machine translation

domain adaptation

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

<table>
<thead>
<tr>
<th>Old Domain (Parliament)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td>monsieur le président, les pêcheurs de homard de la région de l'atlantique sont dans une situation catastrophique.</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td>mr. speaker, lobster fishers in atlantic canada are facing a disaster.</td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
<tr>
<td>mr. speaker, the lobster fishers in atlantic canada are in a mess.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td>comprimés pelliculés blancs pour voie orale.</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td>white film-coated tablets for oral use.</td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
<tr>
<td>white <em>pelliculés</em> tablets to oral.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td>mode et voie(s) d'administration</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td>method and route(s) of administration</td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
<tr>
<td><em>fashion</em> and voie(s) of directors</td>
</tr>
</tbody>
</table>

**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
  - 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

Parallel data

Translation Rules (+context)

traitement ↔ treatment
le traitement ↔ the salary

Context-aware MT
Approach

Parallel data

New domain parallel data

traitement ↔ treatment
le traitement ↔ the salary
Translation Rules (+context)

traitement ↔ processing
Translation Rules (+context)

Context-aware MT
Approach

Parallel data

New domain parallel data

traitement ↔ treatment
le traitement ↔ the salary
Translation Rules (+context)

traitement ↔ processing
Translation Rules (+context)

Context-aware MT
Approach

Parallel data

traitement ↔ treatment
le traitement ↔ the salary
Translation Rules (+context)

Context-aware MT
Parallel data

Translation Rules
(traitement ↔ treatment)
(le traitement ↔ the salary)

Context-aware MT

New domain comparable data

Active Learning

Mining

Phrase Pairs
(traitement ↔ processing)
(+context)

Unk. Sense

traitement ↔ ???
Approach

Parallel data

Translation Rules
(traitement ↔ treatment)
(le traitement ↔ the salary)

Context-aware MT

Active Learning

Phrase Pairs
(traitement ↔ processing)
(+context)

Mining

New domain comparable data

Unk. Sense

traitement ↔ ???
Approach

Parallel data

New domain parallel data

New domain comparable data

Context-aware MT

Translation Rules (+context)

traitement ↔ treatment
le traitement ↔ the salary

Translation Rules (+context)

traitement ↔ processing

Translation Rules (+context)

Active Learning

Mining

Phrase Pairs (+context)

traitement ↔ processing

Unk. Sense

traitement ↔ ???
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.
**EMEA**: 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.
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:

<table>
<thead>
<tr>
<th>SOURCE: French</th>
<th>OLD</th>
<th>INTERSECTION</th>
<th>NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>T A R G E T: English</td>
<td>OLD</td>
<td>INTERSECTION</td>
<td>NEW</td>
</tr>
</tbody>
</table>
Measuring domain overlap (cont’d)

- Multiple possible items to count:

```
<table>
<thead>
<tr>
<th>SOURCE: French</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLD</td>
</tr>
<tr>
<td>INTERSECTION</td>
</tr>
<tr>
<td>NEW</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>TARGET: English</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLD</td>
</tr>
<tr>
<td>INTERSECTION</td>
</tr>
<tr>
<td>NEW</td>
</tr>
</tbody>
</table>
```

PAIRS: OLD, INTERSECTION, NEW
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 $\rightarrow$ EMEA

<table>
<thead>
<tr>
<th></th>
<th>French types</th>
<th>English types</th>
<th>Pair types</th>
<th>French tokens</th>
<th>English tokens</th>
<th>Pair tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLD∩NEW</td>
<td>17845</td>
<td>13743</td>
<td>63087</td>
<td>6124518</td>
<td>5522972</td>
<td>6290162</td>
</tr>
<tr>
<td>NEW-OLD</td>
<td>16779</td>
<td>15920</td>
<td>431877</td>
<td>419575</td>
<td>381324</td>
<td>2002943</td>
</tr>
</tbody>
</table>
Hansards $\mapsto$ Science

<table>
<thead>
<tr>
<th></th>
<th>OLD $\cap$ NEW</th>
<th>NEW - OLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>French types</td>
<td>40016</td>
<td>77653</td>
</tr>
<tr>
<td>English types</td>
<td>32947</td>
<td>81270</td>
</tr>
<tr>
<td>Pair types</td>
<td>135247</td>
<td>879423</td>
</tr>
<tr>
<td>French tokens</td>
<td>4057191</td>
<td>235429</td>
</tr>
<tr>
<td>English tokens</td>
<td>3358471</td>
<td>244328</td>
</tr>
<tr>
<td>Pair tokens</td>
<td>3995699</td>
<td>1179428</td>
</tr>
</tbody>
</table>
SMT quality across domains: Coarse mixture models can help BLEU, sometimes

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

<table>
<thead>
<tr>
<th>Domain</th>
<th>News</th>
<th>EMEA</th>
<th>Science</th>
<th>Subs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>22.61</td>
<td>22.72</td>
<td>21.22</td>
<td>13.64</td>
</tr>
<tr>
<td>New</td>
<td>20.33</td>
<td><strong>34.83</strong></td>
<td>32.49</td>
<td>20.57</td>
</tr>
<tr>
<td>Old+New</td>
<td><strong>23.82</strong></td>
<td>34.76</td>
<td>??</td>
<td>??</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Domain</th>
<th>News</th>
<th>EMEA</th>
<th>Science</th>
<th>Subs</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td>20.27</td>
<td>40.84</td>
<td>32.48</td>
<td><strong>25.50</strong></td>
</tr>
<tr>
<td>Mix LM</td>
<td>21.57</td>
<td>40.95</td>
<td>32.60</td>
<td><strong>25.51</strong></td>
</tr>
<tr>
<td>Mix LM + Mix TM</td>
<td><strong>23.50</strong></td>
<td><strong>41.47</strong></td>
<td><strong>32.78</strong></td>
<td><strong>25.38</strong></td>
</tr>
</tbody>
</table>

Warning: scores not comparable across 2 tables! (different MT systems, larger test sets!)
Analysis: how difficult is lexical choice across domains?

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

- 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

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

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

<table>
<thead>
<tr>
<th>domain</th>
<th>OLD</th>
<th>SEEN ADD</th>
<th>SENSE ADD</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>23.82</td>
<td>23.68</td>
<td>23.93</td>
</tr>
<tr>
<td>Medical</td>
<td>22.62</td>
<td>32.77 (45%)</td>
<td>32.58 (44%)</td>
</tr>
<tr>
<td>Science</td>
<td>21.22</td>
<td>26.36 (24%)</td>
<td>25.58 (21%)</td>
</tr>
<tr>
<td>Subtitles</td>
<td>13.64</td>
<td>16.60 (22%)</td>
<td>17.75 (30%)</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>domain</th>
<th>OLD</th>
<th>SCORE NEW</th>
<th>CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>23.82</td>
<td>23.93</td>
<td>23.82</td>
</tr>
<tr>
<td>Medical</td>
<td>22.62</td>
<td>30.69 (36%)</td>
<td>40.53</td>
</tr>
<tr>
<td>Science</td>
<td>21.22</td>
<td>26.20 (23%)</td>
<td>30.17</td>
</tr>
<tr>
<td>Subtitles</td>
<td>13.64</td>
<td>17.65 (29%)</td>
<td>20.41</td>
</tr>
</tbody>
</table>

*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 (Sanjeeeval)

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

• Sanjeeeval 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 (Sanjeeeval)

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

• Sanjeeeval 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
(Sanjeeeval)

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
(Sanjeeeval)

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)

Percent of reference alignments

Hansard | EMEA | Hansard-32 + EMEA

- Sense-Freebie
- OOV-Freebie
- Correct
- New Sense
- OOV-Wrong
- Score/Search Error
Extrinsic Word-Level Evaluation
(Sanjeeval)

Percent of reference alignments

- Sense-Freebie
- OOV-Freebie
- Correct
- New Sense
- OOV-Wrong
- Score/Search Error

Hansard
Science
Hansard-32 + Science
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 \textit{rapport}

\begin{itemize}
  \item PSD = phrase translation as classification
  \item PSD at test time
    \begin{itemize}
      \item use context to predict correct English translation of French phrase
    \end{itemize}
  \item PSD at train time
    \begin{itemize}
      \item use word alignment to extract training instances
      \item occurrences of French phrases \textbf{in context} are annotated with their English translations
    \end{itemize}
\end{itemize}
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:
    
    $x = \text{le croissant rouge}$
    $y_1 = \text{the red croissant}$
    $y_2 = \text{the croissant red}$
    $y_3 = \text{the croissant}$
    $y_4 = \text{the red}$

    $x = \text{mange}$
    $y_1 = \text{eat}$
    $y_2 = \text{eats}$
    $y_3 = \text{ate}$

  - Different inputs have different #s and definitions of possible “labels,” each with it's own features
  - Define feature space as $X \times Y$ 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
  – \( P(\text{e_phrase}|\text{f_phrase},\text{f_context}) \) added as a feature function to the decoder
    • Tuned along with standard phrase table features such as \( p(\text{e_phrase}|\text{f_phrase}) \) 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?
<table>
<thead>
<tr>
<th>Rep Phrase</th>
<th>Hansard</th>
<th>EMEA</th>
<th>Science</th>
<th>Subs</th>
</tr>
</thead>
<tbody>
<tr>
<td>'état'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>enceinte</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>formation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rapport</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>régime</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Train</th>
<th>Exact</th>
<th>Stem</th>
<th>Synonym</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansard</td>
<td>78.02%</td>
<td>0.85%</td>
<td>1.86%</td>
<td>9.17%</td>
</tr>
<tr>
<td>EMEA</td>
<td>93.16%</td>
<td>0.68%</td>
<td>0.75%</td>
<td>0.86%</td>
</tr>
<tr>
<td>Hansard + EMEA</td>
<td>92.52%</td>
<td>0.45%</td>
<td>0.56%</td>
<td>1.92%</td>
</tr>
</tbody>
</table>

Meteor alignments between representative phrases from Moses output and reference set

Precision of Alignments

<table>
<thead>
<tr>
<th>Train</th>
<th>Synonym</th>
<th>Paraphrase</th>
<th>Either</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansard</td>
<td>0.98</td>
<td>0.47</td>
<td>0.50</td>
</tr>
<tr>
<td>EMEA</td>
<td>0.98</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Hansard + EMEA</td>
<td>0.97</td>
<td>0.68</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Aleš Tamchyna
Il a rédigé un rapport

... 

He wrote a report

... 

Align words

Il a rédigé un rapport

He wrote a report

Extract phrases

rapport report rapport relationship ...

Translate

Il a rédigé à nouveau notre rapport.

He wrote our relationship again.
PSD Pipeline

Standard Pipeline:

Il a rédigé un rapport

He wrote a report

With PSD:

Extract with context

Extract features

rapport | report | a rédigé ...
rapport | relationship | ...

Extract features

rapport | w ll w a w_rédigé ...
report:0 (correct)
relationship:1 (wrong)

Train classifier

PSD Model

Extract features

Translate

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

He wrote our report again.
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^ne_le_savons
Target indicator: p^do_not_know
Source internal: w^ne w^le w^savons
Target internal: w^do w^not w^know
Context: c^0_-1_nous c^1_-1_il c^2_-1_CLS c^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^ne_not p^le_know p^savons_know
Scores: sc^0_1 sc^0_9 sc^0_8 sc^1_1 sc^2_10 ...
Phrase-Based MT: Evaluation

Lexical Selection

▶ Evaluated on representative phrases in Science domain.

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Accuracy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>PSD</td>
</tr>
<tr>
<td>Hansard</td>
<td>—</td>
<td>73.6</td>
</tr>
<tr>
<td>Hansard + Science</td>
<td>69.0</td>
<td>73.7</td>
</tr>
</tbody>
</table>

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 diabetic patient
• confiné dans une enceinte ⇒ hidden in a building
Syntax Based SMT

Input → Parser → Machine translation system → Language model → Output

• 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

- confined 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:
    - $X/X \rightarrow <X$ enceinte, ...$>$
    - $X/X \rightarrow <$patiente $X$, ...$>$
    - $X/X \rightarrow <$patiente $X$ enceinte, ...$>$

  - **For each source side**:
    - *M* target sides:
      - $X/X \rightarrow <X$ enceinte, pregnant $X>$
      - $X/X \rightarrow <X$ enceinte, $X$ building$>$

  - 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

  • \textbf{personne} diabétique enceinte ⇒ pregnant diabetic patient
  • \textbf{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

- \[
  \left( \text{personne}_{\text{NN}} \text{ diabétique}_{\text{ADJ}} \right)_{\text{NP}} \text{ enceinte} \right)_{\text{NP}}
  \Rightarrow \text{pregnant diabetic patient}
\]

- \[
  \left( \text{confiné}_{\text{VPART}} \text{ dans}_{\text{PREP}} \text{ une}_{\text{DET}} \text{ enceinte} \right)_{\text{SENT}}
  \Rightarrow \text{hidden in a building}
\]

- 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

Accuracy vs Training Data Size
4 Iterations of VW

Accuracy

1 out of $2^X$ Examples

Train
Dev

0.6
0.7
0.8
0.9
1

14 12 10 8 6 4

Different Initializations of VW Classifier

Accuracy

1 out of $2^X$ Examples

8 6 4

No Initialization
Hansard Initialization
PSD Domain Adaptation

Damt
EMEA32 Baselines

Accuracy

- Most Frequent Translation: 68.9%
- Old + EMEA32: 75.84%
- EMEA32: 77.74%
- Old: 60.5%
- Random Guess: 32.67%

21% Unambiguous Cases
Science Baselines

Accuracy

- Most Frequent Translation: 70.4
- Old + Science: 75.07
- Science: 75.64
- Old: 67.24
- Chance: 11.6
Old and New Agreement

<table>
<thead>
<tr>
<th>EMEA</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>57%</td>
<td>4%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>21%</td>
<td>18%</td>
</tr>
</tbody>
</table>
## Old and New Agreement

<table>
<thead>
<tr>
<th></th>
<th>EMEA</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Old</strong></td>
<td><strong>Correct</strong></td>
<td><strong>Correct</strong></td>
</tr>
<tr>
<td>Correct</td>
<td>57%</td>
<td>62.7%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>21%</td>
<td>12.9%</td>
</tr>
<tr>
<td><strong>Incorrect</strong></td>
<td>4%</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>18%</td>
<td>19.9%</td>
</tr>
</tbody>
</table>
Domain Adaptation Techniques

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

Rapport: (Hansard)
- 20%
- 80%

Rapport: (Science)
- 20%
- 80%

Aucun rapport!
No relationship!
Il a rédigé un rapport.
He wrote a report.
Le rapport des valeurs
the ratio of values

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

[Blitzer and Daume, ICML 2010]
Frustratingly Easy DA

**Rapport:** (Hansard)
- 20%
- 80%

**Rapport:** (Science)
- 20%
- 80%

Aucun rapport!
No relationship!

Il a rédigé un rapport.
He wrote a report.

le rapport des valeurs
the ratio of values

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

[Blitzer and Daume, ICML 2010]
Feature Augmentation

Old: $x \rightarrow <x, x, 0>$
New: $x \rightarrow <x, 0, x>$

Original Features

Augmented Features

Hansard

- AUCUN RÉDIGÉ
- O_AUCUN O_RÉDIGÉ

Science

- VALEURS RÉDIGÉ
- N_VALEURS N_RÉDIGÉ

[Blitzer and Daume, 2010]
Instance Weighting

Old Data + New Data → Adapted Model
Instance Weighting

Rapport: (Hansard)

20%

Aucun rapport!
No relationship!

0.1

80%

Il a rédigé un rapport.
He wrote a report.

0.6
Instance Weighting

Rapport: (Hansard)

20% 
Aucun rapport!
No relationship!
0.1

80%
Il a rédigé un rapport.
He wrote a report.
0.6

Rapport: (Science)

1%
Aucun rapport!
No relationship!

45%
Il a rédigé un rapport.
He wrote a report.

54%
le rapport des valeurs 
the ratio of values
Old Predictions Feature in New

Old Model

Old Data

New Data

New Data

Adapted Model
Model Interpolation

- Linear
- Log-linear
- Cross Validation
Domain Adaptation Results on EMEA32

- EMEA32: 77.74
- Old + EMEA32: 75.84
- FEDA: 77.33
- Instance Weighting: 78.17
- Old Prediction Feature: 78.09
- Linear Mixture: 77.99
Domain Adaptation Results on Science

- Science: 75.64
- Old + Science: 75.84
- FEDA: 75.89
- Instance Weighting: 75.86
- Old Prediction Feature: 75.41
- Linear Mixture: 75.93
5 MIN BREAK
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
<table>
<thead>
<tr>
<th></th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>$e_{n-1}$</th>
<th>$e_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$p(e_1, f_1)$</td>
<td>$p(e_2, f_1)$</td>
<td>$\ldots$</td>
<td>$p(e_{n-1}, f_1)$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$p(e_1, f_2)$</td>
<td>$\ldots$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{m-1}$</td>
<td>$p(e_1, f_{m-1})$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_m$</td>
<td>$p(e_1, f_m)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Old Domain
French-English Parallel Data
Old Domain
French-English Parallel Data

Fr  En

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>p(e₁,f₁)</td>
<td>p(e₂,f₁)</td>
<td>...</td>
<td>p(eₙ₋₁,f₁)</td>
<td>p(eₙ,f₁)</td>
</tr>
<tr>
<td>f₂</td>
<td>p(e₁,f₂)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>p(e₁,fₘ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>p(e₁,fₘ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

K New Domain
French-English Comparable Document Pairs

Fr  En

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>q(e₁)</td>
<td>q(e₂)</td>
<td>...</td>
<td>q(eₙ₋₁)</td>
<td>q(eₙ)</td>
</tr>
<tr>
<td>f₂</td>
<td>q(e₁)</td>
<td>q(e₂)</td>
<td>...</td>
<td>q(eₙ₋₁)</td>
<td>q(eₙ)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>q(e₁)</td>
<td>q(e₂)</td>
<td>...</td>
<td>q(eₙ₋₁)</td>
<td>q(eₙ)</td>
</tr>
<tr>
<td>fₘ</td>
<td>q(e₁)</td>
<td>q(e₂)</td>
<td>...</td>
<td>q(eₙ₋₁)</td>
<td>q(eₙ)</td>
</tr>
</tbody>
</table>
### Old Domain
- French-English Parallel Data

![Diagram showing parallel data in a table format]

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

![Diagram showing comparable data in a table format]
Old Domain
French-English Parallel Data

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>p(e₁, f₁)</td>
<td>p(e₂, f₁)</td>
<td>...</td>
<td>p(eₙ₋₁, f₁)</td>
<td>p(eₙ, f₁)</td>
</tr>
<tr>
<td>f₂</td>
<td>p(e₁, f₂)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>p(e₁, fₘ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>p(e₁, fₘ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

K New Domain
French-English Comparable Document Pairs

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>q(e₁)</td>
<td>q(e₂)</td>
<td>...</td>
<td>q(eₙ₋₁)</td>
<td>q(eₙ)</td>
</tr>
<tr>
<td>f₂</td>
<td>q(f₁)</td>
<td>q(f₂)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>q(fₘ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>q(fₘ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Old Domain
French-English Parallel Data

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>p(e₁, f₁)</td>
<td>p(e₂, f₁)</td>
<td>...</td>
<td>p(eₙ₋₁, f₁)</td>
<td>p(eₙ, f₁)</td>
</tr>
<tr>
<td>f₂</td>
<td>p(e₁, f₂)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>p(e₁, fₘ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>p(e₁, fₘ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

K New Domain
French-English Comparable Document Pairs

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>q(e₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f₂</td>
<td>q(e₂)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>q(eₙ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>q(eₙ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

q(e₁) q(e₂) ... q(eₙ₋₁) q(eₙ)
Old Domain
French-English Parallel Data

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

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

$$\sum_{e \in E, f \in F} (p(e, f) - \hat{p}(e, f))^2 + \hat{p}(e, f) \cdot (freqw(e, f) + ed(e, f) + wikidist(e, f) + 1)$$
For each comparable document pair...

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

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

[distance from original joint]
For each comparable document pair...

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

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

distance from original joint
For each comparable document pair...

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

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

- distance from original joint
- monolingual relative frequency difference
- string edit distance between $e$ and $f$
For each comparable document pair...

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

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

- distance from original joint
- monolingual relative frequency difference
- difference between wikipedia page distributions
- string edit distance between $e$ and $f$
For each comparable document pair...

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

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

- Distance from original joint
- String edit distance between e and f
- Difference between monolingual relative frequency difference
- Difference between wikipedia page distributions
- Sparsity Penalty
For each comparable document pair...

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

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

Subject to constraints:

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

$$\sum_{e \in E} \hat{p}(e, f) - q(f) < \epsilon$$
For each comparable document pair...

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

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

Subject to constraints:

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>...</th>
<th>$e_{n-1}$</th>
<th>$e_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$p_k(e_1, f_1)$</td>
<td>$p_k(e_2, f_1)$</td>
<td>...</td>
<td>$p_k(e_{n-1}, f_1)$</td>
<td>$p_k(e_n, f_1)$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$p_k(e_1, f_2)$</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{m-1}$</td>
<td>$p_k(e_1, f_{m-1})$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_m$</td>
<td>$p_k(e_1, f_m)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Translation Mining: Evaluation

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>p_{k(e₁,f₁)}</td>
<td>p_{k(e₂,f₁)}</td>
<td>...</td>
<td>p_{k(eₙ₋₁,f₁)}</td>
<td>p_{k(eₙ,f₁)}</td>
</tr>
<tr>
<td>f₂</td>
<td>p_{k(e₁,f₂)}</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>p_{k(e₁,fₘ₋₁)}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>p_{k(e₁,fₘ)}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gold Standard: New Domain French-English Parallel Data

<table>
<thead>
<tr>
<th></th>
<th>e₁</th>
<th>e₂</th>
<th>...</th>
<th>eₙ₋₁</th>
<th>eₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₁</td>
<td>p(e₁,f₁)</td>
<td>p(e₂,f₁)</td>
<td>...</td>
<td>p(eₙ₋₁,f₁)</td>
<td>p(eₙ,f₁)</td>
</tr>
<tr>
<td>f₂</td>
<td>p(e₁,f₂)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ₋₁</td>
<td>p(e₁,fₘ₋₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fₘ</td>
<td>p(e₁,fₘ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Translation Mining:
Evaluation

Gold Standard:
New Domain
French-English Parallel Data

Mean Reciprocal Rank
where is max \( p_{\text{new}}(e|f) \) in
\( p_{\text{learned}}(e|f) \) ranked list over \( f \)
Translation Mining: Evaluation

<table>
<thead>
<tr>
<th></th>
<th>(e_1)</th>
<th>(e_2)</th>
<th>...</th>
<th>(e_{n-1})</th>
<th>(e_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_1)</td>
<td>(p_k(e_1, f_1))</td>
<td>(p_k(e_2, f_1))</td>
<td>...</td>
<td>(p_k(e_{n-1}, f_1))</td>
<td>(p_k(e_n, f_1))</td>
</tr>
<tr>
<td>(f_2)</td>
<td>(p_k(e_1, f_2))</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f_{m-1})</td>
<td>(p_k(e_1, f_{m-1}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f_m)</td>
<td>(p_k(e_1, f_m))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **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**

<table>
<thead>
<tr>
<th></th>
<th>(e_1)</th>
<th>(e_2)</th>
<th>...</th>
<th>(e_{n-1})</th>
<th>(e_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_1)</td>
<td>(p(e_1, f_1))</td>
<td>(p(e_2, f_1))</td>
<td>...</td>
<td>(p(e_{n-1}, f_1))</td>
<td>(p(e_n, f_1))</td>
</tr>
<tr>
<td>(f_2)</td>
<td>(p(e_1, f_2))</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f_{m-1})</td>
<td>(p(e_1, f_{m-1}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(f_m)</td>
<td>(p(e_1, f_m))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Translation Mining: Evaluation

<table>
<thead>
<tr>
<th></th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>...</th>
<th>$e_{n-1}$</th>
<th>$e_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$p_k(e_1, f_1)$</td>
<td>$p_k(e_2, f_1)$</td>
<td>...</td>
<td>$p_k(e_{n-1}, f_1)$</td>
<td>$p_k(e_n, f_1)$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$p_k(e_1, f_2)$</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{m-1}$</td>
<td>$p_k(e_1, f_{m-1})$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_m$</td>
<td>$p_k(e_1, f_m)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th></th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>...</th>
<th>$e_{n-1}$</th>
<th>$e_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$p(e_1, f_1)$</td>
<td>$p(e_2, f_1)$</td>
<td>...</td>
<td>$p(e_{n-1}, f_1)$</td>
<td>$p(e_n, f_1)$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$p(e_1, f_2)$</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{m-1}$</td>
<td>$p(e_1, f_{m-1})$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_m$</td>
<td>$p(e_1, f_m)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Translation Mining:
Evaluation

Domain: EMEA

Hansard-Only
$p(e,f)$ Baseline

results are similar for Science domain
Domain: EMEA

Translation Mining: Evaluation

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

Sense-Freebie
OOV-Freebie*
Correct

New Sense
OOV-Wrong*
Score/Search Error

Percent of reference alignments

* OOV wrt original table

Domain: Science
## Preliminary MT Results

**BLEU**

<table>
<thead>
<tr>
<th></th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansard-Trained</td>
<td>26.08</td>
</tr>
<tr>
<td>Hansard-Trained + Scored OOV Trans.</td>
<td>26.12</td>
</tr>
</tbody>
</table>

**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

\[ Score(w) = \sum_{k \in \text{topics}_{new}} P_{new}(k|w) \times \max_{k' \in \text{topics}_{old}} (P_{old}(k'|w) \times \text{cossim}(k, k')) \]

➢ 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

\[
\text{ngram.score}(w) = \frac{|\text{NEW.WITHOUT.OOV} \setminus (\text{NEW}\cap\text{OLD})|}{|\text{NEW.WITHOUT.OOV}|} \quad (= \frac{|\text{RED}|}{|\text{REDUPURPLE}|})
\]
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:
  - $p_{\text{learned}}(f) > 0$ ?
  - $\max_e p_{\text{old}}(e|f) = p_{\text{learned}}(e|f)$ ?
  - overlap $[\text{top-5}_e p_{\text{old}}(e|f), \text{top-5}_e p_{\text{learned}}(e|f)] / 5$
  - 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**: topics || matchm topics || matchm topics flow
Ann Clifton
Document-Level Info in MT

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

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

Compute lexical probability conditioned on topic distribution:

$$p_{zn}(e, |f) = \frac{e_{zn}(e, f)}{\sum_e e_{zn}(e, f)}$$
Intrinsic Evaluation

Table: Average per-word log likelihood of EMEA data

<table>
<thead>
<tr>
<th></th>
<th>no-topic</th>
<th>doc-topic</th>
<th>word-topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>old-alignment, old topic</td>
<td>-1.78</td>
<td>-0.47</td>
<td>-0.48</td>
</tr>
<tr>
<td>new-domain, new-topic</td>
<td>-1.12</td>
<td>-0.26</td>
<td>-0.26</td>
</tr>
<tr>
<td>old-domain, new-topic</td>
<td>-1.78</td>
<td>-0.27</td>
<td>-0.27</td>
</tr>
</tbody>
</table>
Lexical Weighting in Phrase-Based MT

As feature in phrase-based MT:

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

\[ \sum_p \lambda_p h_p(\bar{e}, \bar{f}) + \sum_{zn} \lambda_{zn} f_{zn}(\bar{e}|\bar{f}) \]
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

<table>
<thead>
<tr>
<th>{0,1,2,3,4,5,6,7,8}</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>{0,1,2,3}</td>
<td>{4,5,6,7}</td>
<td></td>
</tr>
<tr>
<td>{0,1}</td>
<td>{2,3}</td>
<td>{4,5}</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
A Discriminative Topic Model: Hierarchical Topic Features

{0,1,2,3,4,5,6,7,8}
A Discriminative Topic Model: Hierarchical Topic Features

\[
z \in \{0, 1\}
\]

<table>
<thead>
<tr>
<th>Source word</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>0</td>
</tr>
<tr>
<td>régime</td>
<td>0</td>
</tr>
<tr>
<td>français</td>
<td>0</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
z = 0
\]

<table>
<thead>
<tr>
<th>Source word</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>0</td>
</tr>
<tr>
<td>régime</td>
<td>0</td>
</tr>
<tr>
<td>français</td>
<td>-2.5</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>2.5</td>
</tr>
</tbody>
</table>

\[
z = 1
\]

<table>
<thead>
<tr>
<th>Source word</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>0</td>
</tr>
<tr>
<td>régime</td>
<td>0</td>
</tr>
<tr>
<td>français</td>
<td>2.5</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>-2.5</td>
</tr>
</tbody>
</table>
A Discriminative Topic Model: Hierarchical Topic Distributions

\[ z \in \{0, 1\} \]

<table>
<thead>
<tr>
<th>source word</th>
<th>target word</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>régime</td>
<td>diet</td>
<td>0.53</td>
</tr>
<tr>
<td>régime</td>
<td>administration</td>
<td>0.47</td>
</tr>
<tr>
<td>français</td>
<td>French</td>
<td>1</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>grapefruit</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ z = 0 \]

<table>
<thead>
<tr>
<th>source word</th>
<th>target word</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>régime</td>
<td>diet</td>
<td>0.99</td>
</tr>
<tr>
<td>régime</td>
<td>administration</td>
<td>0.01</td>
</tr>
<tr>
<td>français</td>
<td>French</td>
<td>1</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>grapefruit</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ z = 1 \]

<table>
<thead>
<tr>
<th>source word</th>
<th>target word</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>régime</td>
<td>diet</td>
<td>0.01</td>
</tr>
<tr>
<td>régime</td>
<td>administration</td>
<td>0.99</td>
</tr>
<tr>
<td>français</td>
<td>French</td>
<td>1</td>
</tr>
<tr>
<td>pamplemousse</td>
<td>grapefruit</td>
<td>1</td>
</tr>
</tbody>
</table>
A Discriminative Topic Model: Example

(1a) le₁ régime₂ français₃
(1b) the₁ French₃ administration₂

(2a) le₁ régime₂ pamplemousse₃
(2b) the₁ grapefruit₃ diet₂

<table>
<thead>
<tr>
<th></th>
<th>topic 0</th>
<th>topic 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence 1</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>sentence 2</td>
<td>0.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>rapport</th>
<th>report</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>relationship</td>
<td>0.1</td>
</tr>
<tr>
<td>rapport</td>
<td>reporting</td>
<td>0.05</td>
</tr>
<tr>
<td>rapport</td>
<td>values</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport
From Type to Token

Adapt type level translations to token level

<table>
<thead>
<tr>
<th>rapport</th>
<th>report</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>relationship</td>
<td>0.1</td>
</tr>
<tr>
<td>rapport</td>
<td>reporting</td>
<td>0.05</td>
</tr>
<tr>
<td>rapport</td>
<td>values</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport

<table>
<thead>
<tr>
<th>rapport</th>
<th>report</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>reporting</td>
<td>0.3</td>
</tr>
</tbody>
</table>
From Type to Token

Adapt type level translations to token level

<table>
<thead>
<tr>
<th>rapport</th>
<th>report</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>relationship</td>
<td>0.1</td>
</tr>
<tr>
<td>rapport</td>
<td>reporting</td>
<td>0.05</td>
</tr>
<tr>
<td>rapport</td>
<td>values</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport

le rapport des valeurs
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

<table>
<thead>
<tr>
<th>le</th>
<th>rapport</th>
<th>des</th>
<th>valeurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>-0.3</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.2</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The ratio of values

| 0.6 | -0.2 | 0.6 | 0.4 |
| -0.1 | 0.3 | 0.2 | -0.1 |
| 0.2 | 0.4 | -0.1 | 0.6 |
| ... | ... | ... | ... |
Main Idea

Step 2
Replace words with vectors
Vectors capture the word meaning

rapport

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
...

...
### Main Idea

**Step 3**

Token representation is a weighted combination of the context vectors.

### Ratio

<table>
<thead>
<tr>
<th>-0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>...</th>
</tr>
</thead>
</table>

### Rapport

<table>
<thead>
<tr>
<th>0.8</th>
<th>-0.3</th>
<th>0.7</th>
<th>0.5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.2</td>
<td>...</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
**Main Idea**

Step 3
Token representation is a weighted combination of the context vectors

\[
\text{rapport} = w_{la} + w_0 + w_{des} + w_{valeurs}
\]

**ratio**

\[
\begin{align*}
-0.2 & \\
0.3 & \\
0.4 & \\
& ...
\end{align*}
\]
Main Idea

Step 3
Token representation is a weighted combination of the context vectors

\[
\begin{align*}
\text{rapport} = & w_{la} \cdot \text{rapport} + w_0 + w_{\text{des}} + w_{\text{valeurs}} \\
\text{rapport}_\text{token} = & \ldots + \ldots + \ldots + \ldots
\end{align*}
\]
Main Idea

Token representation is a weighted combination of the context vectors.

\[
\forall (f_i, e_i) \quad w_0 \bar{v}(f_i) + \sum_{f_j \in \text{Ctx}(f_i)} w_{f_j} \bar{v}(f_j) \approx \bar{v}(e_i)
\]
Extensions

1. Co-regularization
   Weights are independent of the focus word
   Add dependency but regularize

2.
Co-regularization

\[ w_{la} + w_0 + w_{des} + w_{valeurs} = \]

\[ \text{rapport token} \]
Co-regularization

\[
\text{rapport}_{\text{la}} \cdot \begin{bmatrix} 0.8 \\ 0.2 \\ 0.1 \\ \ldots \end{bmatrix} + \text{rapport}_{\text{w}} \cdot \begin{bmatrix} -0.3 \\ 0.4 \\ 0.2 \\ \ldots \end{bmatrix} + \text{rapport}_{\text{des}} \cdot \begin{bmatrix} 0.7 \\ 0.3 \\ 0.4 \\ \ldots \end{bmatrix} + \text{rapport}_{\text{valeurs}} \cdot \begin{bmatrix} 0.5 \\ -0.2 \\ 0.8 \\ \ldots \end{bmatrix} = \text{rapport}_{\text{token}}
\]
Co-regularization

\[ \text{rapport}_{\text{la}} + \text{rapport}_0 + \text{rapport}_{\text{des}} + \text{rapport}_{\text{valeurs}} = \text{rapport}_{\text{token}} \]

\[ \text{ratio} \]

\[ w_f^{\text{rapport}} = w_f + r_f^{\text{rapport}} \]

\[ \text{add} \ || \sum_f r_f^{\text{rapport}} ||^2 \]
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

<table>
<thead>
<tr>
<th>rapport</th>
<th>report</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>relationship</td>
<td>0.1</td>
</tr>
<tr>
<td>rapport</td>
<td>reporting</td>
<td>0.05</td>
</tr>
<tr>
<td>rapport</td>
<td>values</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport
Intrinsic evaluation

On 7.4K tokens from EMEA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>40.29</td>
</tr>
<tr>
<td>Max Probable -- p(e</td>
<td>f)</td>
</tr>
<tr>
<td>Best Cue-Word</td>
<td>61.85</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport
Intrinsic evaluation

On 7.4K tokens from EMEA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>40.29</td>
</tr>
<tr>
<td>Max Probable -- $p(e</td>
<td>f)$</td>
</tr>
<tr>
<td>Best Cue-Word</td>
<td>61.85</td>
</tr>
<tr>
<td>Simple adaptation</td>
<td>55.21</td>
</tr>
<tr>
<td>Co-regularization</td>
<td>59.15</td>
</tr>
<tr>
<td>Max-Margin</td>
<td>60.21</td>
</tr>
<tr>
<td>Coreg+MaxMargin</td>
<td>??</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Token adapted</th>
<th>rapport</th>
</tr>
</thead>
<tbody>
<tr>
<td>report</td>
<td>rapport</td>
</tr>
<tr>
<td>relationship</td>
<td>rapport</td>
</tr>
<tr>
<td>reporting</td>
<td>rapport</td>
</tr>
<tr>
<td>values</td>
<td>rapport</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport
Intrinsic evaluation

On 7.4K tokens from EMEA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>40.29</td>
</tr>
<tr>
<td>Max Probable -- (p(e</td>
<td>f))</td>
</tr>
<tr>
<td>Best Cue-Word</td>
<td>61.85</td>
</tr>
<tr>
<td>Simple adaptation</td>
<td>55.21</td>
</tr>
<tr>
<td>Co-regularization</td>
<td>59.15</td>
</tr>
<tr>
<td>Max-Margin</td>
<td>60.21</td>
</tr>
<tr>
<td>Coreg+MaxMargin</td>
<td>??</td>
</tr>
<tr>
<td>PSD Classifier</td>
<td>70.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Token adapted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>rapport</td>
<td>report 0.5</td>
</tr>
<tr>
<td>rapport</td>
<td>relationship 0.1</td>
</tr>
<tr>
<td>rapport</td>
<td>reporting 0.05</td>
</tr>
<tr>
<td>rapport</td>
<td>values 0.3</td>
</tr>
</tbody>
</table>

Il a rédigé un rapport
Marine Carpuuat
summary & conclusion

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: 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