Domain adaptation for SMT

* Prioritize translation candidates that are most relevant to a specific task
Domain adaptation for SMT

* Prioritize translation candidates that are most relevant to a specific task
Domain adaptation for SMT

- Prioritize translation candidates that are most relevant to a specific task

Heterogeneous training data

Specific translation task
Domain adaptation for SMT

* Prioritize translation candidates that are most relevant to a specific task

| source      | target             | p(f|e) | p(e|f) |
|-------------|--------------------|-------|-------|
| الحمد ل      | praise be to       | 0.1   | 0.2   |
| الحمد ل      | praise for         | 0.2   | 0.2   |
| الحمد ل      | thank              | 0.1   | 0.2   |
| حبيبيتي ي   | my dear            | 0.2   | 0.1   |
| حبيبيتي ي   | my love            | 0.2   | 0.1   |
| حبيبيتي ي   | my sweetheart     | 0.1   | 0.1   |
Domain adaptation for SMT

* Prioritize translation candidates that are most relevant to a specific task

| source | target      | p(f|e) | p(e|f) |
|--------|-------------|-------|-------|
| الحمد ل | praise be to| 0.1   | 0.2   |
| الحمد ل | praise for  | 0.2   | 0.2   |
| الحمد ل | thank       | 0.1   | 0.2   |
| حبيبيتي ي | my dear    | 0.2   | 0.1   |
| حبيبيتي ي | my love    | 0.2   | 0.1   |
| حبيبيتي ي | my sweetheart | 0.1  | 0.1   |
Domain adaptation for SMT

* Prioritize translation candidates that are most relevant to a specific task

| source | target       | p(f|e) | p(e|f) | ...
|--------|--------------|-------|-------|-----
| الحمد ل | praise be to | 0.1   | 0.2   | ... |
| الحمد ل | praise for   | 0.2   | 0.2   | ... |
| الحمد ل | thank        | 0.1   | 0.2   | ... |
| حبيبيتي ي | my dear    | 0.2   | 0.1   | ... |
| حبيبيتي ي | my love     | 0.2   | 0.1   | ... |
| حبيبيتي ي | my sweetheart | 0.1 | 0.1 | ... |

* What type of domain information to use?
Dimensions of domains

• **Topic** refers to general subject
  ✦ politics, sports, tennis
Dimensions of domains

- **Topic** refers to general subject
  - politics, sports, tennis

- **Genre** refers to function, style, text type
  - editorials, newswire, user-generated text
  - orthogonal to topic
Dimensions of domains

- **Topic** refers to general subject
  - politics, sports, tennis

- **Genre** refers to function, style, text type
  - editorials, newswire, user-generated text
  - orthogonal to topic

- **Provenance** refers to document’s origin
  - LDC2005T13, Europarl, EMEA
The problem with provenance

Provenance information has proven useful for adaptation in SMT, but is it the best representation of a domain?
The problem with provenance

Provenance information has proven useful for adaptation in SMT, but is it the best representation of a domain?

* It’s not an intrinsic text property
The problem with provenance

Provenance information has proven useful for adaptation in SMT, but is it the best representation of a domain?

- It’s not an intrinsic text property
- We might need manual labeling
  - labor-intensive
  - arbitrary
The problem with provenance

Provenance information has proven useful for adaptation in SMT, but is it the best representation of a domain?

* It’s not an intrinsic text property
* We might need manual labeling
  - labor-intensive
  - arbitrary
* Often combines particular **topic** and **genre**
Disentangling topic and genre in SMT*

* Experiments on controlled test set: Gen&Topic

**Culture**
- **News**: The 12 contestants competed during a May 3rd Prime.
- **Comment**: You allowed Barwas to represent Iraq while she sings in Kurdish!!!

**Economy**
- **News**: Yemen is mulling the establishment of 13 industrial zones.
- **Comment**: What development in Yemen are you talking about?

* Van der Wees et al., 2015
Disentangling topic and genre in SMT*

* Experiments on controlled test set: Gen&Topic

* Genre has larger impact on SMT than topic

---

**Culture**

- News: The 12 contestants competed during a May 3rd Prime.
- Comment: You allowed Barwas to represent Iraq while she sings in Kurdish!!

**Economy**

- News: Yemen is mulling the establishment of 13 industrial zones.
- Comment: What development in Yemen are you talking about?

---

* Van der Wees et al., 2015
Disentangling topic and genre in SMT*

* Experiments on controlled test set: Gen&Topic
* Genre has larger impact on SMT than topic
* We want to adapt to different genres in a test corpus!

* Van der Wees et al., 2015

---

**News**
- The 12 contestants competed during a May 3rd Prime.

**Comment**
- You allowed Barwas to represent Iraq while she sings in Kurdish!!!

**Culture**
- Yemen is mulling the establishment of 13 industrial zones.

**Economy**
- What development in Yemen are you talking about?
What information to use for adaptation?

- **Provenance** information
  - manual grouping of sub-corpora
What information to use for adaptation?

✦ **Provenance** information
  ✦ manual grouping of sub-corpora

✦ **Topic** information
  ✦ unsupervised LDA-inferred topics
What information to use for adaptation?

✧ **Provenance** information
  ✦ manual grouping of sub-corpora

✧ **Topic** information
  ✦ unsupervised LDA-inferred topics

✧ **Genre** information
  ✦ determine and exploit intrinsic genre-revealing text features
What information to use for adaptation?

- **Provenance** information
  - manual grouping of sub-corpora

- **Topic** information
  - unsupervised LDA-inferred topics

- **Genre** information
  - determine and exploit intrinsic genre-revealing text features
What information to use for adaptation?

- **Provenance** information
  - manual grouping of sub-corpora

- **Topic** information
  - unsupervised LDA-inferred topics

- **Genre** information
  - determine and exploit intrinsic genre-revealing text features
What information to use for adaptation?

✧ **Provenance** information
  ✦ manual grouping of sub-corpora

✧ **Topic** information
  ✦ unsupervised LDA-inferred topics

✧ **Genre** information
  ✦ determine and exploit intrinsic genre-revealing text features
Genre adaptation: the task

* Arabic-English phrase-based SMT

* [ilps.science.uva.nl/resources/gen-topic/](ilps.science.uva.nl/resources/gen-topic/)
Genre adaptation: the task

* Arabic-English phrase-based SMT

* Two multi-genre evaluation sets:
  ✦ Gen&Topic*:
    • newswire (NW)
    • comments (UG)

* ilps.science.uva.nl/resources/gen-topic/
Genre adaptation: the task

- Arabic-English phrase-based SMT

- Two multi-genre evaluation sets:
  - Gen&Topic*:
    - newswire (NW)
    - comments (UG)
  - NIST:
    - newswire (NW)
    - weblogs (UG)

* ilps.science.uva.nl/resources/gen-topic/
Genre adaptation: the task

- Arabic-English phrase-based SMT

- Two multi-genre evaluation sets:
  - Gen&Topic*:
    - newswire (NW)
    - comments (UG)
  - NIST:
    - newswire (NW)
    - weblogs (UG)

- Translation model adaptation

* [ilps.science.uva.nl/resources/gen-topic/](ilps.science.uva.nl/resources/gen-topic/)
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

* Following Chen et al., 2013
Genre adaptation: general framework

Vector space model (VSM) for translation model adaptation*

| source   | target             | p(f|e) | p(e|f) | ...
|----------|--------------------|-------|-------|----------
| الحمد ل   | praise be to       | 0.1   | 0.2   | ...      
| الحمد ل   | praise for         | 0.2   | 0.2   | ...      
| الحمد ل   | thank              | 0.1   | 0.2   | ...      
| حبيبيتي ي | my dear           | 0.2   | 0.1   | ...      
| حبيبيتي ي | my love           | 0.2   | 0.1   | ...      
| حبيبيتي ي | my sweetheart     | 0.1   | 0.1   | ...      

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source | target        | p(f|e) | p(e|f) | ... | phrase vector |
|--------|---------------|-------|-------|-----|---------------|
| الحمد ل | praise be to  | 0.1   | 0.2   |     | < w₁ ... w₇ > |
| الحمد ل | praise for    | 0.2   | 0.2   |     | < w₁ ... w₇ > |
| الحمد ل | thank         | 0.1   | 0.2   |     | < w₁ ... w₇ > |
| حبيبيتك ي | my dear     | 0.2   | 0.1   |     | < w₁ ... w₇ > |
| حبيبيتك ي | my love      | 0.2   | 0.1   |     | < w₁ ... w₇ > |
| حبيبيتك ي | my sweetheart| 0.1   | 0.1   |     | < w₁ ... w₇ > |

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source       | target              | p(f|e) | p(e|f) | phrase vector |
|--------------|---------------------|-------|-------|--------------|
| الحمد ل       | praise be to        | 0.1   | 0.2   | < w₁ ... wₙ > |
| الحمد ل       | praise for          | 0.2   | 0.2   | < w₁ ... wₙ > |
| الحمد ل       | thank               | 0.1   | 0.2   | < w₁ ... wₙ > |
| حبيبي ي       | my dear             | 0.2   | 0.1   | < w₁ ... wₙ > |
| حبيبي ي       | my love             | 0.2   | 0.1   | < w₁ ... wₙ > |
| حبيبي ي       | my sweetheart      | 0.1   | 0.1   | < w₁ ... wₙ > |

Vector for development set:  < w₁(dev) ... wₙ(dev) >

* Following Chen et al., 2013
## Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source | target            | p(f|e) | p(e|f) | phrase vector | similarity score |
|--------|-------------------|-------|-------|---------------|-----------------|
| الحمد ل | praise be to      | 0.1   | 0.2   | < w₁ ... wₙ > | 0.1             |
| الحمد ل | praise for        | 0.2   | 0.2   | < w₁ ... wₙ > |
| الحمد ل | thank             | 0.1   | 0.2   | < w₁ ... wₙ > |
| حبيبيي ي | my dear          | 0.2   | 0.1   | < w₁ ... wₙ > |
| حبيبيي ي | my love          | 0.2   | 0.1   | < w₁ ... wₙ > |
| حبيبيي ي | my sweetheart    | 0.1   | 0.1   | < w₁ ... wₙ > |

Vector for development set: < w₁(dev) ... wₙ(dev) >

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source  | target        | p(f|e) | p(e|f) | phrase vector | similarity score |
|---------|---------------|-------|-------|--------------|-----------------|
| الحمد ل  | praise be to  | 0.1   | 0.2   | < w₁ ... wₙ > | 0.1             |
| الحمد ل  | praise for    | 0.2   | 0.2   | < w₁ ... wₙ > | 0.2             |
| الحمد ل  | thank         | 0.1   | 0.2   | < w₁ ... wₙ > |                 |
| حبيبيك ي | my dear      | 0.2   | 0.1   | < w₁ ... wₙ > |                 |
| حبيبيك ي | my love      | 0.2   | 0.1   | < w₁ ... wₙ > |                 |
| حبيبيك ي | my sweetheart| 0.1   | 0.1   | < w₁ ... wₙ > |                 |

Vector for development set: < w₁(dev) ... wₙ(dev) >

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source  | target            | p(f|e) | p(e|f) | phrase vector | similarity score |
|---------|-------------------|-------|-------|---------------|-----------------|
| الحمد ل  | praise be to      | 0.1   | 0.2   | <w₁ ... wₙ > | 0.1             |
| الحمد ل  | praise for        | 0.2   | 0.2   | <w₁ ... wₙ > | 0.2             |
| الحمد ل  | thank             | 0.1   | 0.2   | <w₁ ... wₙ > | 0.4             |
| حبيبيي ي | my dear          | 0.2   | 0.1   | <w₁ ... wₙ > |                 |
| حبيبيي ي | my love          | 0.2   | 0.1   | <w₁ ... wₙ > |                 |
| حبيبيي ي | my sweetheart    | 0.1   | 0.1   | <w₁ ... wₙ > |                 |

Vector for development set: <w₁(dev) ... wₙ(dev)>

* Following Chen et al., 2013
Genre adaptation: general framework

- Vector space model (VSM) for translation model adaptation*

| source    | target               | p(f|e) | p(e|f) | phrase vector | similarity score |
|-----------|----------------------|-------|-------|---------------|-----------------|
| الحمد ل    | praise be to         | 0.1   | 0.2   | < w₁ ... wₙ > | 0.1             |
| الحمد ل    | praise for           | 0.2   | 0.2   | < w₁ ... wₙ > | 0.2             |
| الحمد ل    | thank                | 0.1   | 0.2   | < w₁ ... wₙ > | 0.4             |
| حبيبيتي ي  | my dear             | 0.2   | 0.1   | < w₁ ... wₙ > | 0.3             |
| حبيبيتي ي  | my love             | 0.2   | 0.1   | < w₁ ... wₙ > |                 |
| حبيبيتي ي  | my sweetheart       | 0.1   | 0.1   | < w₁ ... wₙ > |                 |

Vector for development set: < w₁(dev) ... wₙ(dev) >

* Following Chen et al., 2013
**Genre adaptation: general framework**

- Vector space model (VSM) for translation model adaptation*

| source   | target            | p(f|e)  p(e|f) … | phrase vector   | similarity score |
|----------|-------------------|-----------------|-----------------|-----------------|
| الحمد    | praise be to      | 0.1 0.2 …       | < w₁ … wₙ >     | 0.1             |
| الحمد    | praise for        | 0.2 0.2 …       | < w₁ … wₙ >     | 0.2             |
| الحمد    | thank             | 0.1 0.2 …       | < w₁ … wₙ >     | 0.4             |
| حبيبيتي  | my dear           | 0.2 0.1 …       | < w₁ … wₙ >     | 0.3             |
| حبيبيتي  | my love           | 0.2 0.1 …       | < w₁ … wₙ >     | 0.4             |
| حبيبيتي  | my sweetheart    | 0.1 0.1 …       | < w₁ … wₙ >     |                 |

Vector for development set:  < w₁(dev) … wₙ(dev) >

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source | target             | p(f|e)  | p(e|f)  | phrase vector  | similarity score |
|--------|--------------------|--------|--------|----------------|-----------------|
| الحمد  | praise be to       | 0.1    | 0.2    | $< w_1 \ldots w_N >$ | 0.1             |
| الحمد  | praise for         | 0.2    | 0.2    | $< w_1 \ldots w_N >$ | 0.2             |
| الحمد  | thank              | 0.1    | 0.2    | $< w_1 \ldots w_N >$ | 0.4             |
| حبيبي  | my dear            | 0.2    | 0.1    | $< w_1 \ldots w_N >$ | 0.3             |
| حبيبي  | my love            | 0.2    | 0.1    | $< w_1 \ldots w_N >$ | 0.4             |
| حبيبي  | my sweetheart     | 0.1    | 0.1    | $< w_1 \ldots w_N >$ | 0.1             |

Vector for development set: $< w_1^{(dev)} \ldots w_N^{(dev)} >$

* Following Chen et al., 2013
Genre adaptation: general framework

* Vector space model (VSM) for translation model adaptation*

| source   | target             | p(f|e) | p(e|f) | phrase vector  | similarity score |
|----------|--------------------|-------|-------|----------------|------------------|
| الحمد ل   | praise be to       | 0.1   | 0.2   | \(< w_1 \ldots w_N >\) | 0.1              |
| الحمد ل   | praise for         | 0.2   | 0.2   | \(< w_1 \ldots w_N >\) | 0.2              |
| الحمد ل   | thank              | 0.1   | 0.2   | \(< w_1 \ldots w_N >\) | 0.4              |
| حبيبيتي ي | my dear           | 0.2   | 0.1   | \(< w_1 \ldots w_N >\) | 0.3              |
| حبيبيتي ي | my love           | 0.2   | 0.1   | \(< w_1 \ldots w_N >\) | 0.4              |
| حبيبيتي ي | my sweetheart     | 0.1   | 0.1   | \(< w_1 \ldots w_N >\) | 0.1              |

Vector for development set: \(< w_1^{(dev)} \ldots w_N^{(dev)} >\)

* Following Chen et al., 2013
How to construct genre-informed vectors?

- Original version: **provenance** information
  - following Chen et al., 2013
How to construct genre-informed vectors?

- Original version: *provenance* information
  - following Chen et al., 2013

- Our version: *intrinsic genre* information
How to construct genre-informed vectors?

- Original version: provenance information
  - following Chen et al., 2013

- Our version: intrinsic genre information
  - document-level genre features borrowed from text classification literature
How to construct genre-informed vectors?

* Original version: **provenance** information
  - following Chen et al., 2013

* Our version: **intrinsic genre** information
  - document-level genre features borrowed from text classification literature
  - directly observable in raw text
How to construct genre-informed vectors?

✦ Original version: *provenance* information
  ✦ following Chen et al., 2013

✦ Our version: *intrinsic genre* information
  ✦ document-level genre features borrowed from text classification literature
  ✦ directly observable in raw text
  ✦ we also test: to what extent can LDA-inferred ‘topics’ distinguish our genres?
### Genre adaptation: genre-revealing features

<table>
<thead>
<tr>
<th>Genre Features</th>
<th>Not used in final VSM genre variant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected for final VSM genre variant</strong></td>
<td><strong>Not used in final VSM genre variant</strong></td>
</tr>
<tr>
<td>First person pronoun count</td>
<td>Third person pronoun count</td>
</tr>
<tr>
<td>Second person pronoun count</td>
<td>Plural pronoun count</td>
</tr>
<tr>
<td>Repeating punctuation count (“...”, “?!”, etc.)</td>
<td>Average word length</td>
</tr>
<tr>
<td>Exclamation mark count</td>
<td>Average sentence length</td>
</tr>
<tr>
<td>Question mark count</td>
<td>Total punctuation count</td>
</tr>
<tr>
<td>Emoticons count</td>
<td>Unique words count</td>
</tr>
<tr>
<td>Numbers count</td>
<td>Long words (&gt; 7 chars) count</td>
</tr>
</tbody>
</table>
Genre adaptation: genre-revealing features

<table>
<thead>
<tr>
<th>Genre Features</th>
<th>Selected for final VSM genre variant</th>
<th>Not used in final VSM genre variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>First person pronoun count</td>
<td></td>
<td>Third person pronoun count</td>
</tr>
<tr>
<td>Second person pronoun count</td>
<td></td>
<td>Plural pronoun count</td>
</tr>
<tr>
<td>Repeating punctuation count (&quot;...&quot;, &quot;?!&quot;, etc.)</td>
<td></td>
<td>Average word length</td>
</tr>
<tr>
<td>Exclamation mark count</td>
<td></td>
<td>Average sentence length</td>
</tr>
<tr>
<td>Question mark count</td>
<td></td>
<td>Total punctuation count</td>
</tr>
<tr>
<td>Emoticons count</td>
<td></td>
<td>Unique words count</td>
</tr>
<tr>
<td>Numbers count</td>
<td></td>
<td>Long words (&gt; 7 chars) count</td>
</tr>
</tbody>
</table>

- Seven most discriminative features between NW and UG are used in final VSM version
Genre adaptation: three hypotheses

The proposed genre-revealing features…
Genre adaptation: three hypotheses

The proposed genre-revealing features…

1. enhance translation performance for NW and UG
   ‣ measured in BLEU
Genre adaptation: three hypotheses

The proposed genre-revealing features…

1. enhance translation performance for NW and UG
   ✦ measured in BLEU
2. can be projected across languages
   ✦ values computed for Arabic and English
Genre adaptation: three hypotheses

The proposed genre-revealing features…

1. enhance translation performance for NW and UG
   ✦ measured in BLEU
2. can be projected across languages
   ✦ values computed for Arabic and English
3. encourage translation consistency
   ✦ since lexical choice is more tailored towards different genres
Enhanced translation performance

- Automatic features can replace manual labels

![Graph showing +BLEU over baseline for different datasets and conditions]

- Manual provenance labels
- Automatic features (genre+LDA)
Projection across languages

* Features can be extracted on either side of the bitext

![Graph showing BLEU score improvements with source-side and target-side genre features for G&T NW, G&T UG, NIST NW, and NIST UG.]
Increased translation consistency*

* Repeated phrase: any phrase that occurs at least twice in a single document

* Following Carpuat and Simard, 2012
Increased translation consistency*

* **Repeated phrase**: any phrase that occurs at least twice in a single document

* If all translations are identical (except for punctuation or stopwords): **consistent** translation

* Following Carpuat and Simard, 2012
Increased translation consistency*

- **Repeated phrase**: any phrase that occurs at least twice in a single document
- If all translations are identical (except for punctuation or stopwords): **consistent** translation

* Following Carpuat and Simard, 2012
Increased translation consistency*

- **Repeated phrase**: any phrase that occurs at least twice in a single document
- If all translations are identical (except for punctuation or stopwords): **consistent** translation

* Following Carpuat and Simard, 2012
Increased translation consistency*

* Repeated phrase: any phrase that occurs at least twice in a single document

* If all translations are identical (except for punctuation or stopwords): consistent translation

* Following Carpuat and Simard, 2012
Increased translation consistency*

* Repeated phrase: any phrase that occurs at least twice in a single document
* If all translations are identical (except for punctuation or stopwords): consistent translation

* Following Carpuat and Simard, 2012
Translation consistency: results

- Adapted system increases translation consistency

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Adapted</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>G&amp;T NW</td>
<td></td>
<td></td>
<td>+4.2</td>
</tr>
<tr>
<td>G&amp;T UG</td>
<td></td>
<td></td>
<td>+2.7</td>
</tr>
<tr>
<td>NIST NW</td>
<td></td>
<td></td>
<td>+0.1</td>
</tr>
<tr>
<td>NIST UG</td>
<td></td>
<td></td>
<td>+2.6</td>
</tr>
</tbody>
</table>
Genre adaptation: some examples

• Genre-adapted system favors:
Genre adaptation: some examples

- Genre-adapted system favors:
  - colloquial translation options for UG
Genre adaptation: some examples

* Genre-adapted system favors:
  ✦ colloquial translation options for UG

<table>
<thead>
<tr>
<th>Source phrase</th>
<th>Baseline translation</th>
<th>Adapted system’s translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>وهذا يدل</td>
<td>and this <em>indicates</em></td>
<td>and this <em>shows</em></td>
</tr>
<tr>
<td>مليار دولار سنويا</td>
<td>billion dollars <em>annually</em></td>
<td>billion dollars <em>a year</em></td>
</tr>
</tbody>
</table>
Genre adaptation: some examples

- Genre-adapted system favors:
  - colloquial translation options for UG
  - formal or concise translation options for NW

<table>
<thead>
<tr>
<th>Source phrase</th>
<th>Baseline translation</th>
<th>Adapted system’s translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>وهذا يدل</td>
<td>and this <em>indicates</em></td>
<td>and this <em>shows</em></td>
</tr>
<tr>
<td>مليار دولار سنويا</td>
<td>billion dollars <em>annually</em></td>
<td>billion dollars <em>a year</em></td>
</tr>
</tbody>
</table>
Genre adaptation: some examples

* Genre-adapted system favors:
  - colloquial translation options for UG
  - formal or concise translation options for NW

<table>
<thead>
<tr>
<th>Source phrase</th>
<th>Baseline translation</th>
<th>Adapted system’s translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>وهذا يدل</td>
<td>and this <em>indicates</em></td>
<td>and this <em>shows</em></td>
</tr>
<tr>
<td>مليار دولار سنويا</td>
<td>billion dollars <em>annually</em></td>
<td>billion dollars <em>a year</em></td>
</tr>
<tr>
<td>القطاع الصحي</td>
<td><em>workers in</em> the health sector</td>
<td>the health sector</td>
</tr>
<tr>
<td>عالميا</td>
<td><em>worldwide</em></td>
<td><em>global</em></td>
</tr>
</tbody>
</table>
In conclusion: what we did and why

* Most approaches to domain adaptation for SMT rely on **provenance** information
In conclusion: what we did and why

* Most approaches to domain adaptation for SMT rely on **provenance** information

* Provenance is not an intrinsic text property and often combines **topic** and **genre**
In conclusion: what we did and why

* Most approaches to domain adaptation for SMT rely on **provenance** information

* Provenance is not an intrinsic text property and often combines **topic** and **genre**

* When disentangling topic and genre, we found that genre differences pose the biggest challenge to SMT
In conclusion: what we did and why

- Most approaches to domain adaptation for SMT rely on **provenance** information
- Provenance is not an intrinsic text property and often combines **topic** and **genre**
- When disentangling topic and genre, we found that genre differences pose the biggest challenge to SMT
- We ask: can we address genre adaptation using only intrinsic text features?
In conclusion: what we learned

* We can eliminate the need for manual provenance information in a flexible adaptation framework
In conclusion: what we learned

- We can eliminate the need for manual provenance information in a flexible adaptation framework
- Our proposed document-level genre features
In conclusion: what we learned

- We can eliminate the need for manual provenance information in a flexible adaptation framework
- Our proposed document-level genre features
  - are simple but powerful
In conclusion: what we learned

• We can eliminate the need for manual provenance information in a flexible adaptation framework

• Our proposed document-level genre features
  ✦ are simple but powerful
  ✦ enhance translation performance
In conclusion: what we learned

- We can eliminate the need for manual provenance information in a flexible adaptation framework

- Our proposed document-level genre features
  - are simple but powerful
  - enhance translation performance
  - can be projected across languages
In conclusion: what we learned

- We can eliminate the need for manual provenance information in a flexible adaptation framework
- Our proposed document-level genre features
  - are simple but powerful
  - enhance translation performance
  - can be projected across languages
  - encourage translation consistency
Thank you!

Translation Model Adaptation Using Genre-Revealing Text Features

Marlies van der Wees       Arianna Bisazza       Christof Monz
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

Abstract

Research in domain adaptation for statistical machine translation (SMT) has resulted in various approaches that adapt system components to specific translation tasks. The concept of a domain, however, is not precisely defined across existing domain adaptation methods. Different domains typically correspond to different subcorpora, in which documents exhibit a particular combination of genre and topic, which is exactly the challenge of domain adaptation in SMT.