Combining Active Learning and Partial Annotation for Domain Adaptation of a Japanese Dependency Parser

Daniel FLANNERY[1] **Shinsuke MORI**[2]

[1]**Vitei Inc. (work at Kyoto University)**
[2]**Kyoto University**

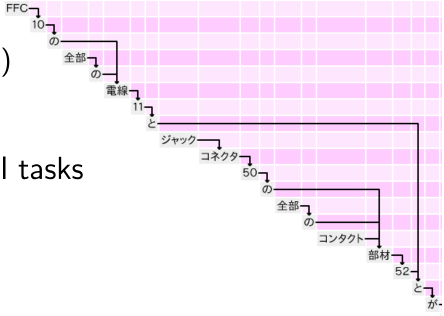
IWPT 2015, July 22nd

IWPT95 at Prague

- My first international presentation!!
  - “Parsing Without Grammar” [Mori 95]
- This is the second!!

Statistical Parsing

- Technology for finding the structure of natural language sentences
- Performed *after* low-level tasks
  - word segmentation (ja, zh, ...)
  - part-of-speech tagging



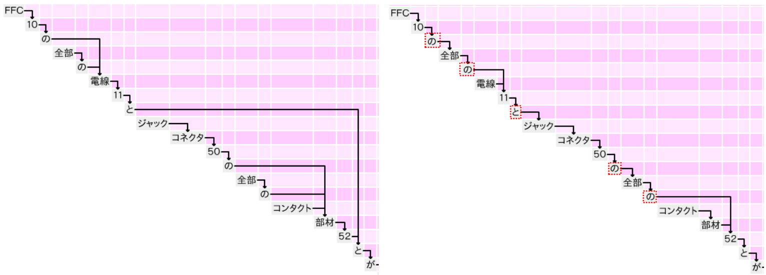
- Parse trees useful for higher-level tasks
  - information extraction
  - machine translation
  - automatic summarization
  - etc.

Portability Problems

- Accuracy drop on a test in a different domain [Petrov 10]
- Need systems for specialized text (patents, medical, etc.)

こう し て プリント 基板 31 は 弾性 部材 32 に 対 し て 位置 決め さ れ る

In this way print plate 31 is positioned against elastic material 32



Parser Overview

- EDA parser: Easily Domain Adaptable Parser [Flannery 12]
  `http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/home-e.html`
  - 1st order Maximum Spanning Tree parsing [McDonald 05]
  - Allows partial annotation: only annotate some words in a sentence
- Use this flexibility for domain adaptation
  - Active learning: Select only informative examples for annotation
  - Goal: Reduce the amount of data needed to train a parser for a new type of text

Pointwise Estimation of Edge Scores

牡蠣 を 広島 に 食べ に 行 く

名詞 助詞 名詞 助詞 動詞 助詞 動詞 語尾

- Choosing a head is an n-class classification problem

$$\sigma(\langle i, d_i \rangle) = p(d_i | \vec{w}, i), \quad (d_i \in [0, n] \wedge d_i \neq i)$$

- Calculate edge scores independently
- Features
  1. Distance between dependent/head
  2. Surface forms/POS of dependent/head
  3. Surface/POS for 3 surrounding words
  4. No surrounding dependencies! (1st order)

Partial and Full Annotation

- Our method can use a partially annotated corpus

牡蠣 を 広島 に 食べ に 行 く

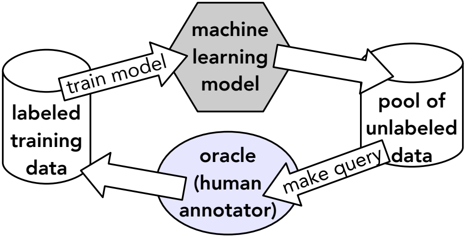
dependent head

  - Only annotate some words with heads
  - Pointwise estimation

- Cf. fully annotated corpus
  - Must annotate all words with heads

牡蠣 を 広島 に 食べ に 行 く

Pool-Based Active Learning [Settles 09]



1. Train classifier **C** from labeled training set $\mathbf{D_L}$
2. Apply **C** to the unlabeled data set $\mathbf{D_U}$ and select **I**, the **n** most informative training examples
3. Ask oracle to label examples in **I**
4. Move training instances in **I** from $\mathbf{D_U}$ to $\mathbf{D_L}$
5. Train a new classifier $\mathbf{C}'$ on $\mathbf{D_L}$
6. Repeat 2 to 5 until stopping condition is fulfilled

Query Strategies

- Criteria used to select training examples to annotate from the pool of unlabeled data
- Should allow for units smaller than full sentences
- Problems
  - Single-word annotations for a sentence are too difficult
  - Realistically, annotators must think about dependencies for some other words in the sentence (not all of them)
- Need to measure actual annotation time to confirm the query strategy's performance!

Tree Entropy [Hwa 04]

- Criterion for selecting sentences to annotate with full parse trees

$$\mathbf{H(V) = -\sum_{v \in V} p(v)\, lg(p(v))}$$

- Models distribution of trees for a sentence
- $\mathbf{V}$ is the set of possible trees, $\mathbf{p(v)}$ is the probability of choosing a particular tree $\mathbf{v}$
- In our case, change the unit from sentences to words and model the distribution of heads for a single word (head entropy)
  - use the edge score $\mathbf{p(d_i|\vec{w}, i)}$ in place of $\mathbf{p(v)}$
- Rank all words in the pool, and annotate those with the highest values (1-Stage Selection)

1-Stage Selection

- Change the selection unit from sentences to words
  - Need to model the distribution of heads for a single word
  - Simple application of tree entropy to the word case
- Instead of probability for an entire tree $\mathbf{p(v)}$, use the edge score $\mathbf{p(d_i|\vec{w}, i)}$ of a word-head pair given by a parsing model
- Rank all words by head entropy, and annotate those with the highest values
- The annotator must consider the overall sentence structure

2-Stage Selection

1. Rank sentences by summed head entropy
2. Rank words in each by head entropy
3. Annotate a fixed fraction
   - partial: annotate top $\mathbf{r = 1/3}$ of words
   - full: annotate all words

Example

- Pool of three sentences

sent.	words			
s1:	A/0.2	B/0.1	C/0.5	D/0.1
s2:	E/0.4	F/0.3	G/0.1	H/0.2
s3:	I/0.4	J/0.2	K/0.3	L/0.2

- 1-stage

C, E, I, F, K, ...

- 2-stage, $\mathbf{r = 1/2}$

sent.	sum	words			
s3:	1.1	I/0.4	J/0.2	K/0.3	L/0.2
s2:	1.0	E/0.4	F/0.3	G/0.1	H/0.2
s1:	0.9	A/0.2	B/0.2	C/0.5	D/0.1

Evaluation Settings

	ID	source	sent.	words /sent.	dep.
	EHJ-train	Dictionary examples	11,700	12.6	136,264
pool	NKN-train	Newspaper articles	9,023	29.2	254,402
	JNL-train	Journal abstracts	322	38.1	11,941
	NPT-train	NTCIR patents	450	40.8	17,928
test	NKN-test	Newspaper articles	1,002	29.0	28,035
	JNL-test	Journal abstracts	32	34.9	1,084
	NPT-test	NTCIR patents	50	45.5	2,225

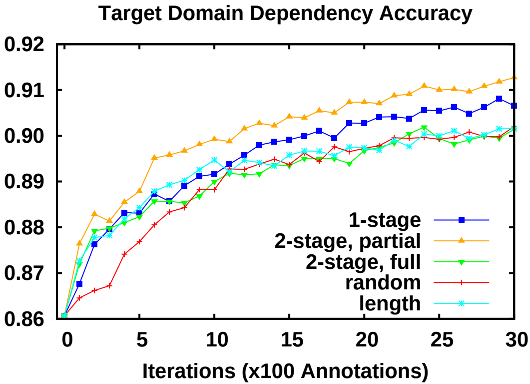
- The initial model: EHJ
- The target domains: NKN, JNL, NPT
  - Manual annotation except for POS by `KyTea`
  - Some are publicly available [Mori 14].

  `http://plata.ar.media.kyoto-u.ac.jp/data/word-dep/home-e.html`

Exp.1: Number of Annotations

- Reduction of the number of in-domain dependencies
- Simulation by selecting the gold standard dependency labels from the annotation pool
- Necessary but not sufficient condition for an effective strategy
- Simple baselines
  - random simply selects words randomly from the pool.
  - length strategy simply chooses words with the longest possible dependency length.
- One iteration:
  1. a batch of one hundred dependency annotations
  2. model retraining
  3. accuracy measurement

EHJ to NKN (Annotations)

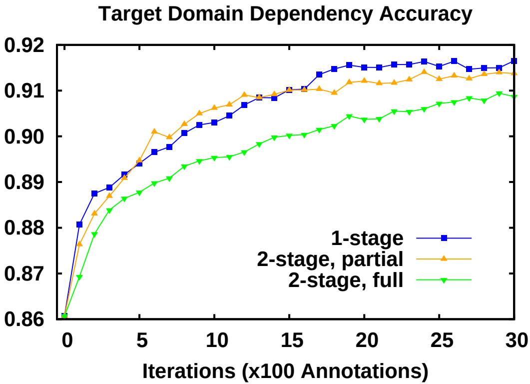


- length and 2-stage-full work good for the first ten iterations but soon begin to falter.
- 2-stage-partial > 1-stage > others

Exp.2: Annotation Pool Size

- NKN annotation pool size $\approx$ **21.3**$\times$ JNL, **14.2**$\times$ NPT
- The total number of dependencies selected is 3k (only 1.2% of NKN-train).
- 2-stage accuracy may suffer when a much larger fraction of the pool is selected.
  - Because the 2-stage strategy chooses some dependencies with lower entropy over competing ones with higher entropy from other sentences in the pool.
- Test a small pool case like JNL or NPT
  - First 12,165 dependencies as the pool

EHJ to NKN with a Small Pool



- After 17 rounds of annotation
  - 1-stage > 2-stage partial > 2-stage full
- The relative performance is influenced by the pool size.
  - 1-stage is robust.
  - 2-stage partial can outperform it for a very large pool.

Exp.3: Time Required for Annotation

- Annotation time for a more realistic evaluation
  - Simulation experiments are still common in active learning
  - Increasing interest in measuring the true costs [Settles 08]
- Settings for annotation time measurement
  - 2-stage strategies
  - Initial model: EHJ-train *plus* NKN-train
  - Target domain: blog in BCCWJ (Balanced Corpus of Contemporary Written Japanese [Maekawa 08])
  - Pool size: 747 sentences
  - One iteration: 2k dependency annotations

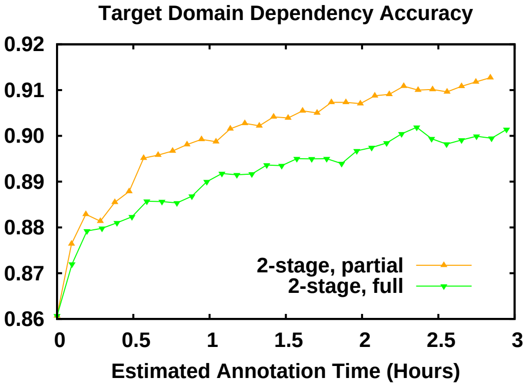
Annotation Time Estimation

- A single annotator, 2-stage partial and full
  - one hour for partial $\Rightarrow$ one hour for full $\Rightarrow$ one hour for partial ...

method	0.25 [h]	0.5 [h]	0.75 [h]	1.0 [h]
partial	226	458	710	1056
full	141	402	756	1018

- After one hour the number of annotations was almost identical
  - For full the annotator was forced to check the annotation standard for subtle linguistic phenomena.
  - partial allows the annotator to delete the estimated heads.
- 1.4k dependencies per hour

EHJ to NKN (Time)



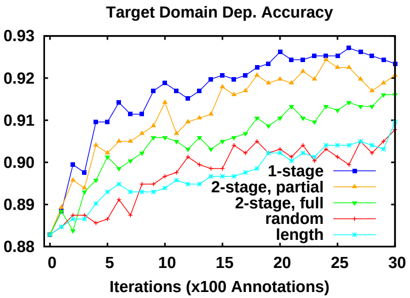
- Applied estimated time by the speeds measured in blog
- 2-stage partial > 2-stage full
- The difference becomes pronounced after 0.5[h].

Results for Additional Domains

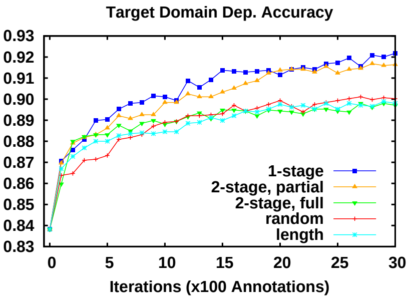
	ID	source	sent.	words /sent.	dep /sent.
	EHJ-train	Dictionary examples	11,700	12.6	136,264
pool	NKN-train	Newspaper articles	9,023	29.2	254,402
	JNL-train	Journal abstracts	322	38.1	11,941
	NPT-train	NTCIR patents	450	40.8	17,928
test	NKN-test	Newspaper articles	1,002	29.0	28,035
	JNL-test	Journal abstracts	32	34.9	1,084
	NPT-test	NTCIR patents	50	45.5	2,225

- Samll pool sizes

To JNL or NPT in (Annotations)



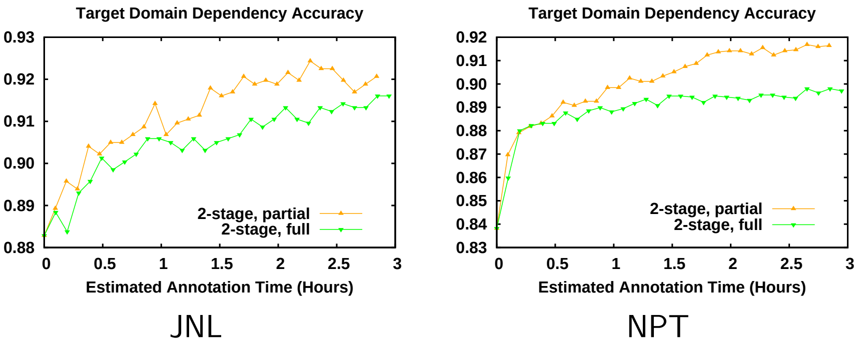
JNL



NPT

- 1-stage > 2-stage partial
  - The pool size is small.
  - 3k dependencies = 25.1% for JNL and 16.7% for NPT
- 2-stage partial > 2-stage full

To JNL or NPT (Time)



- Estimated annotation time
- 2-stage partial > 2-stage full
- The gap is the largest for NPT and the smallest for JNL.

Reduction in In-domain Data

domain	random	full	partial
NKN	3,000	–	1,300
JNL	3,000	1,800	900
NPT	2,700	–	1,500

- random: #annotations needed for the highest accuracy by the random baseline
- full, partial: #annotations needed for the full and partial versions of 2-stage to outperform it
- 2-stage full had mixed results.
- 2-stage partial offers large savings consistently.

Conclusion

- A practical criterion for active learning of a dependency parser
  - Entroy-based
  - Semi-sentence-based
- 2-stage partial: the best when a large size of pool is available
- The corpora and the parser available at `http://plata.ar.media.kyoto-u.ac.jp/home-e.html`
- Future work
  - Combine with a 2nd or 3rd order parser

References

- Flannery, D., Miyao, Y., Neubig, G., and Mori, S.: A Pointwise Approach to Training Dependency Parsers from Partially Annotated Corpora, *Journal of Natural Language Processing*, Vol. 19, No. 3 (2012)
- Hwa, R.: Sample selection for statistical parsing, *Computational Linguistics*, Vol. 30, No. 3, pp. 253–276 (2004)
- Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, in *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102 (2008)
- McDonald, R., Pereira, F., Ribarov, K., and Hajič, J.: Non-projective Dependency Parsing Using Spanning Tree Algorithms, in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 523–530 (2005)

- Mori, S. and Nagao, M.: Parsing Without Grammar, in *Proceedings of the The Forth International Workshop on Parsing Technologies*, pp. 174–185 (1995)
- Mori, S., Ogura, H., and Sasada, T.: A Japanese Word Dependency Corpus, in *Proceedings of the Nineth International Conference on Language Resources and Evaluation*, pp. 753–758 (2014)
- Petrov, S., Chang, P.-C., Ringgaard, M., and Alshawi, H.: Uptraining for Accurate Deterministic Question Parsing, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 705–713 (2010)
- Settles, B., Craven, M., and Friedland, L.: Active Learning with Real Annotation Costs, in *NIPS Workshop on Cost-Sensitive Learning* (2008)

Settles, B.: Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)