Semantic Role Labeling Tutorial: Part 3
Semi-, unsupervised and cross-lingual approaches

Ivan Titov

NAACL 2013
Shortcomings of Supervised Methods

Supervised methods:

- Rely on large expert-annotated datasets (FrameNet and PropBank > 100k predicates)
- Even then they do not provide high coverage (esp. with FrameNet)
  - \( \sim 50\% \) oracle performance on new data [Palmer and Sporleder, 2010]
- Resulting methods are domain-specific [Pradhan et al., 2008]
- Such resources are not available for many languages

How can we reduce reliance of SRL methods on labeled data?

- Transfer a model or annotation from a more resource-rich language (crosslingual transfer / projection)
- Complement labeled data with unlabeled data (semi-supervised learning)
- Induce SRL representations in an unsupervised fashion (unsupervised learning)

Much less mature area than supervised learning for SRL
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
Exploiting crosslingual correspondences: classes of methods

- The set-up:
  - Annotated resources or a SRL model is available for the source language (often English)
  - No or little annotated data is available for the target language

- How can we build a semantic-role labeller for the target language?
  - If we have parallel data, we can project annotation from the source language to the target language (annotation projection)
    - [Pado and Lapata, 2005; Johansson and Nugues, 2006; Pado and Pitel, 2007; Tonelli and Pianta, 2008; van der Plas et al., 2011]
  - If no parallel data, we can directly apply a source SRL model to the target language (direct model transfer)
    - [Kozhevakov and Titov, 2013]
Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair

Peter knows the situation.
Peter kennt die Situation.

Example from Sebastian Pado
Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair
- Label the source sentence

Example from Sebastian Pado

\[
\text{Awareness} \quad \rightarrow \quad \text{Cognizer} \quad | \quad \text{Content}
\]

\[
Peter \text{ knows } \text{the situation.}
\]

\[
Peter \text{ kennt die Situation.}
\]
Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair
- Label the source sentence
- Check if a target predicate can evoke the same frame

Example from Sebastian Pado
Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair
- Label the source sentence
- Check if a target predicate can evoke the same frame
- Project roles from source to target sentence

Example from Sebastian Pado
Word-based projection

- For each source semantic role:

Example from Sebastian Pado
Word-based projection

- For each source semantic role:
  - Follow alignment links

Example from Sebastian Pado
Word-based projection

- For each source semantic role:
  - Follow alignment links
  - Target role spans all the projected words

Example from Sebastian Pado
Word-based projection

For each source semantic role:

- Follow alignment links
- Target role spans all the projected words
- Ensure contiguity

Example from Sebastian Pado

Noisy because of errors and omission in word alignments
Syntax-based projection

- Find alignment between constituents
- For each source semantic role:
  - Identify a set of constituents in the source sentences
  - Label aligned constituents with the semantic role

[Example from Sebastian Pado]

We have an alignment between words, how do we get one for constituents?
Syntax-based projection

- Define semantic alignment as an optimization task on a graph
- Graph for each sentence pair

Nodes are constituents

Edge weights are similarities between constituents (overlap in aligned words)

[Pado and Lapata, 2006]
Syntax-based projection

- Define semantic alignment as an optimization task on a graph
- Graph for each sentence pair

Choose an optimal alignment graph, maybe with some constraints:
- Covers all target constituents (edge cover)
- Edges in the alignment do not have common endpoints (matching)

Nodes are constituents

Edge weights are similarities between constituents (overlap in aligned words)

[Pado and Lapata, 2006]
Evaluation

- English to German, FrameNet-style representations
- Manual syntax (for 2 languages), manual SRL for source, auto alignments

---

The evaluation is limited to sentences where a frame is preserved in translation.

---

The diagram shows the F1 scores for different methods:
- Word-based
- Edge cover
- Matching
- Upper bound

The bar for Upper bound is the tallest, indicating the highest F1 score among the methods compared.
Evaluation

- English to German, FrameNet-style representations
- **Auto** syntax (for 2 languages), **auto** SRL for source, **auto** alignments

The evaluation is limited to sentences where a frame is preserved in translation.

The projected annotation can be used to train a semantic-role labeller for the target language.

For **semantic-role dependency representations** word-based transfer (along with an heuristic for treating prepositions and conjunctions) is more competitive.
Outline

- Crosslingual annotation and model transfer
  - Annotation projection
  - Direct transfer
- Semi-supervised learning
- Unsupervised learning
Direct transfer of models

- Is there a simpler (?) method which does not (directly) require parallel data?

- Direct transfer (DT) of models:
  - Train a model in one language
  - Apply to sentences in another language

- Is this realistic at all?
  - Requires (maximally) language-independent feature representation
  - Have been tried successfully for syntax [Zeman and Resnik, 2008; Tackstrom et al., 2012]
  - Performance depends on how different the languages are
Language independent feature representations

- Instead of words use either
  - cross-lingual word clusters [Tackstrom et al., 2012] or
  - cross-lingual distributed word features [Klementiev et al., 2012]

- Instead of fine-grain part-of-speech (PoS) tags use coarse universal PoS tags [Petrov et al., 2012]

- Instead of rich (constituent or dependency) syntax either use either
  - unlabeled dependencies or
  - transfer syntactic annotation from the source language before transferring semantic annotation and use it
Language independent feature representations

- CoNLL-2009 data (dependency representation for semantics)
- Target syntax is obtained using direct transfer
- Only accuracy on labeling arguments (not identification)

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Direct transfer</th>
<th>Annotation projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>English to Chinese</td>
<td>70.1</td>
<td>69.2</td>
</tr>
<tr>
<td>Chinese to English</td>
<td>65.6</td>
<td>61.3</td>
</tr>
<tr>
<td>English to Czech</td>
<td>50.1</td>
<td>46.3</td>
</tr>
<tr>
<td>Czech to English</td>
<td>53.3</td>
<td>54.7</td>
</tr>
<tr>
<td>English to French</td>
<td>65.1</td>
<td>66.1</td>
</tr>
</tbody>
</table>

For the identification task relative performance between the methods is similar.

DT achieves comparable performance to AP and does not (directly) require parallel data.

A SRL model trained on projected sentences (word-based projection on top of dependencies).
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
Semi-supervised learning: classes of methods

- There are three main groups of semi-supervised learning (SSL) methods considered for SRL:
  - **methods creating surrogate supervision**: automatically annotate unlabeled data and treat it as new labeled data (annotation projection / bootstrapping methods)
  - **parameter sharing methods**: use unlabeled data to induce less sparse representations of words (clusters or distributed representations)
  - **semi-unsupervised learning**: adding labeled data (and other forms of supervision) to guide unsupervised models

We will discuss these methods towards the end of the tutorial
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
  - methods creating surrogate supervision
  - parameter sharing methods
- Unsupervised learning
Creating surrogate supervision

1. Choose examples (sentences) to label from an unlabeled dataset
2. Automatically annotate the examples
3. Add them to the labeled training set
4. Train a classifier on the expanded training set
5. Optional: Repeat

Basic self-training
   - Use the classifier itself to label examples (and, often, its confidence to choose examples at stage 1)
   - Does not produce noticeable improvement for SRL [He and Gildea, 2006]
Monolingual projection: an idea

- Assumptions: sentences similar in their lexical material and syntactic structure are likely to share the same frame-semantic structure

- An example:
  - Labeled sentence: \([\text{His back}]_{\text{Impactor}} [\text{thudded}]_{\text{Impact}} [\text{against the wall}]_{\text{Impactee}}\)
  - Unlabeled sentence: *The rest of his body thumped against the front of the cage*

- An Implementation (roughly):
  - Choose labeled examples which are similar to an unlabeled example (compute scored alignments between them, select pairs with high scores)
  - Use alignments to project semantic role information to the unlabeled sentences

How do we compute these alignments?
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate

The rest of his body thumped against the front of the cage

[Monolingual projection: alignment](Furstenau and Lapata, 2009)
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment

(Furstenau and Lapata, 2009)
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment

Word alignments

Syntactic arc alignments

Use a heuristic to select the alignment domain
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment, with \( \text{Score} = \text{Lexical Score} + \text{Syntactic Score} \)

Using integer linear programming

Word alignments

Syntactic arc alignments

Use a heuristic to select the alignment domain

[Furstenau and Lapata, 2009]
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment, with \( \text{Score} = \text{Lexical Score} + \text{Syntactic Score} \)
- Project annotation

Using integer linear programming

From the one or few closest labeled sentences

Word alignments

Syntactic arc alignments

Use a heuristic to select the alignment domain

[Furstenau and Lapata, 2009]
Evaluation scenario:

- For a verb, we observe in the labeled training set a few seed examples.
- The seed corpora is expanded by selecting $k$ closest unlabeled examples, projecting annotation to them and adding them to training data.
Evaluation scenario:

- For a verb, we observe in the labeled training set a few seed examples.
- The seed corpora is expanded by selecting $k$ closest unlabeled examples, projecting annotation to them and adding them to training data.

Self-training

Gains more than from manually annotating 1 more labeled example per verb.

[36]

Furstenau and Lapata, 2009
Outline

- Crosslingual annotation and model transfer
  - Semi-supervised learning
    - methods creating surrogate supervision
    - parameter sharing methods
  - Unsupervised learning
Reducing sparsity of word representations

- **Lexical features** are crucial for accurate semantic role labeling
  - However, they are problematic as they are sparse
- **Less sparse features** capturing lexical information are needed

Representations can be learnt from unlabeled data in the context of the language model task, for example:

- Brown clusters [Brown et al., 1992]
- Distributed word representations [Bengio et al., 2003]

and then used as features in SRL systems

Challenge: they might not capture phenomena relevant to SRL or not have needed granularity.
Learning lexical representations

Predict if an ngram belongs to the language

Distributed word representations

Can be trained on large unlabeled texts

[Collobert et al., 2011]
Learning lexical representations

Predict a semantic role for the middle word

Distributed word representations

Can be trained only on semantically annotated texts

[Collobert et al., 2011]
Learning lexical representations

Predict if an ngram belongs to the language

Predict a semantic role for the middle word

Word representations are shared across the tasks

Share words representations across tasks and learn simultaneously to be useful for both tasks

[Collobert et al., 2011]
Evaluation on PropBank

Significant boost from semi-supervised learning

Nearly state-of-the-art performance

Supervised NN
Semi-sup NN
Supervised NN + syntax

[Collobert et al., 2011]
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
  - Unsupervised learning
    - agglomerative clustering
    - generative modeling
Defining Unsupervised SRL

- Semantic role labeling is typically divided into two sub-tasks:
  - Identification: identification of predicate arguments
  - Labeling: assignment of their semantic roles

Arguably, the easier sub-task, can be handled with heuristics, e.g. [Lang and Lapata, 2010]

Goal: induce semantic roles automatically from unannotated texts

- Equivalent to clustering of argument occurrences (or “coloring” them)
Before we begin, a note about evaluating unsupervised SRL

We do not have labels for clusters, so we use standard clustering metrics instead

- **Purity** (PU) measures the degree to which each induced role contains arguments sharing the same gold (“true”) role

  \[
  PU = \frac{1}{N} \sum_{i} \max_{j} |G_{j} \cap C_{i}|
  \]

- **Collocation** (CO) evaluates the degree to which arguments with the same gold roles are assigned to a single induced role

  \[
  CO = \frac{1}{N} \sum_{j} \max_{i} |G_{j} \cap C_{i}|
  \]

- Report F1, harmonic mean of PU and CO
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
  - agglomerative clustering [Lang and Lapata, 2011b]
  - generative modeling [Titov and Klementev 2012]

Earlier methods [Swier and Stevenson, 2004; Grenager and Manning 2006] relied on strong linguistic priors / resources for the language in question
Role Labeling as Clustering of Argument Keys

- Associate argument occurrences with syntactic signatures or argument keys
  - Will include simple syntactic cues such as verb voice and position relative to predicate

  ![Diagram](image)

  - Role 1
    - Mary
    - climbed
    - up
    - Mont Ventoux
  - Role 2

- Argument keys are designed to map to a single semantic role as much as possible (for an individual predicate)

- Instead of clustering argument occurrences, the method clusters their argument keys

- Here, we would cluster **ACTIVE:RIGHT:OBJ** and **ACTIVE:RIGHT:PMOD_up** together

- We assume the automatic syntactic analyses are available

  
  Purity of around 90%

  All occurrences with the same key are automatically in the same cluster

[Lang and Lapata, 2011b]
Role Labeling via "Split-Merge" Clustering

- **Agglomerative clustering of arguments**
  - Start with each argument key in its own cluster (high purity, low collocation)
  - Merge clusters together to improve collocation

- **For a pair of clusters score**
  - whether a pair contains lexically similar arguments
  - whether arguments have similar parts of speech
  - whether the constraint that arguments in a clause should be in different roles is satisfied

  *John taught students math*

- **Prioritization**
  - Instead of greedily choosing the highest scoring pair at each step, start with larger clusters and select best match for each of them

[Lang and Lapata, 2011b]
PropBank (CoNLL 08)

Gold syntax

Predicted syntax

F1 (Clustering)

Syntactic baseline

A graph-based method (Lang and Lapata, 2011a)

[Lang and Lapata, 2011b]
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
  - agglomerative clustering
  - generative modeling
A Bayesian model for role labeling

- Idea: propose a generative model for inducing argument clusters
  - As before, clusters are of argument keys, not argument occurrences

- Learning signals are similar to Lang and Lapata (2011a, 2011b), e.g.
  - Selection preferences
    - i.e. distribution of argument fillers is sparse for every role
  - Duplicate roles are unlikely to occur. E.g. this clustering is a bad idea:
    
    John taught students math

- How can we encode these signals in a generative story?
A Bayesian model for role labeling

for each predicate $p = 1, 2, \ldots$

for each occurrence $l$ of $p$

for every role $r \in B_p$

if $[n \sim \text{Unif}(0, 1)] = 1$:

GenArgument($p, r$)

while $[n \sim \psi_{p,r}] = 1$:

GenArgument($p, r$)

for each predicate $p = 1, 2, \ldots$

$B_p \sim \text{CRP}(\alpha)$

for each role $r \in B_p$:

$k_{p,r} \sim \text{Unif}(1, \ldots, |r|)$

$x_{p,r} \sim \theta_{p,r}$

for each predicate $p = 1, 2, \ldots$

for each role $r \in B_p$:

$\theta_{p,r} \sim \text{DP}(\beta, H^{(A)})$

$\psi_{p,r} \sim \text{Beta}(\eta_0, \eta_1)$

At least one argument

Draw first argument

Continue generation

Draw more arguments

Decide on arg key clustering

Draw argument key

Draw argument filler

[Titov and Klementiev, 2012a]
PropBank (CoNLL 08)

Gold syntax

Predicted syntax

[Titov and Klementiev, 2012a]
The approaches we discussed induce roles for each predicate independently. These clusterings define permissible *alternations*. But many alternations are shared across verbs. Can we share this information across verbs?

```
John gave the book to Mary vs John gave Mary the book
Mike threw the ball to me vs Mike threw me the ball
```

Dative alternation

or changes in the syntactic realizations of the argument structure of the verb

[Titov and Klementiev, 2012a]
A Bayesian model for role labeling

- **Idea:** keep track of how likely a pair of argument keys should be clustered
  - Define a similarity matrix (or similarity graph)

```plaintext
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT:RIGHT:OBJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT:LEFT:SBJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASS:RIGHT:LGS-by</td>
<td></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>PASS:LEFT:SBJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Similarity score between `PASS:LEFT:SBJ` and `ACT:RIGHT:OBJ`
A Bayesian model for role labeling

[Titov and Klementiev, 2012a]
A Bayesian model for role labeling

[Titov and Klementiev, 2012a]
A Bayesian model for role labeling

[Titov and Klementiev, 2012a]
A formal way to encode this: dd-CRP

- Can use CRP to define a prior on the partition of argument keys:
  - The first customer (argument key) sits the first table (role)
  - m-th customer sits at a table according to:
    \[
    p(\text{previously occupied table } k | F_{m-1}, \alpha) \propto n_k
    \]
    \[
    p(\text{next unoccupied table } | F_{m-1}, \alpha) \propto \alpha
    \]

- An extension is distance-dependent CRP (dd-CRP):
  - m-th customer chooses a customer to sit with according to:
    \[
    p(\text{different customer } j | D, \alpha) \propto d_{m,j}
    \]
    \[
    p(\text{itself } | D, \alpha) \propto \alpha
    \]
PropBank (CoNLL 08)

Gold syntax

Predicted syntax

[Titov and Klementiev, 2012a]
Qualitative

Looking into induced graph encoding ‘priors’ over clustering arguments keys, the most highly ranked pairs encode (or partially encode)

- Passivization
- Near-equivalence of subordinating conjunctions and prepositions
  - E.g., whether and if
- Benefactive alternation
  - Martha carved a doll for the baby
  - Martha carved the baby a doll
- Dativization
  - I gave the book to Mary
  - I gave Mary the book
- Recovery of unnecessary splits introduced by argument keys

Encoded as (ACTIVE:RIGHT:OBJ_if, ACTIVE:RIGHT:OBJ_whether)
A Bayesian model for role labeling

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

supervised data

- **open**
- **overtake**

[Titov and Klementiev, 2012a]
A Bayesian model for role labeling

[Titov and Klementiev, 2012b]
PropBank (CoNLL 09)

![Graph showing Clustering F1 (all roles) vs Number of Annotated Sentences]

Bayes (Semisupervised)
Bayes
Supervised
SyntF

[Titov and Klementiev, 2012b]
PropBank (CoNLL 09)

[Titov and Klementiev, 2012b]
Generalization of the role induction model

- The model can be generalized for joint induction of predicate-argument structure of an entire sentence
  - start with a (transformed) syntactic dependency graph (~ argument identification)
Generalization of the role induction model

- The model can be generalized for joint induction of predicate-argument structure of an entire sentence
  - start with a (transformed) syntactic dependency graph (~ argument identification)
  - predict decomposition and labeling of its parts
    - label on nodes are frames (or semantic classes of arguments)
    - labels on edges are roles (frame elements)

[Titov and Klementiev, 2011]
Conclusions

- We looked in examples of key directions in exploiting unlabeled data and cross-lingual correspondences
  - a lot of relevant recent work has not been covered
- Still a new direction with a lot of ongoing work
  - research in the related area of information extraction should also closely watched

- Many thanks to Alex Klementiev, Hagen Furstenau, Sebastian Pado for their help
References


H. Fürstenau and M. Lapata. Semisupervised semantic role labeling. EACL.


M. Kozhevnikov and I. Titov. 2013. Crosslingual Transfer of Semantic Role Models. ACL.


J. Lang and M. Lapata. 2011a. Unsupervised semantic role induction with graph partitioning. EMNLP.


S. Pado and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. TALN.
S. Pado and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. TALN.


S. Petrov, D. Das, and R. McDonald. 2012. A universal part-of-speech tagset. LREC.

L. van de Plas, P. Merlo, and J. Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. ACL.


R. Swier and S. Stevenson. Unsupervised Semantic Role Labeling. EMNLP.

O. Tackstrom, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. NAACL.


I. Titov and A. Klementiev. 2012b. Crosslingual Induction of Semantic Roles, ACL.
