Unsupervised Induction of Semantic Roles within a reconstruction-error minimization framework

Ivan Titov and Ehsan Khoddam

University of Amsterdam
From Syntax to Semantics

- Emergence of robust syntactic parsers [Collins 1999, Charniak 2001, Petrov and Klein 2006, McDonald 2005, Titov and Henderson 2007] for many languages has been one of the key successes of statistical NLP in recent years.

- However, syntactic analyses are a long way from representing the meaning of sentences.

Specifically, they do not define **Who** did What to **Whom** (and **How**, **Where**, **When**, **Why**, …)

- In other words, they do not specify the underlying **predicate argument structure**.
Semantic Role Labeling (SRL)

- Identification of arguments and **their semantic roles**
- Example: predicate *open*

  Jack opened **the lock** with a paper clip

**Semantic Roles (PropBank-style):**

- **PROTO-AGENT (A0)** – an initiator/doer in the event [Who?]
- **PROTO-PATIENT (A1)** - an affected entity [to Whom / to What?]
- **INSTRUMENT (A3)** – the entity manipulated to accomplish the goal
Though syntactic and lexical representations are often predictive of the predicate argument structure, this relation is far from trivial, consider *alternations*:

1. John broke the window
2. The window broke
3. The window was broken by John

**Semantic Roles:**

- **AGENT** – an initiator/doer in the event [Who?]
- **PATIENT** – an affected entity [to Whom / to What?]
Approaches to SRL

- **Supervised learning approaches** (e.g., [Gildea and Jurafsky, 02; Johansson, 08])
  - Rely on large expert-annotated datasets (e.g., PropBank ~40k sentences)
  - Even then they provide very low coverage and are domain dependent
  - Annotated data is not available for many languages

- **Semi-supervised methods** – combine labeled and unlabeled data
  - Have relatively limited success so far (e.g., Furstenau and Lapata [09]; Deschacht and Moens [09])

- **Unsupervised methods**
  - Grenager and Manning ['06], Lang and Lapata ['10, '11], Titov and Klementiev ['12]

Main contributions:
- an efficient method achieving state-of-the-art results on En and De without language-specific feature engineering
- a framework for inducing semantics building on ideas from statistical relational learning (tensor factorization) and supervised SRL
Outline

- **Motivation**: why we need unsupervised feature-rich models
- **Framework**: role-induction with reconstruction-error minimization
- **Empirical evaluation**: experiments and discussion
Unsupervised role induction

- State-of-the art models rely on (agglomerative) clustering [Lang and Lapata '10, '11,...] or generative modeling [Titov and Klementiev '12,...]

- These models rely on very restricted sets of features
  - not very effective in the semi-supervised set-up, and not very appropriate for languages with freer order than English

- ... over-rely on syntax
  - e.g., they do not generalize nominal predicates

- ... use language-specific priors
  - a substantial drop in performance if no adaptation

How can we induce frames in a less restrictive feature-rich framework and tackle other challenges along the way?
Feature-rich models of semantic roles

Consider a predicate realization

The police charged the demonstrators with their batons

\[ a = (a_1, \ldots, a_n) \] - arguments (police, the demonstrators, their batons)

\[ v \] - predicate (charge)

\[ r = (r_1, \ldots, r_n) \] - roles (Perpetrator, Victim, Instrument)
Feature-rich models of semantic roles

Consider a predicate realization

\[ \mathbf{a} = (a_1, \ldots, a_n) \] - arguments (police, the demonstrators, their batons)

\[ \nu \] - predicate (charge)

\[ \mathbf{r} = (r_1, \ldots, r_n) \] - roles (Perpetrator, Victim, Instrument)

How can we define a feature-rich model for unsupervised induction of roles?
The police charged the demonstrators with their batons

Consider a predicate realization

\[
\text{Hypothesis: semantic roles are the latent representation which helps to reconstruct arguments}
\]
Argument reconstruction

Consider a predicate realization

The police charged the demonstrators with their batons

Argument prediction ( = Reconstruction)

"Argument prediction" model

$ p(a_i | a_{-i}, r, v, \theta) $

Charge(Agent: police, Patient: demonstrator, Instrument: baton)

Semantic role labeling ( = Encoding)

Feature-rich model

$ p(r | x, w) $

Feature representation of "The police charged... " (x)

Any existing supervised role labeler would do

How do the components look like and how do we estimate them jointly?
Outline

- Motivation: why we need unsupervised feature-rich models
- **Framework:** role-induction with reconstruction-error minimization
- **Empirical evaluation:** experiments and discussion
Reconstruction-error minimization

**Neural autoencoders** [Hinton '99, Vincent et al. 08]:

- Input \( x \in \mathbb{R}^m \)
- Latent representation \( y \in \mathbb{R}^p \)
- Encoding
- Reconstruction
- Reconstructed input \( \tilde{x} \in \mathbb{R}^m \)

Trained to minimize the reconstruction error, for example, e.g., \( ||x - \tilde{x}||_2 \)

but

- ... applicable not only to neural models
- ... reconstruction and encoding components can belong to different model families
- ... no need to reconstruct the entire input

See also Ammar et al. [NIPS '14] and also Daumé [ICML '09]
Argument reconstruction

Consider a predicate realization

Similar to inferring missing entries in databases (*statistical relational learning*)

**Tensor factorization**

- **Argument prediction** (= Reconstruction)
  - "Argument prediction" model
    - \( p(a_i|a_{-i}, r, v, \theta) \)

- **Semantic role labeling** (= Encoding)
  - **Feature-rich model**
    - \( p(r|x, w) \)

**A (structured) linear model**

- **Feature representation of "The police charged... "** (\( x \))

![Diagram showing the predicate realization process](image-url)
Component 1: argument reconstruction

Distributed vectors:
- \( \mathbf{u}_a \in \mathbb{R}^d \) encode semantic properties of argument \( a \)
- \( C_{v,r} \mathbf{u}_a \in \mathbb{R}^k \) encode expectations about other argument given that \( a \) is assigned to role \( r \) of predicate \( v \)

A role-specific projection matrix

The reconstruction model:

\[
p(a_i | \mathbf{a}_{-i}, r, v) = \frac{\exp(\mathbf{u}^T_{a_i} C^T_{v,r_i} \sum_{j \neq i} C_{v,r_j} \mathbf{u}_{a_j})}{Z(r, v, i)}
\]
Component 1: argument reconstruction

Distributed vectors:

\[ u_a \in \mathbb{R}^d \]
- encode semantic properties of argument \( a \)

\[ C_{v,r} u_a \in \mathbb{R}^k \]
- encode expectations about other argument given that \( a \) is assigned to role \( r \) of predicate \( v \)

Intuitively, score argument tuples according to the factorization:

\[ \sum_{i \neq j} u_{a_i}^T C_{v,r_i}^T C_{v,r_j} u_{a_j} \]

Parallels to work on relation modeling (e.g., Bordes et al.,'11), distributional semantics (e.g., Mikolov et al.,'13) or (coupled) tensor factorization (e.g., Yilmiz et al.,'11)

May encode that demonstrators are similar to protestors

If Agent of Assault is the police, then Patient can be demonstrators or protestors
Component 2: role prediction

The role labeling model: \( p(r|x, w) \propto \exp(w^T g(x, v, r)) \)

It can be any model as long as role posteriors \( p(r_i|x, w) \) can be computed (or approximated)

The majority of supervised SRL models; we used (Johansson and Nugues, '08)

A feature-rich representation encoding syntax-semantics interface
Joint learning

For every structure, we aim to optimize the expectation of the argument prediction quality given roles and frames:

$$\sum_{1}^{N} \log \sum_{r} p(a_{i} | a_{-i}, r, v, C, u) p(r | x, w)$$

Not very tractable in this exact form, usual 'tricks' are needed:
- 'mean-field': substituting posterior means instead of marginalization
- negative sampling (as, e.g., in Mikolov et al '13) instead of 'softmax'

Training can be quite efficient as all models are linear (or bilinear)
Outline

- Motivation: why we need unsupervised feature-rich models
- Framework: role-induction with reconstruction-error minimization
- Empirical evaluation: experiments and discussion
Experiments

- Only role labeling: rely on gold standard argument identification
- Evaluate on a dataset annotated with roles (PropBank for En, SALSA for De)
- Compare against previous models evaluated in this set-up
  - use clustering evaluation measures (purity, collocation, F1)
- Rich and language-agnostic feature set (around 50,000 features for English) [Johansson and Nugues, 2008]

We replicate previous evaluation: datasets are fairly small (e.g., ~90,000 predicate-argument structures for English)

May not be the optimal set-up for our expressive model
Previous approaches evaluated in the same setting

Optimal deterministic mapping from syntactic relations

Performs on par with best methods (without language-specific priors)

Induces fewer roles than most other approaches but under certain regimes, roles start to capture verb senses

The feature-rich model

Logistic: Lang and Lapata (’10)
GraphP: Lang and Lapata (’11a)
Linking: Fürstenau and Rambow (’12)
Aggl: Lang and Lapata (’11b)
Order: Garg and Henderson (’12)
Aggl+: Lang and Lapata (’14)
Bayes: Titov and Klementiev (’12)
German (F1)

Bayes: Titov and Klementiev ('12a)
Bayes (De): Titov and Klementiev (12b)

Perform on par with the best method without language-specific engineering

Optimal deterministic mapping from syntactic relations

Bayes modified for German

The feature rich model
Conclusions

- A new framework for inducing roles
  - Allows us to combine ideas from relational modeling and supervised role labeling
  - more data, more languages, other factorizations, …

- The framework naturally supports:
  - Semi-supervised learning

In principle, the reconstruction objective can be easily extended with the conditional likelihood objective on labeled data

- Learning for inference

Link argument across texts and define reconstruction as inferring new facts
Thank you!

- Special thanks to Diego Marcheggiani, Dipanjan Das, Mike Kozhevnikov, Alexis Palmer, Manfred Pinkal for suggestions …

- Research is supported by Google Focused Award 2013