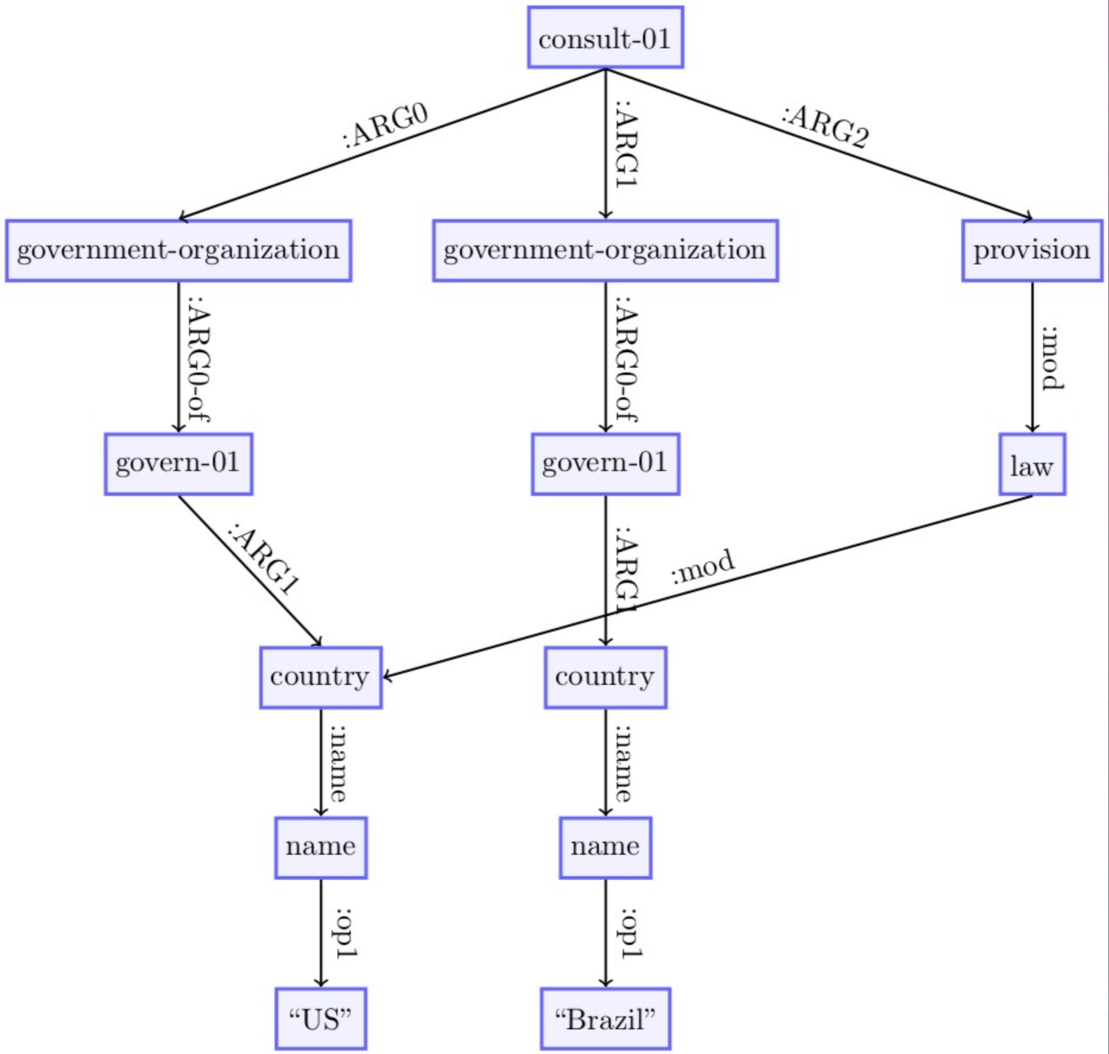
REFERENCELESS PARSING-BASED EVALUATION OF AMR-TO-ENGLISH GENERATION

EMMA MANNING & NATHAN SCHNEIDER

GEORGETOWN UNIVERSITY

INTRODUCTION

- Evaluating NLG is notoriously difficult: one-to-many problem
- Reference-based metrics are popular but flawed
- We explore a referenceless alternative for AMR-to-English generation: *parsing-based evaluation*
- Also suggested by Opitz & Frank (EACL 2021); we evaluate the approach in new ways:
  - Comparison to human judgments
  - Manual editing experiment



The US government has consulted the Brazilian Government about the provisions of US law.

Real generated sentences:

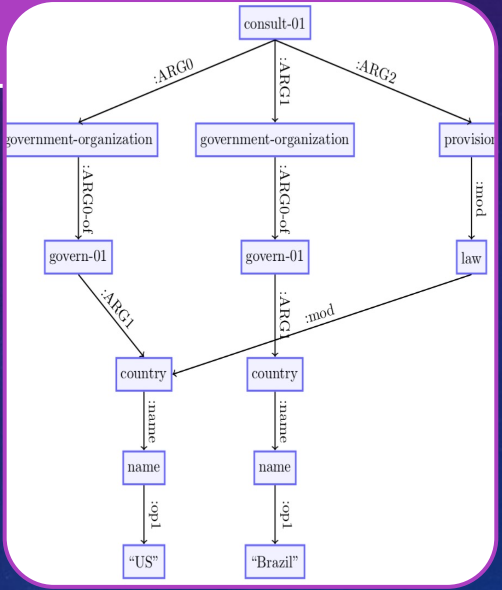
- the US government has consulted *with* the *Brazil* government *over* the provisions of US law .
- the us government has consulted the brazilian government **as** *a provision* of **brazilian** law
- the us government has consulted *with* the *brazil* government **for** the provisions of *the* **south korean** law .
- the us government *will* consult the brazilian government **with** a **canadian** *law provision* .
- Government **organizations governing** US *consulting* government **organizations governing** Brazil law provisions .

TYPICAL MODEL: REFERENCE-BASED EVALUATION

Original Sentence (Reference)

- The US government has consulted the Brazilian Government about the provisions of US law.

Gold AMR



Generated Sentence

- the US government has consulted with the Brazil government over the provisions of US law .

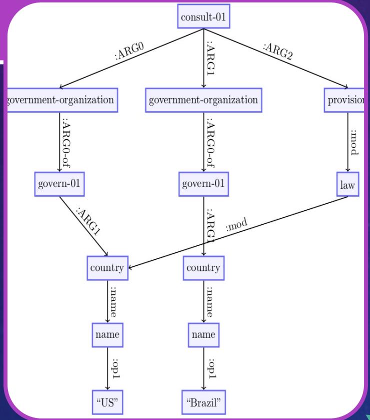
Compare

NEWER IDEA: PARSING-BASED EVALUATION

Original Sentence (Reference)

- The US government has consulted the Brazilian Government about the provisions of US law.

Gold AMR

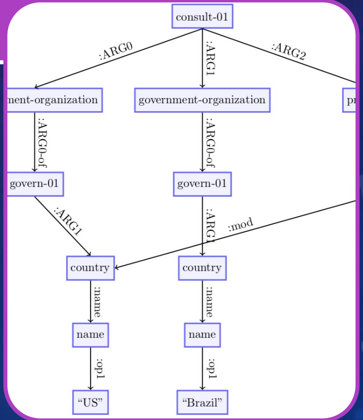


Generated Sentence

- the US government has consulted with the Brazil government over the provisions of US law .

Requires: parser

Parsed AMR



Compare

Requires: similarity metric

METHODS: DATA

- Human judgment data produced in previous study (Manning et al. 2020)
- Judgments on 100 sentences each from 5 generation systems (+ references)
- Scalar judgments of *fluency* and *adequacy*
    - For this we're most interested in adequacy!

REFERENCE-BASED BASELINES

- Correlations of popular RBMs with our human judgments

	Fluency	Adequacy
BLEU$_\uparrow$	0.40	0.52
METEOR$_\uparrow$	0.41	0.57
TER$_\downarrow$	−0.33	−0.43
CHRF++$_\uparrow$	0.32	0.47
BERTScore$_\uparrow$	0.47	0.60
BLEURT$_\uparrow$	**0.60**	**0.69**

EXPERIMENT 1: AUTOMATIC PARSING

What should we use for the parser and similarity metric? How much does it matter?

- Tried 3 automatic parsers:
  - Baseline: JAMR (Flanigan et al., 2014, 2016)
  - Medium: Lyu & Titov (2018)
  - Best: Cai & Lam (2020)
- And 3 similarity metrics:
  - Standard: Smatch (Cai & Knight, 2013)
  - Minor Variant: Smatch$_{100}$+seed – Smatch, but more reliable & reproducible
  - Bigger Variant: S^{2}match (Opitz et al., 2020)

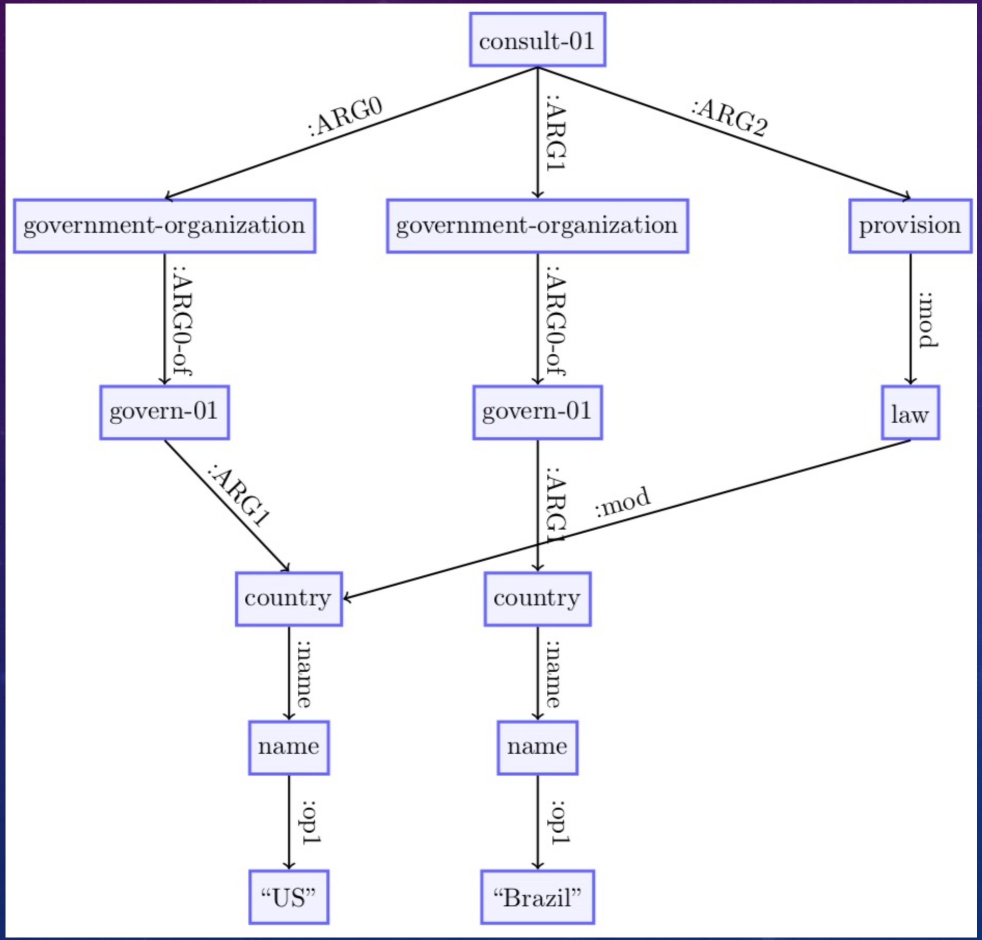
EXPERIMENT 1: RESULTS

- Better parser -> better results!
- Similarity metric doesn't matter much
- Correlation with adequacy lower than for most automatic metrics 😔

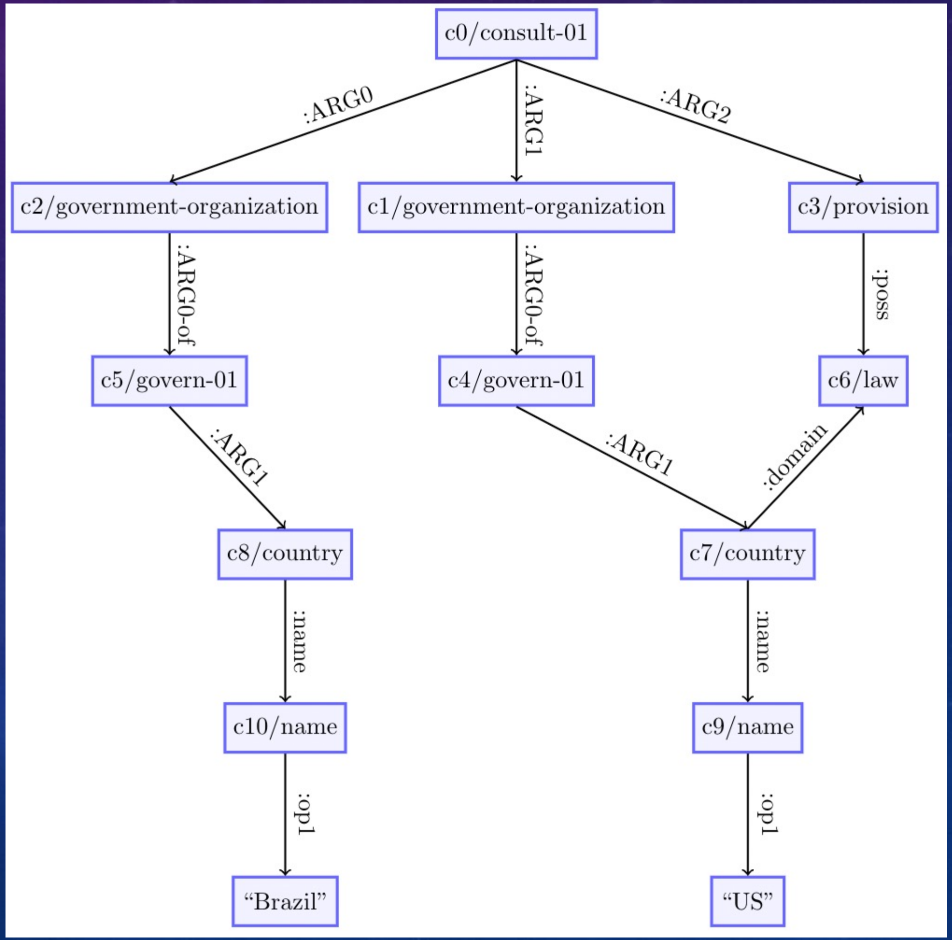
	Smatch_4	Smatch_{100}+seed	S^2match
JAMR	0.358	0.356	0.362
Lyu & Titov	0.462	0.460	0.465
Cai & Lam	0.495	0.492	0.494

WHAT WENT WRONG?

The US government has consulted the Brazilian Government about the provisions of US law.



the US government has consulted with the Brazil government over the provisions of US law .

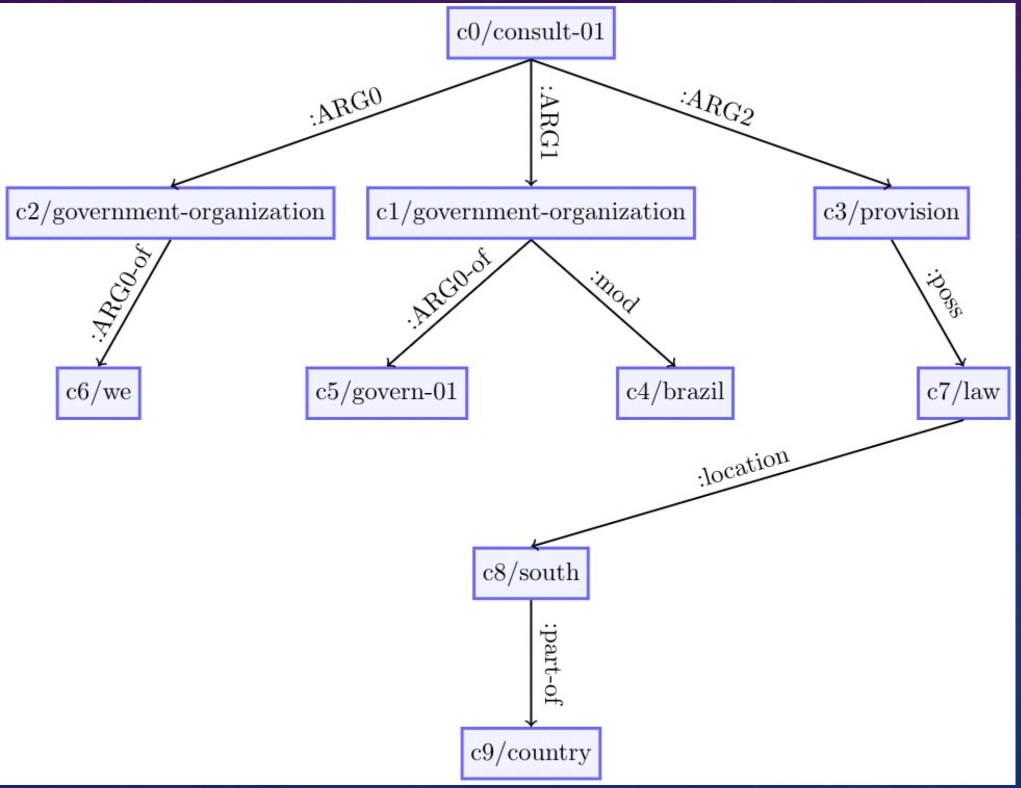


EXPERIMENT 2: MANUAL EDITING

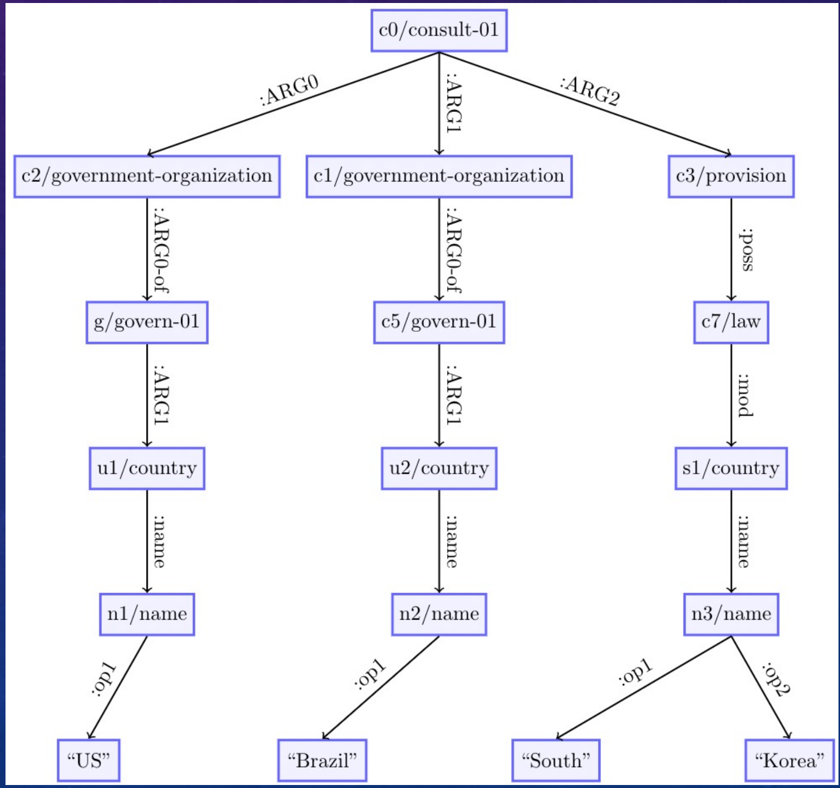
- State-of-the-art AMR parser introduces a lot of errors for this data!
- How well could this work with an even better parser?
- Idea: Correct the parses myself to approximate an upper bound!

EDITING EXAMPLE

the us government has consulted with the brazil government for the provisions of the south korean law .



Automatic Parse



Edited Parse

MANUAL EDITING RESULTS

- After editing, correlation with adequacy increases to 0.66 🥳
- Indicates that this will work better automatically as parsers continue to improve

Metric	Correlation w/ Adequacy
BLEU$_\uparrow$	0.52
METEOR$_\uparrow$	0.57
TER$_\downarrow$	-0.43
CHRF++$_\uparrow$	0.47
BERTScore$_\uparrow$	0.60
BLEURT$_\uparrow$	**0.69**
Automatic Parsing	0.49
Edited Parsing	0.66

CONCLUSION: MAIN TAKEAWAYS

- Of existing automatic reference-based metrics, BLEURT and BERTScore look pretty good for AMR generation
  - Please stop relying on BLEU!
  - But: concerns about transparency, bias, etc.
- Parsing-based referenceless evaluation has potential, but is currently limited by parser accuracy