Syntax-based MT Evaluation with Expected Dependency Pair Match

Jeremy G. Kahn
Mari Ostendorf
UW (SSLI Lab)

Brian Roark
OHSU

Overview

Expected Dependency Pair Match

- Straightforward intuition
- Uses syntactic and lexical information
- Does well* at predicting HTER

No:

  - Synonym or paraphrase tables
  - Dev-set tuning (much)
  - Fuzzy approximations (except in the parser)

Building EDPM: the F measure

- F-measure is intuitively appealing
  - Hard to game
  - Bag-of-words has easy intuition.
  - Multiple references? Match vs. any ref bag
- What's in the bag?
  - Words?
  - Word-sequences → n-grams

Building DPM: Outsourcing adjacency

Why insist on adjacency? *Ne ... pas* skip-n-gram is perfectly good

- But *which* non-adjacent n-grams?

"If only we had a tool for deciding which words in a sentence were related!"

- With dependency parse, we can F-measure
  - *syntactically-local n-grams* [Liu and Gildea]
  - *Labeled dep-head links* [Roark et al. SParseval]

Salutary side effect: heads "overcounted"

Building EDPM: partial match

- Don't have to use whole link.
  - We have n outbound links (dependent→link→)
  - We have n-1 inbound links (→link→head)
  - (also n unigrams, n-1 bigrams)
- Prev. work with LFG dependencies [Owczarzak et al.] found that partial-link was better vs. human judgments

Note: still a single F-measure (all 4 subclasses have different signatures)

Building EDPM: Mistrusting the parse

Parses are hidden, even on reference

- Use n-best lists on reference and hypothesis
- Use weighted counts (based on parser probabilities)
- Mistrust parser probabilities (flatten with γ)
  - $\gamma=0$: uniform distribution, $\gamma=1$: no change

EDPM: to review

Free parameters:

- Which graph fragments?
  - Unigrams, bigrams, etc
  - head+inbound link
  - dependent+outbound link
  - dependent+link+head
  - dependent+head [no link!]
- Number of n-best parses to include
- γ parse confidence trustworthiness parameter

Implementation

- Dependency forest extraction
  - Charniak parser in n-best mode
  - Head-finding table [tweaked with semantic heads]
  - Arc-labels from lowest-over-highest constituents
- The rest is in Perl

Experiments [Chinese MTC]

Set up experiments against MTC judgments

- Similar to Owczarzak 2007 experiments

Key results in r correlation vs fluency+adequacy:

- Full-link-alone F-measure ≈ BLEU-4, TER
- Improved by using partial rather than full links
- Using 1- and 2-grams ≈ inbound and outbound word+link (≫ BLEU and TER)
- Including 1g, 2g, inbound, outbound better still
- small jump from 1- to 50-best
- γ = 0.25 is good setting.

Experiments [GALE 2.5]

Compared EDPM measures (same settings) to TER, BLEU-4 on docs, sentences of GALE 2.5

- Correlations between Δ(score) & ΔHTER

Measure *m*	all-Arabic	all-Chinese	all
TER	0.51	0.19	0.39
$BLEU_4$	0.40	0.19	0.32
EDPM	**0.61**	**0.25**	**0.47**

Discussion and Future Work [1]

Internal weight tuning?

- Weight relative contribution of 1g, 2g, out-bound & inbound links.
- Introduces only 3 free parameters; no need for additional parsing when tuning.

Very different strategy from (e.g.) METEOR and TER. Combination approaches seem fertile

- Question: cross-correlation among metrics?

Discussion and Future work [2]

Defers to [expensive!] parser for syntactic info

- Better labeling? Better parser
- New target language? New parser
- More candidates? Longer n-best lists

Cherry & Quirk (2008) discriminative parsing

- vs. a better [Viterbi] parser?

L. Huang's [2008] packed forests

- Better than longer n-best lists?

Directly get dependency parse?

- We really want dist. of likely heads, arcs for each word

Thank you!

Conversations and comments from:

- My co-authors, Mari Ostendorf and Brian Roark
- Two anonymous reviewers
- Kevin Knight, Kevin Duh, Matt Snover, Michel Galley
- GALE team-members and Karolina Owczarzak

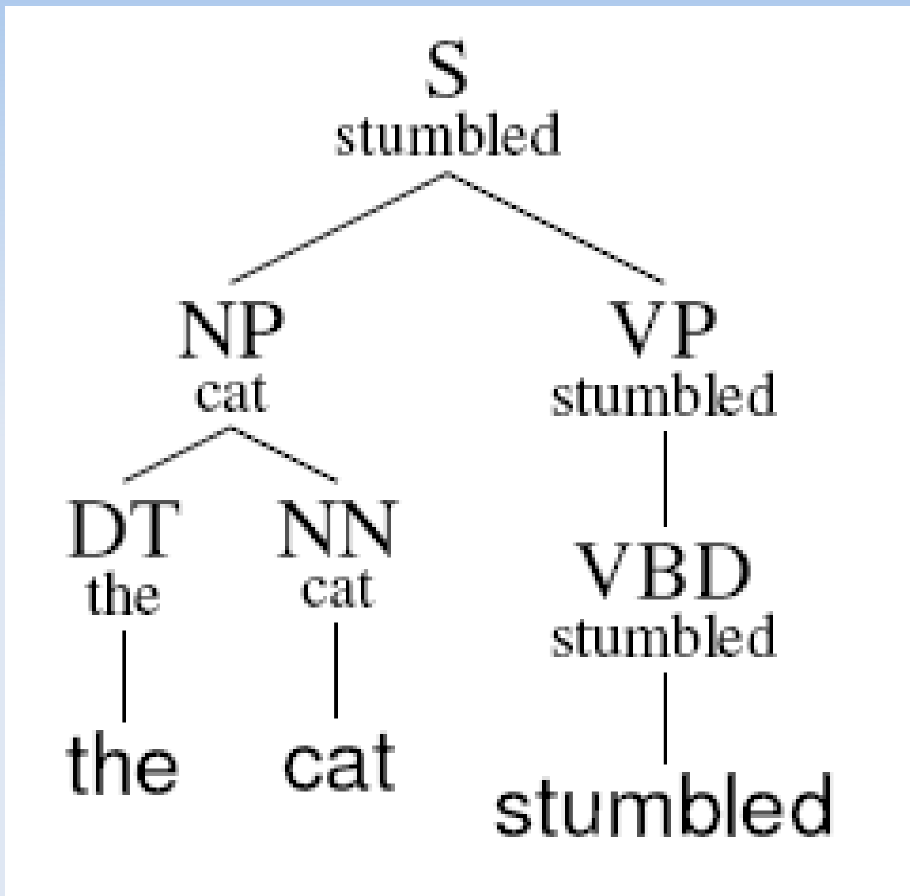
Sparseval example [1]

	string	dependency-pairs $\langle$dependent, relation, head$\rangle$
hyp	the red furry dog	$\langle \text{the}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{red}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{furry}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{dog}, \overset{\text{ROOT}}{\longrightarrow}, \text{ROOT} \rangle$
ref	the furry red dog	$\langle \text{the}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{furry}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{red}, \overset{\text{nmod}}{\longrightarrow}, \text{dog} \rangle$ $\langle \text{dog}, \overset{\text{ROOT}}{\longrightarrow}, \text{ROOT} \rangle$

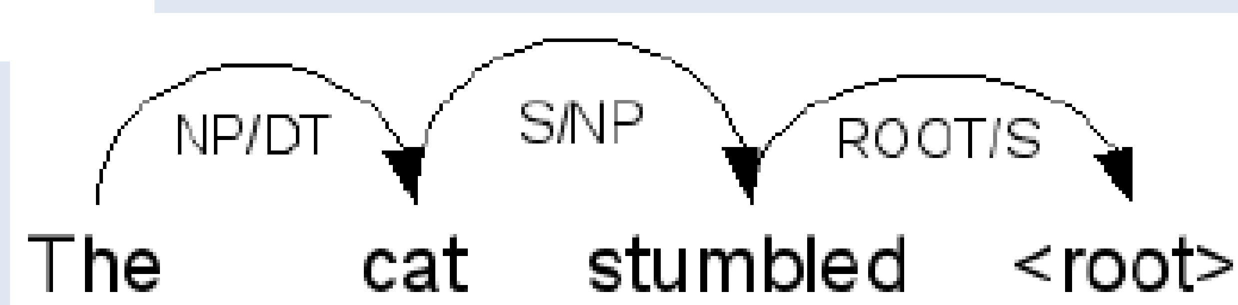
Sparseval example [2]

	string	dependency-pairs
hyp	White House spokesman	$\langle \text{White}, \stackrel{\texttt{nmod}}{\longrightarrow}, \text{House} \rangle$ $\langle \text{House}, \stackrel{\texttt{nmod}}{\longrightarrow}, \text{spokesman} \rangle$ $\langle \text{spokesman}, \stackrel{\texttt{ROOT}}{\longrightarrow}, \text{ROOT} \rangle$
ref	House spokesman White	$\langle \text{House}, \stackrel{\texttt{nmod}}{\longrightarrow}, \text{spokesman} \rangle$ $\langle \text{spokesman}, \stackrel{\texttt{nmod}}{\longrightarrow}, \text{White} \rangle$ $\langle \text{White}, \stackrel{\texttt{ROOT}}{\longrightarrow}, \texttt{ROOT} \rangle$

Extracting dependency trees



- Charniak PCFG with WSJ default training
- Head-finding with modified Charniak rules
- Arc-labels are Gov/MaxProj



Correlation improvements [MTC]

Measure	r
$\mathrm{DPM}_{dl,lh}$ (~**d_var**)	0.226
1+BLEU$_4$	0.218
DPM_{dlh} (~**d**)	0.185
TER	−0.173

Measure	r
$\mathrm{DPM}_{1g,2g,dl,lh}$	0.237
$\mathrm{DPM}_{1g,dl,lh}$	0.234
$\mathrm{DPM}_{1g,2g}$(≡ bag-of-ngrams(2))	0.227
$\mathrm{DPM}_{dl,lh}$	0.226
$\mathrm{DPM}_{1g,dl,dlh}$	0.227
1+BLEU$_4$	0.218
DPM_{dlh}	0.185
TER	−0.173

GALE 2.5 by-genre document correlations with HTER

Measure m	bc	bn	nw	wb
	Arabic			
TER	0.59	**0.24**	0.22	0.26
$BLEU_4$	0.50	0.10	**0.30**	0.31
EDPM	**0.80**	0.10	**0.31**	**0.33**
	Chinese			
TER	0.06	0.13	0.35	0.14
$BLEU_4$	0.01	0.22	**0.36**	0.07
EDPM	**0.14**	**0.30**	**0.37**	**0.16**