Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning

Guillaume Wisniewski       Nicolas Pécheux
Souhir Gahbiche-Braham     François Yvon

Université Paris-Sud & LIMSI-CNRS

October 28, 2014
Context

- Supervised Machine Learning techniques have established new performance standards for many NLP tasks
- Success crucially depends on the availability of annotated in-domain data
- Not so common situation (e.g. under-resourced languages)

- What can we do then?
Unsupervised learning

Crawl data (e.g. Wiktionary)
Cross-lingual transfer (weakly supervised learning)
Context

- Cross-lingual transfer (weakly supervised learning)

Example

Making a Market for Scientific Research

Un marché pour la recherche scientifique
In most cases this only results in partially annotated data. Alternative ML techniques need to be designed.

- Partially observed CRF [Täckström et al., 2013]
- Posterior regularization [Ganchev and Das, 2013]
- Expectation maximization [Wang and Manning, 2014]
Contributions

1. We cast this problem in the framework of ambiguous learning [Bordes et al., 2010, Cour et al., 2011]
2. We present a novel method to learn from ambiguous supervision data
3. We show significant improvements over prior state of the art
4. We conduct a detailed analysis that allows us to identify the limits of transfer-based methods and their evaluation
Part I

Projecting Labels across Aligned Corpora
Hypothesis

- In this work we focus on **POS tagging**

**Strong assumption**

Syntactic categories in the source language can be directly related to the ones in the target one

**Universal tagset [Petrov et al., 2012]**

\{ NOUN, VERB, ADJ, ADV, PRON, DET, ADP, NUM, CONJ, PRT, ‘.’, X \}

- All annotations are mapped to this universal tagset
Transfer-based methods only deliver \textit{partial} and \textit{noisy} supervision

\begin{itemize}
  \item Heuristic filtering rules [Yarowsky et al., 2001]
  \item Graph-base projection [Das and Petrov, 2011]
  \item Combine with monolingual information [Täckström et al., 2013]
\end{itemize}

**Type and token constraints [Täckström et al., 2013]**

1. \textit{type} constraints from a dictionary
2. \textit{token} constraints projected through alignment links
Type constraints

From tag dictionaries

- Automatically extracted from Wiktionary
Type constraints

From tag dictionaries

- Automatically extracted from Wiktionary

- Build from the projected labels across the aligned corpora

```
market  walked  ...  ⇒  market
...  marché  marché  ...
```

```
market  marché
...  marché
```
Type constraints

From tag dictionaries

- Automatically extracted from Wiktionary
- Build from the projected labels across the aligned corpora

```
market  \(\rightarrow\) marché  \(\rightarrow\) marché
walked \(\rightarrow\) marché
```

We use the intersection of the two above
1. Use the type constraints
Token constraints

2. Use the alignment links from the parallel corpora

Making a Market for Scientific Research

Un marché pour la recherche scientifique
3. Tag the source side (resource-rich)

Making a Market for Scientific Research

Un marché pour la recherche scientifique
4. Project labels if licensed by type constraints

Making a Market for Scientific Research

Un marché pour la recherche scientifique
Part II

Modeling Sequences under Ambiguous Supervision
Problem

Un marché pour la recherche scientifique

- Gold labels: a set of possible labels of which only one is true
- How to learn from ambiguous supervision?
- Can be cast in the framework of ambiguous learning [Bordes et al., 2010, Cour et al., 2011]
History-based model: inference

\[ x: \text{Un marché pour la \ldots} \]

\[ y: \text{DET NOUN ADP ?} \]

\[ y_i^* = \]

**Principle**

- Structured prediction is reduced to a sequence of multi-classification problems
History-based model: inference

\[ y_i^* = \arg \max_{y \in \{\text{NOUN, VERB, ...}\}} F(x, y, y_{i-1}, y_{i-2}, ...) \]

Principle

- Structured prediction is reduced to a sequence of multi-classification problems
- At each step, the decision is taken based on the input structure and the so far partially tagged sequence
History-based model: training

- Linear classifier \( y^*_i = \arg \max_{y \in \mathcal{Y}} \mathbf{w}^T \phi(x, i, y, h_i) \)
- Perceptron update

**Full supervision**

If \( y^*_i \neq \hat{y}_i \) then

\[
\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \phi(x, i, y^*_i, h_i) + \phi(x, i, \hat{y}_i, h_i)
\]

- Heighten the gold label score at the cost of the wrongly predicted one
History-based model: training

- Linear classifier $y_i^* = \arg\max_{y \in \mathcal{Y}} w^T \phi(x, i, y, h_i)$
- Perceptron-like update

Ambiguous supervision

if $y_i^* \not\in \hat{\mathcal{Y}}_i$ then

$w_{t+1} \leftarrow w_t - \phi(x, i, y_i^*, h_i) + \sum_{\hat{y}_i \in \hat{\mathcal{Y}}_i} \phi(x, i, \hat{y}_i, h_i)$

- Heighten the gold labels score at the cost of the wrongly predicted one
History-based model: training

- Linear classifier $y_i^* = \arg \max_{y \in \mathcal{Y}} w^T \phi(x, i, y, h_i)$
- Perceptron-like update

Ambiguous supervision

\[ \text{if } y_i^* \notin \hat{\mathcal{Y}}_i \text{ then} \]

\[ w_{t+1} \leftarrow w_t - \phi(x, i, y_i^*, h_i) + \sum_{\hat{y}_i \in \hat{\mathcal{Y}}_i} \phi(x, i, \hat{y}_i, h_i) \]

- Heighten the gold labels score at the cost of the wrongly predicted one
- Theoretical guarantees for similar problems under mild assumptions [Bordes et al., 2010, Cour et al., 2011]
Part III

Experiments
Experimental setup

- Experiments on 10 languages from different families
- English as the source side

Our method needs

- Parallel corpora: Europarl, NIST, Open Subtitle
- English POS tagger: Wapiti
- Crawled dictionary: Wiktionary
- Labeled test data: CoNLL’07, UDT v2.0, Treebanks
- Standard feature set
## Results

|      | CRF  | HBAL | \(\Delta\) | [1]  | [2]  | [3]  | Unsupervised [1] |
|------|------|------|------------|------|------|------|------------------|
| ar   | 33.9 | 27.9 | -6.0       | 49.9 | —    | —    | —                |
| cs   | 11.6 | 10.4 | -1.2       | 19.3 | 18.9 | —    | —                |
| de   | 12.2 | 8.8  | -3.4       | 9.6  | 9.5  | 14.2 | 18.7             |
| el   | 10.9 | 8.1  | -2.8       | 9.4  | 10.5 | 20.8 | 28.2             |
| es   | 10.7 | 8.2  | -2.5       | 12.8 | 10.9 | 13.6 | 18.7             |
| fi   | **12.9** | 13.3 | **+0.4**  | —    | —    | —    | —                |
| fr   | 11.6 | **10.2** | -1.4  | 12.5 | 11.6 | —    | —                |
| id   | 16.3 | **11.3** | -5.0  | —    | —    | —    | —                |
| it   | 10.4 | **9.1** | -1.3  | 10.1 | 10.2 | 13.5 | 31.9             |
| sv   | 11.6 | **10.1** | -1.5  | 10.8 | 11.1 | 13.9 | 29.9             |

**CRF** Partially supervised CRF baseline
[Täckström et al., 2013]

**HBAL** Our History-based model

[1] [Ganchev and Das, 2013]
[2] [Täckström et al., 2013]
[3] [Li et al., 2012]
Part IV

Discussion
Discussion

Closer look on Spanish results:

| State of the art | 10.9% 😞 |
|------------------|---------|

Our model still falls short of a fully supervised model!
Discussion

Closer look on Spanish results:

|                  |               |
|------------------|---------------|
| State of the art | 10.9%         |
| Our model HBAL   | 8.2%          |

Our method still falls short of a fully supervised model!
## Discussion

### Closer look on Spanish results:

|                          |       |
|--------------------------|-------|
| State of the art         | 10.9% |
| Our model HBAL           | 8.2%  |
| Our model trained on supervised data (HBSL) | **2.4%** |
Discussion

Closer look on Spanish results:

|                           |        |
|---------------------------|--------|
| State of the art          | 10.9%  |
| Our model HBAL            | 8.2%   |
| Our model trained on supervised data (HBSL) | 2.4%   |

Our method still falls short of a fully supervised model!
Why such a large gap?

Noisy constraints

- Type constraints precision on test data is 94%
- I.e. using our type constraints as hard constraints at decoding time yields at least 6% of errors
- In this setting HBSL gets 7.3%
- Noisy dictionaries
Why such a large gap?

Noisy constraints

- Type constraints precision on test data is 94%
- I.e. using our type constraints as hard constraints at decoding time yields at least 6% of errors
- In this setting HBSSL gets 7.3%
- Noisy dictionaries...not only?
Why such a large gap?

Noisy constraints

- Type constraints precision on test data is 94%
- I.e. using our type constraints as hard constraints at decoding time yields at least 6% of errors
- In this setting HBSL gets 7.3%
- Noisy dictionaries...not only?

Out-of-domain evaluation

Train  \neq  Test

1. **tokenization** differs
2. **domain** differs
3. **annotation conventions** differ
Why such a large gap?

Noisy constraints

- Type constraints precision on test data is 94%
- I.e. using our type constraints as hard constraints at decoding time yields at least 6% of errors
- In this setting HBSL gets 7.3%
- Noisy dictionaries...not only?

Out-of-domain evaluation

Train ≠ Test

1. tokenization differs
2. domain differs
3. annotation conventions differ
The annotation convention problem

- Several independently designed information sources are combined
- They follow conflicting annotation conventions

Example
Impact of annotation and train/test mismatches

Fixing some annotation mismatches in type constraints

|       | ar | cs | de | el | es | fi | fr | id | it | sv |
|-------|----|----|----|----|----|----|----|----|----|----|
| HBAL  | 27.9 | 10.4 | 8.8 | 8.1 | 8.2 | 13.3 | 10.2 | 11.3 | 9.1 | 10.1 |
| HBAL + match | 24.1 | 7.6 | 8.0 | 7.3 | 7.4 | 12.2 | 7.4 | 9.8 | 8.3 | 8.8 |
| Δ     | -3.8 | -2.8 | -0.8 | -0.8 | -0.8 | -1.1 | -2.8 | -1.5 | -0.8 | -1.3 |

Supervised experiments for Spanish

| train                  | train labels           | test error rate |
|------------------------|------------------------|-----------------|
| UDT                    | manual                 | 2.4%            |
| Europarl               | HBSL                   | 4.2%            |
| Europarl               | FREE Ling              | 6.1%            |
| Europarl Cross-lingual transfer (ambiguous) | | 8.2% |
Part V

Conclusion
We introduce a new, simple and efficient learning criterion.

Performance surpasses best reported results.

Results close to the best achievable performance?

Evaluation of such settings much be taken with great care.

Additional gains might be more easily obtained by fixing systematic biases than by designing more sophisticated weakly supervised learners.
Thank you for your attention

Questions?

Tools and resources available from http://perso.limsi.fr/wisniews/weakly
References

Bordes, A., Usunier, N., and Weston, J. (2010).
Label ranking under ambiguous supervision for learning semantic correspondences.
In ICML, pages 103–110.

Cour, T., Sapp, B., and Taskar, B. (2011).
Learning from partial labels.
Journal of Machine Learning Research, 12:1501–1536.

Das, D. and Petrov, S. (2011).
Unsupervised part-of-speech tagging with bilingual graph-based projections.
In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT ’11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ganchev, K. and Das, D. (2013).
Cross-lingual discriminative learning of sequence models with posterior regularization.
In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1996–2006, Seattle, Washington, USA. Association for Computational Linguistics.

Li, S., Graça, J. a. V., and Taskar, B. (2012).
Wiki-ly supervised part-of-speech tagging.
In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL ’12, pages 1389–1398, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012).
A universal part-of-speech tagset.
In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey. European Language Resources Association (ELRA).

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013).
Token and type constraints for cross-lingual part-of-speech tagging.