Training a Parser for Machine Translation Reordering

Jason Katz-Brown Slav Petrov Ryan McDonald Franz Och

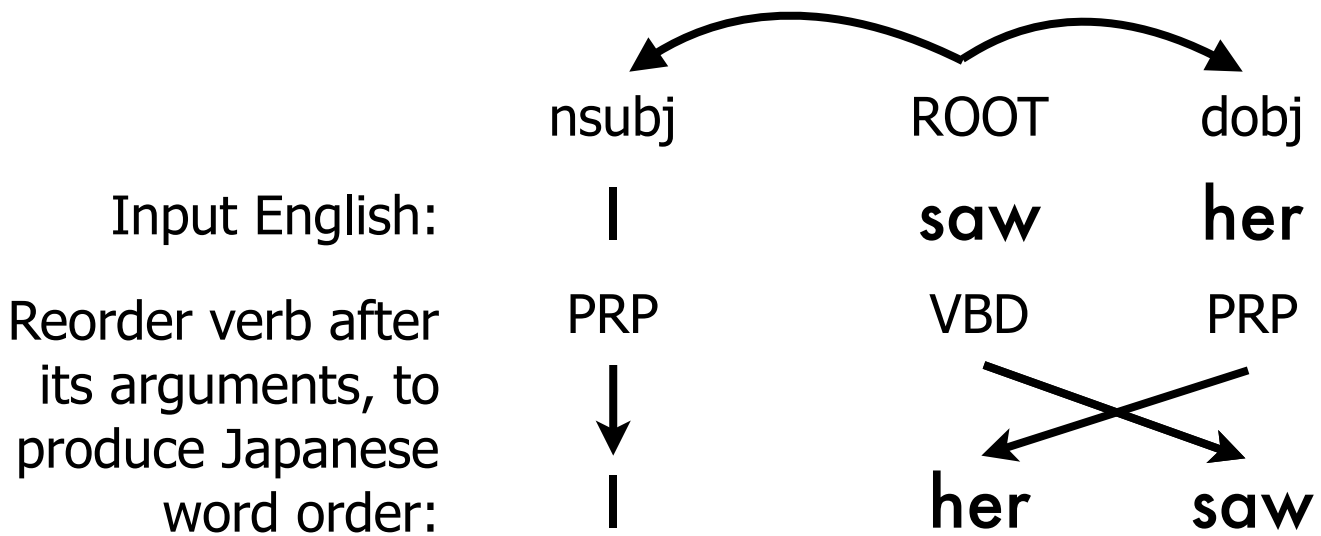
David Talbot Hiroshi Ichikawa Masakazu Seno Hideto Kazawa

Two juicy bits

1. Method for improving performance of any parser on any extrinsic metric
2. Experimental results in training parsers for machine-translation reordering

Our reordering system

- Syntactic reordering during preprocessing at training and decoding time (Collins et al., 2005)
- Translation quality sensitive to reordering quality
- Reordering quality sensitive to parse quality



Evaluating reordering quality

Source	Wear sunscreen that has an SPF of 15 or above
Japanese-y Reference	*15 or above of an SPF has that sunscreen Wear*

References made by bilingual non-linguist annotators
→Cheap
→Fast

Evaluating reordering quality

Source	Wear sunscreen that has an SPF of 15 or above
Japanese-y Reference	*15 or above of an SPF has that sunscreen Wear*
Experiment	*15 or above of an SPF has that* | *Wear* | *sunscreen*

Experiment reordering score: 1 - (3-1)/(11-1) = 0.8

More detail: Talbot et al., EMNLP-WMT 2011

Let’s make a magical treebank!

Let’s make a magical treebank!

- Reordering quality is predictive of parse quality

Let’s make a magical treebank!

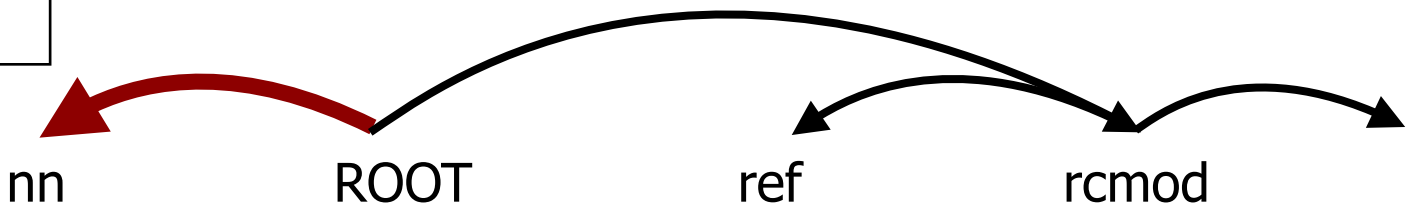
- Reordering quality is predictive of parse quality
- So, what if we make a treebank that contains only parse trees that get a good reordering score?

Let’s make a magical treebank!

- Reordering quality is predictive of parse quality
- So, what if we make a treebank that contains only parse trees that get a good reordering score?
- We can search through a large n-best list of parses from a baseline parser for the one with the best reordering score.

Targeting good reordering

Parse #1:

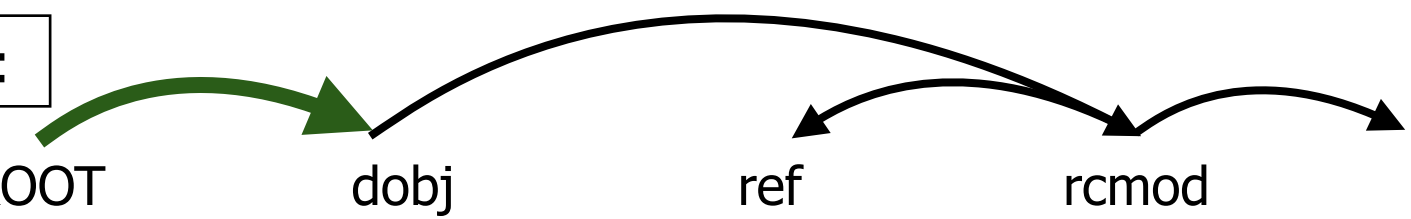


Wear	sunscreen	that	has	…
NN	NN	WDT	VBZ	

Reordered: … *has that Wear sunscreen*

Score: 0.8 ("Wear" is out of place)

Parse #23:



Wear	sunscreen	that	has	…
VB	NN	WDT	VBZ	

Reordered: … *has that sunscreen Wear*

Score: 1.0 (matches reference)

Sweet

- This could work for any extrinsic metric!

Self-Training

- McClosky et al. (2006) demonstrated self-training to improve implicit parse accuracies e.g. labeled attachment score (LAS)
- For every sentence in some new corpus:
    1. Parse sentence with a baseline parser.
    2. Add parse to baseline parser's training data.

Targeted Self-Training

- We **target** the self-training to incorporate an extrinsic metric
- For every sentence in some new corpus:
    1. Parse sentence with a baseline parser **and produce a ranked n-best list**.
    2. **Select parse with highest extrinsic metric**.
    3. Add parse to baseline parser's training data.

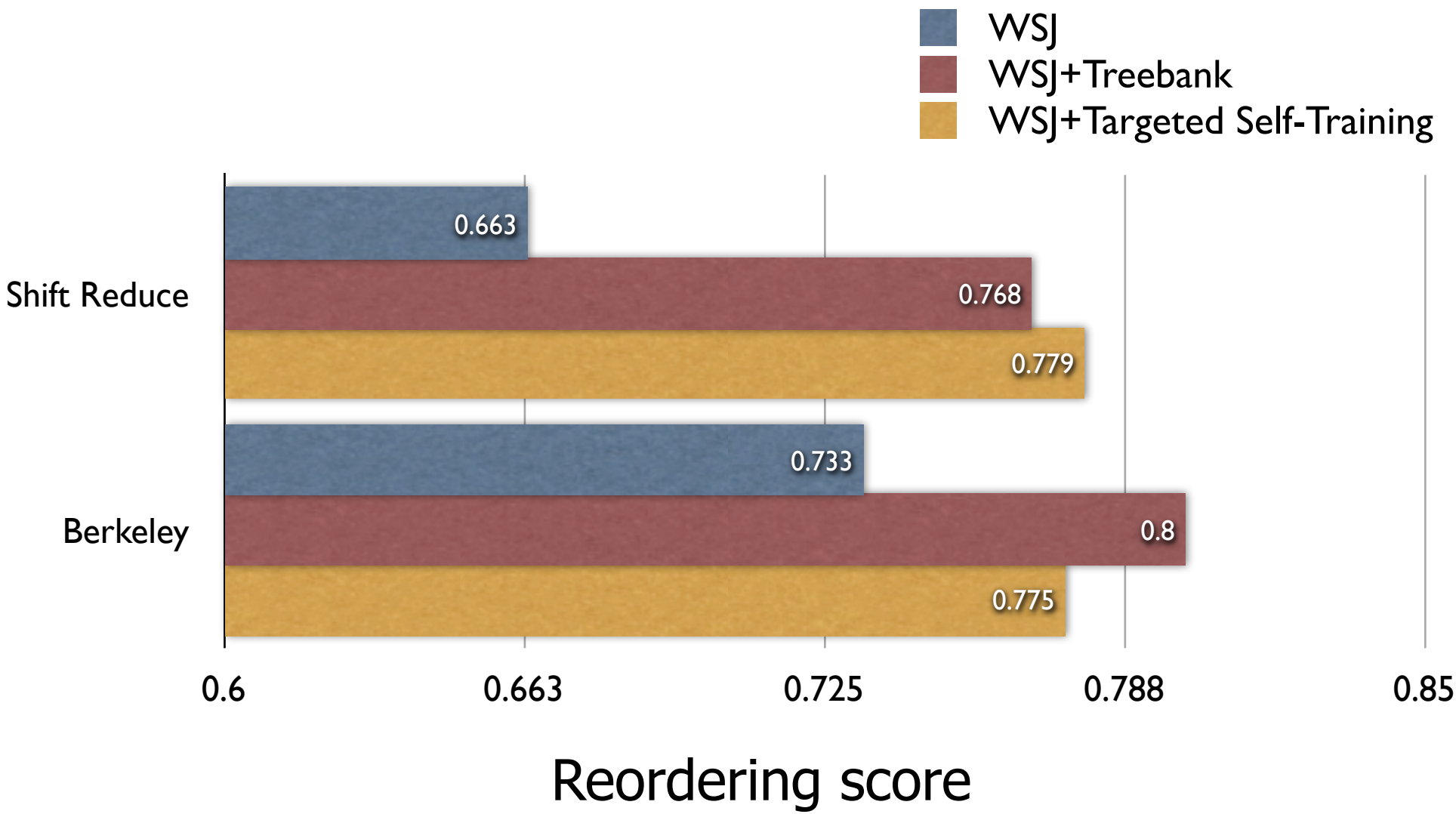
Experimental Setup

- Two parser models
  - Deterministic Shift-Reduce Parser (Nivre et al.) + TnT Tagger (Brants)
  - Latent Variable Parser (BerkeleyParser; Petrov et al. '06)
- English→SOV reordering rules of Xu et al. (2009)
  - Precedence rules for how to reorder the children of a dependency tree node
  - Suitable for English to Japanese, Korean, Turkish, Hindi, ...
- Stanford converter to convert constituency trees to dependency trees (Marnaffe et al., 2006)

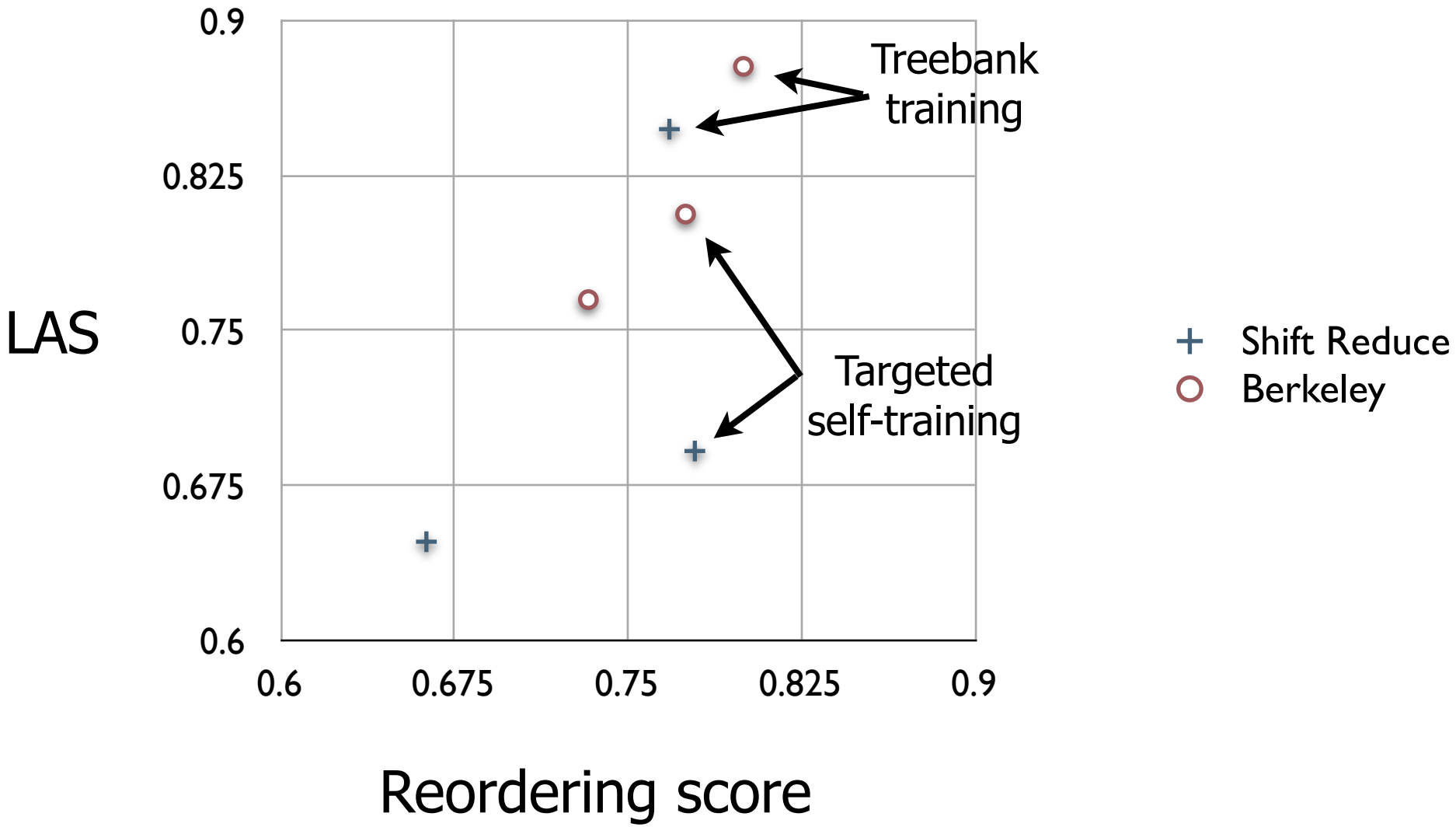
Training on questions

- QuestionBank (Judge et al. '06)
- 1000-sentence train set, 1000-sentence test set
- Compare training on:
  1. WSJ
  2. WSJ + 1000 QuestionBank trees
  3. WSJ + 1000 reference reorderings of same sentences

Question reordering scores



Question attachment scores



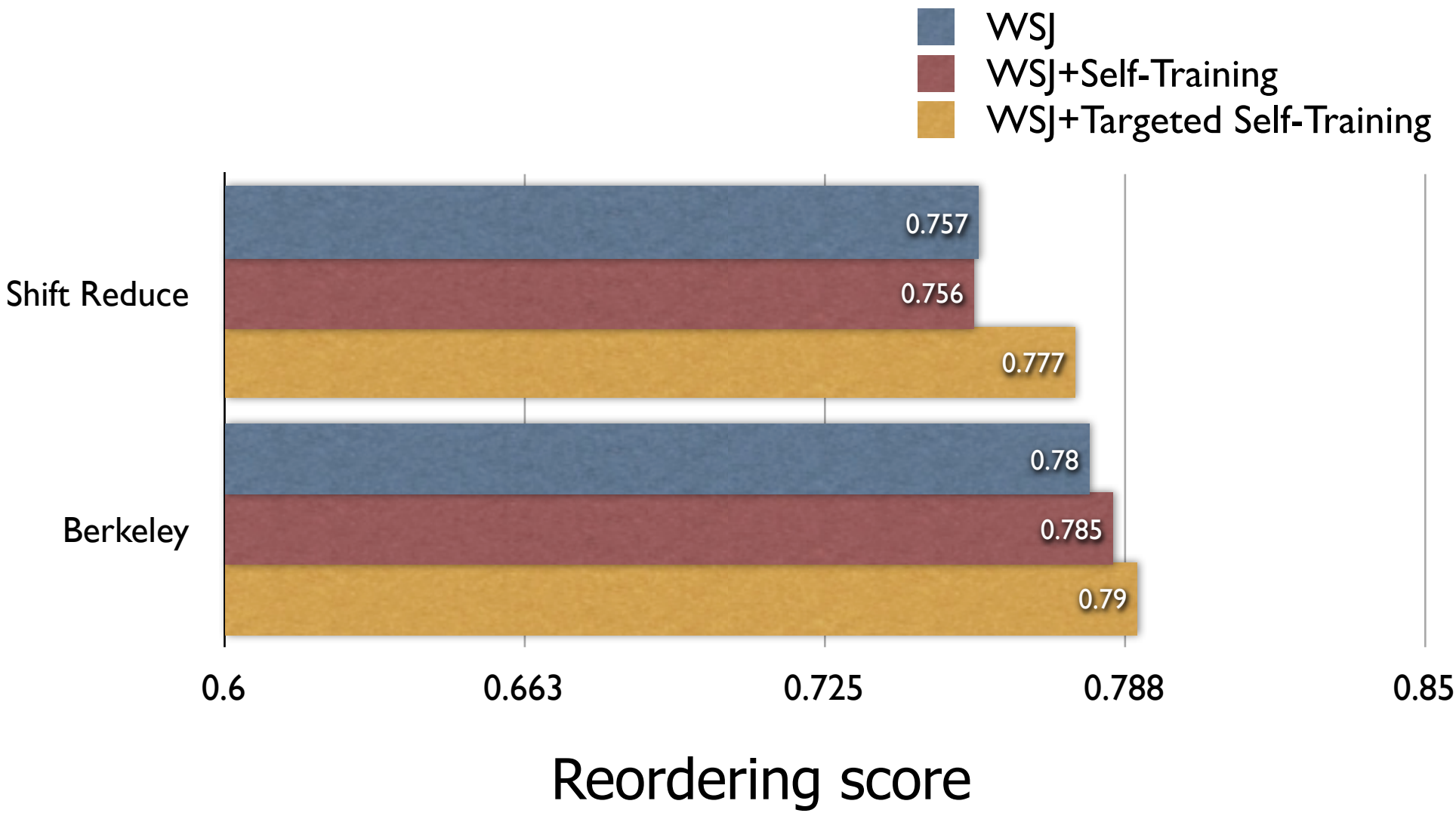
Question translation evaluations

English to	BLEU		Human eval (scores 0-6)		
	WSJ-only	Targeted	WSJ-only	Targeted	Sig. diff?
Japanese	0.2379	0.2615	2.12	2.94	yes

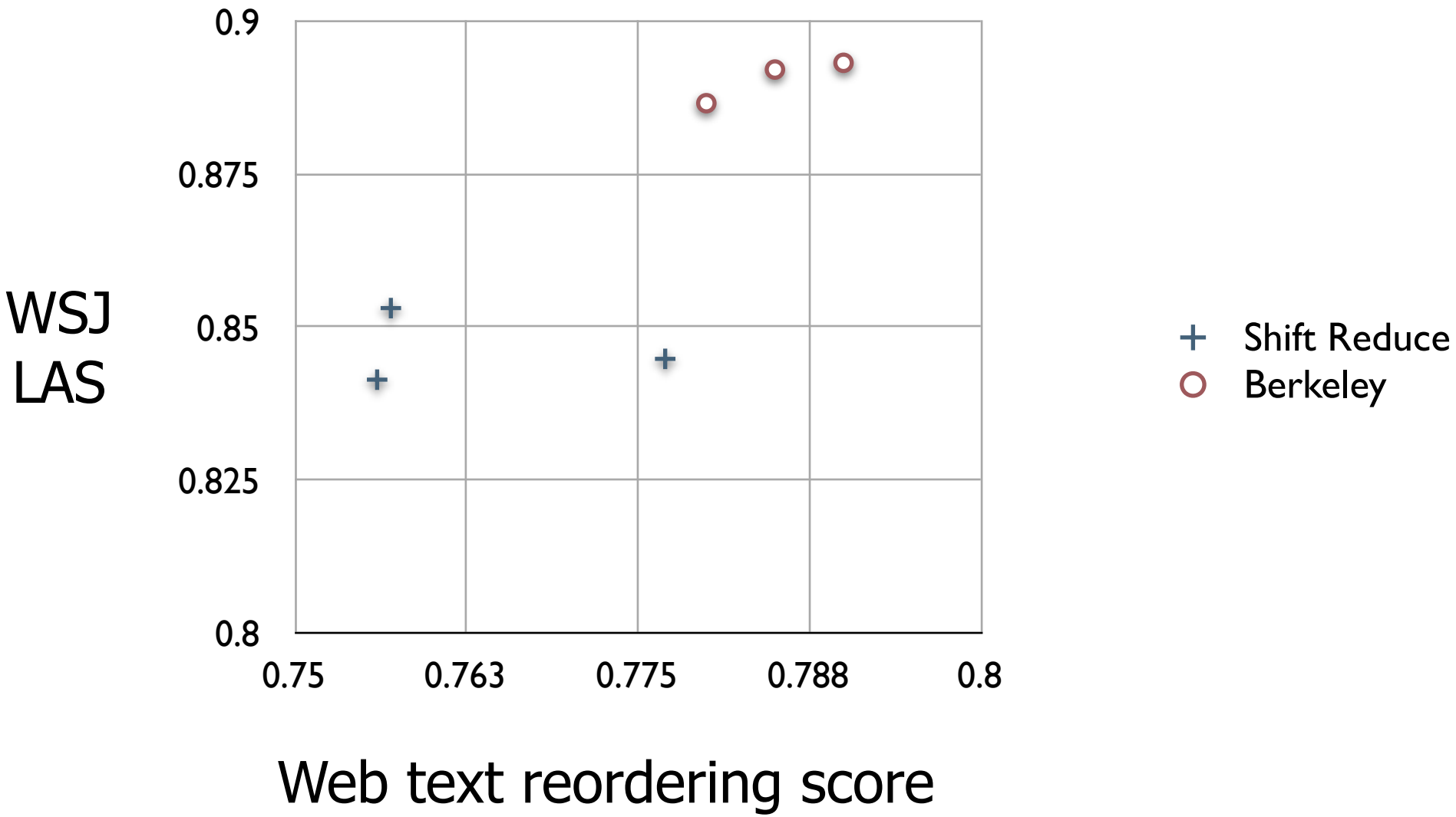
Training on web text

- 13595 English sentences sampled from web
- 6268-sentence train set, 7327-sentence test set
- Compare training on:
  - WSJ + self-training on 6268 sentences
  - WSJ + targeted self-training on 6268 reference reorderings

Web text reordering scores



Web text attachment scores



Web text translation evaluations

English to	BLEU		Human eval (scores 0-6)		
	WSJ-only	Targeted	WSJ-only	Targeted	Sig. diff?
Japanese	0.1777	0.1802	2.56	2.69	yes (95%)
Turkish	0.3229	0.3259	2.61	2.70	yes (90%)
Korean	0.1344	0.1370	2.10	2.20	yes (95%)

Related work

- Perceptron model for end-to-end MT system, updating alignment parameters from n-best list based on which leads to best BLEU score (Liang et al., 2006)
- Weakly or distantly supervised structured prediction (e.g. Chang et al., 2007)
- Re-training with bilingual data jointly parsed with multiview learning objective (Burkett et al., 2010)
- Up-training (Petrov et al., 2010)
- *"Training dependency parsers by jointly optimizing multiple objectives"* (Hall et al., EMNLP 2011)

Contributions

- Introduced targeted self-training which
  - Adapts any parser to any extrinsic metric
  - Improves translation quality significantly when applied to word reordering