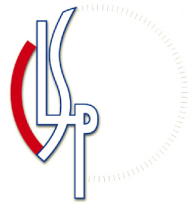


The Center for Language and Speech Processing
at the Johns Hopkins University

Monolingual Sentence Rewriting as Machine Translation

Courtney Napoles

PhD Defense, 18 June 2018

Thesis Committee: Chris Callison-Burch, Philipp Koehn, and Benjamin Van Durme

Monolingual Sentence Rewriting

Input: English sentence

Output: Another English sentence

Constraints: Same meaning as input

Grammatical

Some task-specific constraint

- Also called *text-to-text generation (T2T)*

Monolingual Sentence Rewriting

What types of constraints?

- **Length**
  length(O) < length(I)
- **Complexity**
  readability(O) > readability(I)
- **Error-free**
  mistakes in input corrected in output

Sentence compression

Text simplification

Grammatical error correction

Sentence compression

Instead of bounding along in celebration after Hirving Lozano delivered the 34th-minute goal that would ultimately give Mexico an unlikely 1-0 win over Germany here on Sunday night, Osorio sat down on his team's bench and adopted a Zen-like pose — a deep breath, in through his nose, out through his mouth, then again, and again — as he plotted his team's next moves.

Instead of celebrating after Lozano scored, Osorio plotted the team's next moves.

Source: New York Times, 17 June 2018

Text simplification

Instead of bounding along in celebration after Hirving Lozano delivered the 34th-minute goal that would ultimately give Mexico an unlikely 1-0 win over Germany here on Sunday night, Osorio sat down on his team's bench and adopted a Zen-like pose — a deep breath, in through his nose, out through his mouth, then again, and again — as he plotted his team's next moves.

Mexico beat Germany with a score of 1-0 on Sunday night. Hirving Lozano scored the only goal. Osorio did not celebrate after the goal. He planned what his team would do next.

Grammatical error correction (GEC)

They were awesome in the first half. Credit to them for their attitude, not many teams out there with the balls to take the game to germans like that.

They were awesome in the first half. I give them credit ~~to them~~ for their attitude. There are not many teams out there with the balls to take the game to the Germans like that.

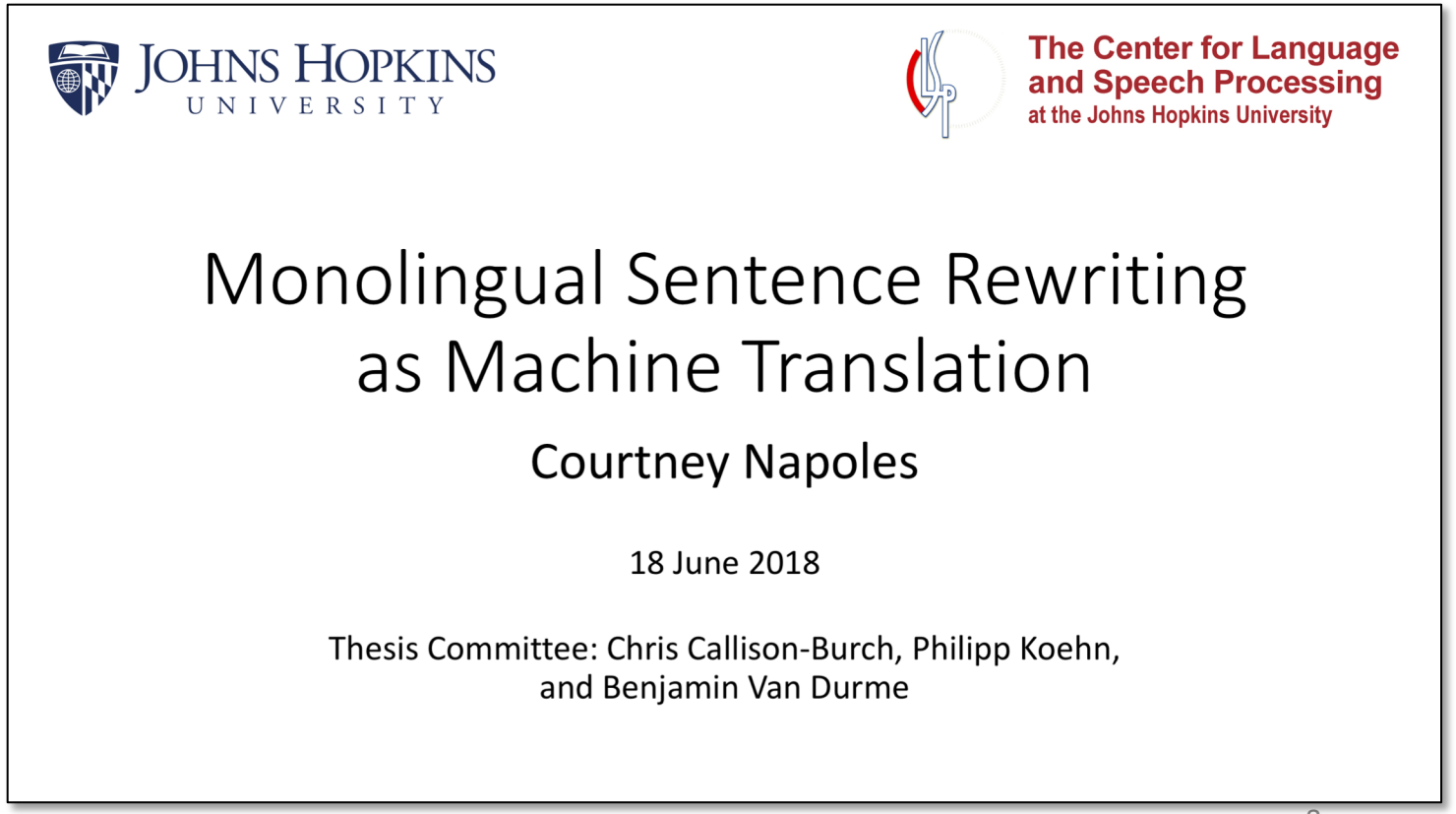
Source: Reddit /r/soccer, 17 June 2018

Monolingual Sentence Rewriting

What?

- **Sentence compression**
- **Text simplification**
- **Grammatical error correction**

How?



Monolingual Sentence Rewriting

What?

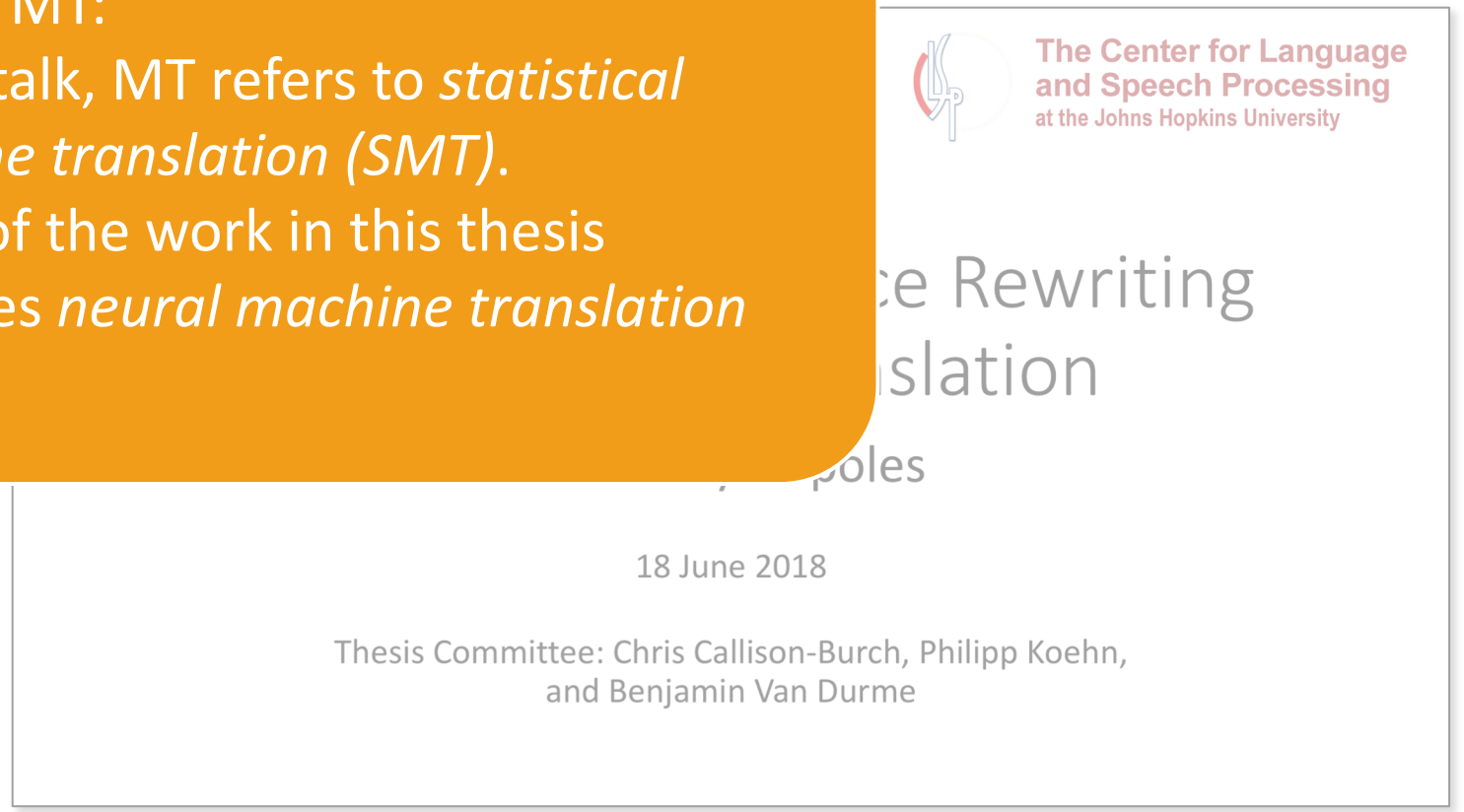
Sentence compression

Text simplification

Grammatical error correction

A note on MT:

- In this talk, MT refers to *statistical machine translation (SMT)*.
- Much of the work in this thesis predates *neural machine translation (NMT)*.



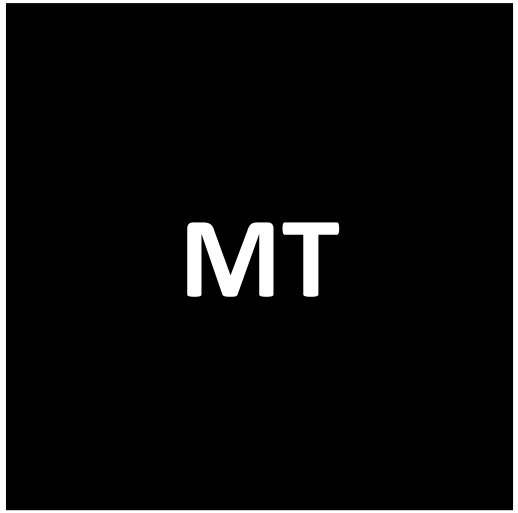
This is easy!

Machine translation (MT) toolkits

- Moses (Koehn et al., 2007)
- Joshua (Li et al., 2009)
- …

MT decoders

- cdec (Dyer et al., 2010)
- Phrasal (Green et al., 2014)
- …

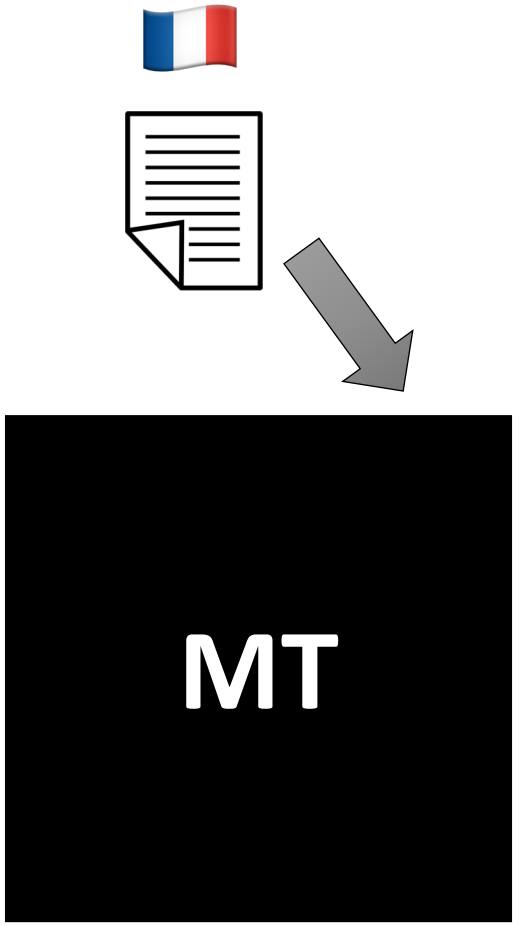


Bilingual translation

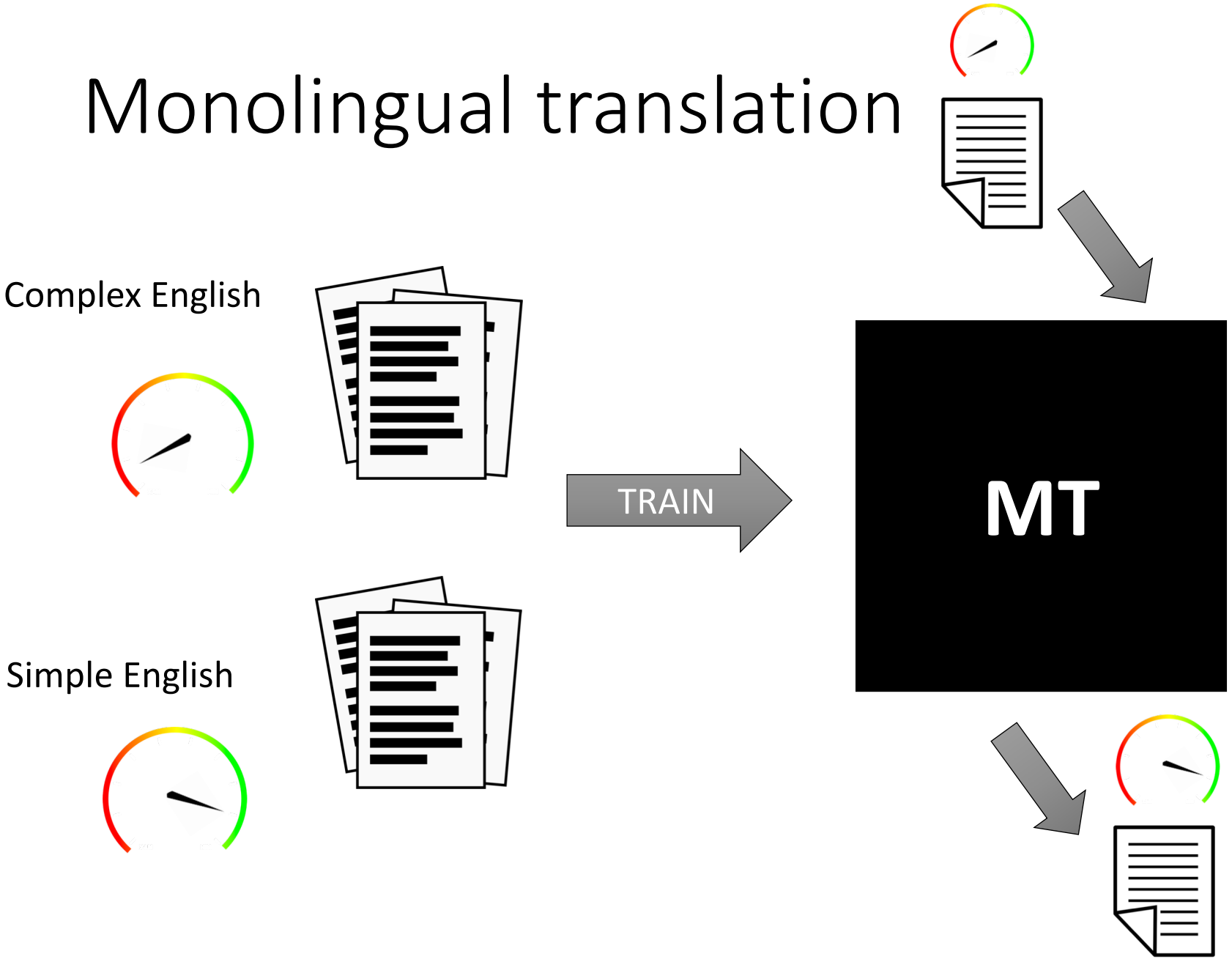
Original “source” language

TRAIN

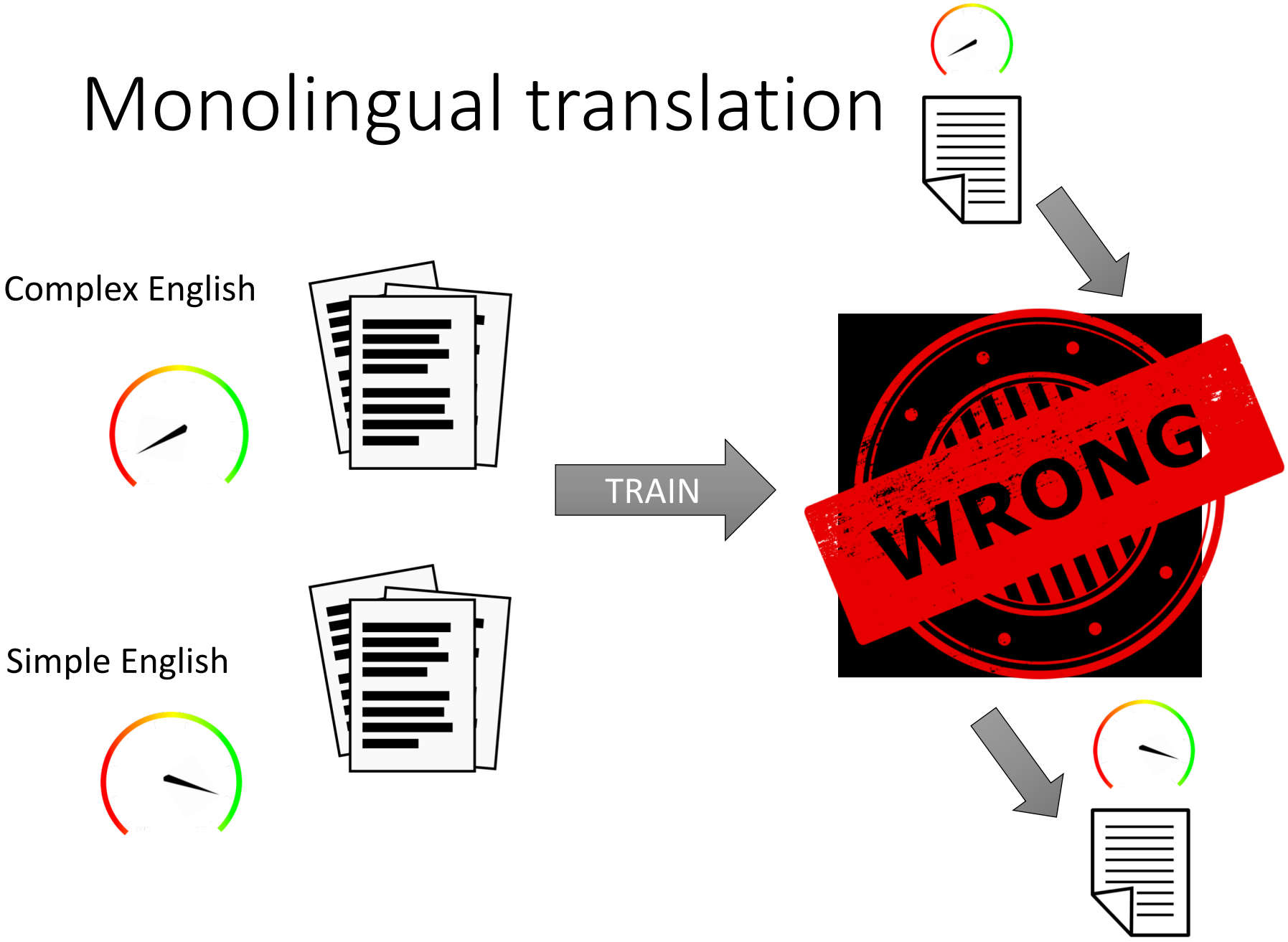
Output “target” language



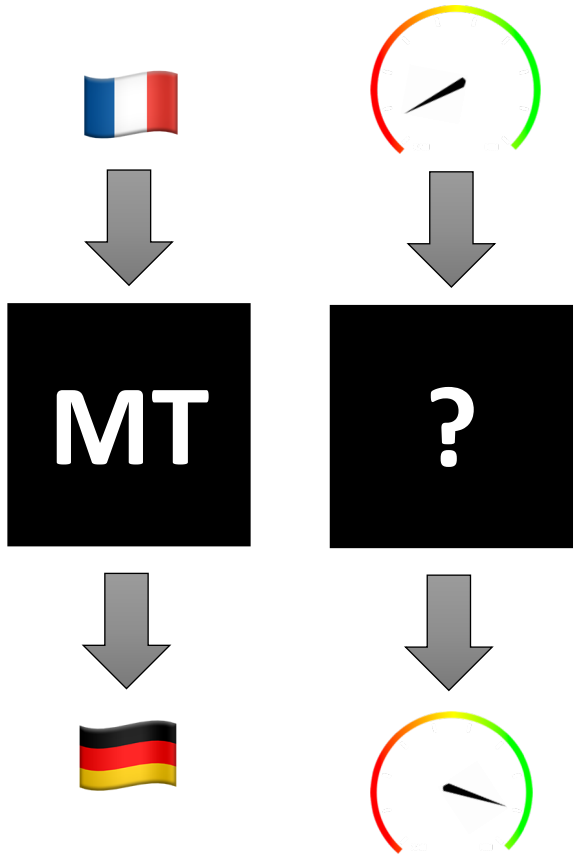
Monolingual translation



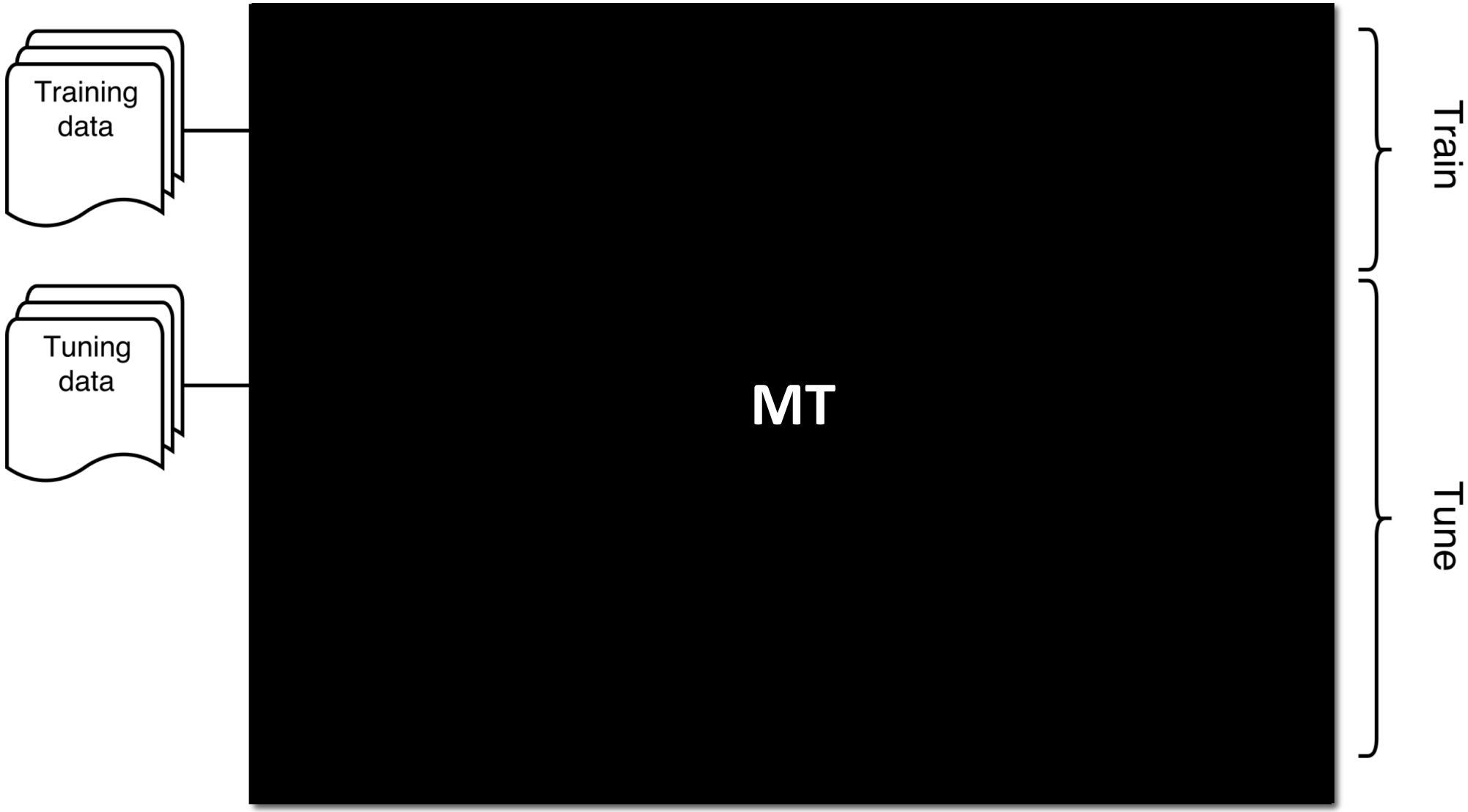
Monolingual translation

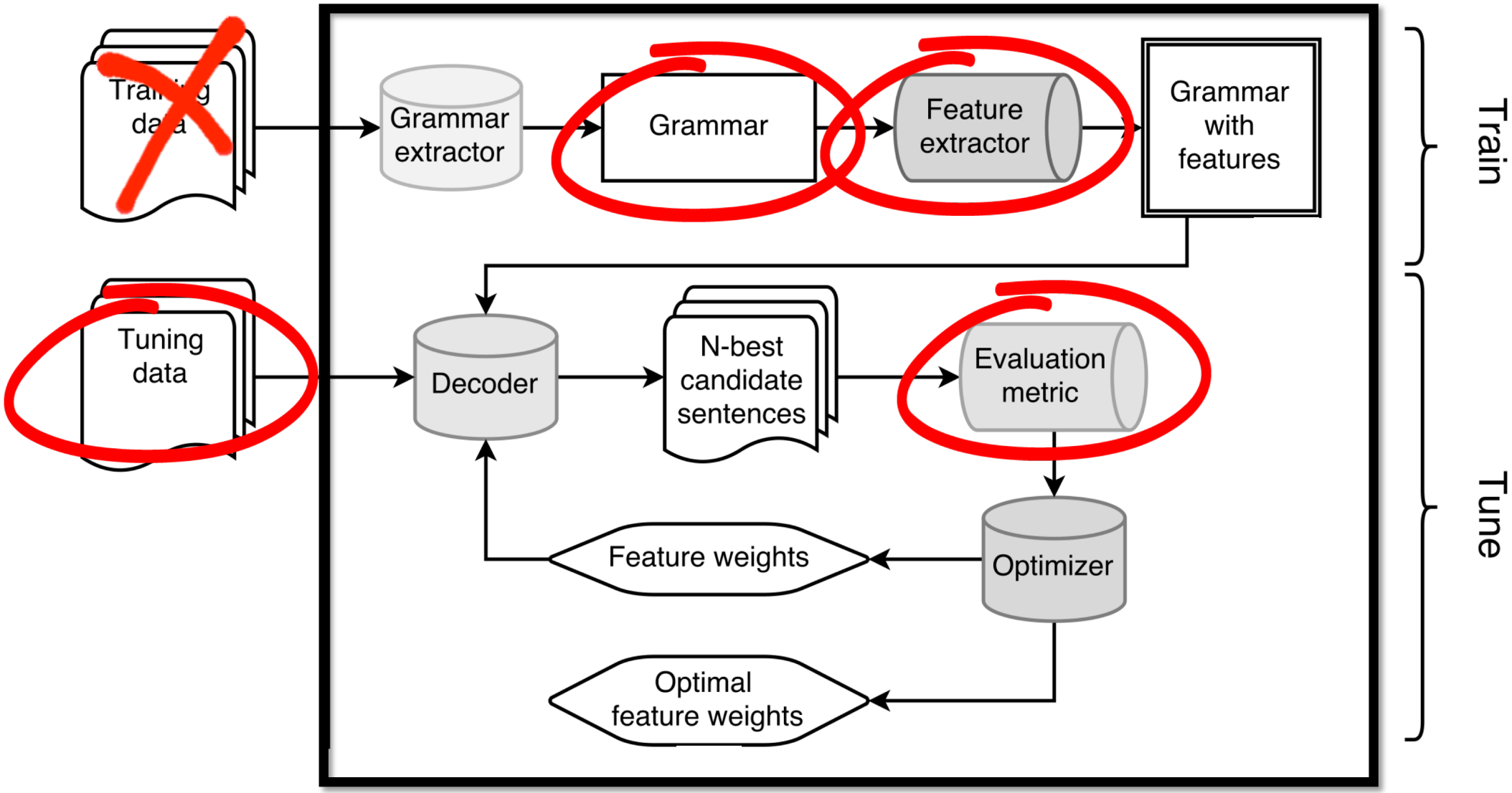


Why black-box MT doesn't work



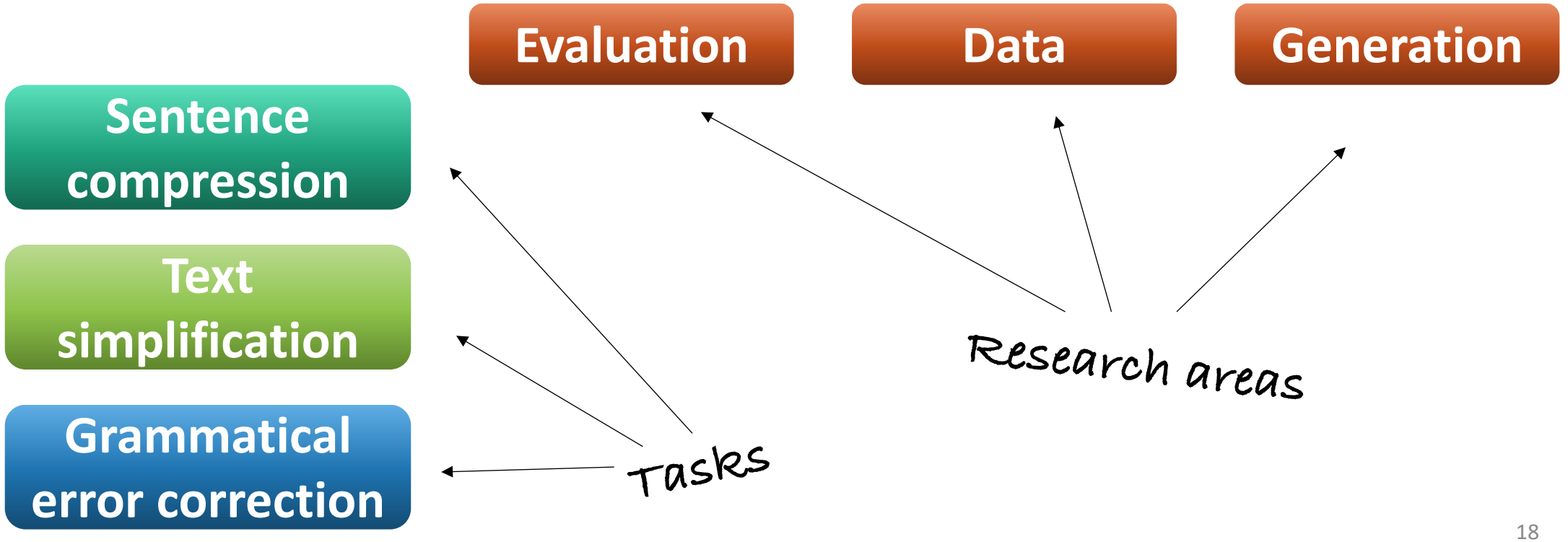
- Many have tried this (Wubben et al., 2012; Coster and Kauchak, 2013; Felice et al., 2014; etc.)
- Lack of large-scale parallel training data
- MT models learn to copy in monolingual tasks
- MT metrics reward copying


Training data
Tuning data
MT
Train
Tune


Training data
Grammar extractor
Grammar
Feature extractor
Grammar with features
Tuning data
Decoder
N-best candidate sentences
Evaluation metric
Feature weights
Optimizer
Optimal feature weights
Train
Tune

This Thesis

- Adapt MT for monolingual T2T tasks
- Address two core issues: (1) Flawed evaluation. (2) Lack of parallel data.



This Thesis: Contributions

1. Evaluation

- Identify flaws & biases in existing methods
- Propose new evaluation methods
- New evaluation is robust and allows paraphrasing

2. Data

- New tuning sets
- Multiple references for evaluation
- Data that has paraphrasing
- Validated to meet goals of task

3. Generation

- Adapt MT for tasks of increasing complexity
- More expressive output due to paraphrasing
- Less task-specific parallel data needed

Sentence Compression

Joint work with Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison Burch

Publications

Napoles, Van Durme, and Callison-Burch
Evaluating Sentence Compression: Pitfalls and Suggested Remedies
Monolingual T2T Workshop, 2011a

Napoles, Callison-Burch, Ganitkevitch, and Van Durme
Paraphrastic Sentence Compression with a Character-based Metric
Monolingual T2T Workshop, 2011b

Ganitkevitch, Callison-Burch, **Napoles**, and Van Durme
Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation
EMNLP, 2011

Background

Given an English sentence *s*, create a new English sentence *c* such that

- *c* is grammatical
- *c* conveys most important meaning of *s*
- length(*c*) < length(*s*)

Background

Two types of sentence compression

1. **Extractive** (just deletion)

2. **Abstractive** (deletion + paraphrasing)

Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.

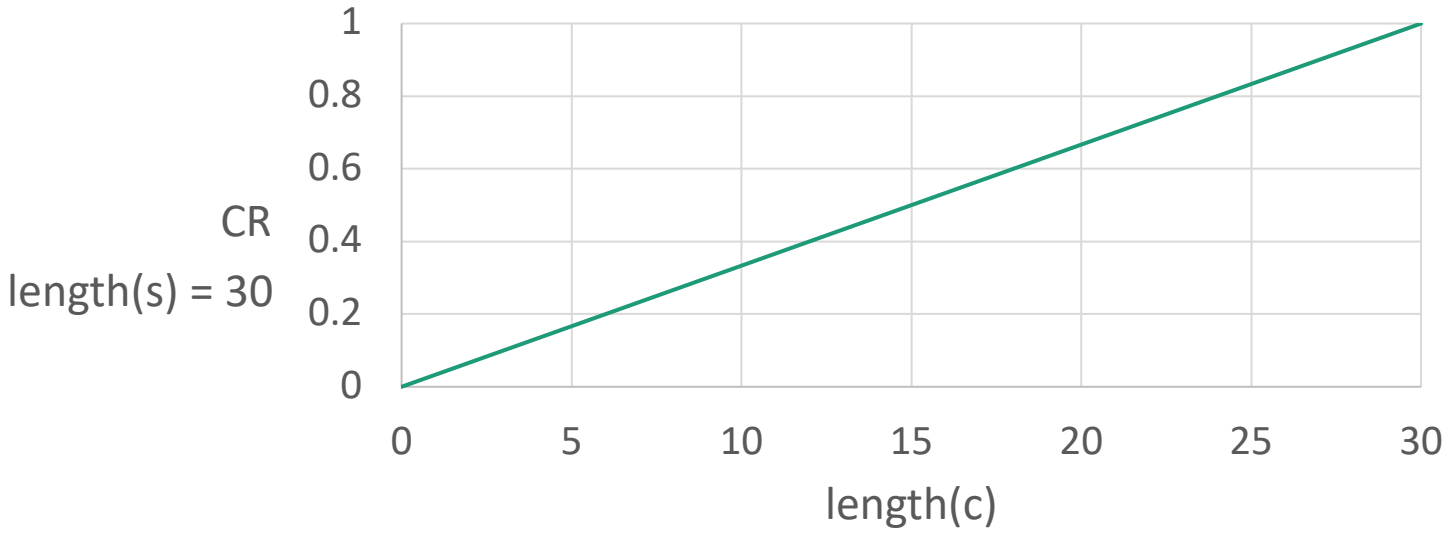
Congress made a **final** agreement **Fri.** to **avoid government shutdown**, after days of haggling and tense hours of **brinkmanship**.

Background

Compressions rate (really, it's the "compression ratio")

$$CR = \frac{\text{length}(c)}{\text{length}(s)}$$

Longer output → higher compression rate



Evaluation

Data

Generation

Sentence compression

Unbiased evaluation

Text simplification

Grammatical error correction

Biased evaluation

Global Inference for Sentence Compression An Integer Linear Programming Approach
Clarke and Lapata, JAIR 2008 (Based on EMNLP 2017 Best Paper)

Models	CompR	F-score
LM	52.0	25.4
Sig	60.9	30.4
McD	68.6	47.6
LM+Constr	49.5	34.8*
Sig+Constr	78.4	48.7*†
McD+Constr	68.5	50.1†
Gold	76.1	—

Their system (Sig+Constr) outperforms McD+Constr (McDonald, 2006)

Biased evaluation

Models	CompR	F-score
LM	52.0	25.4
Sig	60.9	30.4
McD	68.6	47.6
LM+Constr	49.5	34.8*
Sig+Constr	78.4	48
McD+Constr	68.5	50.1†
Gold	76.1	—

9.9 point difference in CR.
Sig+Constr is 10% longer!

Effect of compression rate

Hypothesis: Compression rate is correlated with sentence quality.

New data:

- 3 native speakers manually compressed 50 sentences at 10 CRs, $10 \leq CR \leq 100$.
- 1500 compressions total

Effect of compression rate

Original:

> Bob Dole pledged to the American people that in return for millions of taxpayer dollars, he would obey the law.

Compressions at a single rate:

CR	Compression
77	Dole pledged to people that in return for taxpayer dollars, he would obey the law.
77	Dole pledged to American people that in return for dollars, he would obey the law.
73	Dole pledged that in return for millions of taxpayer dollars, he would obey the law.

Effect of compression rate

Original:

Bob Dole pledged to the American people that in return for millions of taxpayer dollars , he would obey the law .

Compressions at multiple rates:

CR	Compression
86	Bob Dole pledged to people that in return for millions of taxpayer dollars he would obey the law.
59	Dole pledged that for taxpayer dollars he would obey the law.
41	Bob Dole pledged he would obey the law
18	Dole would obey .

Napoles et al., 2011a. Evaluating Sentence Compression

Effect of compression rate

- Human evaluation on Mechanical Turk
- Rated compressions on 2 5-point scales:
  - How grammatical is the compression?
  - How much of the important meaning is preserved?

Napoles et al., 2011a. Evaluating Sentence Compression

Judge the quality of shortened sentences

Please read the groups of sentences below. Each pair contains an original sentence along with some automatically shortened versions of it. Your job is to grade the quality of the shortened sentences with two 5-point scales: **meaning** and **grammar**. In this HIT you should ignore capitalization mistakes.

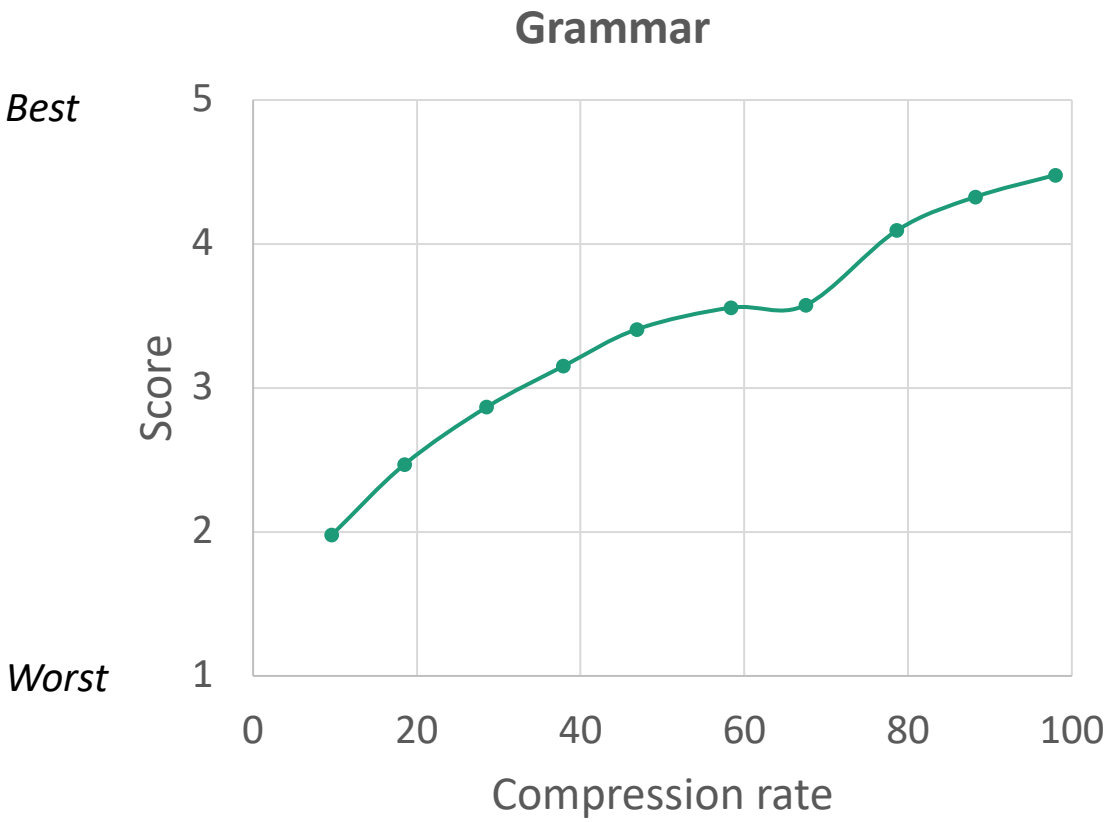
Meaning

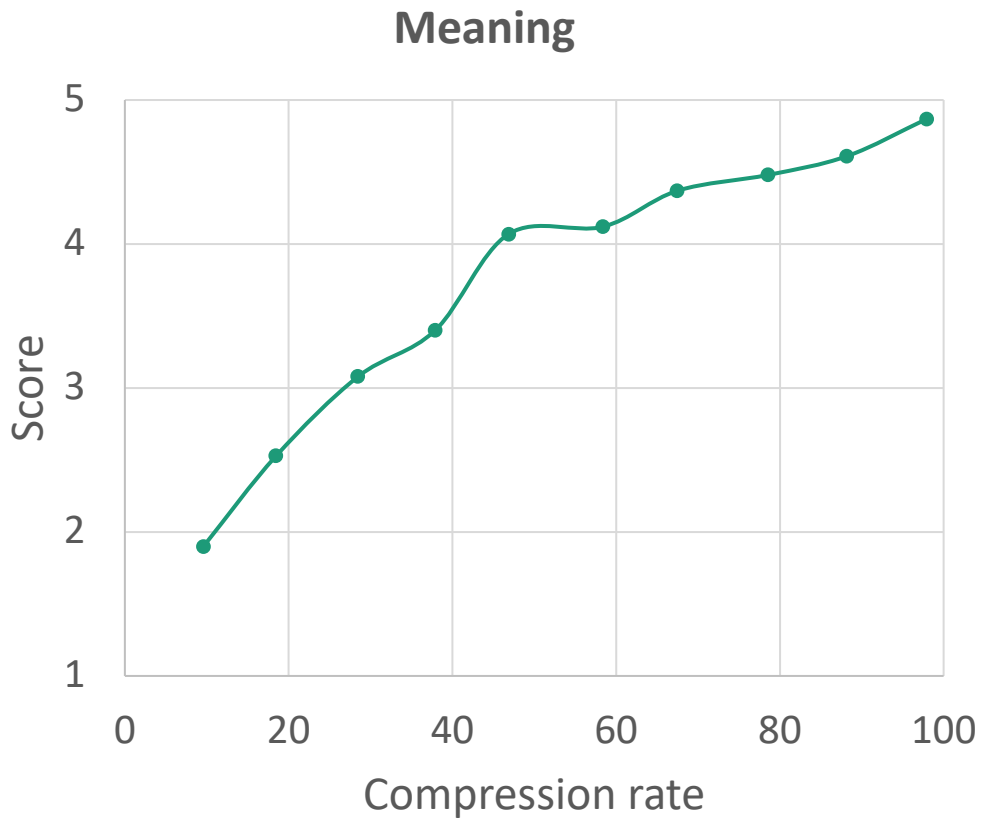
- **5 - perfect**: All of the meaning of the original sentence is retained, and nothing is added
- **4 - minor differences**: The meaning of the original sentence is retained, although some minor information may be deleted or added without too great a difference in meaning
- **3 - moderate differences**: Some of the meaning of the original sentence is retained, although a non-trivial amount of information was deleted or added
- **2 - substantially different**: Substantial amount of the meaning is different
- **1 - completely different**: The shortened sentence doesn't mean anything close to the original sentence

Grammar

- **5 - perfect**: The shortened sentence is perfectly grammatical
- **4 - ok but awkward**: The sentence is grammatical, but might sound slightly awkward
- **3 - one error**: The sentence has an error (such as an agreement error between subject and verb, a plural noun with a singular determiner, or the wrong verb form)
- **2 - many errors**: The sentence has multiple errors or omits words that would be required to make it grammatical
- **1 - ungrammatical**: The sentence is totally ungrammatical

Biased evaluation: Effect of compression rate

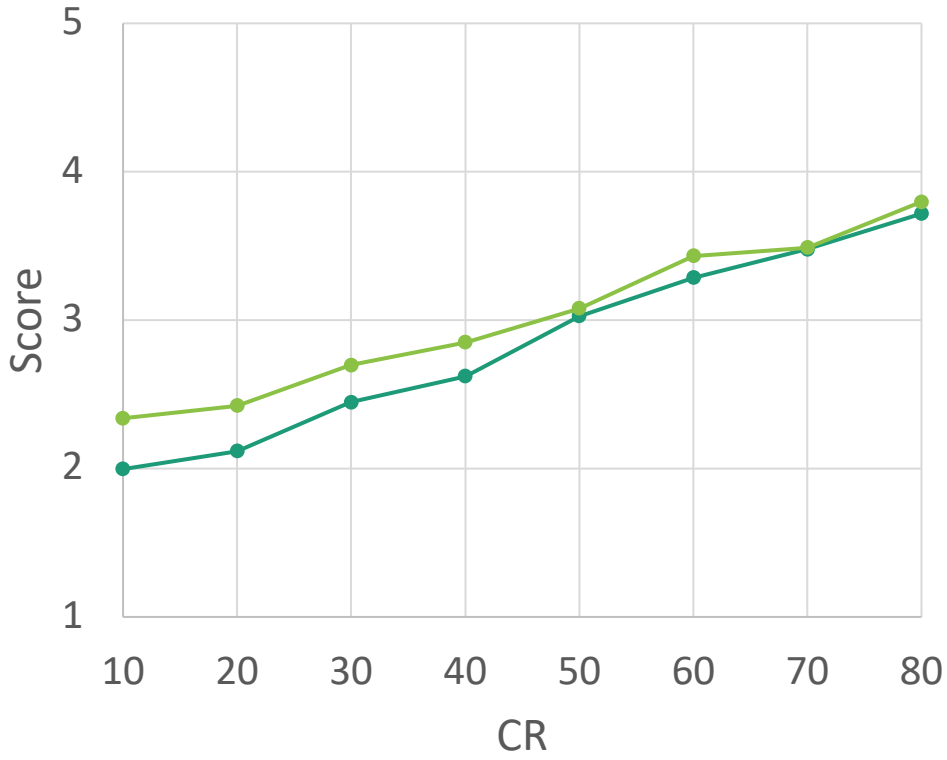




Napoles et al., 2011a. Evaluating Sentence Compression

Effect of compression rate on C&L 2008

- Our implementation of their model
- Systematically varied target CR
- Ran same human evaluation as before



Biased evaluation: Effect of compression rate

- We implemented the best paper system (C&L) and re-evaluated the output when the CR was fixed

	System	Meaning	Grammar	CR
Original published result	C&L	**3.8**	**3.5**	78.4
	McD	3.5	3.2	68.5

	System	Meaning	Grammar	CR
New unbiased result	C&L	3.8	3.7	64
	McD	**3.9**	**3.9**	64

System ranking is reversed when CR is fixed!

Napoles et al., 2011a. Evaluating Sentence Compression

Biased evaluation: Effect of compression rate

It is only valid to report that System A is better System B if

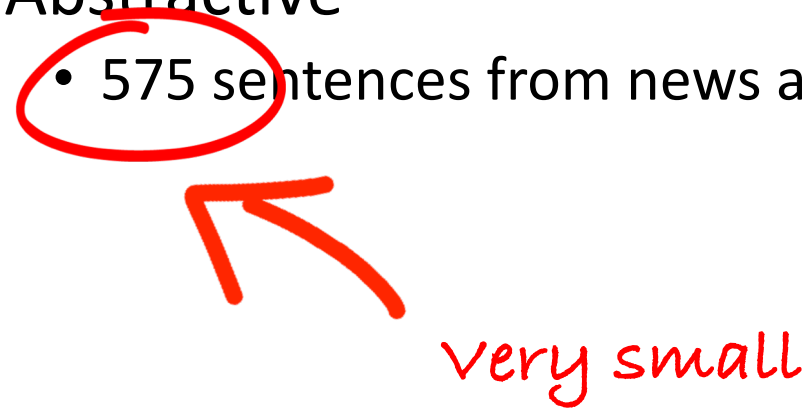
- Judges favor A **and**
  - The outputs of A and B have the same compression rate
    or
  - System A has shorter output.

Napoles et al., 2011a. Evaluating Sentence Compression

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	
Text simplification			
Grammatical error correction			

Corpora for sentence compression

- Extractive:
  - 1k sentences from article/abstracts (Ziff-Davis; Knight and Marcu, 2002)
  - 3k sentences from news articles and transcripts (Clarke and Lapata, 2008)
- Abstractive
  - 575 sentences from news articles (Cohn and Lapta, 2007)



New abstractive corpus

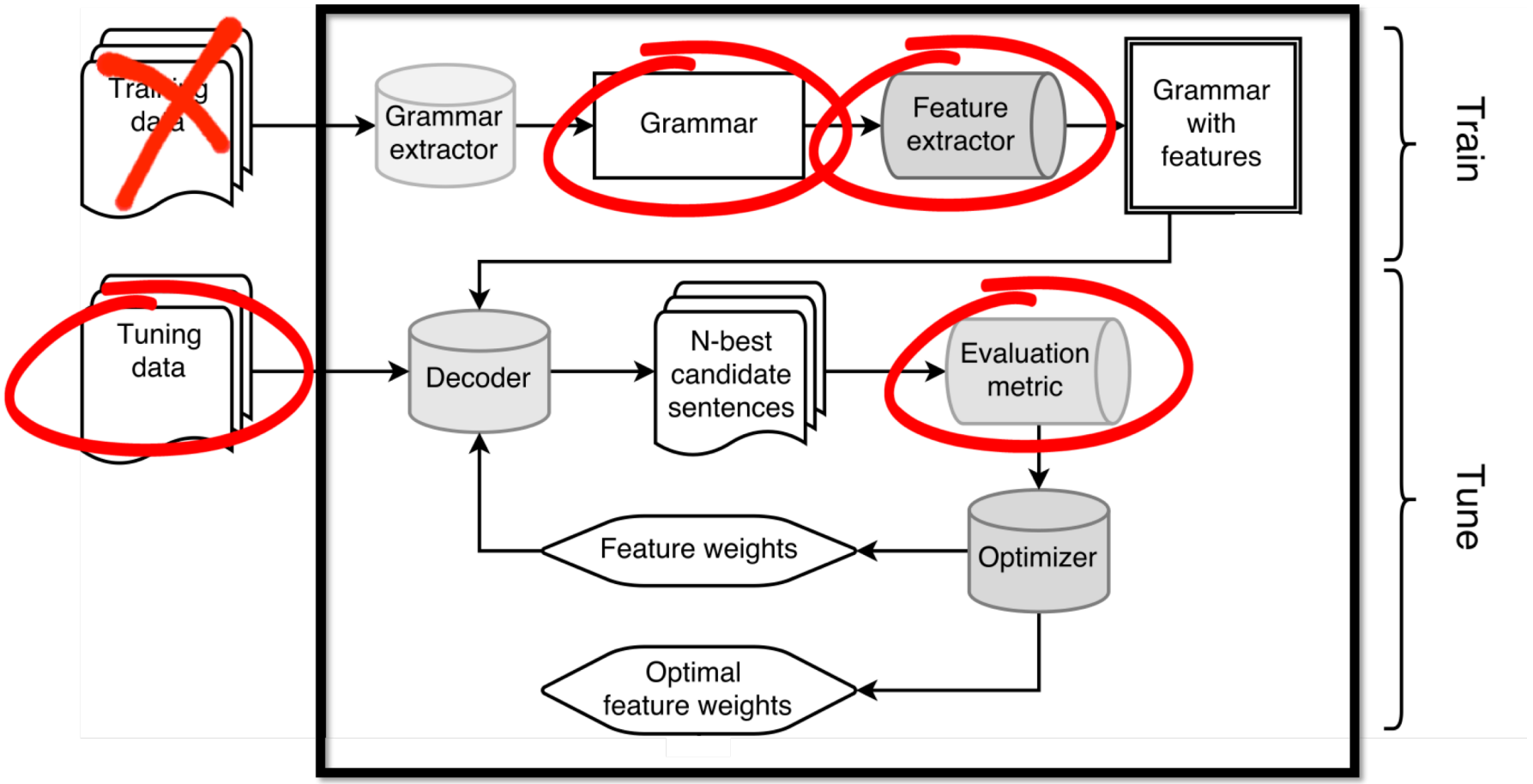
- “Naturally” occurring: Multiple-reference translations
- Multiple-Translation Chinese Corpus (LDC2002T01)
- Pair the longest and the shortest references
- 1500 sentence pairs with 50 < CR ≤ 80

Longest	Later, some soldiers arrived in an armored personnel carrier and rescued the seriously wounded man.
Shortest	The severely wounded man was later rescued by an armored carrier.

Napoles et al., 2011b. *Paraphrastic Sentence Compression with a Character-based Metric*

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification			
Grammatical error correction			

Generating abstractive compressions



Generating abstractive compressions

Adaptations

- **Grammar:** Large, general-purpose paraphrase corpus (precursor to PPDB, Ganitkevitch et al., 2013)
- **Task-specific features**
- **Objective function:** Penalize sentences above target CR

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Paraphrase corpus

- Instead of extracting a grammar from parallel corpus, we use a large paraphrase corpus (23M paraphrases).
- Corpus represents many types of transformations, and we calculate task-specific features to select appropriate paraphrases for this task.

Original *o*	**Paraphrase candidates *c***
ancient	antique ancestral old age-old longstanding archaic

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Paraphrase corpus

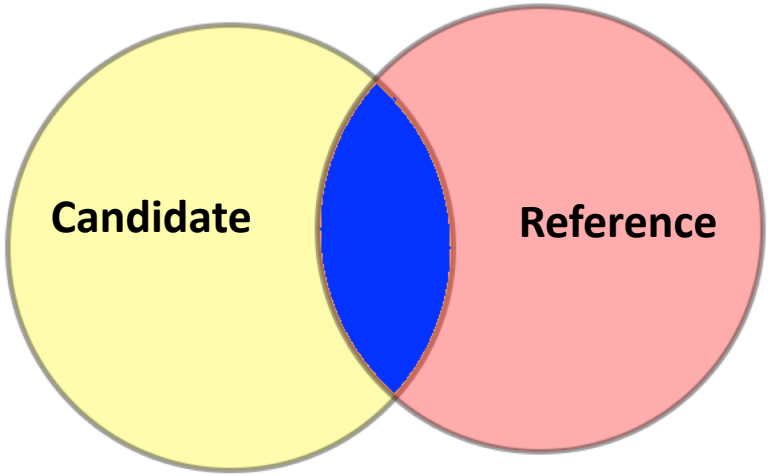
- Instead of extracting a grammar from parallel corpus, we use a large paraphrase corpus (23M paraphrases).
- Corpus represents many types of transformations, and we calculate task-specific features to select appropriate paraphrases for this task.

		Features		
Original *o*	**Paraphrase candidates *c***	**length(*o*)**	**length(*c*)**	**length(*c*) – length(*o*)**
ancient	antique	7	7	0
	ancestral	7	9	2
	old	7	3	-4
	age-old	7	7	0
	longstanding	7	12	5
	archaic	7	7	0

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Objective Function

- BLEU (Papineni et al., 2002)
- BLEU compares the overlap of the generated candidate sentences to reference “gold-standard” sentences.
- Captures *fluency* and *adequacy*



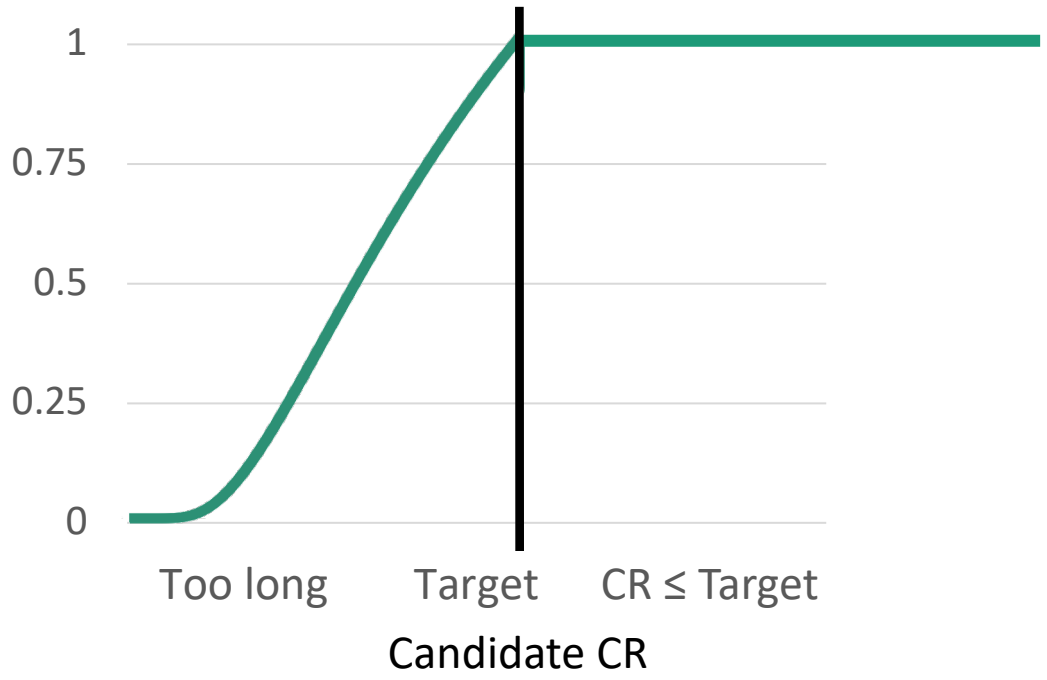
Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Objective Function

- New metric: PRÉCIS
- Add a length penalty to BLEU

PRÉCIS penalty

$$e^{\lambda(\varphi-\frac{c}{i})}$$



Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Generating Compressions

- Syntax-based statistical MT
- Joshua Toolkit (Li et al., 2010)
- Tune on 1k sentences from our abstractive corpus
- Test on 500 sentences from our corpus

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Other systems

Random deletion baseline

- Randomly delete 1/2 words

Integer linear programming (ILP; Clarke and Lapata, 2008)

- Uses a set of constraint features to find best deletions via ILP

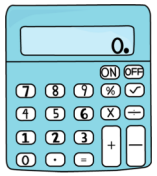
Tree-to-tree transducer (T3; Cohn & Lapata, 2007)

- Uses a synchronous tree-substitution grammar to delete constituents

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Evaluation

3 dimensions to evaluate:

1. Is the output shorter?  *Automatic metric*

2. Is it grammatical?  *Human judgments*

3. Does it retain the most important meaning? 

Human Evaluation

Two 5 point scales:

- Grammar

  How grammatical is this sentence?

- Meaning

  How well does this sentence retain the meaning of the original?

3 judges rated each sentence.

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Judge the quality of similar sentences

Please read the groups of sentences below. Each group contains an original sentence along with five different versions of it, which are created automatically by a computer. Your job is to grade the quality of the automatically generated sentences with two 5-point scales: **meaning** and **grammar**.

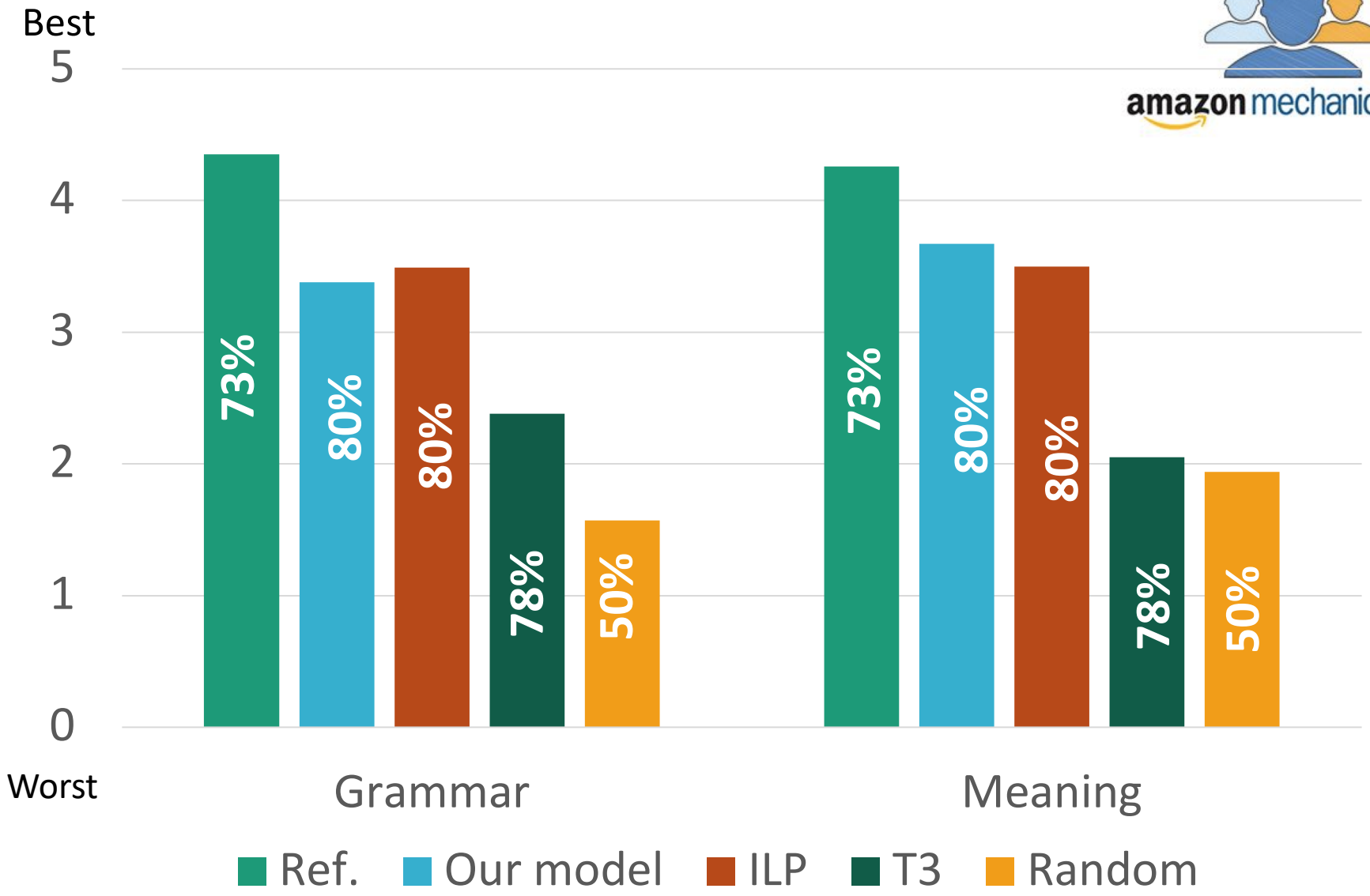
Meaning

- **5 - perfect**: All of the meaning of the original sentence is retained, and nothing is added
- **4 - minor differences**: The meaning of the original sentence is retained, although some minor information may be deleted or added without too great a difference in meaning
- **3 - moderate differences**: Some of the meaning of the original sentence is retained, although a non-trivial amount of information was deleted or added
- **2 - substantially different**: Substantial amount of the meaning is different
- **1 - completely different**: The new sentence doesn't mean anything close to the original sentence

Grammar

- **5 - perfect**: The new sentence is perfectly grammatical
- **4 - ok but awkward**: The sentence is grammatical, but might sound slightly awkward
- **3 - one error**: The sentence has an error (such as an agreement error between subject and verb, a plural noun with a singular determiner, or the wrong verb form)
- **2 - many errors**: The sentence has multiple errors or omits words that would be required to make it grammatical
- **1 - ungrammatical**: The sentence is totally ungrammatical

Results

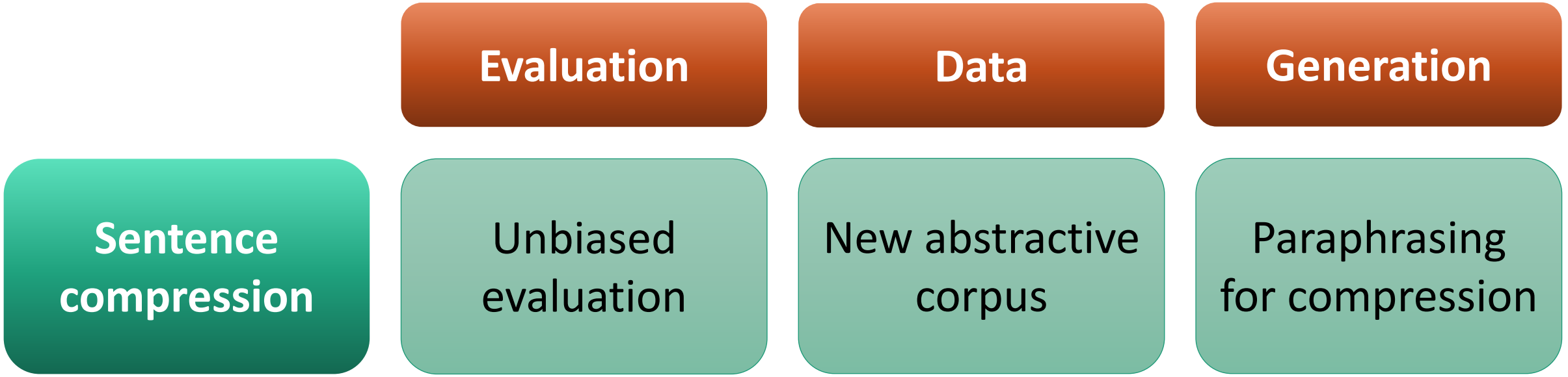


Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*

Example output

Source	In this war which has carried on for the last 12 days around 700 Palestinians which include a large number of women and children, have died
Reference	About 700 Palestinians mostly women and children have been killed in the Israeli offensive over the last 12 days
Paraphrase	In this war [redacted] has done for the last 12 days [redacted] around 700 Palestinians, including [redacted] women and children, [redacted] died.
ILP	In this war which has carried [redacted] for the [redacted] days [redacted] Palestinians which include a [redacted] number of women and children [redacted] died.

Ganitkevitch et al., 2011. *Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation*



Summary

- Compression rate correlates with quality.
- Earlier evaluation has been biased.
- Compression with paraphrasing retains more meaning than just deletions.

Text Simplification

Joint work with Chris Callison-Burch and Benjamin Van Durme

Publications

Napoles and Dredze
Learning Simple Wikipedia
CL&W Workshop, 2010

Napoles
Computational Approaches to Shortening and Simplifying Text
JHU Qualifying Project, 2012

Xu, **Napoles**, Pavlick, Chen and Callison-Burch
Optimizing Statistical Machine Translation for Text Simplification
TACL, 2016

Background

Transform text with simplifying syntactic and lexical transformations so that it is easier to read.

Many operations:

- Lexical substitution
- Sentence splitting
- Syntactic substitution
- Some compression

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric		
Grammatical error correction			

Readability metric

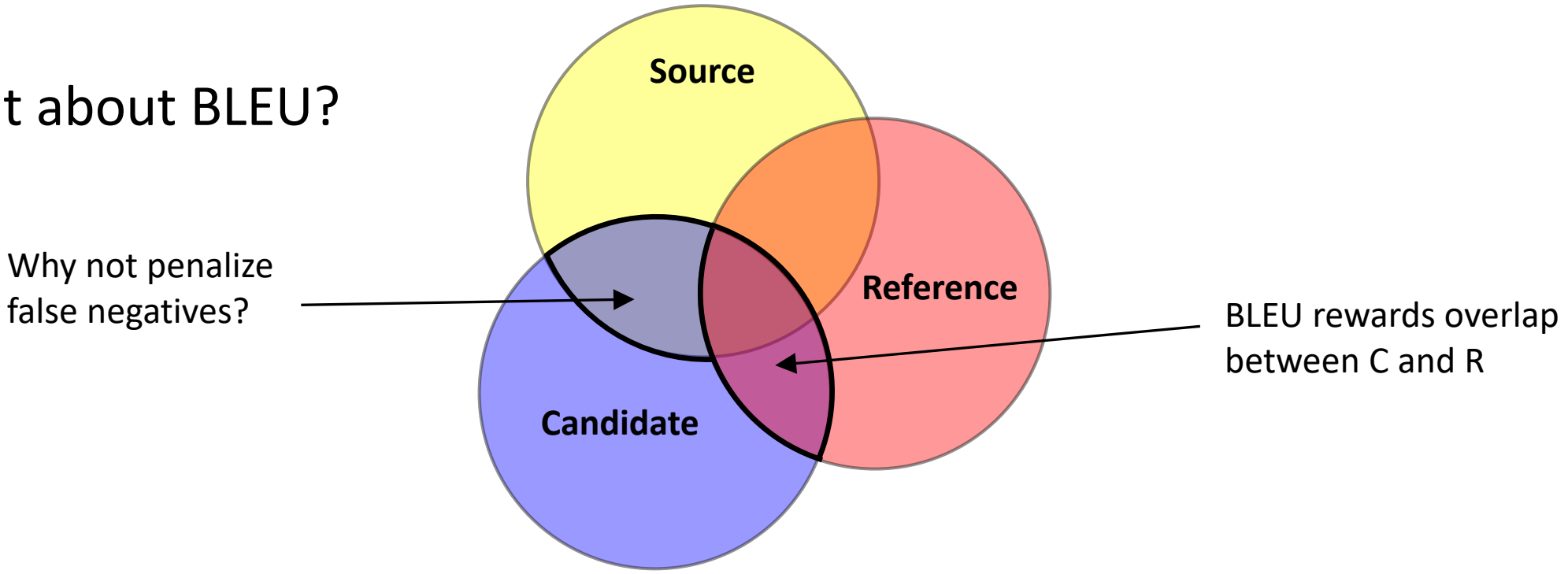
- Flesch-Kincaid Grade Level (Kincaid et al., 1975)
- Grade-level of education necessary to understand a text

$$\text{FKGL} = 0.39\left(\frac{\text{number words}}{\text{number sentences}}\right) + 11.8\left(\frac{\text{number syllables}}{\text{number words}}\right)$$

- Parameters set based on test results of 530 Navy personnel

Objective function

- What about BLEU?



- There is too much copying between the source and the candidate
- For monolingual tasks, evaluation should account for Reference **and Source**.

Napoles, 2012. Computational Approaches to Shortening and Simplifying Text

New objective function

Grade-Level iBLEU (GLiB)

- Take insights from BLEU (fluency and adequacy)
- Adapt it to penalize overlap with source (input-aware BLEU, *iBLEU; Sun and Zhao, 2012)*)

$$i\text{BLEU}(S, R, C) = \alpha\text{BLEU}(C, R) - (1 - \alpha)\text{BLEU}(C, S)$$

- Penalize input that isn't more readable

$$\text{GLiB} = \begin{cases} \text{iBLEU} & \text{if FKGL}(C) < \lambda\text{FKGL}(S) \\ 0 & \text{otherwise} \end{cases}$$

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	

Data for Simplification

Simple English
WIKIPEDIA

- Naturally occurring data is abundant
- Simple English Wikipedia contained 60k articles in 2010
- Parallel WiKiPedia corpus (PWKP, Zhu et al., 2010)
  - 190k automatically aligned sentence pairs from English and Simple English Wikipedia
- After this work, the Newsela corpus was released
  - Each sentence is rewritten at 4 different grade levels
  - Xu, Callison-Burch, and **Napoles**, TACL 2015

newsela

Simple English Wikipedia: Not so simple…

- 4000 sentence pairs from PWKP
- Compare readability of sentence on Mechanical Turk
- Human judges chose which sentence from each pair was more simple.
- 3 judges compared each sentence pair

Which sentence is more simple?

In this HIT, you will read pairs of sentences. Your task is to decide which of the two sentences is more "simple", or easier to read and understand. The options are:

- Sentence 1 is more simple (easier to read) than Sentence 2.
- Both sentences are the same difficulty.
- Sentence 1 is more difficult than Sentence 2.

In this HIT, the two sentences should be describing the same thing and one is a simplified version of the other. However, some sentence pairs may describe **different** things. If you think that this is the case, check the box next to "Neither sentence is a simplification of the other".

Simple English Wikipedia: Not so simple…

- In 64% of sentence pairs, the "simplified" version was rated **at least as difficult** as the original sentence.
- We created a **verified** tune and test set from the remaining 1400 sentences.

	Original		Verified	
Data	**FKGL**	**Length (words)**	**FKGL**	**Length (words)**
Simple Wikipedia	10.5	21	9.5	16
English Wikipedia	12.5	18	13.3	24
Difference	**-2**	**+3**	**-3.8**	**-8**

Napoles, 2012. Computational Approaches to Shortening and Simplifying Text

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification

Simplification grammars

- Other work uses paraphrase rules learned from Simple Wikipedia
  - RevILP (Woodsend and Lapata, 2011) – uses rules learned from edit history
  - TSM (Zhu et al., 2010) – tree-based translation model, rules from PWKP
- This work: general paraphrase corpus we used for compression

Grammar Source	# Rules
PWKP	700k
Paraphrase	5M

Simplification grammars

PWKP grammar:

ancient ⇒ old

Paraphrase grammar:

ancient ⇒ old

ancient ⇒ classic

ancient ⇒ out-dated

ancient ⇒ age-old

…

Modifications for MT

1. Grammar
2. Tuning set
   - 1400 verified PWKP sentences
3. Task-specific features
   - Counts of “Basic English” words (Ogden, 1930)
   - Character, word, and syllable lengths
   - Gigaword and Simple English LM scores
4. Objective function
   - GLiB

Joshua MT Toolkit for decoding and optimization

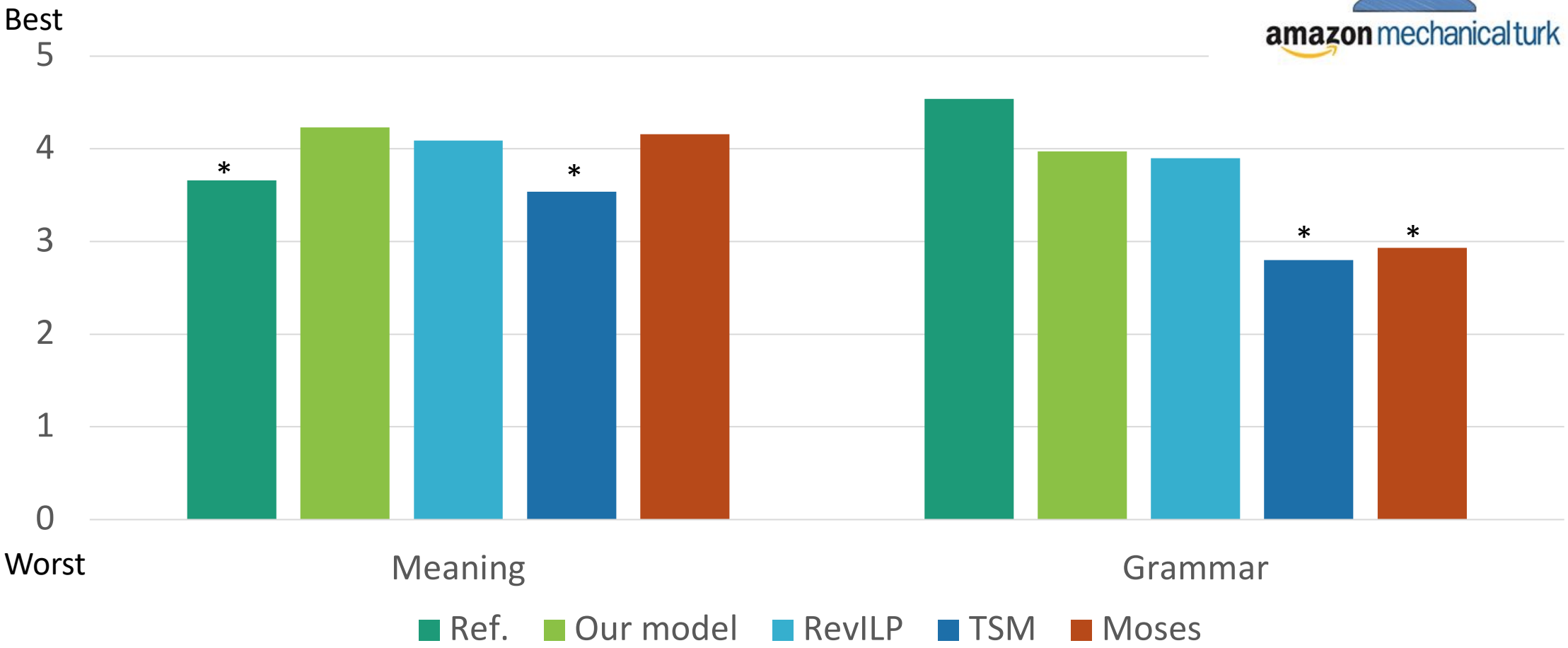
Other systems

- RevILP (Woodsend and Lapata, 2011)
  - Constraint-based system rewarding shorter sentences, with transformations extracted from Simple Wikipedia edit history
- TSM, Tree-based simplification model (Zhu et al., 2010)
  - 3 types of transformations extracted from PWKP
- Baseline: Moses (black box)
  - Data: PWKP
  - Metric: BLEU

Evaluation

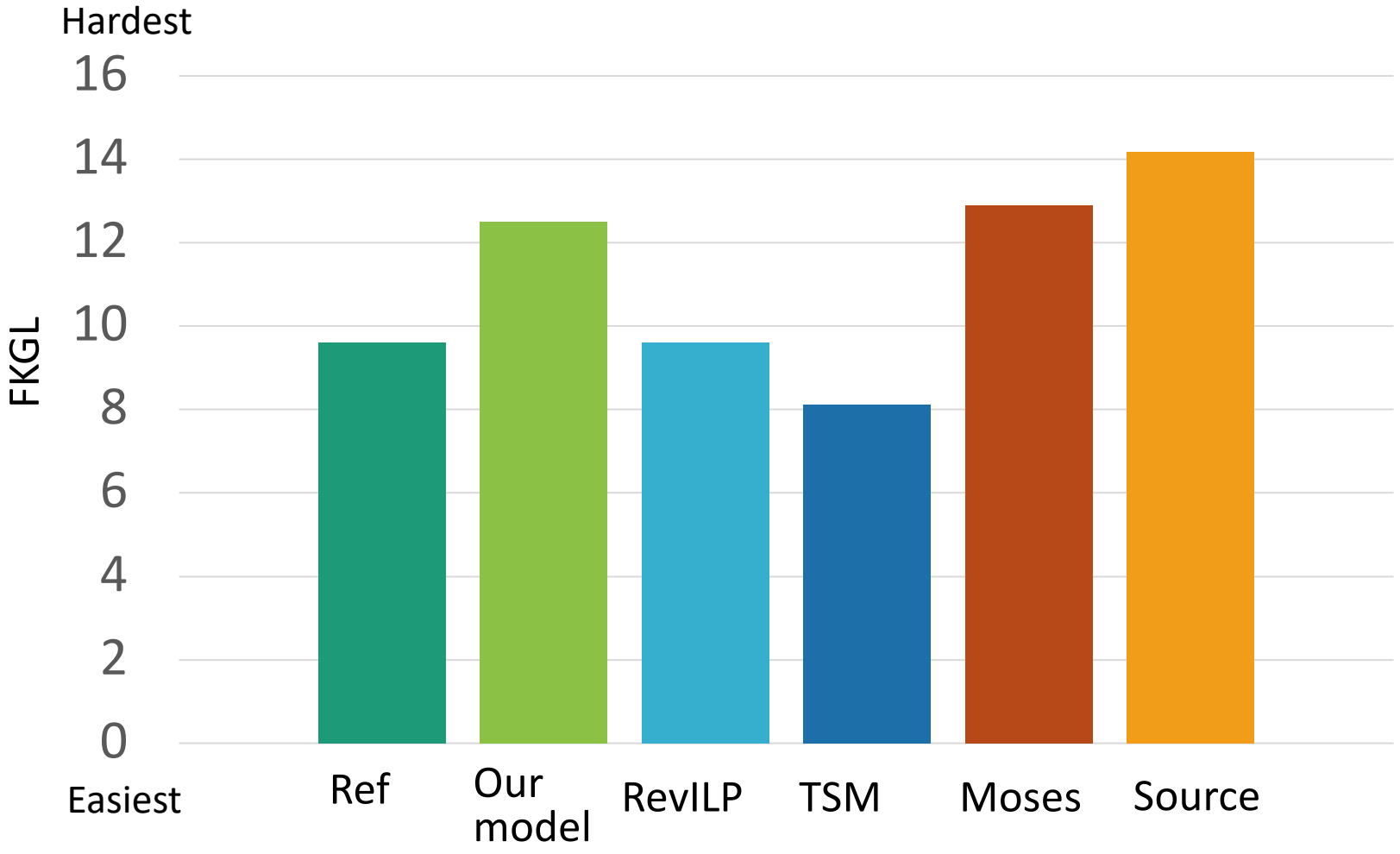
- We collected 2 sets of human judgments comparing the outputs to the original sentence:
  1. Judge the grammar and meaning of simplified sentences.
  2. Rank the simplicity of sentences.
- 100-sentence PWKP test set
- Mechanical Turk
- 3 ratings for each judgment

Human Evaluation



* Our model is significantly better

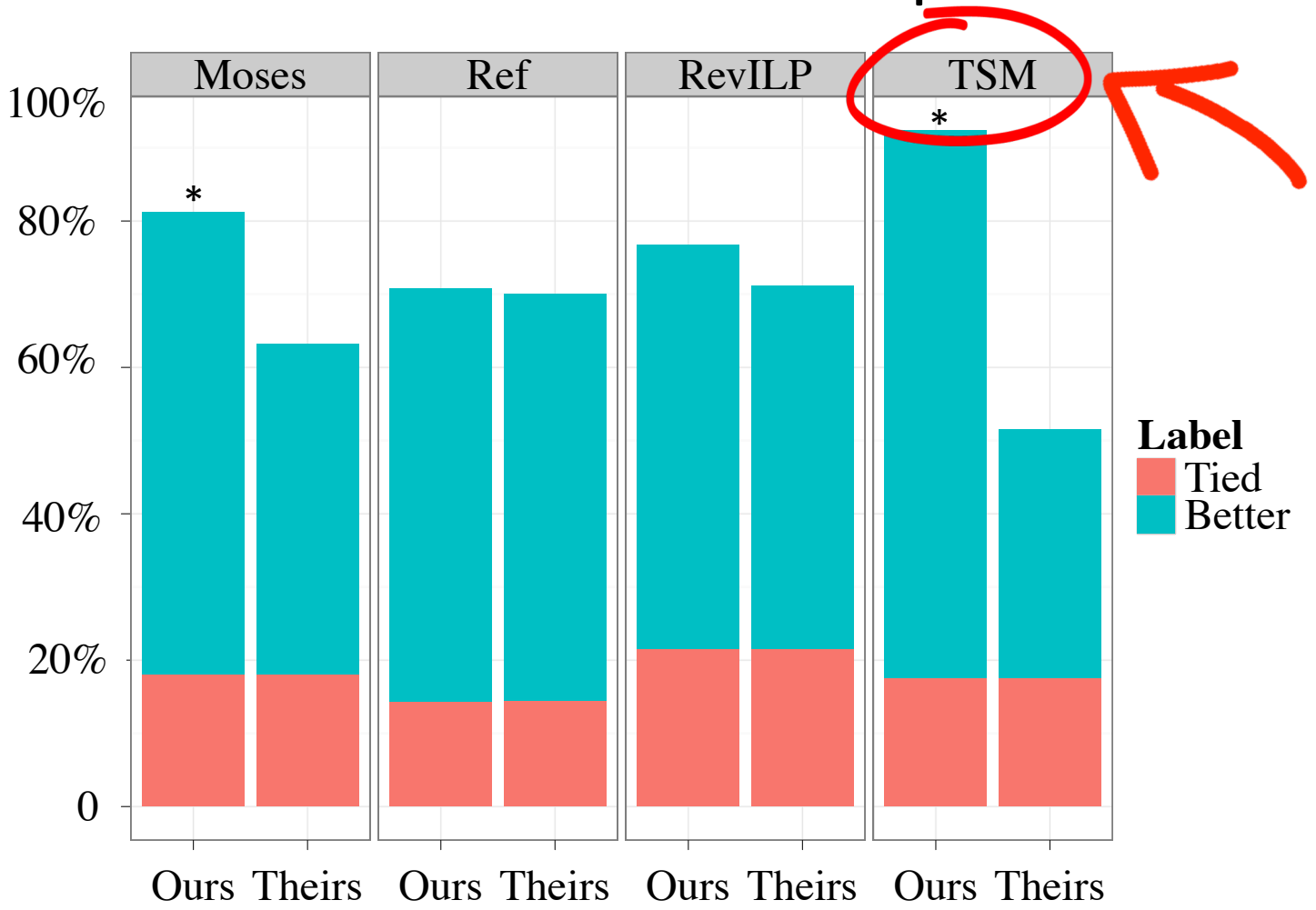
Automatic Metric (FKGL)



Our Paraphrase model (PP) and Moses have higher FK because they **do not** split sentences.

Napoles, 2012. Computational Approaches to Shortening and Simplifying Text

Which sentence is more simple?



TSM's FKGL is 3 grades lower but our model is significantly more simple.

Napoles, 2012. Computational Approaches to Shortening and Simplifying Text

Example output

Source	The combination of new weapons and tactics have caused many historians to consider this battle the beginning of the end of chivalry.
Paraphrase	The mix of [redacted] weapons and tactics have caused many historians to see this battle the beginning of the end of chivalry.
RevILP	It was The combination of new weapons and tactics. It have caused many historians to consider this battle the start of the end of chivalry.
TSM	The combination of [redacted] weapons and tactics have caused many historians. To consider this battle the beginning of the end of knight.

	Evaluation	Data	Generation
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification

Summary

- Created verified development corpus
- Existing readability metrics not ideal for evaluation.
  - Fewer sentences does not mean more readable.
- Our model is as good or better as alternate models.
  - Does not require large parallel corpus.

Grammatical error correction

Joint work with Keisuke Sakaguchi, Joel Tetreault, Matt Post, and Chris Callison-Burch

Publications

Napoles, Sakaguchi, Post, and Tetreault (ACL 2015)
Ground Truth for Grammatical Error Correction Metrics

Sakaguchi, **Napoles**, Post, and Tetreault (TACL 2016)
Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality

Napoles, Sakaguchi, and Tetreault (EMNLP 2016)
There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction

Napoles, Sakaguchi, Post, and Tetreault (EACL 2017)
JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction

Napoles and Callison-Burch (BEA Workshop 2017)
Systematically Adapting Machine Translation for Grammatical Error Correction

What is the goal of GEC?

Historically, GEC aims to correct small targeted errors

- E.g., preposition, punctuation, or determiner mistakes.

Supported by data, metrics, and shared tasks:

- Corpora
  - Cambridge Learner Corpus (CLC; Nicholls et al., 2003)
  - CLC First Certificate in English (FCE; Yannakoudakis et al., 2011)
  - NUS Corpus of Learner English (NUCLE; Dahlmeier et al., 2013)
- Shared tasks:
  - HOO 2011, HOO 2012, CoNLL 2013 targeted limited error types (Dale and Kilgariff, 2011; Dale et al., 2012; Ng et al., 2013).

Until 2014: CoNLL Shared Task included *all* 27 error types (Ng et al., 2014).

What is the goal of GEC?

Even after these mistakes are corrected, the sentences are not *fluent*.

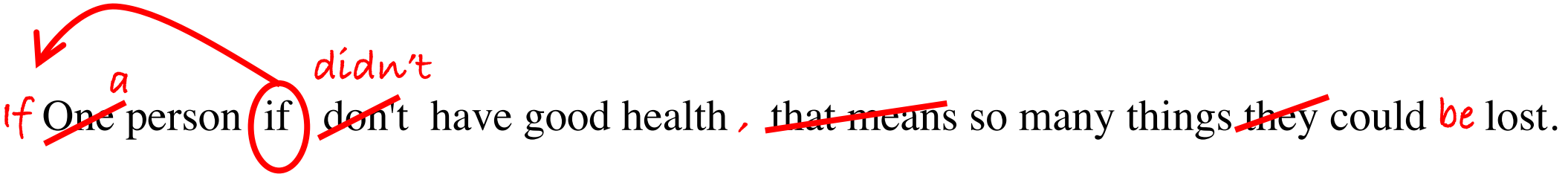
Original

One person if don't have good health that means so many things they could lost.

Correction

One person, if they don't have good health, that means so many things ~~they~~ could be lost.

Better, fluent correction



Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

What is the goal of GEC?

We argue GEC should correcting sentences so they have native-sounding ***fluency***, instead of just making ***minimal edits***.

Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

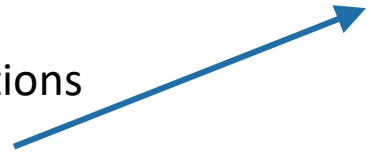
	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification
Grammatical error correction	Fluency evaluation		

GEC Evaluation M

Designed for “minimal-edi

- Existing annotated corpo

Sample annotations from NUCLE



```
<ANNOTATION teacher_id="8">
<MISTAKE start_par="1" start_off="42" end_par="1" end_off="46">
<TYPE>ArtOrDet</TYPE>
<CORRECTION></CORRECTION>
</MISTAKE>
<MISTAKE start_par="1" start_off="118" end_par="1" end_off="125">
<TYPE>Nn</TYPE>
<CORRECTION>diseases</CORRECTION>
</MISTAKE>
<MISTAKE start_par="1" start_off="620" end_par="1" end_off="627">
<TYPE>Trans</TYPE>
<CORRECTION>However,</CORRECTION>
</MISTAKE>
<MISTAKE start_par="1" start_off="740" end_par="1" end_off="751">
<TYPE>Mec</TYPE>
<CORRECTION>diagnosed</CORRECTION>
</MISTAKE>
<MISTAKE start_par="1" start_off="751" end_par="1" end_off="754">
<TYPE>Prep</TYPE>
<CORRECTION></CORRECTION>
</MISTAKE>
<MISTAKE start_par="1" start_off="776" end_par="1" end_off="783">
<TYPE>Nn</TYPE>
<CORRECTION>diseases</CORRECTION>
</MISTAKE>
```

GEC Evaluation Metrics

- Designed for “minimal-edit” corpora
  - Existing annotated corpora: CLC, FCE, and NUCLE.
- Metrics in shared tasks:
  - F-score (HOO 2012)
  - M^2 (CoNLL 2013, CoNLL 2014)
    - $F_{0.5}$-score over alignment lattice
- Recently proposed metric: I-measure (Felice and Briscoe, 2015)
  - Reflects how much edits improve/degrade grammaticality of source

GEC Evaluation Metrics

None of these metrics work for fluency edits.

- They require aligned text and error spans.
- Fluency edits contain rewrites and phrasal movements.

None of these metrics have been validated.

A new metric for GEC

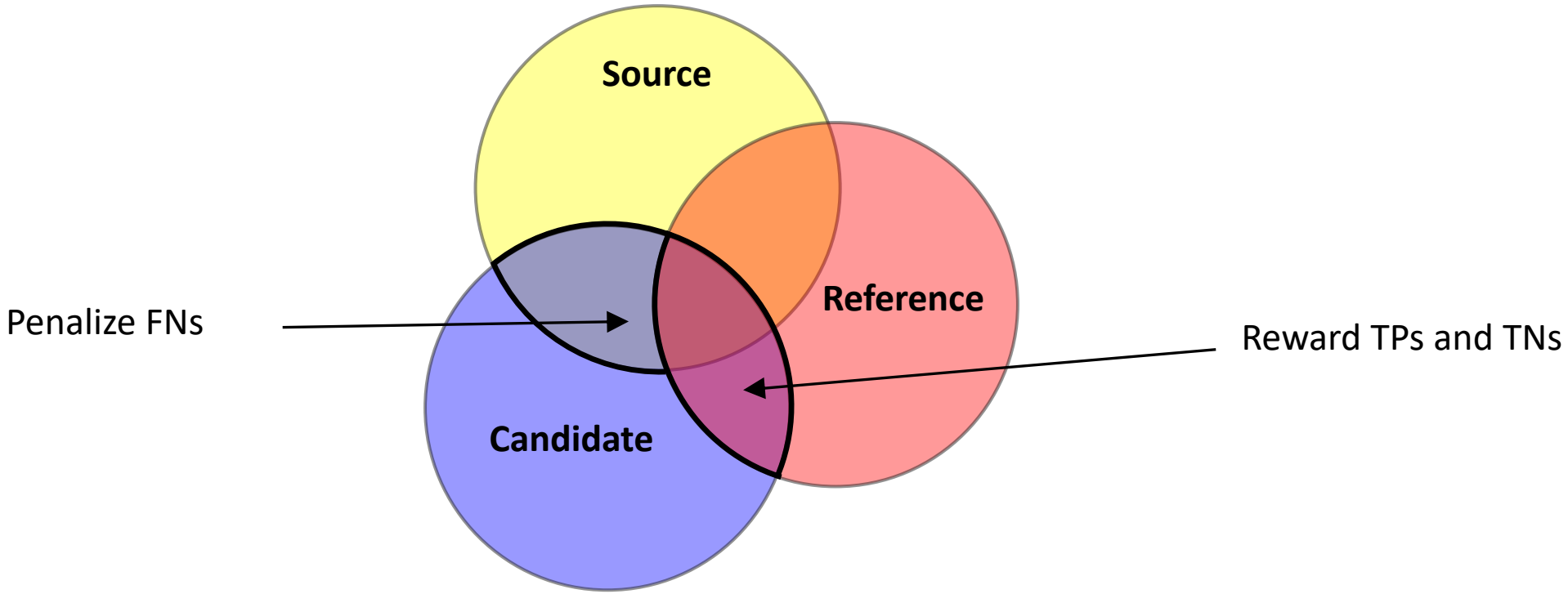
GLEU: Generalized Language Evaluation Understanding

- Like BLEU, calculate precision over n-grams present in the reference and the candidate.
- Extra weight to n-grams present in the reference and the candidate but *not* the source.
- Penalize n-grams present in the source and the output *but not* the reference.

Napoles, et al., 2015. Ground Truth for Grammatical Error Correction Metrics

A new metric for GEC

$$p_n^* = \frac{\left(\sum_{ngram \in \{C \cap R\}} count_{C,R}(ngram) - \sum_{ngram \in \{C \cap S\}} \max\left[0, count_{C,S}(ngram) - count_{C,R}(ngram)\right] \right)}{\sum_{ngram \in \{C\}} count(ngram)}$$



Napoles, et al., 2015. Ground Truth for Grammatical Error Correction Metrics

Which metric is best?

- System outputs from CoNLL-2014 GEC shared task released.
  - 1312 sentences, 13 systems
- We ran WMT-style evaluation of CoNLL-2014 system outputs.
  - Sentences from the NUCLE corpus (Dahlmeier et al., 2013)
- Authors ranked system outputs.
- Inferred absolute system ranking from pairwise judgments with TrueSkill (Herbrich et al., 2006; Sakaguchi et al., 2014.)

Human Evaluation

Napoles, et al., 2015. Ground Truth for Grammatical Error Correction Metrics

System rankings

Human	BLEU	I-measure	M^2	GLEU
CAMB	UFC	UFC	CUUI	CUUI
AMU	source	source	CAMB	AMU
RAC	IITB	IITB	AMU	UFC
CUUI	SJTU	SJTU	POST	CAMB
source	UMC	CUUI	UMC	source
POST	CUUI	PKU	NTHU	IITB
UFC	PKU	AMU	PKU	SJTU
SJTU	AMU	UMC	RAC	PKU
IITB	IPN	IPN	SJTU	UMC
PKU	NTHU	POST	UFC	NTHU
UMC	CAMB	RAC	IPN	POST
NTHU	RAC	CAMB	IITB	RAC
IPN	POST	NTHU	source	IPN

Napoles, et al., 2015. Ground Truth for

Which metric is best?

- Calculate correlation of metric scores with human “score”

Metric	Pearson’s	Spearman’s
BLEU	-0.12	-0.23
I-measure	-0.05	-0.01
M^2	0.36	0.43
GLEU	**0.54**	**0.56**

Napoles, et al., 2015. Ground Truth for Grammatical Error Correction Metrics

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification
Grammatical error correction	Fluency evaluation	New fluency corpora	

Fluency references

Fluency metric is better than minimal-edit metrics.

What about fluency *references*?

Only 1 parallel fluency corpus exists, Lang-8 (Mizumoto et al., 2011)

- Very noisy
- Social media site, users correct other users' text
- Automatically aligned

Fluency references

We collected 8 additional references for the CoNLL-14 test sentences.

 # New references collected

Annotator	Fluency	Min Edit	
Expert	1312 * 2	1312 * 2	
Non-expert	1312 * 2	1312 * 2	
Total references	**5248**	**5248**	**10,496**

- Experts: Native-speaking authors
- Non-experts: Mechanical Turk with quality controls

Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

Fluency references

Fluency rewrites

Sentence Rewriting Challenge: Instructions

Please read the following instructions carefully!

Please correct the following sentences **to make them sound more natural to a native speaker of English**. You should also fix grammatical mistakes, but focus on changing the sentences to remove awkward phrases and to follow standard written usage conventions. Not all sentences require corrections.

The sentences are taken from student essays. For context, the previous sentence in the essay is also presented, greyed out.

Minimal edits

Please correct grammatical mistakes in the following sentences. In many cases, there are lots of ways to correct a sentence; please prefer corrections that require **the smallest number of changes to the original sentence**, even if the resulting sentence is slightly awkward or non-native sounding. Not all sentences require corrections.

Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

Evaluating with fluency references

We scored the CoNLL-14 system outputs with every combination of metric and reference set.

References

NUCLE

Bryant and Ng, 2015 (10 min-ed refs)

Expert min-ed

Expert fluency

Non-expert min-ed

Non- fluency

Metrics

BLEU

GLEU

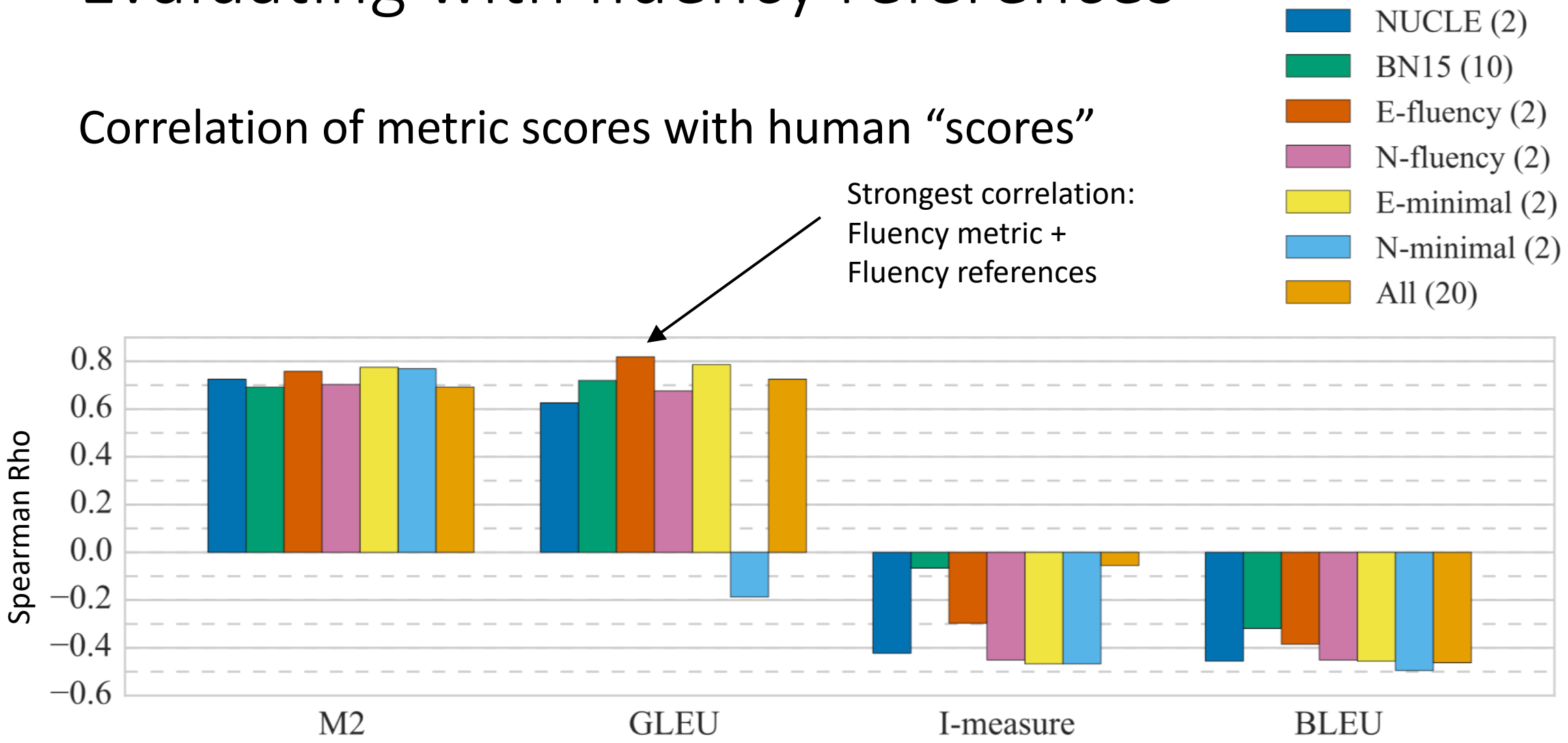
I-measure

M^2

Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

Evaluating with fluency references

Correlation of metric scores with human “scores”



Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error

Fluency corpora for GEC

- The best automatic evaluation uses fluency references and fluency metric.
- We created a new, multiple-reference dataset with fluency corrections.
  Sentences are from NUCLE and used as the CoNLL-2014 test set

Sakaguchi, et al., 2016. Reassessing the Goals of Grammatical Error Correction

A representative fluency corpus

- NUCLE is only representative of a small percentage of English text.
  - Students from National University of Singapore
  - English language learners
  - Student essays over constrained topics
- We constructed a new corpus for development and testing:

  JFLEG: JHU FLuency-Extended GUG corpus
- GUG Corpus (Heilman et al., 2014)
  - Sentences written by English language learners
  - Variety of essay topics, L1s, and English proficiency levels

Napoles et al., 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction

A representative fluency corpus

For 1500 sentences in GUG dev/test set, we collected 4 reference corrections on Mechanical Turk.

- Annotators passed a qualifying test and were monitored for quality.

Sentence Rewriting Challenge: Instructions

Please read the following instructions carefully!

Please correct the following sentences **to make them sound more natural to a native speaker of English**. You should also fix grammatical mistakes, but focus on changing the sentences to remove awkward phrases and to follow standard written usage conventions. Not all sentences require corrections.

The sentences are taken from student essays. For context, the previous sentence in the essay is also presented, greyed out.

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification
Grammatical error correction	Fluency evaluation	New fluency corpora	SMT + artificial rules

SMT for GEC

- We have shown that SMT with a paraphrase grammar outperforms other approaches with less task-specific training data.
- Does it work for GEC?

Related work:
Junczys-Dowmunt and Grundkiewicz, 2016

Approach

Paraphrasing corpus does not capture error corrections.

1. Extract grammar from parallel corpus (Lang-8)
2. Augment grammar
3. New features based on edit analysis
4. Parameter optimization with GLEU fluency metric
5. Tune on fluency data (JFLEG)

Augment grammar

Artificially generated rules

- Spelling corrections (py-enchant)
  - Weighted by Gigaword frequency
- Morphological transformations (RASP)
  - Nouns: switch plurality
  - Verbs: switch inflected forms
  - Uniform weight

detailled → detailed

dog → dogs

dogs → dog

come → comes

come → came

come → coming

Napoles and Callison-Burch, 2017. Systematically Adapting Machine Translation for Grammatical Error Correction

Edit features

Automatically characterize the transformations made

- # spelling errors and corrections
- # deleted, inserted, and substituted words
- # inflection changes and changed lemmas
- Edit distance and length ratio

Edit features

Example rule learned from Lang-8:

Source: argued that

Target: may argue that

Transformation	Value
VB substitution	1
MD insertion	1
Verb error	2
Inflection change	1
Edit distance	5
…	

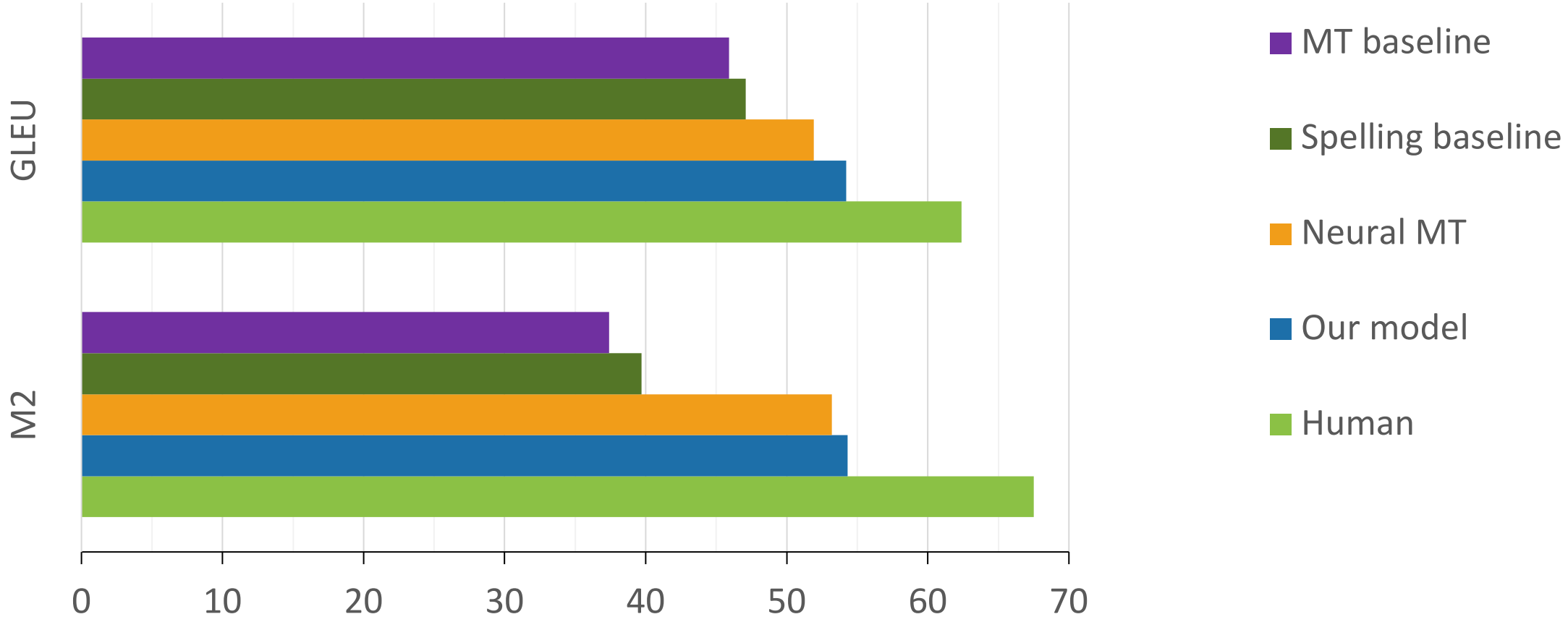
Verb error example:
* In the debate tomorrow, she **argued that**...
In the debate tomorrow, she **may argue that**...

Experiments

Compared 4 systems on JFLEG test set:

- Joshua phrase-based MT (black box) trained on Lang-8 v.1 (1M sents)
- Spelling baseline (py-enchant)
- Neural MT (Yuan and Briscoe, 2016) trained on Lang-8 v.2 (2M sents)
- Our model trained on Lang-8 v.1, using Joshua toolkit

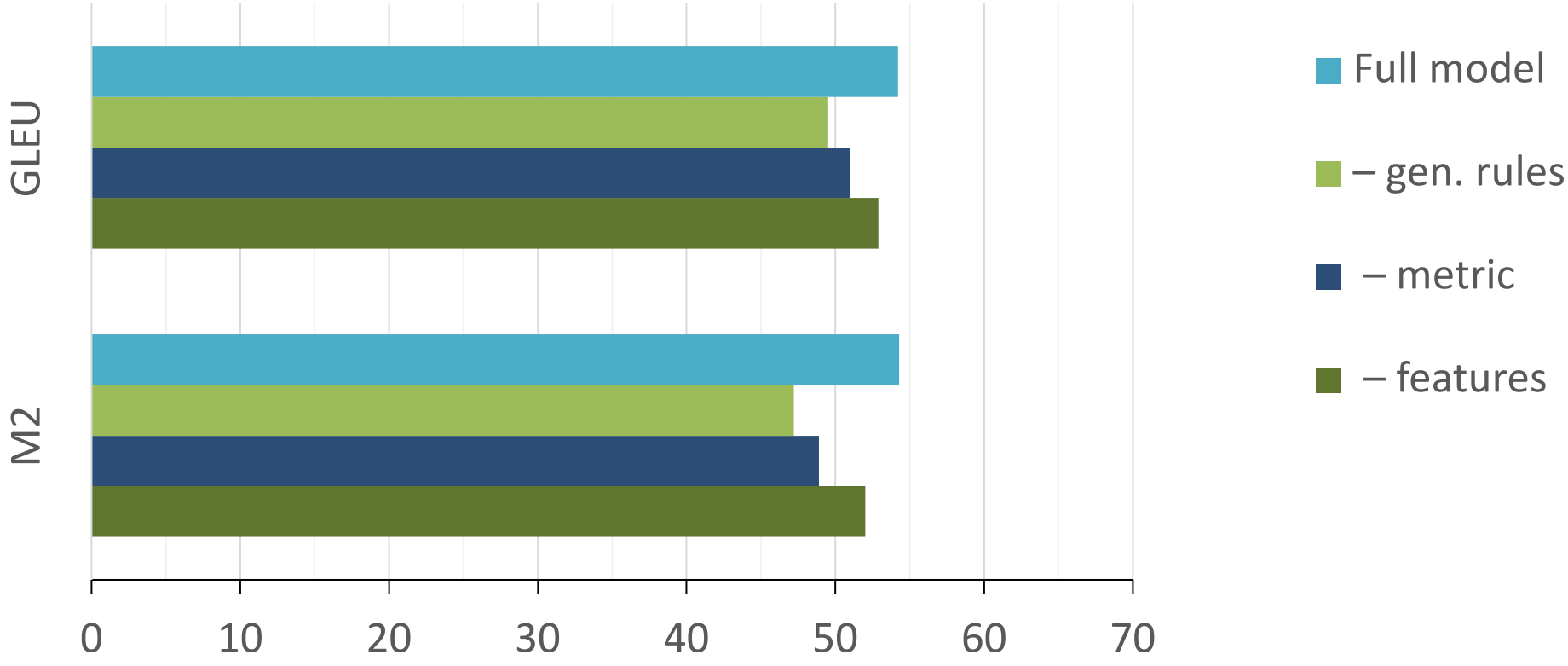
Results



Our model is better than NMT and needs ½ the data.

Napoles and Callison-Burch, 2017. Systematically Adapting Machine Translation for Grammatical Error Correction

Impact of modifications

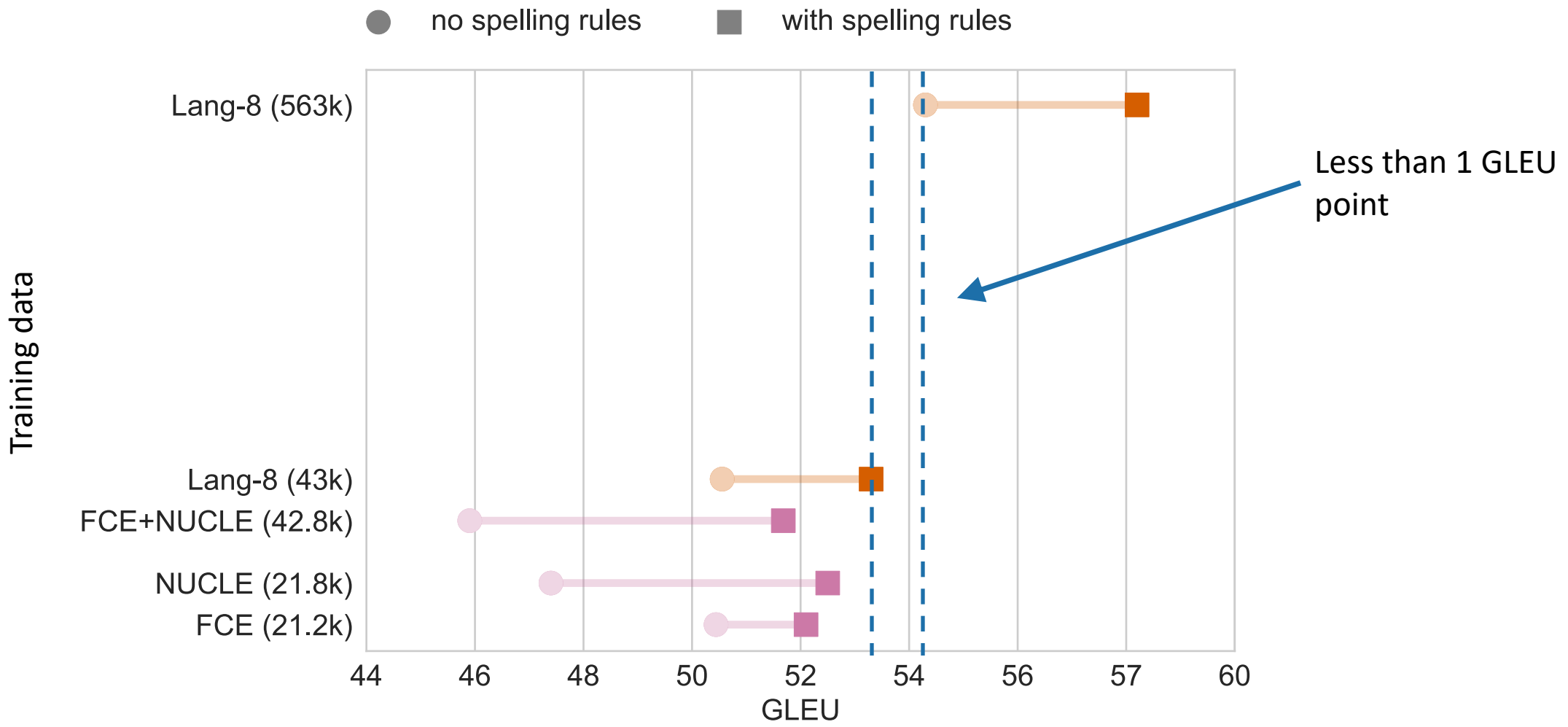


Napoles and Callison-Burch, 2017. Systematically Adapting Machine Translation for Grammatical Error Correction

Impact of spelling rules

- Performance drops 35% without spelling rules.
- Why: Spelling rules decrease sparsity in grammar (lots of unseen mispelligns).
- How many parallel sentences are equivalent to generated spelling rules?

Impact of spelling rules



Napoles and Callison-Burch, 2017. Systematically Adapting Machine Translation for Grammatical Error Correction

Example output

Source	Unforturntly, almost older people can not use internet, in spite of benefit of internet.
Reference	Unfortunately, most older people can not use the internet, in spite of benefits of the internet.
Full model	Unfortunately, most older people can not use the internet, in spite of the benefits of the internet .
– metric	Unfortunately, almost older people can not use internet, in spite of benefit of internet.
– gen. rules	Unforturntly, [redacted] older people can not use the internet, in spite of the benefits of the internet .

	Evaluation	Data	Generation
Grammatical error correction	Fluency evaluation	New fluency corpora	SMT + artificial rules

Summary

- Fluency corrections are the future of GEC.
  - MT advancements support more difficult rewriting operations.
- We have provided a new fluency metric and 2 new fluency datasets.
- For generation, our model has better performance than NMT and uses 1/2 the data.
- Artificial rules provide a big performance gain, though all GEC modifications help.

	Evaluation	Data	Generation
Sentence compression	Unbiased evaluation	New abstractive corpus	Paraphrasing for compression
Text simplification	Better metric	Reliable corpus	Paraphrasing for simplification
Grammatical error correction	Fluency evaluation	New fluency corpora	SMT + artificial rules

Summary and Conclusion

This work makes fundamental contributions to monolingual T2T.

Specifically, we

- Provide a rigorous examination of evaluation methodologies and proposes solutions to shortcomings.
- Verify and develop new datasets appropriate for the given tasks.
- Demonstrate how to exploit existing technologies and monolingual resources to reduce the need for task-specific data.

Thank you!

- Advisors — Chris Callison-Burch and Benjamin Van Durme
- Committee Member — Philipp Koehn
- Collaborators — Joel Tetreault, Keisuke Sakaguchi, Matt Post, Juri Ganitkevitch, Mark Dredze, Aoife Cahill, Nitin Madnani, Aasish Pappu, Matt Gormley, Wei Xu
- Colleagues @ JHU — Ann Irvine, Svitlana Volkova, Ruth Scally, Katie Henry, TeCHo and BLAB members
- NSF Graduate Research Fellowship