UNIVERSITY OF BIRMINGHAM

Dependency Language Models for Transition-based Dependency Parsing

Juntao Yu (juntao.cn@gmail.com)

Bernd Bohnet

Overview

1. Semi-supervised techniques for parsing
2. Dependency language models
3. Experimental set-up
4. Empirical results
5. Conclusions

Outline of the problem

Expensive and time consuming

Labeled Data

VS

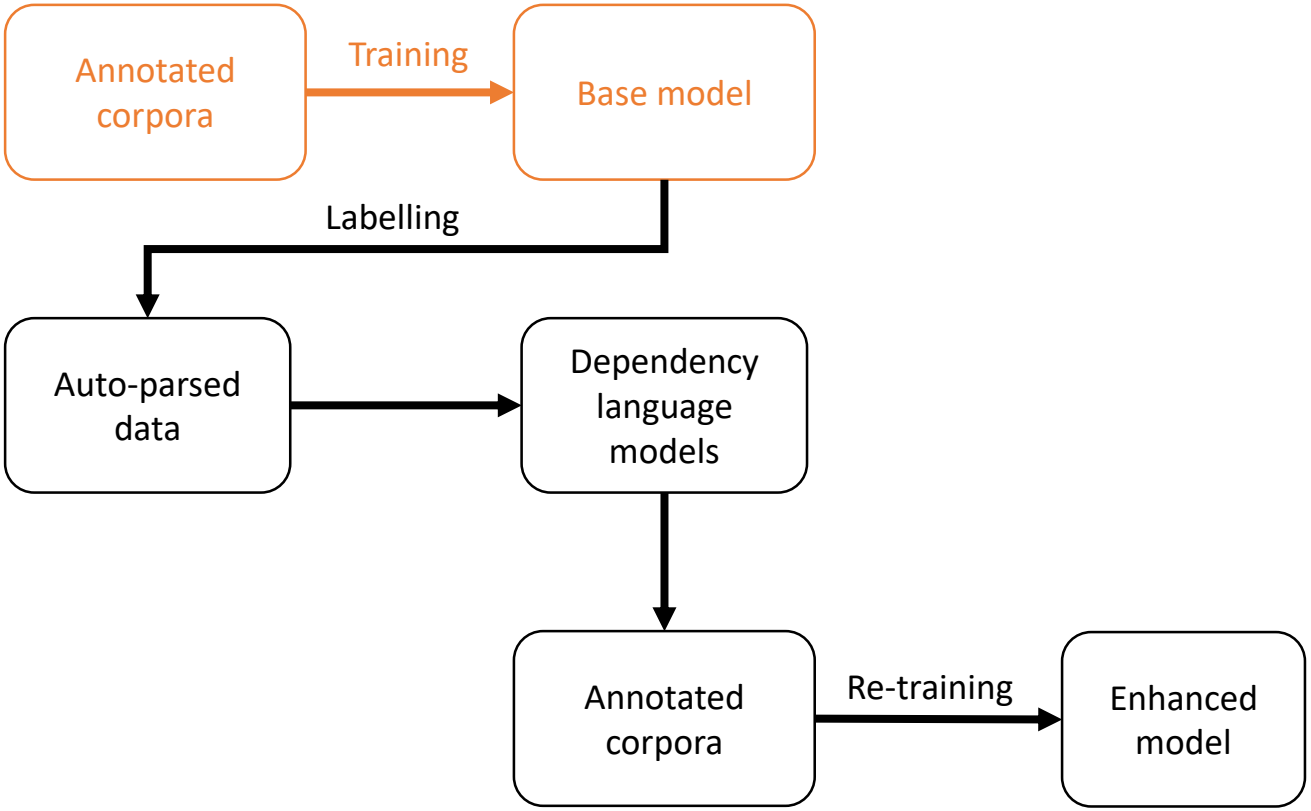
Unlabeled Data

Cheap and widely accessible

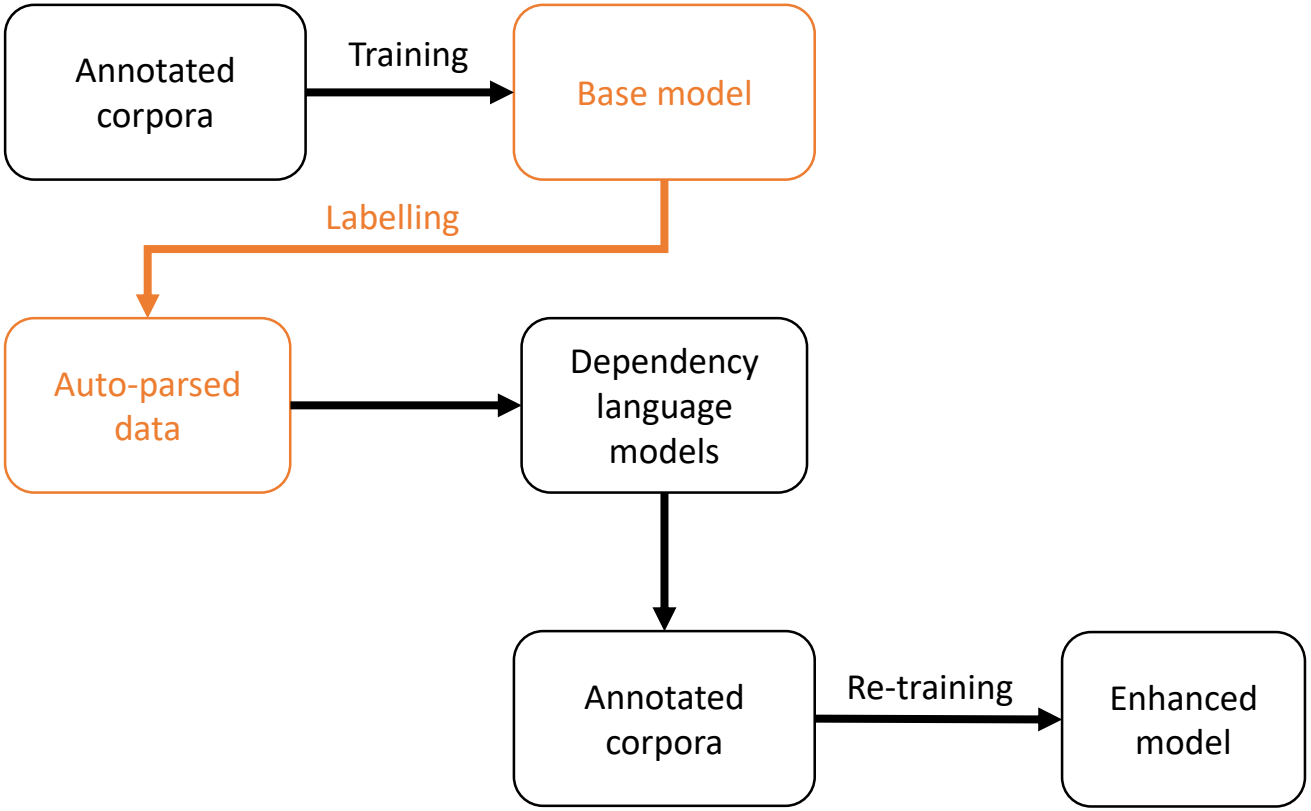
Semi-supervised approaches

- Boosting the training set
  - co-training (Sarkar, 2001; Steedman et al., 2003; Sagea and Tsujii, 2007)
  - self-training (McClosky et al., 2006a; Reichart and Rappoport,2007; Petrov and McDonald, 2012)
- Using features learned from unlabeled data
  - word clustering (Koo et al., 2008; Cerisara, 2014)
  - word embeddings (Chen and Manning, 2014; Weiss et al., 2015; Dozat and Manning, 2017)

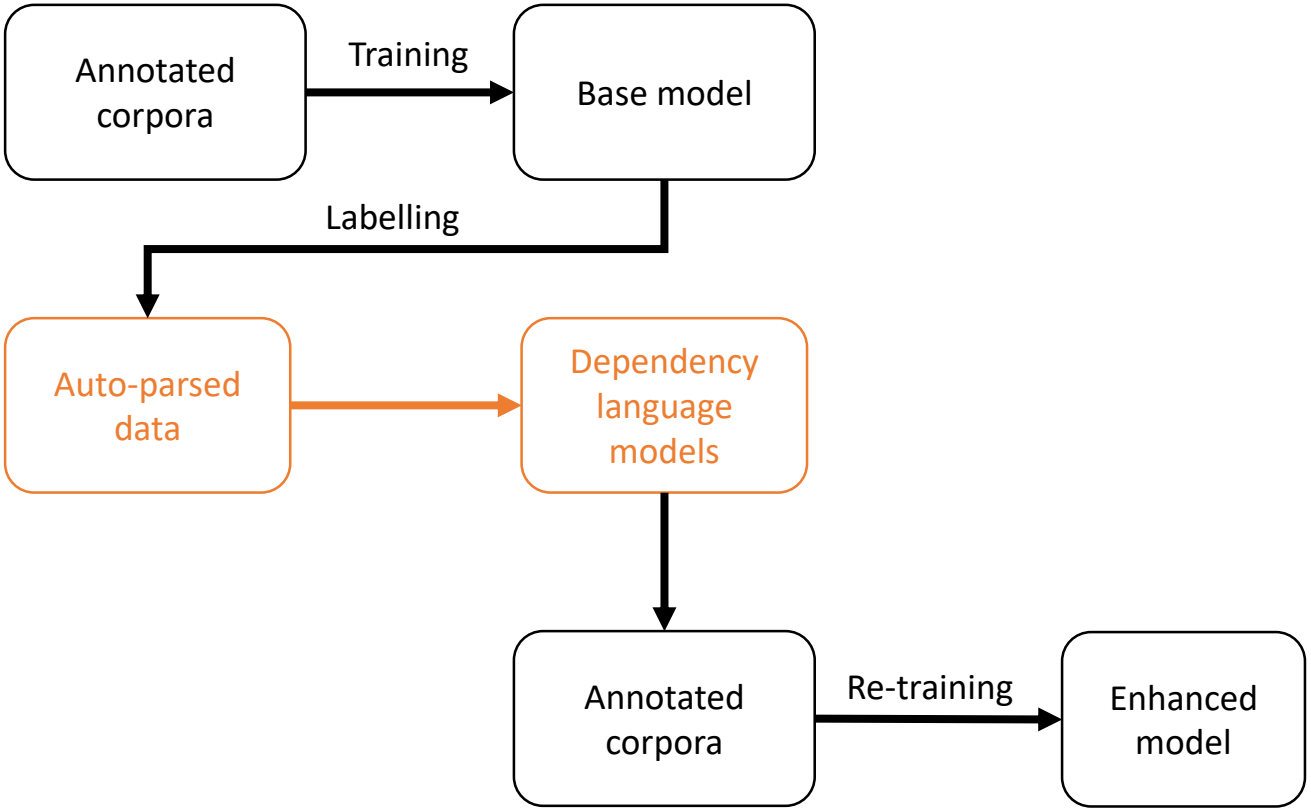
Dependency language models



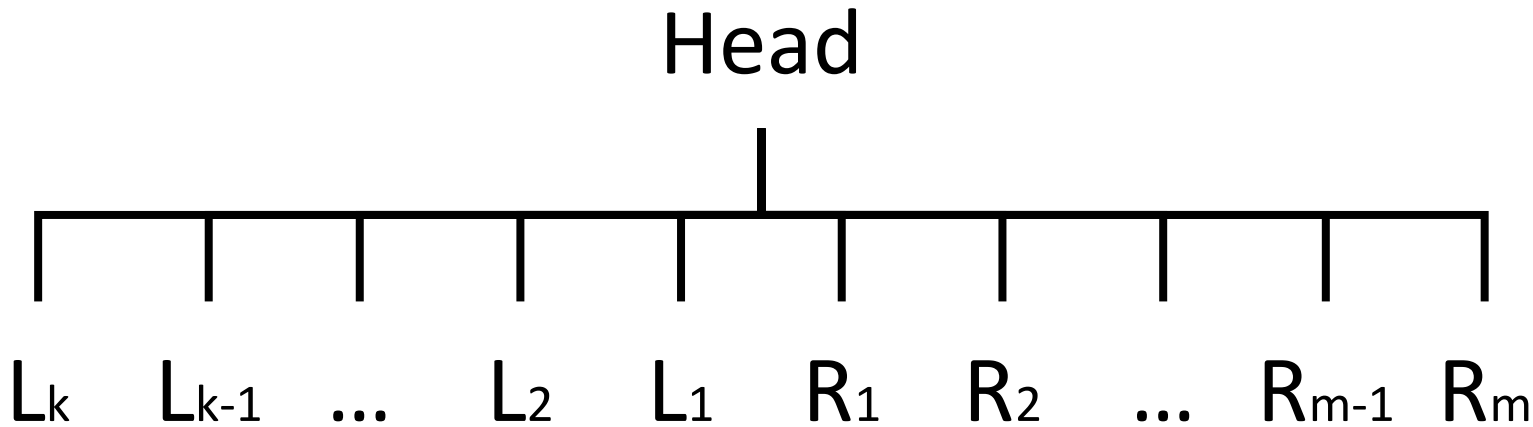
Dependency language models



Dependency language models



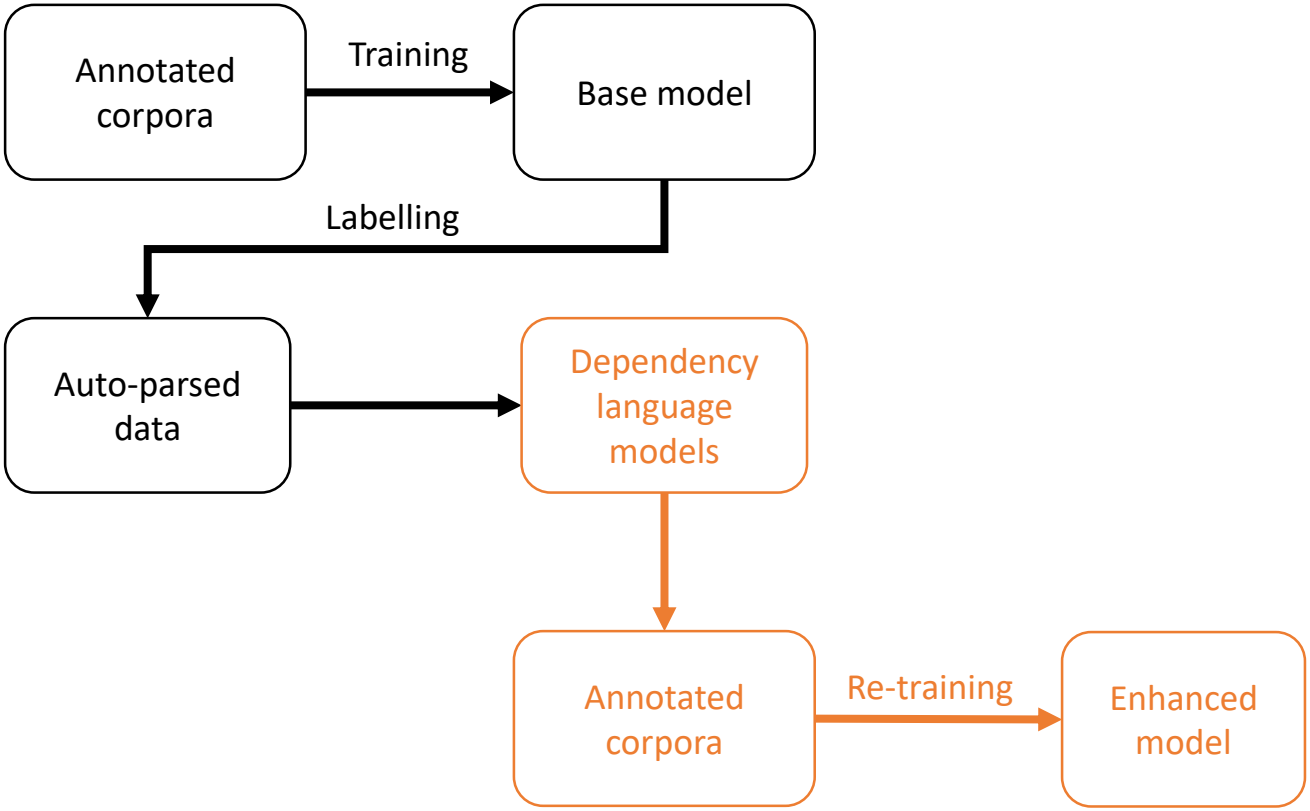
Dependency language models



$$P_u(x_{ch}|H) = \frac{count(x_{ch}, H)}{\sum_{x'_{ch}} count(x'_{ch}, H)}$$

PH 10%
PM 10%-30%
PL 30%-100%
PO Unknown

Dependency language models



Dependency language models

- Differences from Chen et al. (2012)

Graph-based	VS	Transition-based
(92.10 UAS)		(93.38 UAS)

Dependency language models

- Differences from Chen et al. (2012)

Unlabeled parsing VS Labeled parsing

Dependency language models

- Differences from Chen et al. (2012)

Pipeline approach for tagging and parsing

VS

Joint tagging and parsing

Dependency language models

- Differences from Chen et al. (2012)

Single DLM

VS

Single and multiple DLMs

Dependency language models

- Differences from Chen et al. (2012)

Single parsed text

VS

Single and double parsed text

Dependency language models

- Differences from Chen et al. (2012)

$$
\begin{array}{l}
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_word > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_1\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_1\_word > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_pos, s_1\_pos > \\
< NO_{DLM}, \phi(P_u(s_0)), \phi(P_u(s_1)), label, s_0\_word, s_1\_word >
\end{array}
$$

Experimental set-up:

- Mate transition-based parser (Bohnet et al., 2013)
- DataSet (PTB and CTB5):
  - English: Stanford v3.3.0
  - Chinese: Zhang and Nivre (2011)
  - Unlabeled set:
    - English: Chelba et al. (2013) 30M sentences (800M words)
    - Chinese: Gigaword v5.0 20M sentences (450M words)

	train	**dev**	**test**
PTB	2-21	22	23
CTB5	001-815, 1001-1136	886-931, 1148-1151	816-885, 1137-1147

Evaluating on dev set:

m	**0**	**1**	**2**	**3**	**1-2**	**1-3**	**1-4**
English	91.05	**91.43**	91.14	91.22	91.27	91.26	N/A
Chinese	78.95	79.85	79.42	79.06	79.97	**80.11**	79.73

Table 3: Effects (LAS) of different number of DLMs for English and Chinese. $m = 0$ refers the baseline.

Evaluating on dev set:

7M English high quality sentences: 91.56

Size	0	5	10	20	30
English	91.05	**91.43**	91.38	91.13	91.28
Chinese	78.95	80.11	**80.15**	79.72	N/A

Table 4: Effects (LAS) of DLMs extracted from different size (in million sentences) of corpus. Size = 0 refers the baseline.

Evaluating on test set (English):

System	Beam	POS	LAS	UAS
Zhang and Nivre (2011)	32	97.44	90.95	93.00
Bohnet and Kuhn (2012)	80	97.44	91.19	93.27
Martins et al. (2013)	N/A	97.44	90.55	92.89
Zhang and McDonald (2014)	N/A	97.44	91.02	93.22
Chen and Manning (2014)†	1	N/A	89.60	91.80
Dyer et al. (2015)†	1	97.30	90.90	93.10
Weiss et al. (2015)†	8	97.44	92.05	93.99
Andor et al. (2016)†	32	97.44	92.79	94.61
Dozat and Manning (2017)†	N/A	N/A	94.08	95.74
Liu and Zhang (2017)†	N/A	N/A	**95.20**	**96.20**
Our Baseline	40	97.36	90.95	93.08
	80	97.34	91.05	93.28
	150	97.34	91.05	93.29
Our DLM	40	97.38	91.41	93.59
	80	97.39	91.47	93.65
	150	97.42	91.56	93.74

Table 5: Comparing with top performing parsers on English.

Evaluating on test set (Chinese):

System	Beam	POS	LAS	UAS
Hatori et al. (2011)	64	93.94	N/A	81.33
Li et al. (2012)	N/A	94.60	79.01	81.67
Chen et al. (2013)	N/A	N/A	N/A	83.08
Chen et al. (2015)	N/A	93.61	N/A	82.94
Our Baseline	40	93.99	78.49	81.52
	80	94.02	78.48	81.58
	150	93.98	78.96	82.11
Our DLM	40	94.27	79.42	82.51
	80	94.39	79.79	82.79
	150	94.40	**80.21**	**83.28**

Table 6: Comparing with top performing parsers on Chinese.

Conclusions:

- Dependency language models significantly improved both English and Chinese accuracies (both LAS and UAS) for a strong transition-based parser.
  - 0.5% improvements on English.
  - 1% improvements on Chinese and achieved the state-of-the-art results.
- 0.4% improvements on Chinese part-of-speech tagging accuracy and constantly better performance on English part-of-speech tagging.

Questions?