Transfer Learning for Low-resource Natural Language Analysis

Yuan Zhang
January 30, 2017
Low-resource Problem

- Top-performing systems need large amounts of annotated data

Dependency Parsing Accuracy on English

<table>
<thead>
<tr>
<th>Size of Training Data</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6k tokens</td>
<td>74.2</td>
</tr>
<tr>
<td>950k tokens</td>
<td>90.3</td>
</tr>
</tbody>
</table>
Low-resource Scenarios

Low-resource Languages:

- **Malagasy** annotations
  - ~1,000 tokens

- English annotations
  - > 1 million tokens
Low-resource Scenarios

Low-resource Languages:

- Malagasy annotations
  ~1,000 tokens

- English annotations
  > 1 million tokens

Low-resource Domains:

- PubMed
  Medical: ~ 500 sentences

- The Wall Street Journal
  News articles: > 100k sentences
Our Work: Transfer Learning

- Use rich resources in related **source** tasks to improve **target** performance
Challenges in Transfer: Multilingual

• Part-of-speech (POS) tagging: different vocabulary

Source: English

<table>
<thead>
<tr>
<th>DET</th>
<th>ADJ</th>
<th>NOUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>red</td>
<td>apple</td>
</tr>
</tbody>
</table>

Target: French

<table>
<thead>
<tr>
<th>DET</th>
<th>NOUN</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>une</td>
<td>pomme (apple)</td>
<td>rouge (red)</td>
</tr>
</tbody>
</table>
Challenges in Transfer: Multilingual

• Part-of-speech (POS) tagging: different vocabulary
• Dependency parsing: different word ordering
Challenges in Transfer: Monolingual

• Domain transfer: different writing-style

Source: Restaurant reviews

The fries were undercooked

Target: Hotel reviews

The room rained water from above
Challenges in Transfer: Monolingual

• Domain transfer: different writing-style

  Source: Restaurant reviews
  The fries were undercooked

  Target: Hotel reviews
  The room rained water from above

• Aspect transfer: different aspects in the same domain

  Source Aspect: IDC
  Target Aspect: LVI

  FINAL DIAGNOSIS: BREAST (LEFT) … INVASIVE DUCTAL CARCINOMA (IDC) Tumor size: num x num x num cm  Grade: 3. Lymphatic vessel invasion (LVI): Not identified. Blood vessel invasion: Suspicious. Margin of invasive carcinoma …
General Setup: Low-resource Transfer

• No annotations for the target task

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✔️</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

• No parallel data, or a few word translation pairs

• Low level of human effort
  ✦ Existing external resources
  ✦ No feature engineering
General Setup: Low-resource Transfer

• No annotations for the target task

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

• No parallel data, or a few word translation pairs

• Low level of human effort
  ✦ Existing external resources
  ✦ No feature engineering

Contribution: Improve low-resource transfer in multilingual and monolingual scenarios
Our Approach

**Multilingual Transfer:**

- **Hierarchical tensors** for dependency parsing
  - *Prior knowledge incorporation without feature engineering*

- Multilingual embeddings for POS tagging

**Monolingual Transfer:**

- Adversarial networks for aspect transfer
Multilingual Transfer for Dependency Parsing

**Train on Source Languages**

- English
- Spanish

* sentences are non-parallel

**Test on Target Language**

- French

```
Je   mange   une   pomme   rouge
(l)   (eat)   (a)   (apple)   (red)
```

Dependency Parser

```
Je   mange   une   pomme   rouge
(l)   (eat)   (a)   (apple)   (red)
```
Non-lexical Transfer via **Universal POS**

**Train on Source Languages**

English

... ➔

PRON   VERB   DET   ADJ   NOUN

Spanish

... ➔

PRON   VERB   DET   NOUN   ADJ

**Test on Target Language**

French

```
PRON   VERB   DET   NOUN   ADJ
```

Dependency Parser

```
PRON   VERB   DET   NOUN   ADJ
```
Challenge: Different Word Ordering

Train on Source Languages

- **English**
  
  ... PRON VERB DET ADJ NOUN ...

- **Spanish**
  
  ... PRON VERB DET NOUN ADJ ...

Test on Target Language

- **French**
  
  PRON VERB DET NOUN ADJ

Dependency Parser

  
  PRON VERB DET NOUN ADJ
Solution: Linguistic Typology

• Form of typological features

<table>
<thead>
<tr>
<th>Typological Feature</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>87A: Order of Noun and Adjective</td>
<td>ADJ-NOUN</td>
<td>NOUN-ADJ</td>
</tr>
</tbody>
</table>

• Idea of selective transfer

English: 87A=ADJ-NOUN

French: 87A=NOUN-ADJ

Spanish: 87A=NOUN-ADJ
Utilizing Typology Knowledge

Knowledge Utilization

Manual  Engineering Effort  Automatic
Utilizing Typology Knowledge

Traditional approach: manual feature engineering
Utilizing Typology Knowledge

Traditional approach: manual feature engineering

Tensor scoring: invalid features violating prior knowledge
Utilizing Typology Knowledge

Traditional approach: manual feature engineering

Our approach: hierarchical tensor with prior knowledge

Tensor scoring: invalid features violating prior knowledge

Manual  Engineering Effort  Automatic

Knowledge Utilization

High  Low
Traditional Approach: Feature Engineering

- Manually conjoin standard parsing features with typological features (Täckström et al., 2013)

\[ f_{100}(\cdot) = \mathbb{1}\{\text{head POS=\textsc{noun}, modifier POS=\textsc{adj}, direction=\text{Right}, 87A=\textsc{noun-adj}}\} \]

\[ \text{• 87A: code of noun-adjective typological feature} \]
Traditional Approach: Feature Engineering

- Manually conjoin **standard parsing features** with **typological features** (Täckström et al., 2013)

\[
f_{100}(\cdot) = \mathbb{I}\{\text{head POS=\textsc{NOUN}, modifier POS=\textsc{ADJ}, direction=Right, } 87A=\textsc{NOUN-ADJ}\}
\]

* 87A: code of noun-adjective typological feature

- Features are selectively shared

**English:** \(87A=\text{ADJ-NOUN}\)

\[
f_{100}(\text{\textsc{NOUN} \rightarrow \text{ADJ}}) = 0
\]

**French:** \(87A=\text{NOUN-ADJ}\)

\[
f_{100}(\text{\textsc{NOUN} \rightarrow \text{ADJ}}) = 1
\]

**Spanish:** \(87A=\text{NOUN-ADJ}\)

\[
f_{100}(\text{\textsc{NOUN} \rightarrow \text{ADJ}}) = 1
\]
Traditional Approach: Feature Engineering

• Manually conjoin standard parsing features with typological features (Täckström et al., 2013)

\[ f_{100}(\cdot) = \mathbb{1}\{\text{head POS=\text{NOUN}, modifier POS=\text{ADJ}, direction=Right, 87A=\text{NOUN-ADJ}}\} \]

* 87A: code of noun-adjective typological feature

• Features are selectively shared

English: 87A=ADJ-NOUN

\[ f_{100}(\text{NOUN} \rightarrow \text{ADJ}) = 0 \]

French: 87A=NOUN-ADJ

\[ f_{100}(\text{NOUN} \rightarrow \text{ADJ}) = 1 \]

Spanish: 87A=NOUN-ADJ

\[ f_{100}(\text{NOUN} \rightarrow \text{ADJ}) = 1 \]

• In practice, need to manually construct hundreds of features
Tensor Scoring Method

- Represent arc features in a tensor view (e.g., 4-way tensor)
- Automatically capture all possible feature combinations
Low-rank Feature Representation

- Avoid parameter explosion via low-rank factorization
- Learn feature mappings to a low-rank representation

\[
\text{head POS feature vector} \times \text{parameter matrix} = \text{low-rank representation}
\]

\[
1 \times d \times d \times r = 1 \times r
\]
Low-rank Feature Representation

head POS  modifier POS  direction  typology
Low-rank Feature Representation

- Compute low-rank representation of an arc via element-wise product.

```
head POS  
modifier POS    
direction     
typology
```

```
low-rank representation of an arc

element-wise product
```
Low-rank Feature Representation

- Compute low-rank representation of an arc via element-wise product
- Compute arc score as:

\[ S(h \rightarrow m) = e_0 + e_1 + e_2 + \cdots + e_r \]
Issue of Tensor Methods

- Capture **invalid** feature combinations and assign non-zero weights
Issue of Tensor Methods

- Capture invalid feature combinations and assign non-zero weights.
- Should avoid directly taking tensor-product between typology and others.

Validity of Tensor Methods

Capture invalid feature combinations and assign non-zero weights.

Avoid directly taking tensor-product between typology and others.

{VERB, NOUN, LEFT, ADJ-NOUN}
Invalid Combination
Avoid Product Operation

• typology

head POS  modifier POS  direction
Target Feature Combination

- Union of different feature groups

Diagram:
- Head POS
- Modifier POS
- Direction
- Typology
- Not combined
Solution: Hierarchical Structure

- Element-wise sum operation over different representations of the same set of atomic features

- Traditional representation over head, modifier and direction

- Typology representation over head, modifier and direction

- Element-wise sum

- head POS

- modifier POS

- direction

- typology
Solution: Hierarchical Structure

- Element-wise sum operation over different representations of the same set of atomic features

Mixed representation over head, modifier and direction

Traditional representation over head, modifier and direction

Typology representation over head, modifier and direction

\[ \text{Mixed representation} \oplus \text{Typology representation} = \text{Traditional representation} \]

= element-wise sum

head POS \quad \text{modifier POS} \quad \text{direction}
Solution: Hierarchical Structure
Solution: Hierarchical Structure

Representation over head, modifier, direction and label

- head POS
- modifier POS
- direction
- typology

label
Solution: Hierarchical Structure

**Representation over head, modifier, direction and label**

**Typology** representation over head, modifier, direction and label. E.g. subject-verb
Solution: Hierarchical Structure

low-rank representation of an arc

head POS

modifier POS

classifier POS

head POS

modifier POS

direction

label typology

typology

label
Algebraic Interpretation

- Algebraically equal the sum of three multiway tensors with shared parameters
- Capture three groups of feature combinations
Algebraic Interpretation

- Algebraically equal the sum of three multiway tensors with shared parameters
- Capture three groups of feature combinations
Avoid Invalid Features

• Exclude the combination of *typology* with head, modifier and direction

• Assign zero weights to invalid features

  • Weight of \{head POS=VERB, mod POS=NOUN, typology=ADJ-NOUN\} is 0
Parameter Initialization and Learning

Algebraic view:
Compute the gradient for each multiway tensor and take the sum

Tensor initialization:
Use iterative power methods

Parameter learning:
Adopt online learning with passive-aggressive algorithm

Other details:
Follow previous work (Lei et al., 2015)
Experimental Setup

**Dataset:** Universal Dependency Treebank v2.0
- 10 languages
- Universal POS tags (12 tags)
- Stanford dependency labels (40 labels)

**Baselines:**
- Direct transfer (McDonald et al., 2005)
- Feature-based transfer (Täckström et al., 2013)
- Traditional multiway tensor
Unsupervised Results

Averaged Unlabeled Attachment Score (UAS)

- Setting: no annotations in the target language
Unsupervised Results

Averaged Unlabeled Attachment Score (UAS)

- Direct Transfer: 67.8
- Ours: 72.6
Unsupervised Results

Averaged Unlabeled Attachment Score (UAS)

- Direct Transfer: 67.8
- NT-Select: 71.5
- Ours: 72.6

NT-Select: our model without the tensor component, corresponding to prior feature-based method (Täckström et al., 2013)
Unsupervised Results

Averaged Unlabeled Attachment Score (UAS)

- Direct Transfer: 67.8
- NT-Select: 71.5
- Multiway: 72
- Ours: 72.6

• **Multiway**: traditional multiway tensor without hierarchical structure
Semi-supervised Results

Averaged Unlabeled Attachment Score (UAS)

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Transfer</td>
<td>73.4</td>
</tr>
<tr>
<td>Sup50</td>
<td>75.6</td>
</tr>
<tr>
<td>NT-Select</td>
<td>76.2</td>
</tr>
<tr>
<td>Multiway</td>
<td>76.9</td>
</tr>
<tr>
<td>Ours</td>
<td>77.9</td>
</tr>
</tbody>
</table>

- **Setting**: 50 annotated sentences in the target language
- **Sup50**: trained only on the 50 sentences in the target language
Summary

• **Modeling:** we present a hierarchical tensor that effectively uses linguistic prior knowledge

• **Performance:** our model outperforms state-of-the-art approach and traditional tensors

• **Limitation:** our model heavily relies on non-lexical transfer via universal POS tags

Next part: lexical-level multilingual transfer
Our Approach

Multilingual Transfer:

• Hierarchical tensors for dependency parsing

• Multilingual embeddings for POS tagging
  - Effective multilingual transfer with ten translation pairs

Monolingual Transfer:

• Adversarial networks for aspect transfer
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

98.2

Supervised 700k tokens
(Brants, 2000)
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

- Multilingual Transfer
  - 2m parallel sentences
  - (Das and Petrov, 2011)
  - 82.8

- Supervised
  - 700k tokens
  - (Brants, 2000)
  - 98.2
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype-driven</td>
<td>25.5</td>
<td>14 prototypes (Haghighi et al., 2006)</td>
</tr>
<tr>
<td>Multilingual Transfer</td>
<td>82.8</td>
<td>2m parallel sentences (Das and Petrov, 2011)</td>
</tr>
<tr>
<td>Supervised</td>
<td>98.2</td>
<td>700k tokens (Brants, 2000)</td>
</tr>
</tbody>
</table>
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

- Prototype-driven
  14 prototypes
  (Haghighi et al., 2006)
  - 25.5

- Multilingual Transfer
  Ten Translation Pairs
  No parallel sentences
  - ?

- Multilingual Transfer
  2m parallel sentences
  (Das and Petrov, 2011)
  - 82.8

- Supervised
  700k tokens
  (Brants, 2000)
  - 98.2
Multilingual Transfer of POS Tagging

Tagging Accuracy on German

Prototype-driven
14 prototypes
(Haghighi et al., 2006)

Multilingual Transfer
Ten Translation Pairs
No parallel sentences

Multilingual Transfer
2m parallel sentences
(Das and Petrov, 2011)

Supervised
700k tokens
(Brants, 2000)

25.5

82.8

98.2

How little parallel data is necessary to enable multilingual transfer?
Our Work

• Task: multilingual transfer of part-of-speech (POS) tagging

• Data:

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Our Work

- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

(non-parallel data)

Ten Translation Pairs

. || .                       und || and
, || ,                       dem || the
der || the                    von || from
die || the                   - || -
in || in                      zu || to
Our Work

- Task: multilingual transfer of part-of-speech (POS) tagging
- Data:

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled</td>
<td>✓</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>✓</td>
</tr>
</tbody>
</table>

(non-parallel data)

Ten Translation Pairs

- . || .
- , || ,
- der || the
- die || the
- in || in
- und || and
- dem || the
- von || from
- - || -
- zu || to

POS Accuracy on German

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype (Haghighi et al., 2006)</td>
<td>Ours</td>
</tr>
<tr>
<td>25.5</td>
<td>68.7</td>
</tr>
</tbody>
</table>
Our Two-step Method

1. Learn **coarse mapping** between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via **unsupervised learning** on the target language
Coarse Mapping between Embeddings

• Goal: find a **linear transformation** from target to source embedding space
• Objective: **minimize the distance** between translation pairs

**Monolingual Embedding**

**Source (English)**
- red
- cat
- is

**Target (German)**
- Katze (cat)
- ist (is)
- Hund (dog)
- rot (red)
Coarse Mapping between Embeddings

- Goal: find a **linear transformation** from target to source embedding space
- Objective: **minimize the distance** between translation pairs

**Monolingual Embedding**

**Source (English)**
- dog
- cat
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

**Monolingual Embedding**

**Source (English)**
- red
- cat
- is

**Target (German)**
- Katze (cat)
- Hund (dog)
- ist (is)
- rot (red)

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot

**Too many degrees of freedom**
- dimension: 20
- # pairs: 10
- degree of freedom: 10
Coarse Mapping between Embeddings

- Goal: find a linear transformation from target to source embedding space
- Objective: minimize the distance between translation pairs

**Monolingual Embedding**

**Source (English)**
- dog
- cat
- is
- red

**Target (German)**
- Hund (dog)
- Katze (cat)
- ist (is)
- rot (red)

**Translation Pairs**
- dog | | Hund
- cat | | Katze
- red | | rot

**Too many degrees of freedom**
- dimension: 20
- # pairs: 10
- degree of freedom: 10

**Solutions need to be constrained!**
Our Solution: Isometric Constraints

• Transformation $P$ is an isometric (orthonormal) matrix
• Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

**Isometric Constraints**

$$P^T P = I$$

**Translation Pairs**

- dog $\|\|\|$ Hund
- cat $\|\|\|$ Katze
- red $\|\|\|$ rot

**Monolingual Embedding**

**Source (English)**
- red
- cat
- is

**Target (German)**
- Katze (cat)
- ist (is)
- Hund (dog)
- rot (red)
Our Solution: Isometric Constraints

• Transformation $P$ is an isometric (orthonormal) matrix
• Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations
  \[ \cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle, \quad \cos\langle\text{dog, red}\rangle \approx \cos\langle\text{Hund, rot}\rangle \]

Monolingual Embedding

Source(English)

red

dog

cat

is

Target(German)

Katze (cat)

Hund (dog)

ist (is)

rot (red)

Isometric Solution

\[ P^T P = I \]

Translation Pairs

dog || Hund

cat || Katze

red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

**Monolingual Embedding**

**Source (English)**
- red
- cat
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- rot (red)
- ist (is)

**Isometric Constraints**

$P^T P = I$

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations

### Monolingual Embedding

**Source (English)**
- red
- cat
- is

**Target (German)**
- Katze (cat)
- Hund (dog)
- ist (is)
- rot (red)

### Isometric Solution

**Isometric Constraints**

$$P^T P = I$$

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Our Solution: Isometric Constraints

- Transformation $P$ is an isometric (orthonormal) matrix
- Transformation preserves angles and lengths (cosine similarity) of word vectors, thus preserving semantic relations
- Use the steepest descent algorithm (Abrudan et al., 2008)

**Monolingual Embedding**

**Source (English)**
- dog
- red
- cat
- is

**Target (German)**
- Hund (dog)
- Katze (cat)
- rot (red)
- ist (is)

**Isometric Constraints**

$$P^T P = I$$

**Translation Pairs**
- dog || Hund
- cat || Katze
- red || rot
Validation of Isometric Constraints

• Validation for $\cos\langle \text{cat}, \text{dog} \rangle \approx \cos\langle \text{Katze}, \text{Hund} \rangle$
• Verify whether nearest neighbors are preserved after translations

- For 50% of word pairs, $k \leq 2$

**English: nearest neighbor**

- dog
- cat

**German: $k$-th ($k \leq 2$) nearest neighbor?**

- Katze (cat)
- Hund (dog)
Validation of Isometric Constraints

- Validation for $\cos\langle\text{cat, dog}\rangle \approx \cos\langle\text{Katze, Hund}\rangle$
- Verify whether nearest neighbors are preserved after translations

For 50% of word pairs, $k \leq 2$

For 90% of word pairs, $k \leq 10$
Direct Transfer Model

- Supervised source language HMM
  - Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - Word embeddings as emission features
Direct Transfer Model

- Supervised source language HMM
  - Feature-based HMM (Berg-Kirkpatrick et al., 2010)
  - Word embeddings as emission features

Coarse mapping is not accurate
Our Two-step Method

1. Learn **coarse mapping** between embeddings via ten translation pairs

2. Refine embedding transformations and model parameters via **unsupervised learning** on the target language
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.

- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$

  $$p(x|y) \propto \exp\{\mathbf{v}_x^T PM \mu_y + \theta_{x,y}\}$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

$$p(x|y) \propto \exp\{\nu_x^T P M \mu_y + \theta_{x,y}\}$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

\[ p(x|y) \propto \exp\{v_x^T PM \mu_y + \theta_{x,y}\} \]
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$

$$p(x|y) \propto \exp\{v_x^T PM \mu_y + \theta_{x,y}\}$$
Unsupervised Target Language HMM

- Use the direct transfer model (based on the coarse mapping) to initialize and regularize the unsupervised tagger on the target language.
- Refine mapping via global linear transformation $M$ and local non-linear adjustment $\theta_{x,y}$.

$$p(x|y) \propto \exp\{v^T_x PM \mu_y + \theta_{x,y}\}$$
Experimental Setup

• **Datasets:** Universal Dependency Treebank v1.2
  ✦ Source: English
  ✦ Target (Indo-European): Danish, German, Spanish
  ✦ Target (non-Indo-European): Finnish, Hungarian, Indonesian

• **Universal tagset:** 14 tags (noun, verb, adjective etc.)

• **Word embeddings:** 20-dimension vectors trained on Wiki dumps using word2vec
Indo-European Results

Averaged Accuracy on Indo-European Languages

- Prototype (Haghigi et al., 2006): 31.8%
- Direct Transfer: 60.9%
- Ours Full: 72.9%
Non-Indo-European Results

Averaged Accuracy on non-Indo-European Languages

- Prototype (Haghigi et al., 2006): 27.6%
- Direct Transfer: 57.7%
- Ours Full: 62.1%
Prediction of Linguistic Typology

- Task: predict whether a language is verb-object or object-verb (five typological properties)
- Features: bigrams and trigrams of POS tags
Impact of Amount of Supervision

- Ours Full with 10 pairs = 150 prototypes
Impact of Amount of Supervision

- Ours Full with 10 pairs = 150 prototypes
- Prototype improves with large amount of annotations

Accuracy on German

![Graph showing the impact of amount of supervision on accuracy]
Summary

• **Modeling:** ten translation pairs are sufficient to enable multilingual transfer for POS tagging

• **Performance:** our model significantly outperforms the direct transfer and the prototype-driven method
Our Approach

**Multilingual Transfer:**

- Hierarchical tensors for dependency parsing
- Multilingual embeddings for POS tagging

**Monolingual Transfer:**

- Adversarial networks for aspect transfer
  - Joint aspect-driven encoding and domain adversarial training
Pathology report:

FINAL DIAGNOSIS: BREAST (LEFT) ... INVASIVE DUCTAL CARCINOMA (IDC) Tumor size: num x num x num cm  Grade: 3. Lymphatic vessel invasion (LVI): Not identified. Blood vessel invasion: Suspicious. Margin of invasive carcinoma ...

Diagnosis results:

IDC: Positive  LVI: Negative

Transfer:

Source: IDC  Target: LVI
Challenge

**Same report; Different key sentences**

Source Aspect: IDC  
Target Aspect: LVI

**FINAL DIAGNOSIS: BREAST (LEFT) … INVASIVE DUCTAL CARCINOMA (IDC)**  
Tumor size: num x num x num cm  

- Traditional methods will fail because they always induce the same representation for the same input
Available Supervision

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Data</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Unlabeled Data</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Relevance Rules</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Relevance rules: common names of aspects
  - ALH: Atypical Lobular Hyperplasia, ALH
  - IDC: Invasive Ductal Carcinoma, IDC
Transfer Assumption: Aspects Are Related

• Different aspects share the same label set: positive/negative

  IDC: Positive  LVI: Negative
Transfer Assumption: Aspects Are Related

- Different aspects share the same label set: positive/negative
  IDC: Positive \hspace{1cm} LVI: Negative

- Common words are directly transferrable

Invasive Carcinoma is present
Label: Positive

\hspace{1.5cm} \rightarrow \hspace{1.5cm}

Lymphatic vessel invasion: present
Label: Positive
Transfer Assumption: Aspects Are Related

- Different aspects share the same label set: positive/negative
  
  IDC: Positive \hspace{2cm} LVI: Negative

- Common words are directly transferrable

  Invasive Carcinoma is present \hspace{1cm} Lymphatic vessel invasion: present
  Label: Positive \hspace{1cm} Label: Positive

- Aspect-specific words are not directly transferrable
  - Goal: map them to invariant representations

  Invasive Ductal Carcinoma

  Lymphatic Vessel Invasion
Key Idea: Aspect-driven Encoding

• Leverage relevance rules to learn to identify key sentences
• Learn differential representations for different aspects from the same input
Key Idea: Aspect-driven Encoding

- Leverage relevance rules to learn to identify key sentences
- Learn differential representations for different aspects from the same input

Reduce aspect transfer to standard domain adaptation
Key Idea: Domain-Adversarial

- Jointly train a domain classifier
- Use domain-adversarial training for learning invariant representations
  - Objective: Not separable by the domain classifier
Overall Framework: Three Components
Overall Framework: Three Components

Pathology report → Document encoder → Document representation → Label predictor → document label y

Domain classifier → domain label d
Sentence Embedding

• Apply a CNN to each sentence

... ductal carcinoma is identified ...

sentence embeddings

max-pooling

\[ h_1 \quad \cdots \quad h_2 \]

\[ x_0 \quad x_1 \quad x_2 \quad x_3 \]
Sentence Embedding

• Apply a CNN to each sentence

• Improve adversarial training by reconstruction

sentence embeddings

max-pooling

\[ \hat{x}_2 = \tanh(W^c h_2 + b^c) \]

... ductal carcinoma is identified ...
Aspect-relevance Prediction

- Predict relevance score based on sentence embeddings
- Train on relevance rules (e.g., names of IDC, LVI)

Pathology report

INVASIVE CARCINOMA
Tumor size ... Grade: 3.

Lymphatic vessel invasion: Not identified.

Predicted relevance score

Sentence embeddings

$\text{r} = 1.0$

$\square \text{r} = 0.0$
Aspect-relevance Prediction

• Predict relevance score based on sentence embeddings
• Train on relevance rules (e.g., names of IDC, LVI)
Aspect-driven Document Encoding

- Combine sentence vectors based on relevance weights

Pathology report

INVASIVE CARCINOMA
Tumor size ... Grade: 3.

Lymphatic vessel invasion: Not identified.

Predicted relevance score

Document representation

Weighted combination

Sentence embeddings

$r = 1.0$

$r = 0.0$
Aspect-driven Document Encoding

- Combine sentence vectors based on relevance weights
- Add a transformation layer at the end

Pathology report

INVASIVE CARCINOMA
Tumor size ... Grade: 3.

Lymphatic vessel invasion: Not identified.

Predicted relevance score

$r = 1.0$

$r = 0.0$

Document representation

Sentence embeddings

Weighted combination
Document Label Predictor

- Share for both source and target aspects
- Train on labeled data in the source aspect

Objective: predict labels
Domain Classifier and Adversary

- Learn domain-invariant representations
- Train on both labeled and unlabeled data

Objective: predict labels

ReLU

Softmax

document label $y$

Document representation

Pathology report

Adversary objective: fail the domain classifier

Objective: predict domains

ReLU

Softmax

domain label $d$
Pathology Dataset

- Aspect-transfer on breast cancer pathology reports from hospitals such as MGH

Source: IDC ➔ Target: LCIS

FINAL DIAGNOSIS: BREAST (LEFT) ... INVASIVE DUCTAL CARCINOMA Grade: 3. Lobular Carcinoma In-situ: Not identified. Blood vessel invasion: Suspicious. ...
Pathology Dataset

- Aspect-transfer on breast cancer pathology reports from hospitals such as MGH

Source: IDC  \[\rightarrow\]  Target: LCIS

**FINAL DIAGNOSIS:** BREAST (LEFT)  ...  INVASIVE DUCTAL CARCINOMA Grade: 3.  Lobular Carcinoma In-situ: Not identified.  Blood vessel invasion: Suspicious.  ...

- Statistics and relevance rules:

<table>
<thead>
<tr>
<th>Aspects</th>
<th>#Labeled</th>
<th>#Unlabeled</th>
<th>Relevance Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCIS</td>
<td>23.8k</td>
<td></td>
<td>DCIS, Ductal Carcinoma In-Situ</td>
</tr>
<tr>
<td>LCIS</td>
<td>10.7k</td>
<td>96.6k</td>
<td>LCIS, Lobular Carcinoma In-Situ</td>
</tr>
<tr>
<td>IDC</td>
<td>22.9k</td>
<td></td>
<td>IDC, Invasive Ductal Carcinoma</td>
</tr>
<tr>
<td>ALH</td>
<td>9.2k</td>
<td></td>
<td>ALH, Atypical Lobular Hyperplasia</td>
</tr>
</tbody>
</table>

- 500 reports for testing
Review Dataset

- Domain transfer for sentiment analysis: positive or negative
- Common words (e.g. excellent) are directly transferrable, but domain-specific words are not

Source: Hotel (TripAdvisor)

- This place was **excellent**!
- In the second bedroom it literally **rained water from above** …

Target: Restaurant (Yelp)

- **Excellent** food.
- The fries were **undercooked** and **thrown haphazardly** into the sauce holder …
Review Dataset

• Domain transfer for sentiment analysis: positive or negative
• Common words (e.g. excellent) are directly transferrable, but domain-specific words are not

Source: Hotel (TripAdvisor)
- This place was excellent!
- In the second bedroom it literally rained water from above …

Target: Restaurant (Yelp)
- Excellent food.
- The fries were undercooked and thrown haphazardly into the sauce holder …

• Statistics and relevance rules:

<table>
<thead>
<tr>
<th>Domains</th>
<th>#Labeled</th>
<th>#Unlabeled</th>
<th>Relevance Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel</td>
<td>100k</td>
<td>100k</td>
<td>Five aspects, 290 keywords (Wang et al., 2011)</td>
</tr>
<tr>
<td>Restaurant</td>
<td>-</td>
<td>200k</td>
<td>(only one overall aspect)</td>
</tr>
</tbody>
</table>

✦ 2k reviews for testing
Results on Pathology Dataset

Averaged accuracy over 6 transfer scenarios

- mSDA: marginalized stacked denoising autoencoder (Chen et al., 2012)
Results on Pathology Dataset

Averaged accuracy over 6 transfer scenarios

- **mSDA**: 67.1
- **Ours-NA**: 81.3
- **Ours-Full**: 94.1

- **Ours-NA**: our model without adversarial training
Results on Pathology Dataset

Averaged accuracy over 6 transfer scenarios

- **mSDA**: 67.1
- **Ours-NA**: 81.3
- **Ours-NR**: 69.8
- **Ours-Full**: 94.1

- **Ours-NR**: our model without aspect-relevance scoring
Results on Pathology Dataset

Averaged accuracy over 6 transfer scenarios

- mSDA: 67.1
- Ours-NA: 81.3
- Ours-NR: 69.8
- Ours-Full: 94.1
- In-domain: 96.9

**In-domain**: supervised training with in-domain annotations
Results on Review Dataset

Averaged accuracy over 5 transfer scenarios

- mSDA: 81.6
- Ours-NA: 83.9
- Ours-NR: 87.3
- Ours-Full: 86.4
- In-domain: 93.4

• Ours-NR and Ours-Full are the two best performing systems
• Relevance scoring has little impact because aspects are highly correlated
Impact of Reconstruction

Average accuracy on the pathology dataset

- The same observation on the review dataset
Reason behind Improvement

• Heat-map: each row corresponds to a document vector
  - Top: source domain; Bottom: target domain

• Adversarial training removes lots of information

- + adversarial, -reconstruction
- + adversarial, + reconstruction
- - adversarial, -reconstruction
Reason behind Improvement

- Heat-map: each row corresponds to a document vector
  - Top: source domain; Bottom: target domain

- Adversarial training removes lots of information

- The reconstruction loss improves both the richness and diversity of the learned representations

Heat-map: each row corresponds to a document vector
- Top: source domain; Bottom: target domain

Adversarial training removes lots of information

The reconstruction loss improves both the richness and diversity of the learned representations
Case Study of Learned Representations

Restaurant Reviews

• the fries were undercooked and thrown haphazardly into the sauce holder. the shrimp was over cooked and just deep fried. … even the water tasted weird.
Case Study of Learned Representations

Restaurant Reviews

• the fries were **undercooked** and **thrown haphazardly** into the sauce holder. The shrimp was **over cooked** and just **deep fried**. … even the water tasted weird.

Nearest Hotel Reviews by **Ours-Full: learns to map domain-specific words**

• the room was **old**. … we didn’t like the night shows at all. …
• however, the decor was **just fair**. … in the second bedroom it literally **rained water from above**.

✦ distance measured by cosine similarity between representations
Case Study of Learned Representations

Restaurant Reviews

• the fries were undercooked and thrown haphazardly into the sauce holder. the shrimp was over cooked and just deep fried. … even the water tasted weird.

Nearest Hotel Reviews by Ours-Full: learns to map domain-specific words

• the room was old. … we didn’t like the night shows at all. …
• however, the decor was just fair. … in the second bedroom it literally rained water from above.

Nearest Hotel Reviews by Ours-NA: only captures common sentiment phrases

• rest room in this restaurant is very dirty. …
• the only problem i had was that … i was very ill with what was suspected to be food poison

* distance measured by cosine similarity between representations
Summary

• **Modeling:** an aspect-augmented adversarial network for cross-aspect and cross-domain transfer tasks.

• **Performance:** our model significantly improves over the mSDA baseline and our model variants on a pathology and a review dataset.
Contributions

Multilingual Transfer:

- Hierarchical tensors for dependency parsing
  - Prior knowledge incorporation without feature engineering

- Multilingual embeddings for POS tagging
  - Effective multilingual transfer with ten translation pairs

Monolingual Transfer:

- Adversarial networks for aspect transfer
  - Joint aspect-driven encoding and domain adversarial training
Thank you!
Contributions

Multilingual Transfer:

• Hierarchical tensors for dependency parsing
  - Prior knowledge incorporation without feature engineering

• Multilingual embeddings for POS tagging
  - Effective multilingual transfer with ten translation pairs

Monolingual Transfer:

• Adversarial networks for aspect transfer
  - Joint aspect-driven encoding and domain adversarial training
Backup Slides
Typological Features

**Word ordering:** five features, e.g.

- Order of Subject and Verb (82A)
- Order of Adjective and Noun (87A)

**Typological feature templates:** eight templates, e.g.

- direction, 87A, head POS=NOUN, modifier POS=ADJ
- direction, 82A, head POS=VERB, modifier POS=NOUN, label=SUBJ
Feature Weights of Multiway Model

Weights of valid features:

head POS=**NOUN**, mod POS=**ADJ**, 87A=**ADJ-NOUN** \( \quad 2.24 \times 10^{-3} \)

Weights of invalid features:

head POS=**VERB**, mod POS=**NOUN**, 87A=**ADJ-NOUN** \( \quad 8.88 \times 10^{-4} \)

head POS=**NOUN**, mod POS=**NOUN**, 87A=**ADJ-NOUN** \( \quad 9.48 \times 10^{-4} \)

Multiway model assigns non-zero weights to invalid features
Impact of Embedding Dimensions and Window Size

- Train embeddings with different dimensions and context window size
- Small window size favors POS tagging
Impact of Embedding Dimensions and Window Size

- Train embeddings with different dimensions and context window size
- Small window size favors POS tagging
- Performance drops with either smaller or larger dimensions