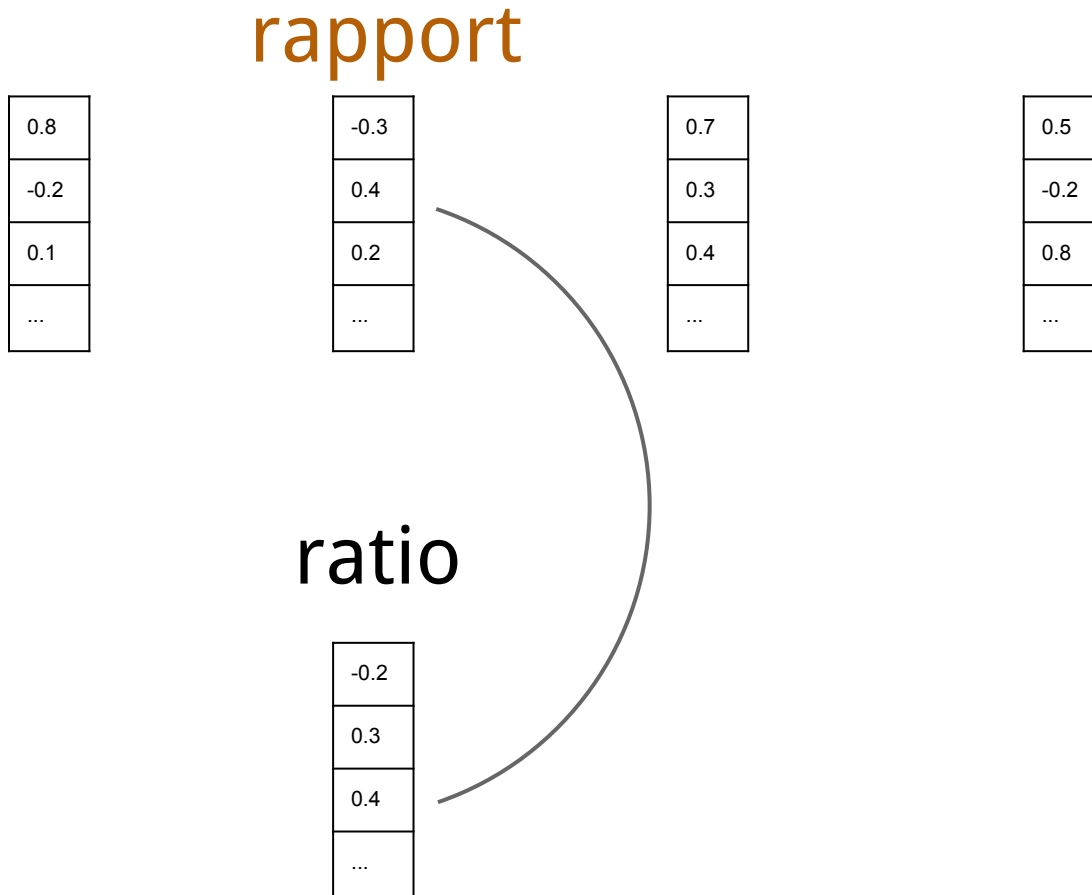


# Main Idea

## Step 2

Replace words with vectors

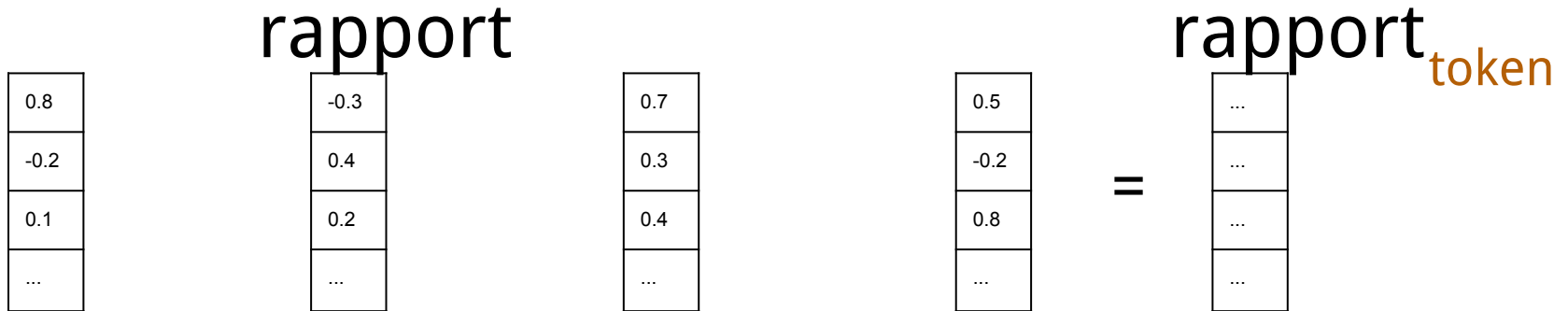
Vectors capture the word meaning



# Main Idea

## Step 3

Token representation is a weighted combination of the context vectors



ratio

-0.2
0.3
0.4
...

# Main Idea

## Step 3

Token representation is a weighted combination of the context vectors

rapport

$$\begin{matrix} w_{la} \\ \begin{matrix} 0.8 \\ -0.2 \\ 0.1 \\ \dots \end{matrix} \end{matrix} + \begin{matrix} w_0 \\ \begin{matrix} -0.3 \\ 0.4 \\ 0.2 \\ \dots \end{matrix} \end{matrix} + \begin{matrix} w_{des} \\ \begin{matrix} 0.7 \\ 0.3 \\ 0.4 \\ \dots \end{matrix} \end{matrix} + \begin{matrix} w_{valeurs} \\ \begin{matrix} 0.5 \\ -0.2 \\ 0.8 \\ \dots \end{matrix} \end{matrix} = \begin{matrix} \text{rapport} \\ \text{token} \\ \begin{matrix} \dots \\ \dots \\ \dots \\ \dots \end{matrix} \end{matrix}$$

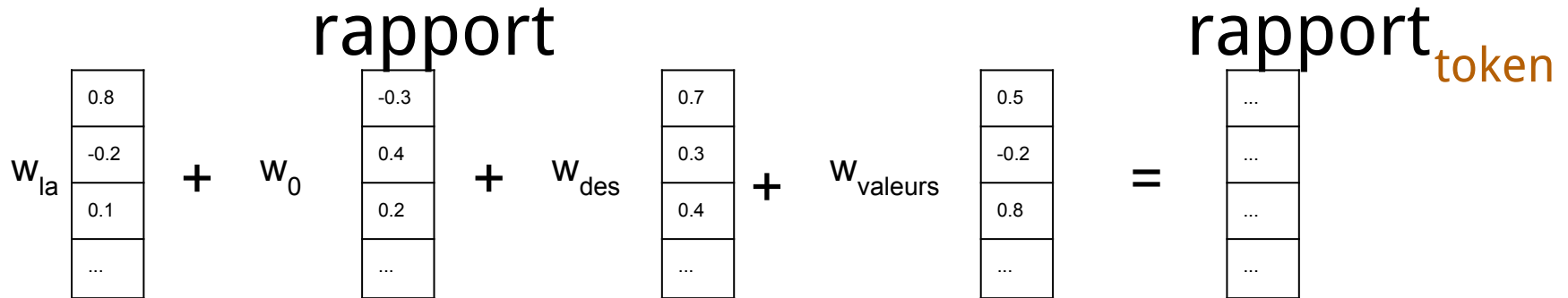
ratio

-0.2
0.3
0.4
...

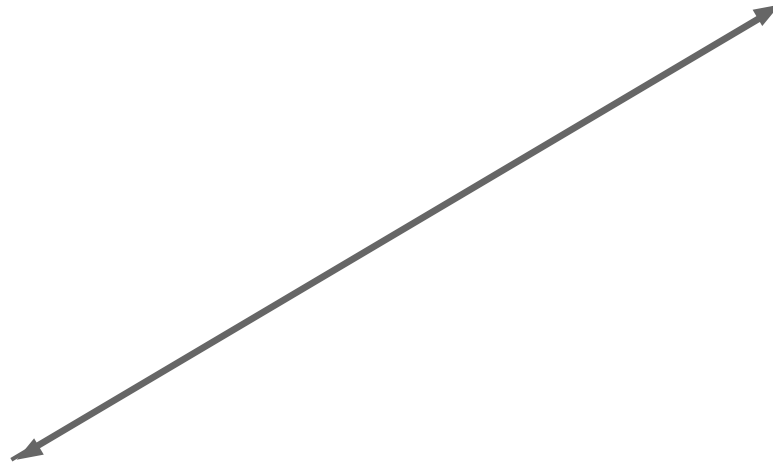
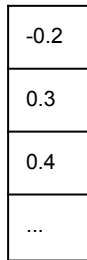
# Main Idea

## Step 3

Token representation is a weighted combination of the context vectors



ratio



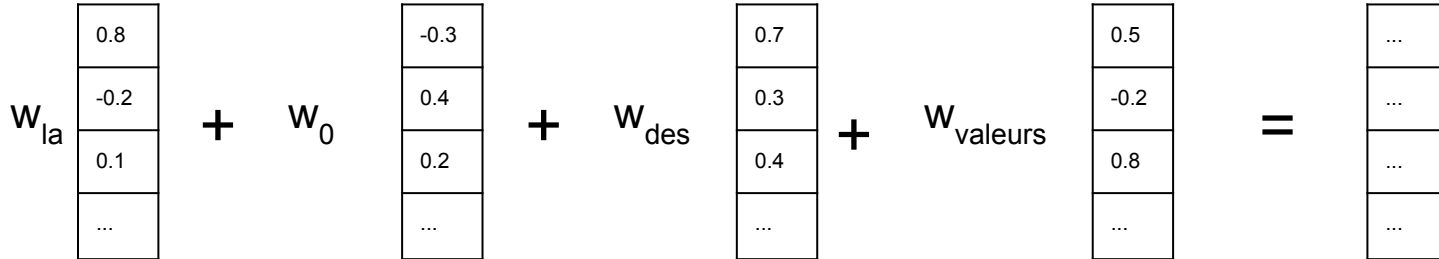
# Main Idea

## Step 3

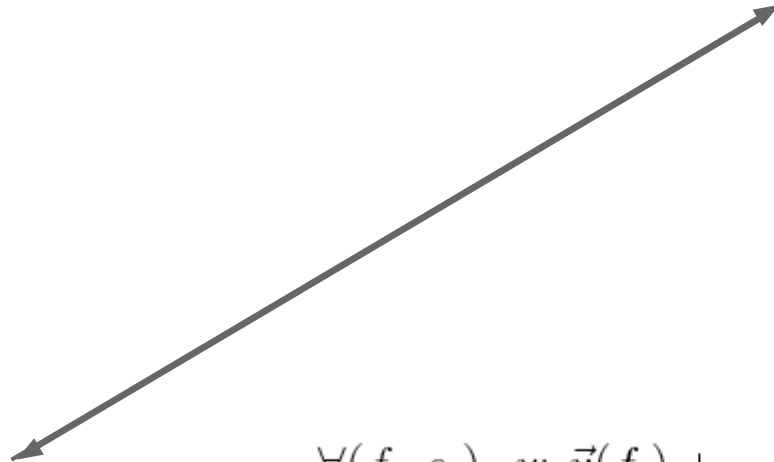
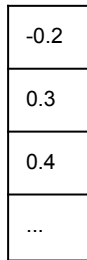
Token representation is a weighted combination of the context vectors

rapport

rapport<sub>token</sub>



ratio



$$\forall (f_i, e_i) \quad w_0 \vec{v}(f_i) + \sum_{f_j \in \text{Ctx}(f_i)} w_{f_j} \vec{v}(f_j) \approx \vec{v}(e_i)$$



# Extensions

## 1. Co-regularization

Weights are independent of the focus word  
Add dependency but regularize

2.

# Co-regularization

$W_{la}$   $+$   $W_0$   $+$   $W_{des}$   $+$   $W_{valeurs}$   $=$   $W_{token}$

rappor

0.8
-0.2
0.1
...

-0.3
0.4
0.2
...

0.7
0.3
0.4
...

0.5
-0.2
0.8
...

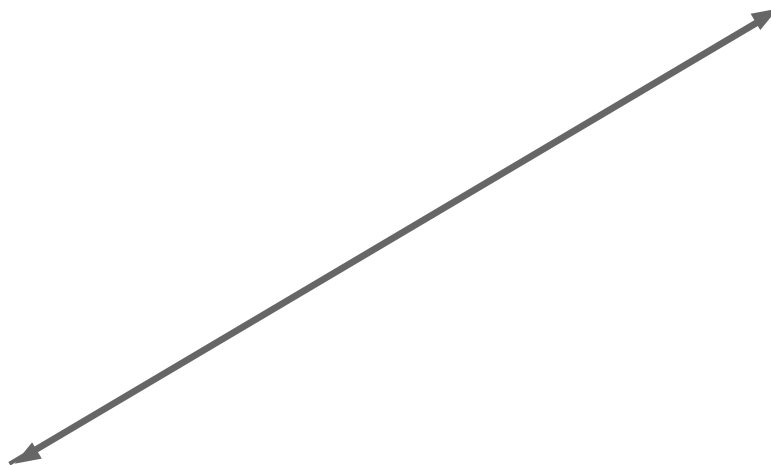
...
...
...
...

rappor

token

ratio

-0.2
0.3
0.4
...



# Co-regularization

rapport  $W_{la}$  + rapport  $W_0$  + rapport  $W_{des}$  + rapport  $W_{valeurs}$  = rapport  $token$

0.8
-0.2
0.1
...

+

-0.3
0.4
0.2
...

+

0.7
0.3
0.4
...

+

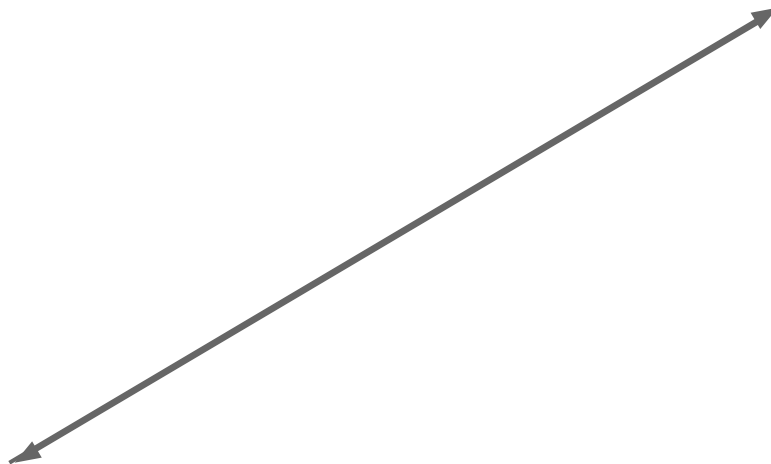
0.5
-0.2
0.8
...

=

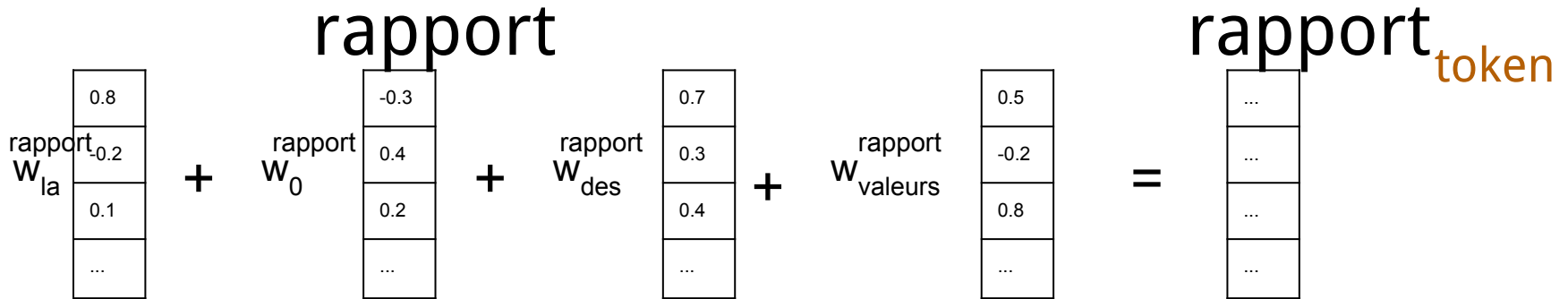
...
...
...
...

ratio

-0.2
0.3
0.4
...



# Co-regularization

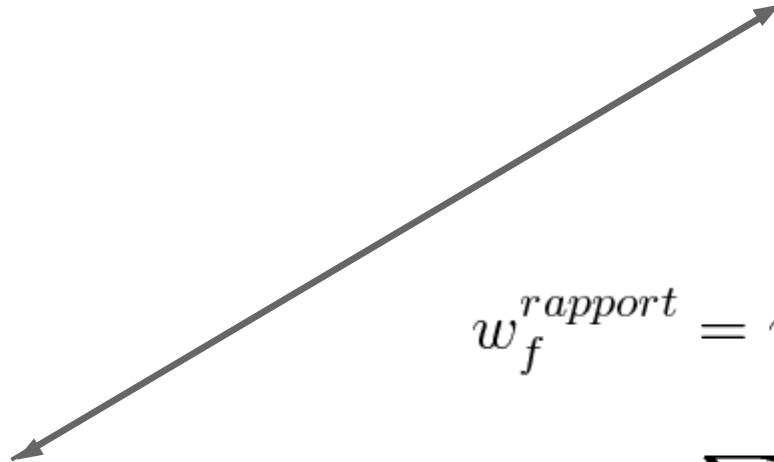


ratio

-0.2
0.3
0.4
...

$$w_f^{rapport} = w_f + r_f^{rapport}$$

$$\text{add } \left\| \sum_f r_f^{rapport} \right\|^2$$



# Extensions

## 1. Co-regularization

Weights are independent of the focus word  
Add dependency but regularize

## 2. Maximum-margin style model

Ignores the candidate translations  
Favor the correct translation  
But move away from the other candidates

# Intrinsic evaluation

On 7.4K tokens from EMEA

rapport	report	0.5
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**

# Intrinsic evaluation

On 7.4K tokens from EMEA

Method	Accuracy Top
Random	40.29
Max Probable -- $p(e   f)$	57.84
Best Cue-Word	61.85

rapport	report	0.5
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**

# Intrinsic evaluation

On 7.4K tokens from EMEA

Method	Accuracy Top
Random	40.29
Max Probable -- $p(e   f)$	57.84
Best Cue-Word	61.85

Token adapted

Simple adaptation	55.21
Co-regularization	59.15
Max-Margin	60.21
Coreg+MaxMargin	??

rapport	report	0.5
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

Il a rédigé un **rapport**



# Intrinsic evaluation

On 7.4K tokens from EMEA

rapport	report	0.5
rapport	relationship	0.1
rapport	reporting	0.05
rapport	values	0.3

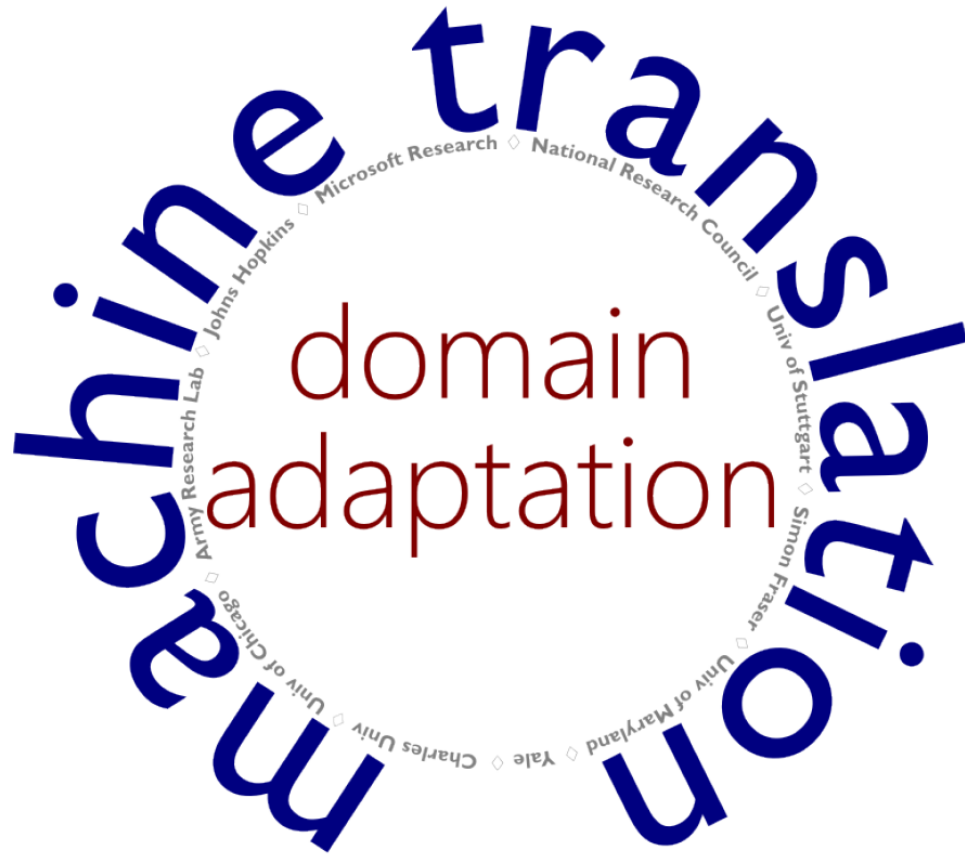
Il a rédigé un **rapport**

Method	Accuracy Top
Random	40.29
Max Probable -- $p(e   f)$	57.84
Best Cue-Word	61.85

Token adapted

Simple adaptation	55.21
Co-regularization	59.15
Max-Margin	60.21
Coreg+MaxMargin	??
PSD Classifier	70.10

Marine Carpuat



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

summary & conclusion

# Summary: Analysis of domain effects

- **Not uniform across domains**
  - Starting OLD domain = Hansard
  - News does not significantly benefit from NEW domain data
  - All other domains benefit substantially from NEW data
- **Baseline adaptation methods are only sometimes effective**
  - Concatenating OLD and NEW data often harms both
  - Linear or log-linear mixtures are a better starting point
  - But there is large room for improvement
- **Errors are distributed amongst SEEN, SENSE and SCORE**
  - In most NEW domains
  - Contextual information can substantially improve translation quality

# Summary: Phrase Sense Disambiguation for DAMT

- Discriminative context-dependent translation lexicon
- Can model lexical choice across domains
  - Context model can fix lexical choice errors
  - But adaptation algorithms not useful yet
    - ~90% accuracy at domain detection with current representation
    - Simple adaptation methods target hard-to-distinguish domains
- Integrated in Moses
  - Fast fully-automated experiment pipeline
  - but still buggy...

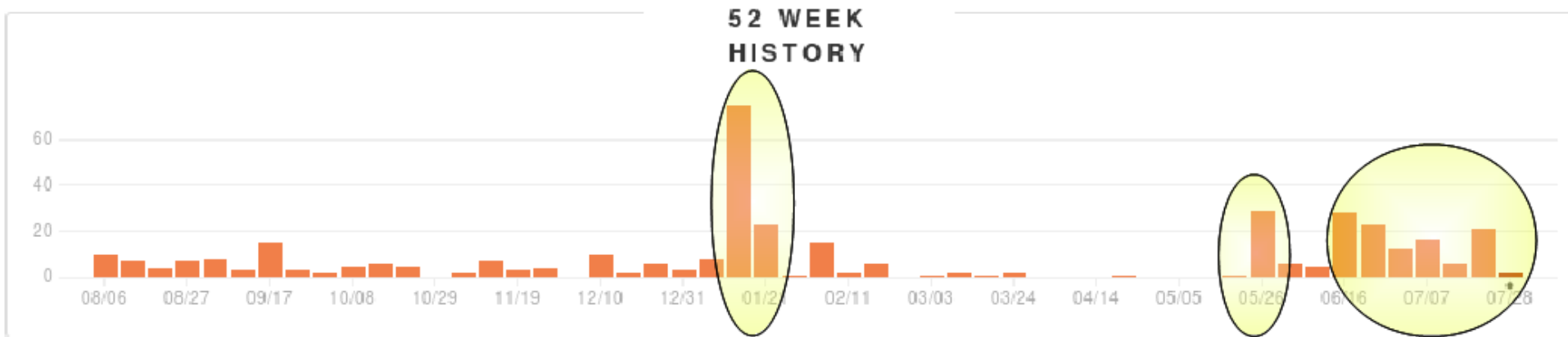
# Summary: Mining new senses and their translations

- We can detect new senses
  - improved from mid 60s AUC to 70+ on EMEA and Science
  - lots of successful feature exploration: ngram, topic, marginal matching, LM perplexity and others
- We can mine some useful translations for OOVs from comparable/parallel data
  - Using new document pair marginal matching
  - Using low-dimensional embeddings
- We can learn topic distinctions targeted at MT

# Contributions: VW

- From stand-alone tool to linkable library
- Extended core classifier
  - label dependent features
  - cost-sensitive classification
  - support for complex feature interaction
- Many bug fixes!

# Contributions: VW



Lines of code committed to VW over the past year



# Contributions: Moses

- Parallelized significance-based phrase-table pruning, many optimizations
- Improved experiment management system
- Many bug fixes
- 247 commits to github, 6917 lines of code added

# Contributions: VW-Moses integration

- First general purpose classifier in Moses
- Tight solid integration
  - can be built and run out-of-the-box, extended with new features, etc
  - Fast: 180% run time of standard Moses, and fully parallelized
- Both in Phrase-based and Hiero Moses
  - Common interface
  - Consistent feature definitions

# Contributions: methodology

- Defined MT domain adaptation tasks
  - On multiple domains: Medical, Science, Subtitles
  - Controlled conditions
- Defined translation lexical choice tasks
  - Translation disambiguation & new sense detection
  - On same data as MT test sets
  - Target domain-relevant vocabulary
- Experiment management system for automatic evaluation of new features
- Everything will be freely available online

# Contributions: new techniques

- Complex classifier integration in SMT decoder
  - feature extraction framework shared between Hiero and Phrase-based decoding frameworks
- New discriminative topic modeling
  - domain-specific
  - translation-aware
- New document-pair marginal matching for translation mining
- Dictionary mining at the token level

# Future work: next steps

- Debug extrinsic PSD
- Improve DA representation
- Extend soft-syntactic features for Hierarchical Moses further
- Integrate mined translation examples and topic models into MT and PSD
- Package up data and software for release
  - Moses+VW already available!
- Final report

# Future work: longer term directions

- Non-lexical domain divergence issues
  - promising preliminary results using syntax
- Other language pairs and directions
  - More distant language
  - Into morphologically richer languages
- Less structured text/genre
  - informal communication
- Scale topic models to really large heterogeneous corpora
  - toward web translation

# Thanks to

- George Foster, Colin Cherry and the Portage team @NRC
- John Langford
- Moses-support
- Cameron Macdonald, Patrik Lambert, Holger Schwenk
- Vlad Eidelman, Kristy Hollingshead, Wu Ke, Gideon Maillette de Buy Wenniger, Ferhan Ture
  
- Dan Povey
- Sanjeev, Monique, Ruth, Lauren, Mani\*, and CLSP
  
- NSF, Google, DOD

# machine translation

domain adaptation

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

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