Towards a Practically Useful Text Simplification System

Reno Kriz
Advisors: Chris Callison-Burch, Marianna Apidianaki
Thesis Committee: Mitch Marchus (chair), Dan Roth, Lyle Ungar, Eleni Miltsakaki (external)

Penn
UNIVERSITY of PENNSYLVANIA

Goal of Text Simplification

Goal of Text Simplification

- Help someone more easily learn about a topic

Goal of Text Simplification

- Help someone more easily learn about a topic

- Amount of help needed varies
  - Prior background knowledge
  - Ability to learn new concepts

Current Formulation

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, retail fashion giant Abercrombie & Fitch has agreed to change a controversial policy dictating employee dress and grooming in response to discrimination lawsuits filed by two San Francisco Bay Area women.

Current Formulation

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, retail fashion giant Abercrombie & Fitch has agreed to change a controversial policy dictating employee dress and grooming in response to discrimination lawsuits filed by two San Francisco Bay Area women.

Retailer Abercrombie & Fitch has agreed to change controversial rules about how its employees can dress. The change comes after two San Francisco Bay Area women sued the company.

Current Formulation

- Substitution

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, **retail fashion giant** Abercrombie & Fitch has agreed to change **a controversial policy** dictating employee dress and grooming **in response to** discrimination lawsuits filed by two San Francisco Bay Area women.

↓

Retailer Abercrombie & Fitch has agreed to change **controversial rules** about how its employees can dress. The change comes **after** two San Francisco Bay Area women sued the company.

Current Formulation

- Substitution
- Reordering

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, retail fashion giant Abercrombie & Fitch has agreed to change a controversial policy dictating employee dress and grooming in response to **discrimination lawsuits filed by two San Francisco Bay Area women.**

⬇

Retailer Abercrombie & Fitch has agreed to change controversial rules about how its employees can dress. The change comes after **two San Francisco Bay Area women sued** the company.

Current Formulation

- Substitution
- Reordering
- Deletion

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, retail fashion giant Abercrombie & Fitch has agreed to change a controversial policy dictating employee dress and grooming in response to discrimination lawsuits filed by two San Francisco Bay Area women.

Retailer Abercrombie & Fitch has agreed to change controversial rules about how its employees can dress. The change comes after two San Francisco Bay Area women sued the company.

Current Formulation

- Substitution
- Reordering
- Deletion
- Sentence Splitting

Finally moving to end a dispute that besmirched its hip, all-American brand with charges of religious intolerance, **retail fashion giant Abercrombie & Fitch has agreed to change a controversial policy dictating employee dress and grooming in response to discrimination lawsuits filed by two San Francisco Bay Area women.**

↓

Retailer Abercrombie & Fitch has agreed to change controversial rules about how its employees can dress. The change comes after two San Francisco Bay Area women sued the company.

Limitations of Current Formulation

Limitations of Current Formulation

- "Complexity" is not a binary notion
  - What is complex often varies by individual

4	Bear-human conflicts turn serious in Central Florida community.
3	Black bears are moving in on Central Florida neighborhoods.
2	Encounters with black bears on the rise in Central Florida.
1	Central Florida has a big problem with bears.
0	Bears are a problem in Central Florida.

Limitations of Current Formulation

- “Complexity” is not a binary notion
  - What is complex often varies by individual

- Many expressions cannot be satisfactorily transformed into simpler language
  - E.g. *algorithm*, *machine learning*

Thesis Statement

In this thesis, we claim that the textual complexity at any level is not a static value, but instead is influenced both by the surrounding text and especially the knowledge of a particular reader. Thus, in order to create a practically useful simplification system, it is critical to be able to adjust the outputs of our models depending on the situation.

Thesis Outline

Thesis Outline

1. Simplification using Paraphrases and Context-based Lexical Substitution

Thesis Outline

1. Simplification using Paraphrases and Context-based Lexical Substitution

2. Modeling and Evaluation of Sentence Simplification

Thesis Outline

1. Simplification using Paraphrases and Context-based Lexical Substitution
2. Modeling and Evaluation of Sentence Simplification
3. Recasting Text Simplification as a Retrieval Task

Thesis Outline

1. **Simplification using Paraphrases and Context-based Lexical Substitution**
2. Modeling and Evaluation of Sentence Simplification
3. Recasting Text Simplification as a Retrieval Task

Lexical Simplification

- The task of replacing difficult words in a text with simpler ones

Lexical Simplification

- The task of replacing difficult words in a text with simpler ones

- Involves two main processes
    - Identifying complex words

The museum's **director** said there has never been such an **exhibition** of Dutch **portraits**.

Lexical Simplification

- The task of replacing difficult words in a text with simpler ones

- Involves two main processes
  - Identifying complex words
  - Suggesting simpler substitutes that are meaning-preserving

head
manager

The museum's **director** said there has never been such an **exhibition** of Dutch **portraits**.

gallery
show

pictures
sketches
images

Lexical Substitution: Method

- For each identified complex word, extracted simplifying rules from SimplePPDB (Pavlick and Callison-Burch, 2016)

Lexical Substitution: Method

- For each identified complex word, extracted simplifying rules from SimplePPDB (Pavlick and Callison-Burch, 2016)

- Ranked synonyms by adapting the AddCos word embedding model (Melamud et al., 2015)

Lexical Substitution: Method

- For each identified complex word, extracted simplifying rules from SimplePPDB (Pavlick and Callison-Burch, 2016)

- Ranked synonyms by adapting the AddCos word embedding model (Melamud et al., 2015)

- Incorporated several additional features
  - PPDB 1.0 and 2.0 scores (Pavlick et al., 2015)
  - SimplePPDB score

Lexical Substitution: Baselines

WordNet Frequency

WordNet substitutes (Miller et al., 1990) ranked based on Google *n*-gram frequency

Lexical Substitution: Baselines

WordNet Frequency

WordNet substitutes (Miller et al., 1990) ranked based on Google *n*-gram frequency

SimplePPDB Score

SimplePPDB substitutes (Pavlick et al., 2016) ranked based on SimplePPDB score

Lexical Substitution: Baselines

WordNet Frequency

WordNet substitutes (Miller et al., 1990) ranked based on Google *n*-gram frequency

SimplePPDB Score

SimplePPDB substitutes (Pavlick et al., 2016) ranked based on SimplePPDB score

SimplePPDB Score

PPDB substitutes (Ganitkevitch et al., 2013) ranked based on AddCos model

Lexical Substitution: Baselines

WordNet Frequency

WordNet substitutes (Miller et al., 1990) ranked based on Google *n*-gram frequency

SimplePPDB Score

SimplePPDB substitutes (Pavlick et al., 2016) ranked based on SimplePPDB score

SimplePPDB Score

PPDB substitutes (Ganitkevitch et al., 2013) ranked based on AddCos model

RoBERTa SimplePPDB

SimplePPDB substitutes ranked based on cosine similarity of RoBERTa embeddings

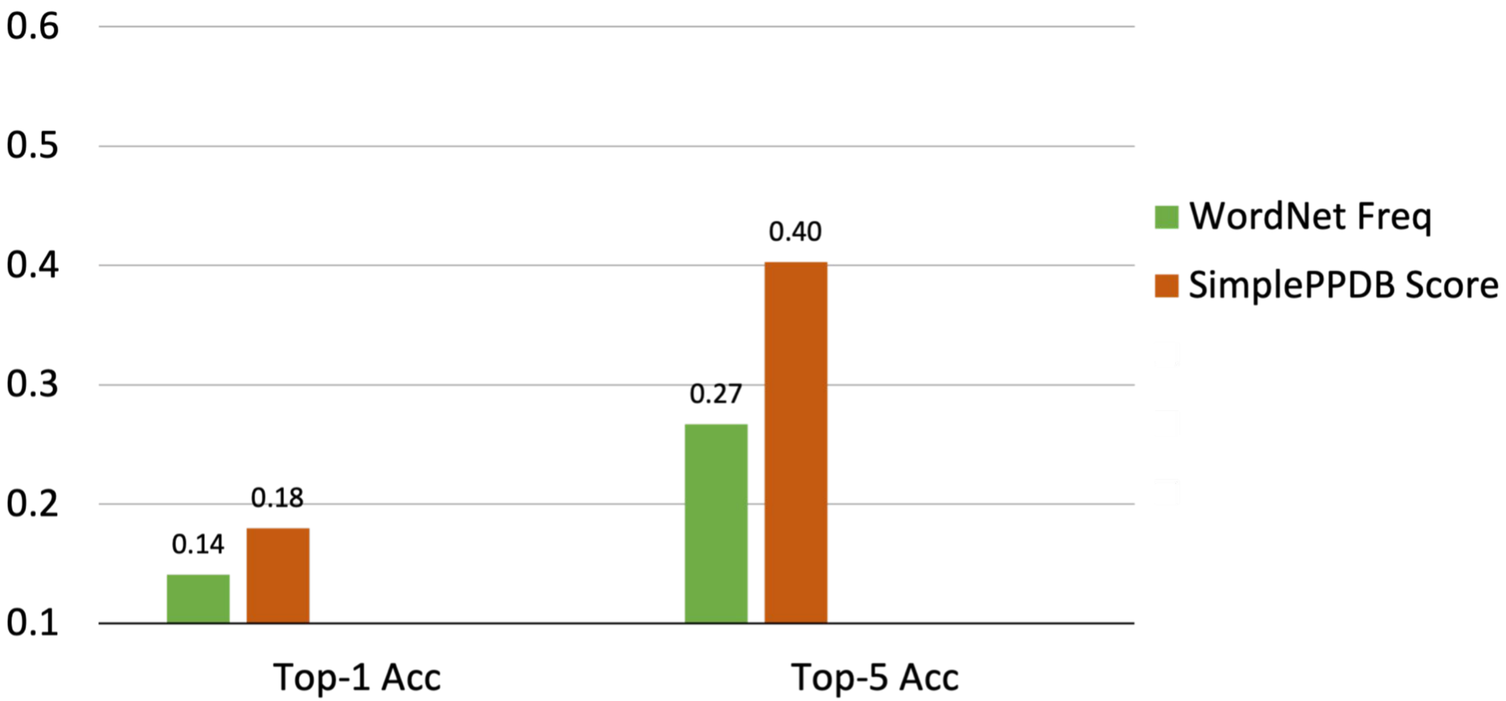
Lexical Substitution: Evaluation Data

- Considered aligned sentence pairs from the Newsela corpus
  - Contains 1,882 articles manually re-written at five complexity levels
  - From these, 141,582 aligned parallel sentence (Xu et al., 2015)

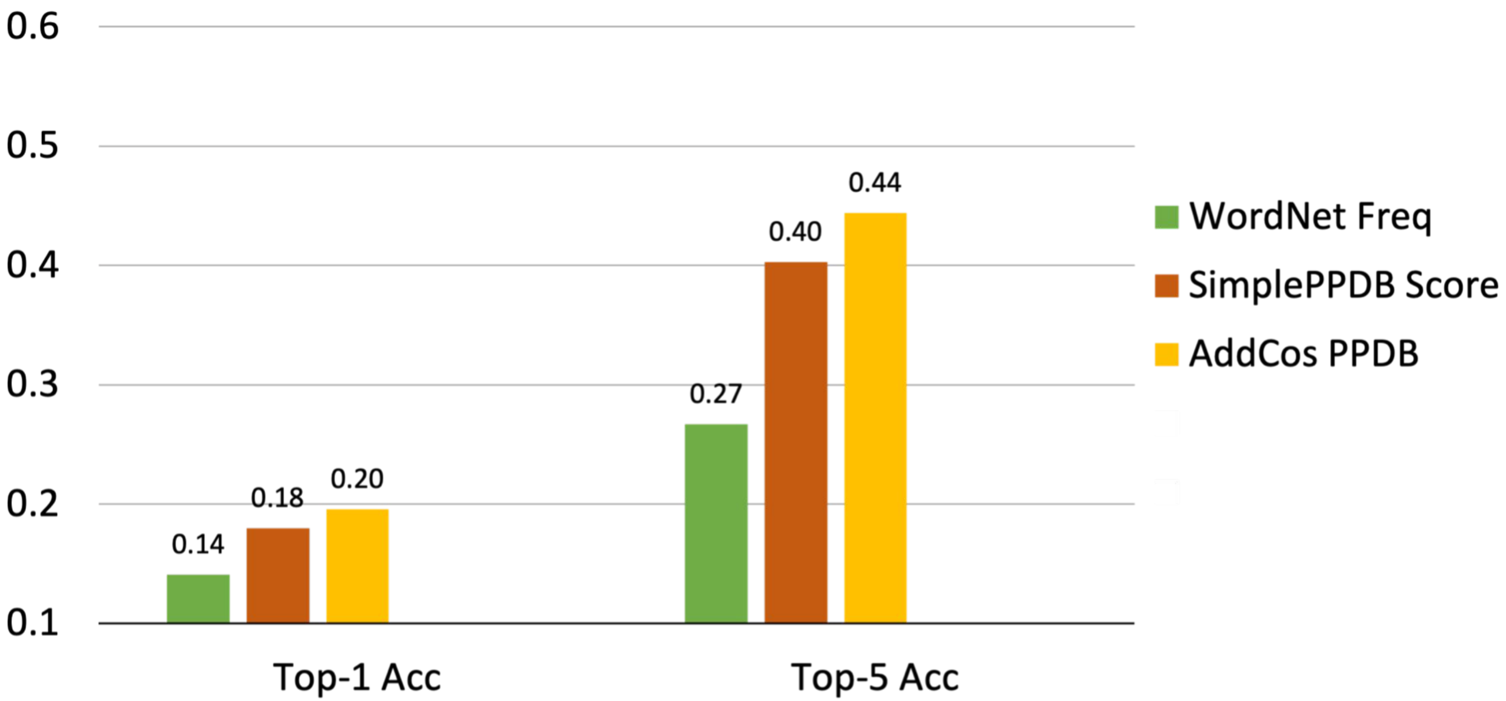
- Aligned parallel sentences using a monolingual word alignment software

- Result: 14,436 complex/simple word pairs

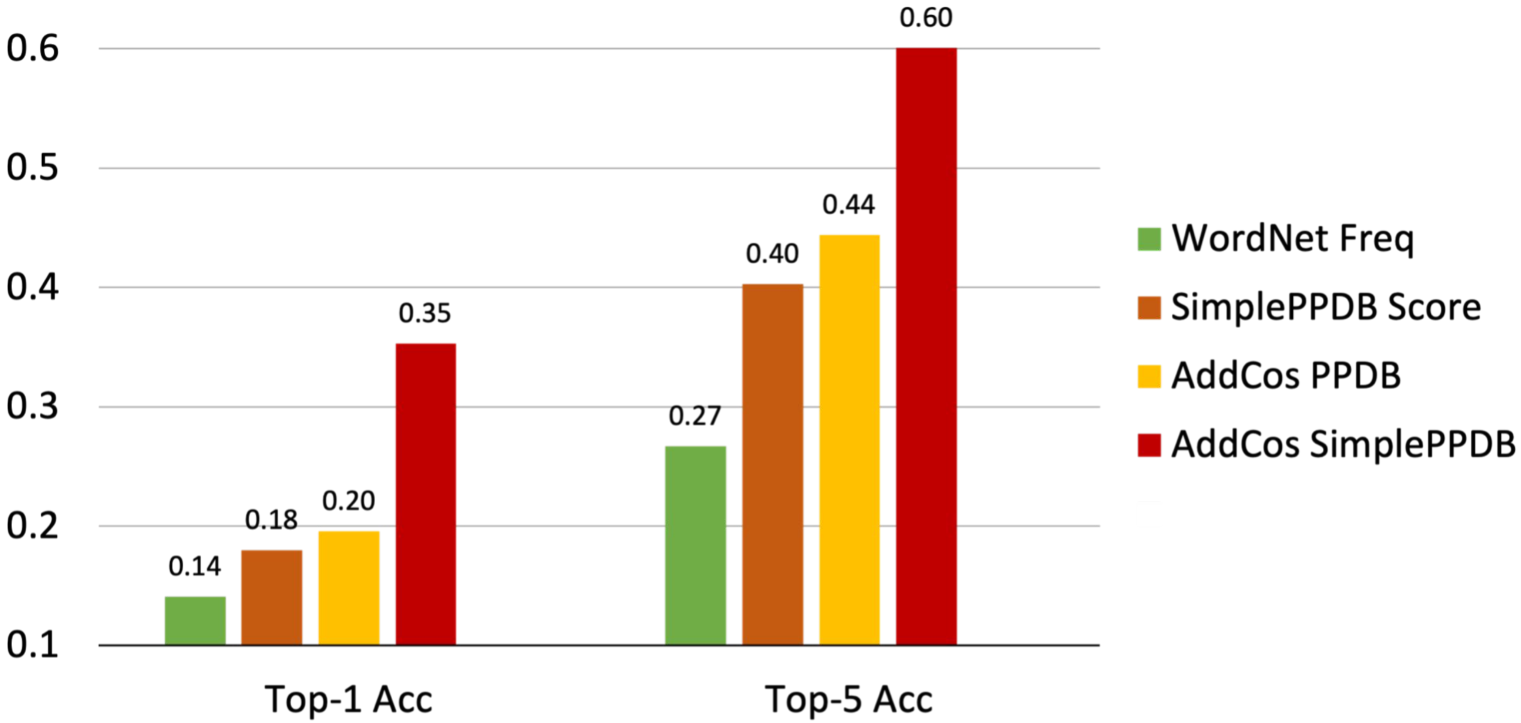
Lexical Substitution: Results



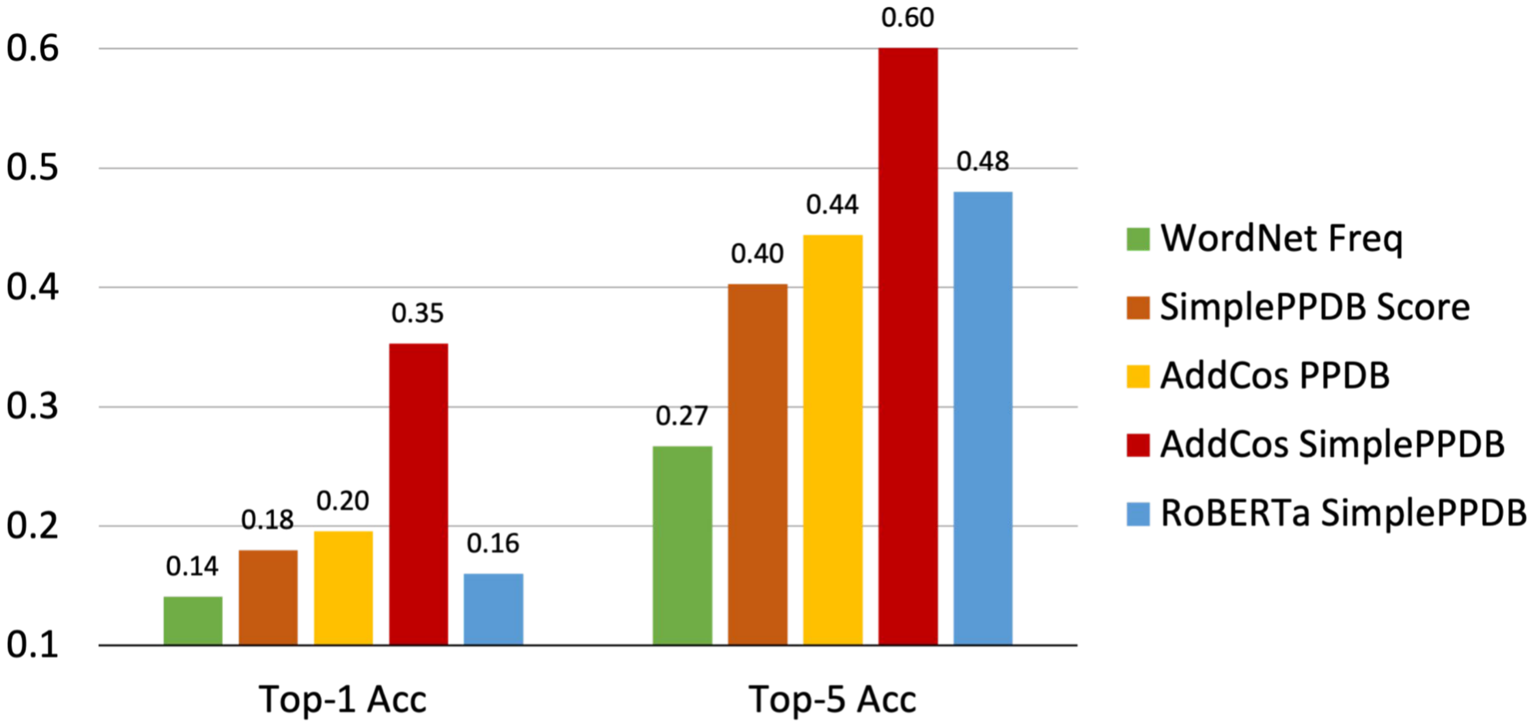
Lexical Substitution: Results



Lexical Substitution: Results



Lexical Substitution: Results



Error Analysis

The identified complex word is part of a phrase and cannot be simplified on its own

Error Analysis

The identified complex word is part of a phrase and cannot be simplified on its own

Word in Context Officials will offer the lunch program at elementary schools.

Bad Substitute basic

Error Analysis

Complex word has no good simpler substitute

Error Analysis

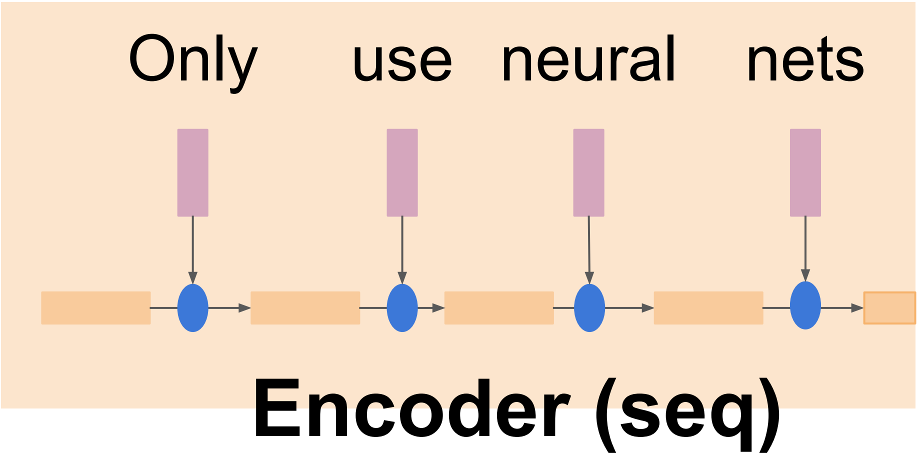
Complex word generally has no good simpler substitute

Word in Context	Although the calculus may be different with Syrian **refugees**, the parallel for me is politics.
Bad Substitutes	life, right, return, shelter, million

Thesis Outline

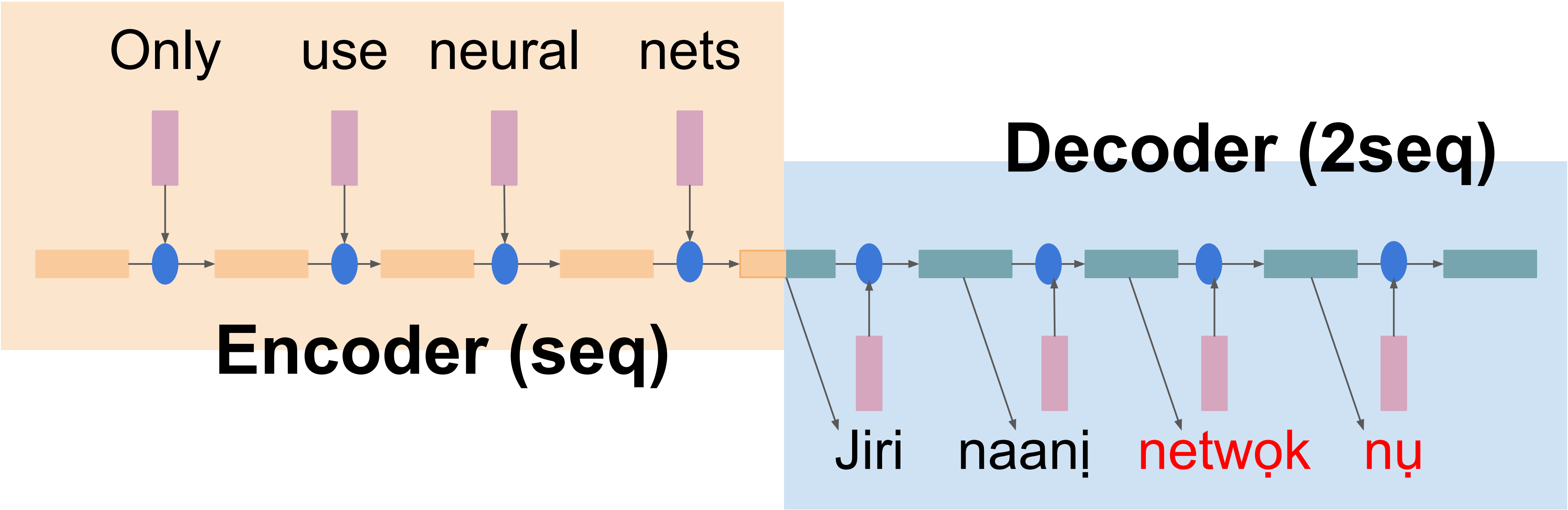
1. Simplification using Paraphrases and Context-based Lexical Substitution
2. **Modeling and Evaluation of Sentence Simplification**
3. Recasting Text Simplification as a Retrieval Task

Sequence-to-Sequence (Seq2Seq) Models



Slide from University of Pennsylvania Lecture for CIS 530 by John Hewitt and Reno Kriz.

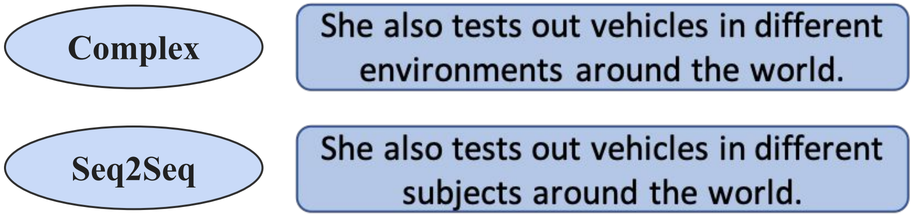
Sequence-to-Sequence (Seq2Seq) Models



Slide from University of Pennsylvania Lecture for CIS 530 by John Hewitt and Reno Kriz.

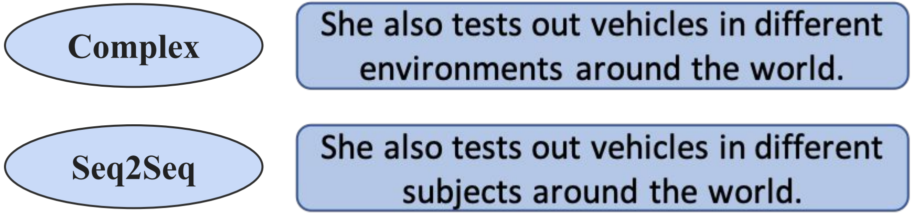
Problems with Seq2Seq

- Copying directly from the original sentence

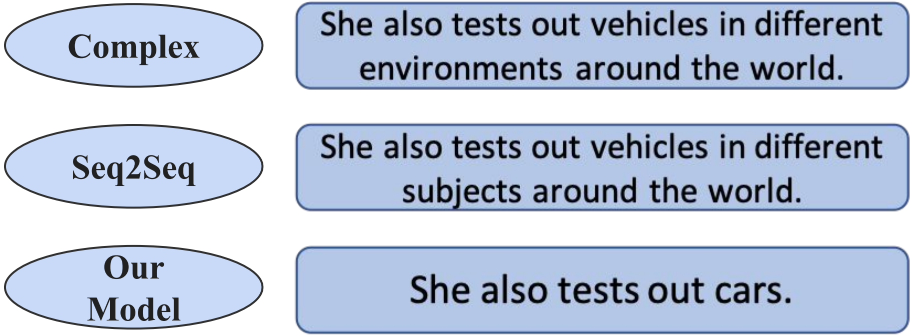


Problems with Seq2Seq

- Copying directly from the original sentence
- Previous solutions
  - Reinforcement learning (Zhang and Lapata, 2017)
  - Memory augmentation (Zhao et al., 2018)

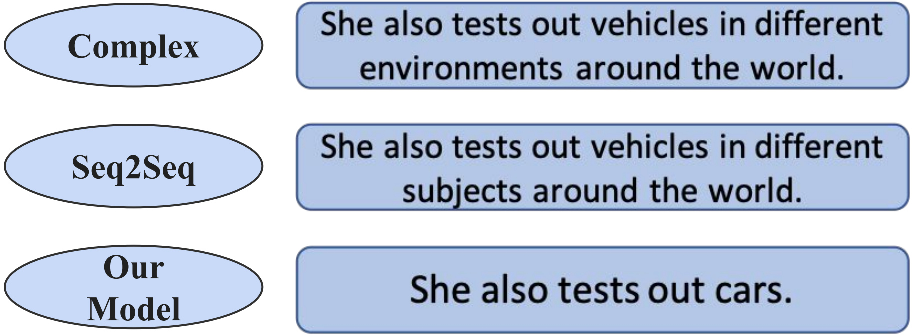


Proposed Solutions



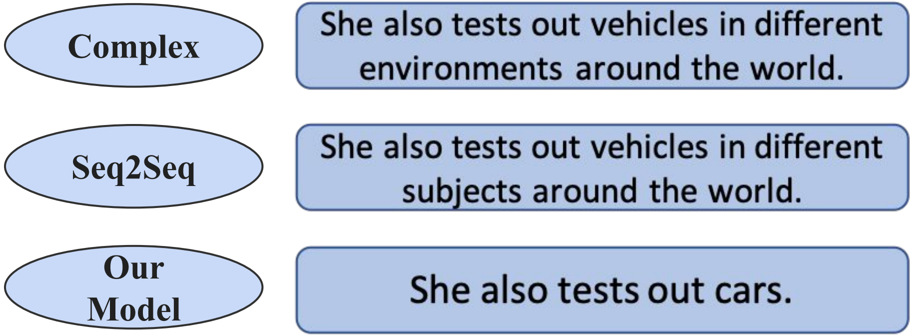
Proposed Solutions

- Accounting for word complexity during training



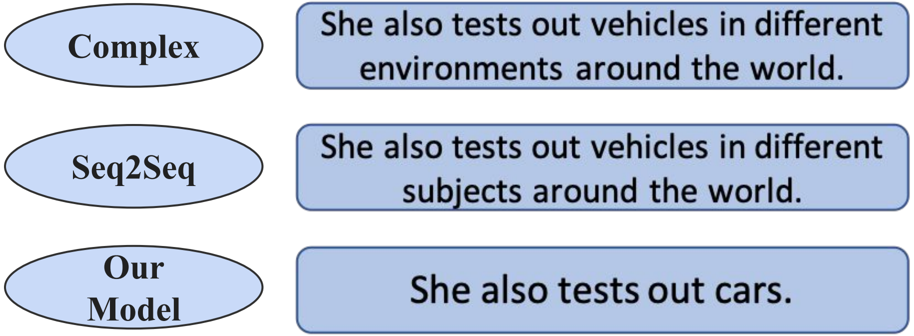
Proposed Solutions

- Accounting for word complexity during training
- Generating multiple diverse output sequences



Proposed Solutions

- Accounting for word complexity during training
- Generating multiple diverse output sequences
- Re-ranking candidate simplifications



Modified Loss Function

Modified Loss Function

- Propose a complexity-weighted loss function (CWL)
  - Places more weight on getting the simpler words correct

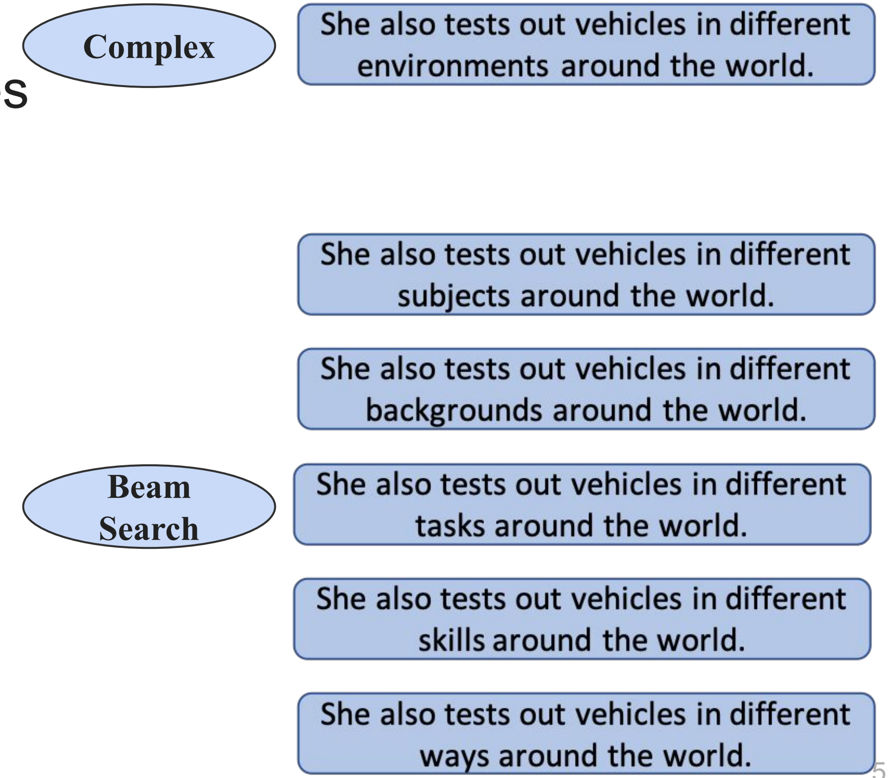
Modified Loss Function

- Propose a complexity-weighted loss function (CWL)
  - Places more weight on getting the simpler words correct
- Predict complexity with a linear regression model

Complexity Level	Examples
0	swim, dog, sleep
1	thin, carvings, ballerina
2	unique, appealed, violators
3	geologic, apparatus, renovation
4	symptomatic, jurisdiction, mitigation

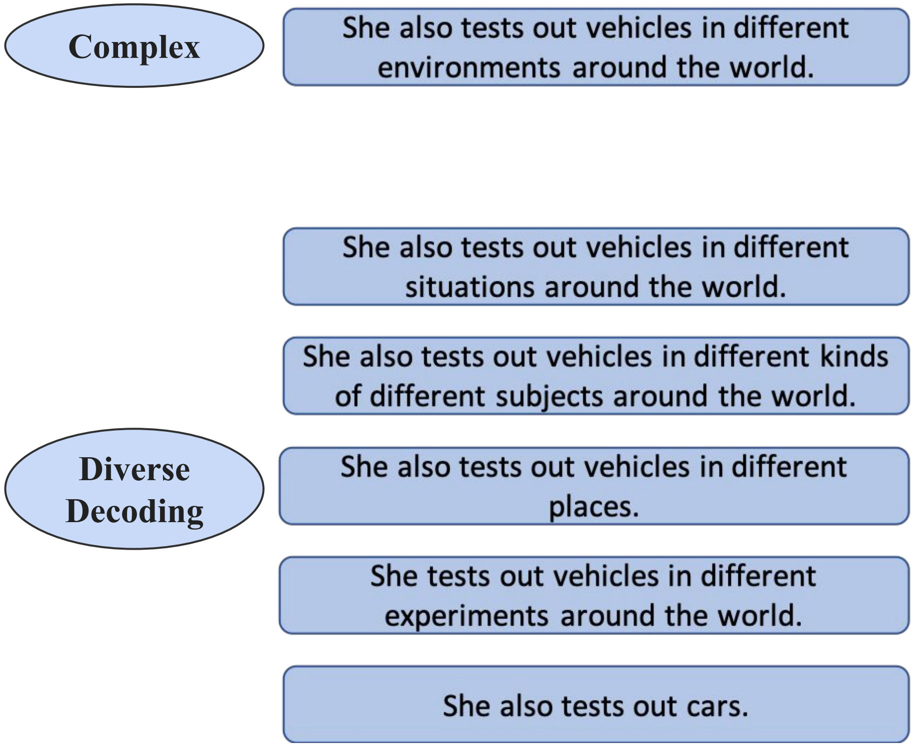
Diverse Decoding

- Beam search generates candidates with only minor differences (Vijayakumar et al., 2016)



Diverse Decoding

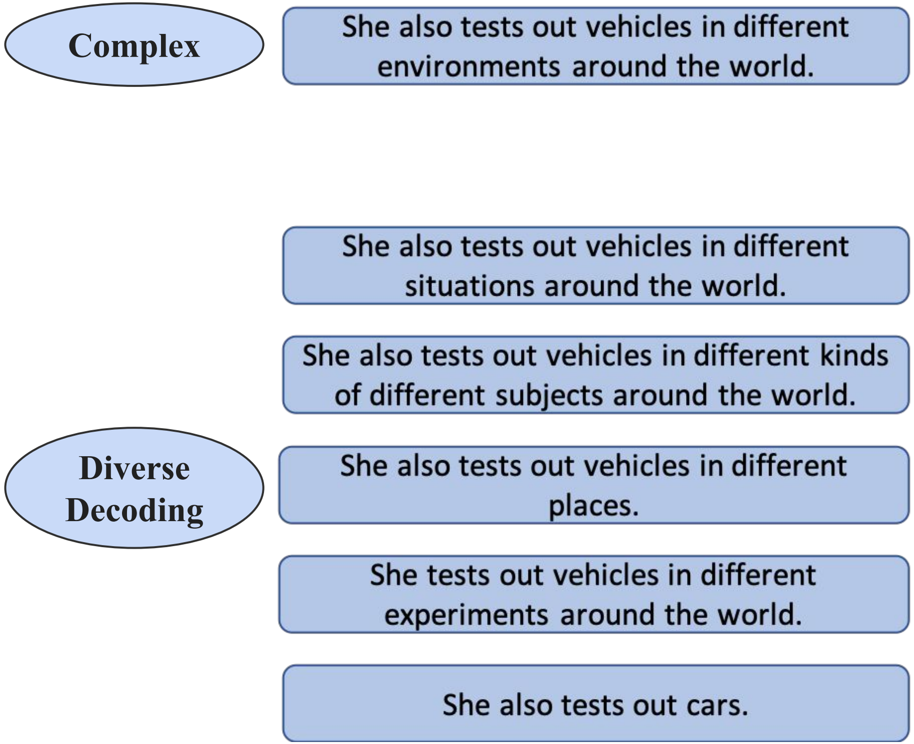
- Augment beam search scores to penalize candidates from the same parent subsequence (Li et al, 2016) (DeDiv)



Diverse Decoding

- Augment beam search scores to penalize candidates from the same parent subsequence (Li et al, 2016) (DeDiv)

- Propose a method for clustering similar outputs after decoding (PDC)



Post-decoding Re-ranking (PDR)

Post-decoding Re-ranking (PDR)

Fluency

Perplexity based on a 5-gram language model

Post-decoding Re-ranking (PDR)

Fluency — Perplexity based on a 5-gram language model

Adequacy — Cosine similarity between sentence representations of input and output

Post-decoding Re-ranking (PDR)

Fluency — Perplexity based on a 5-gram language model

Adequacy — Cosine similarity between sentence representations of input and output

Complexity — Trained CNN to predict the sentence-level complexity of the output

Baseline Comparisons

Seq2Seq Vanilla Seq2Seq model (Nisioi et al., 2017)

Baseline Comparisons

Seq2Seq Vanilla Seq2Seq model (Nisioi et al., 2017)

Hybrid Performs splitting and deletion before applying a phrase-based machine translation system (Narayan and Gardent, 2014)

Baseline Comparisons

Seq2Seq Vanilla Seq2Seq model (Nisioi et al., 2017)

Hybrid Performs splitting and deletion before applying a phrase-based machine translation system (Narayan and Gardent, 2014)

DRESS Seq2Seq with reinforcement learning (Zhang and Lapata, 2017)

Baseline Comparisons

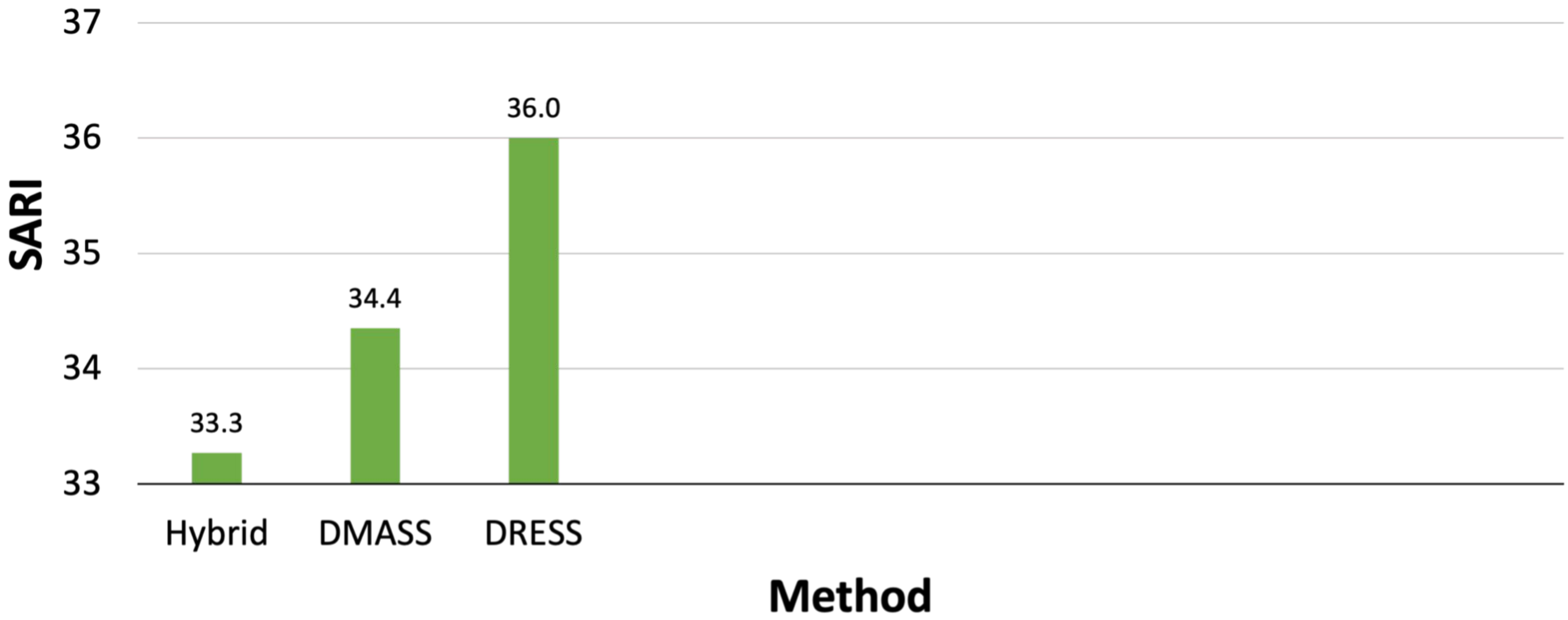
Seq2Seq Vanilla Seq2Seq model (Nisioi et al., 2017)

Hybrid Performs splitting and deletion before applying a phrase-based machine translation system (Narayan and Gardent, 2014)

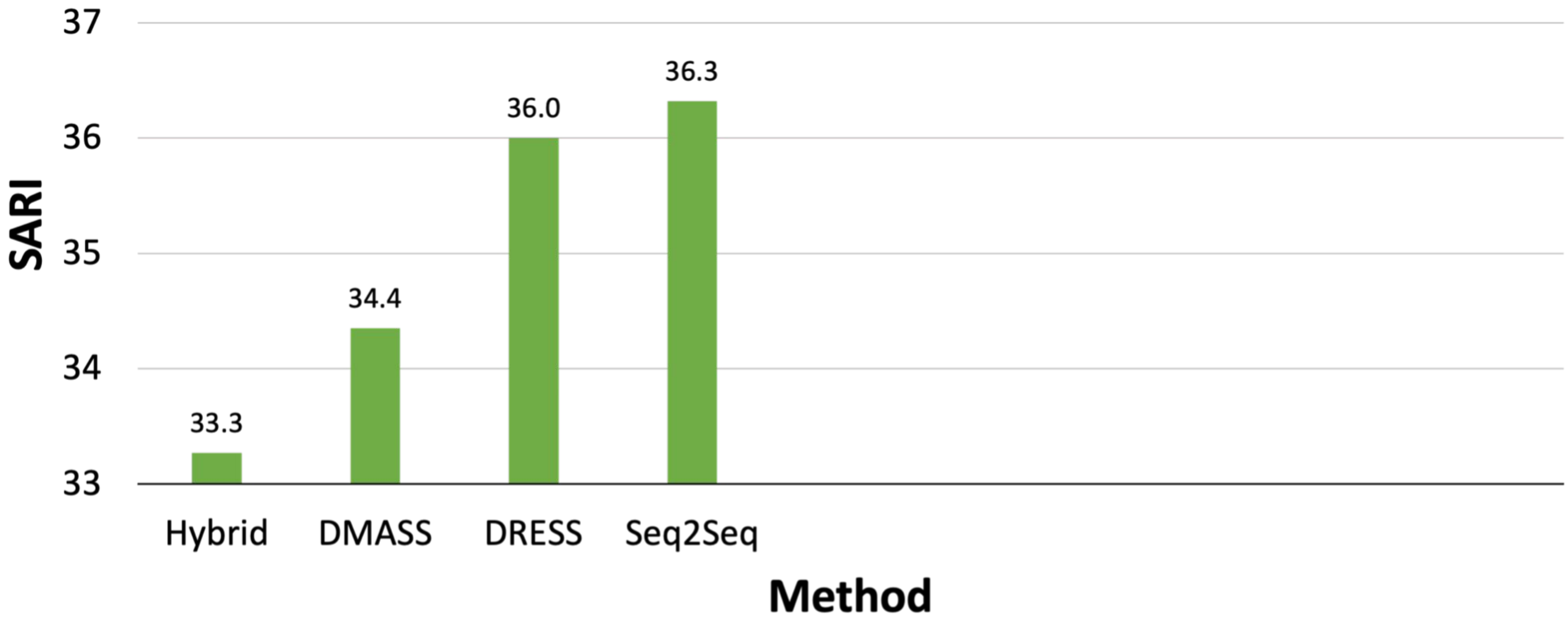
DRESS Seq2Seq with reinforcement learning (Zhang and Lapata, 2017)

DMASS Transformer with memory augmentation (Zhu et al., 2018)

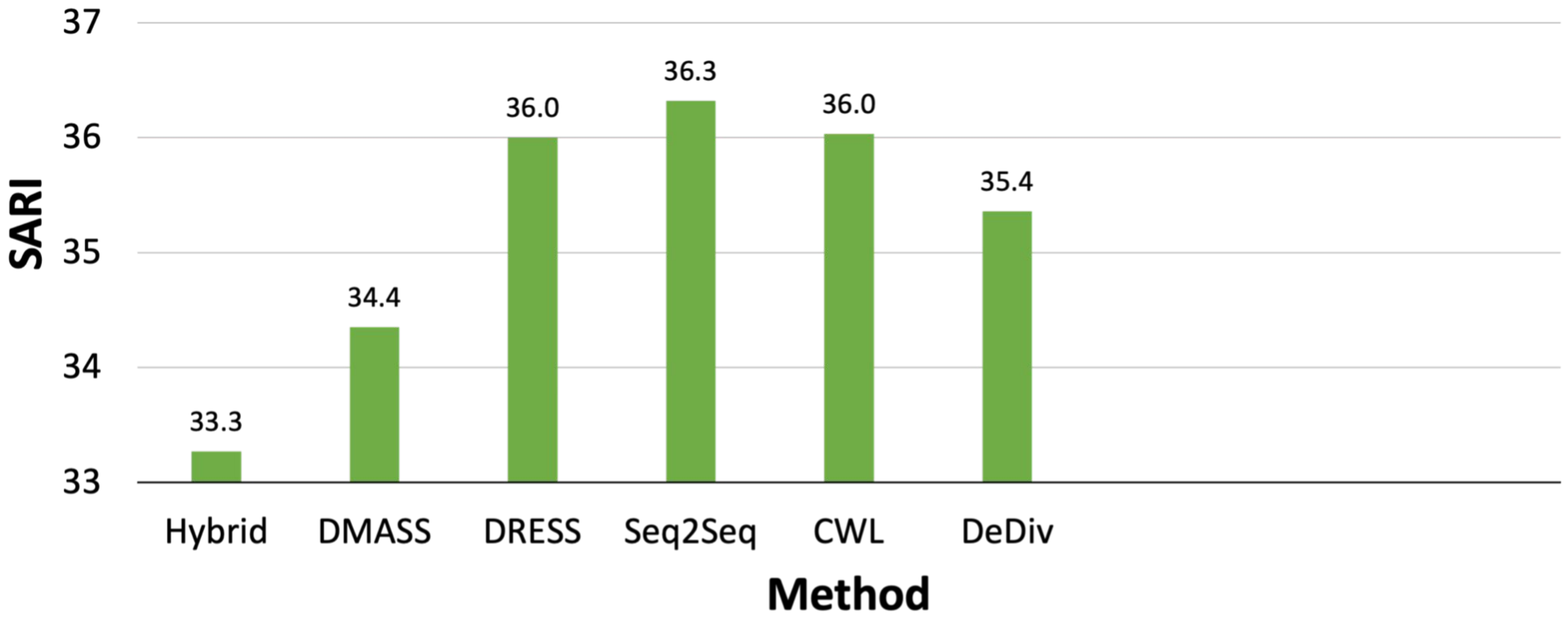
Results: Automatic Evaluation



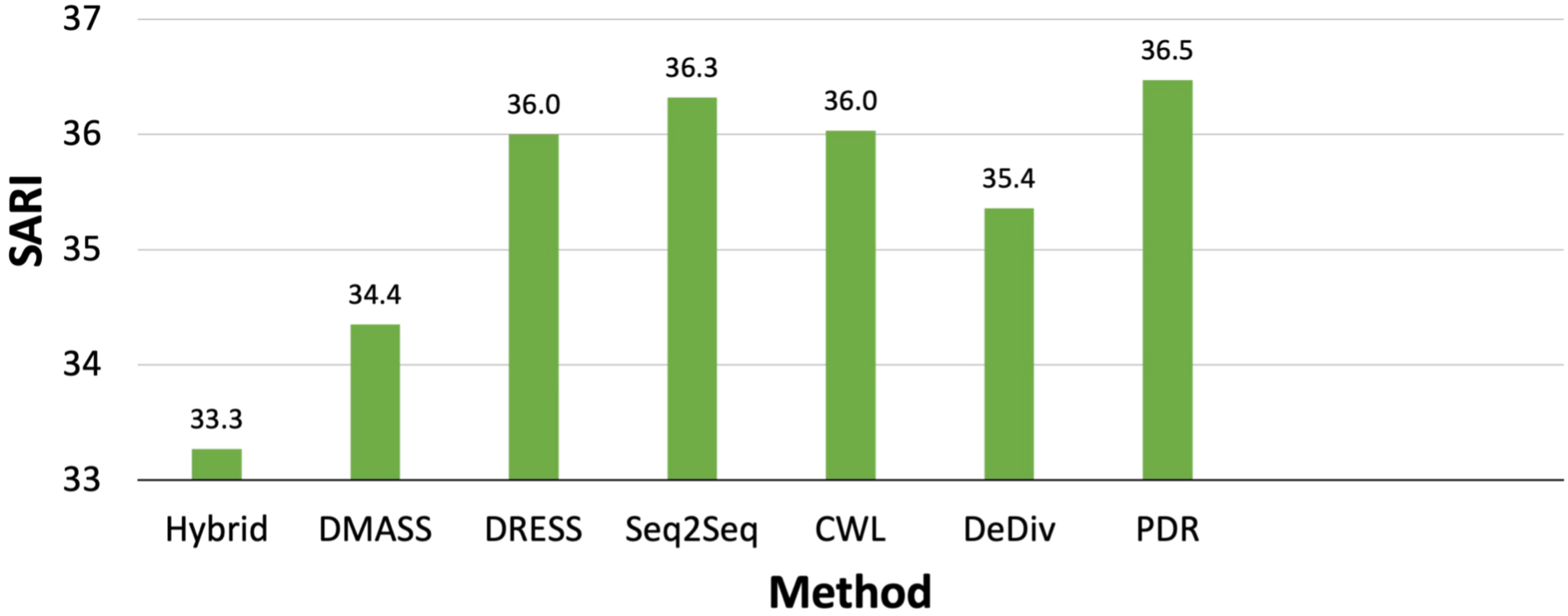
Results: Automatic Evaluation



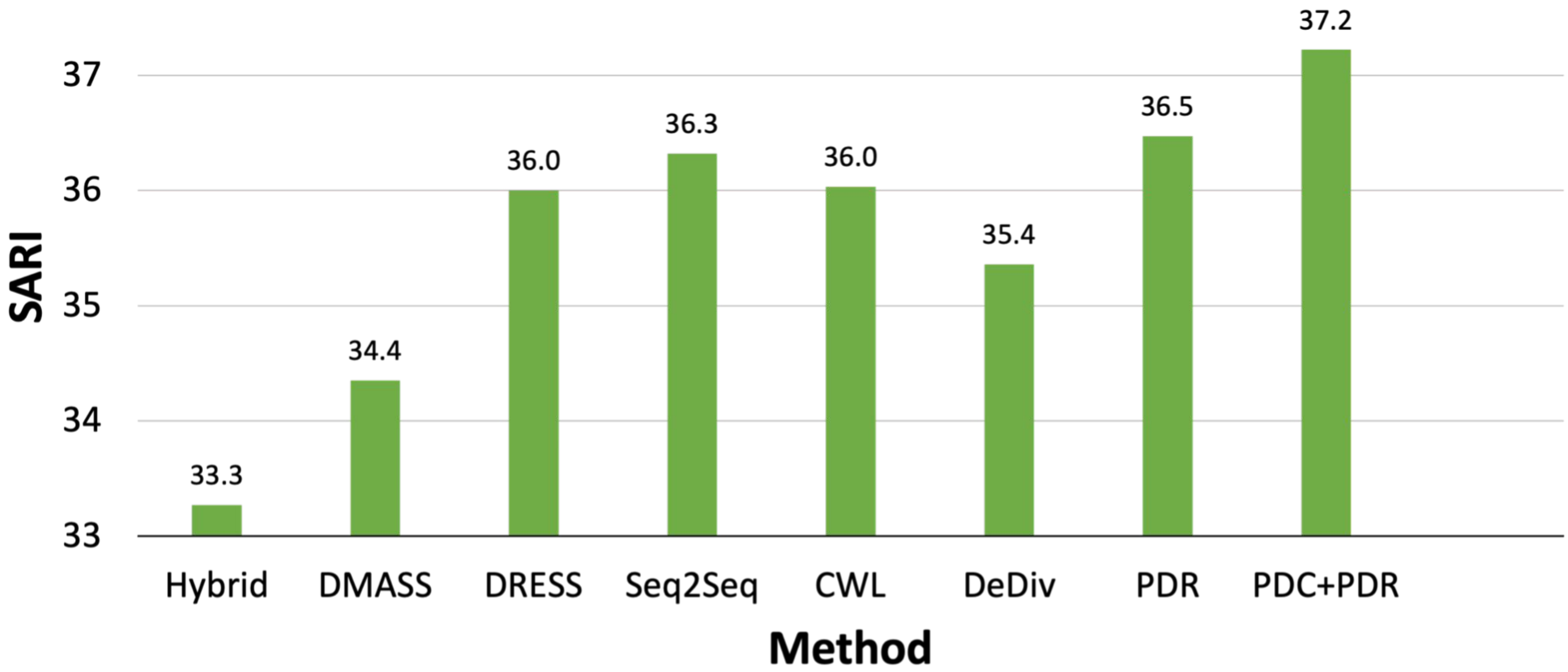
Results: Automatic Evaluation



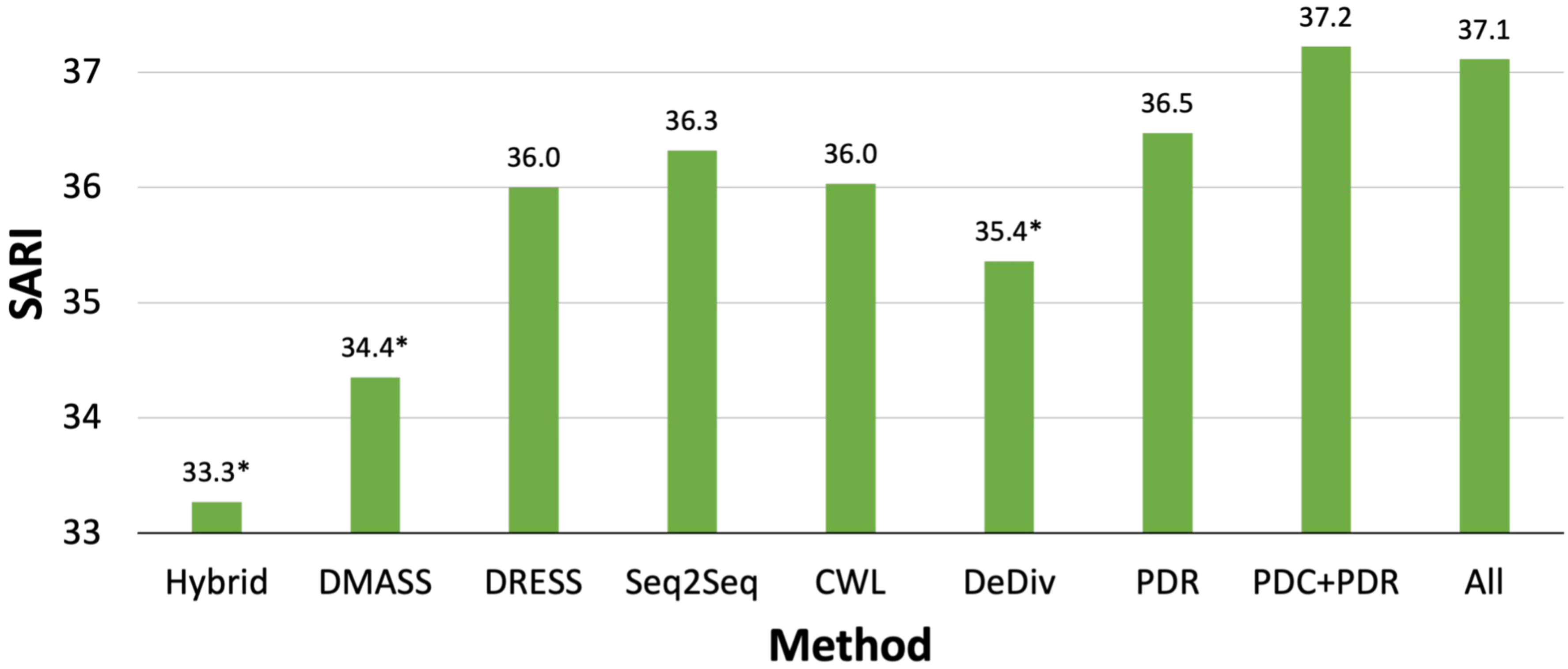
Results: Automatic Evaluation



Results: Automatic Evaluation

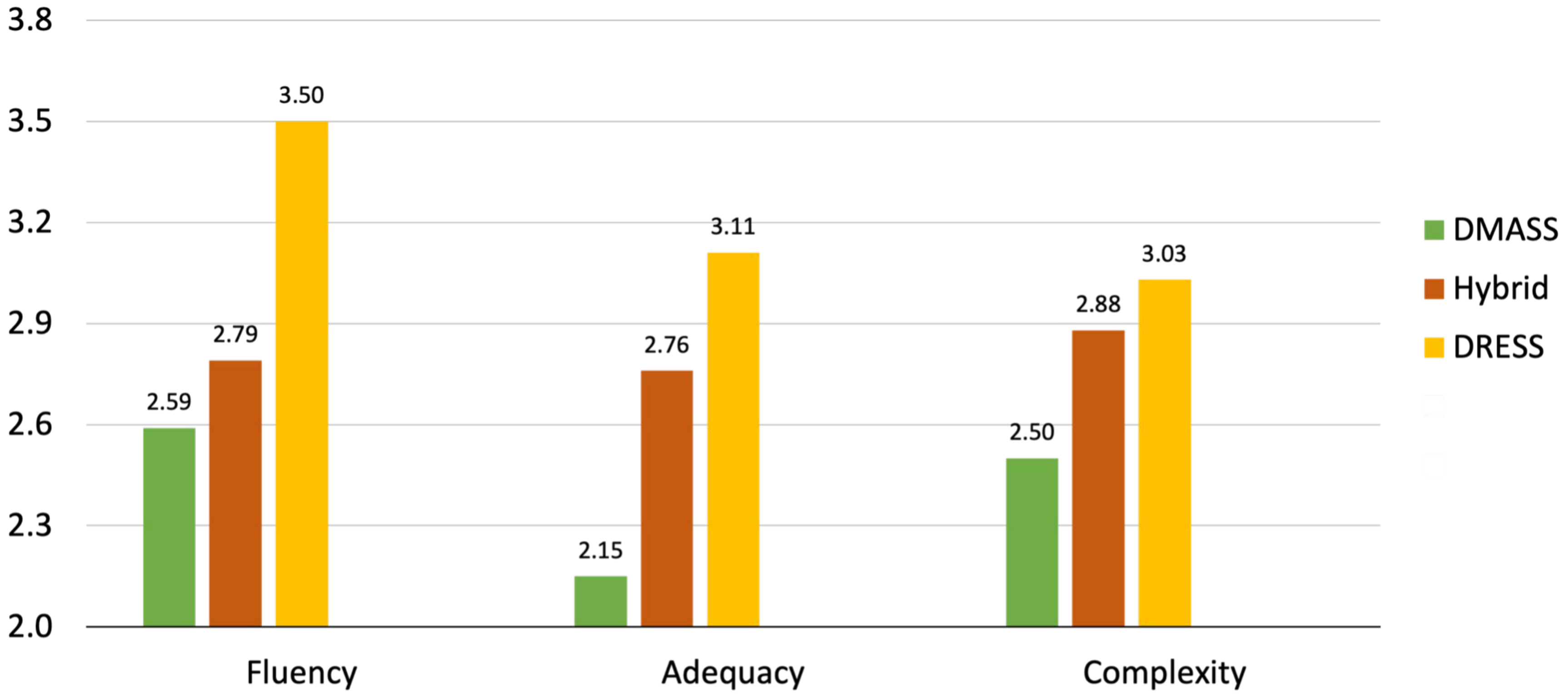


Results: Automatic Evaluation

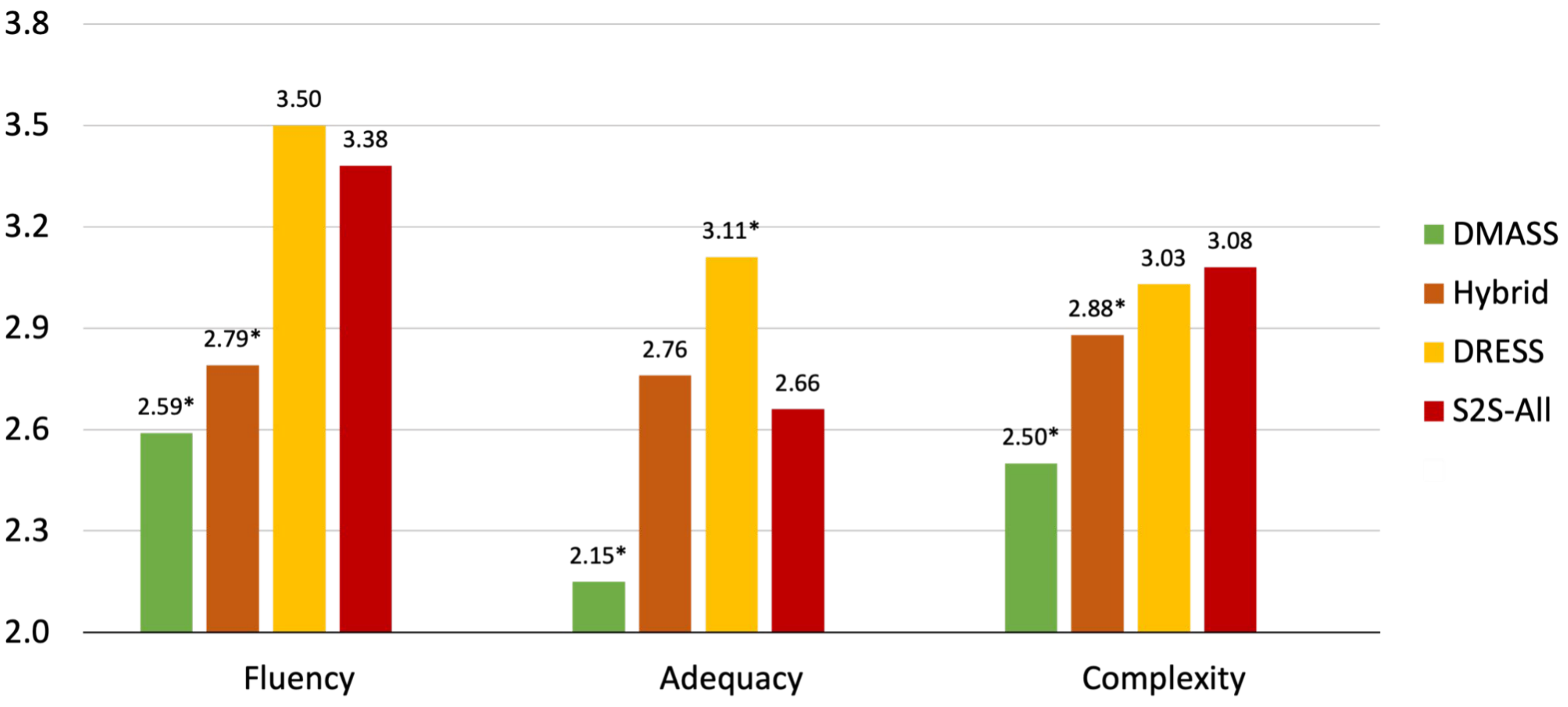


Results: Human Evaluation

Results: Human Evaluation



Results: Human Evaluation

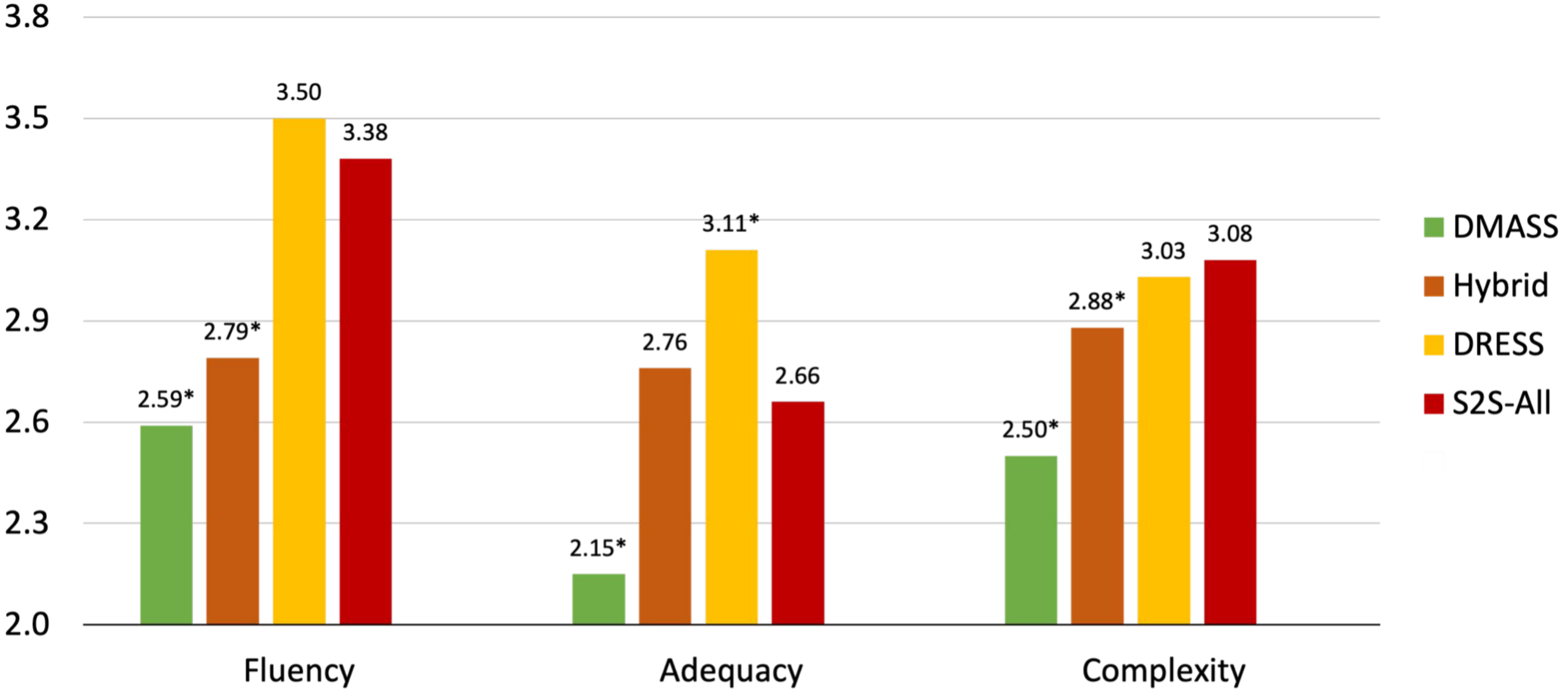


Results: Human Evaluation

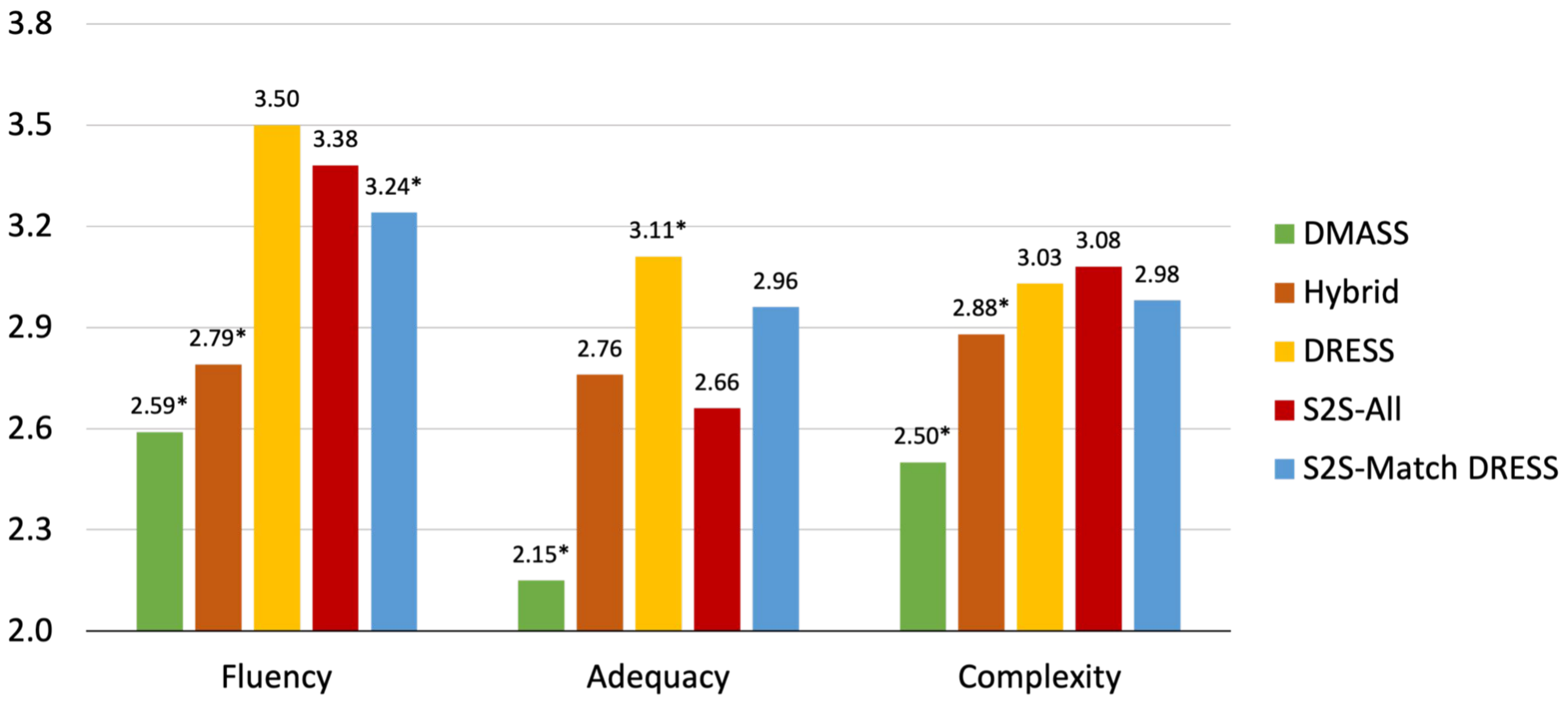
DRESS length: 14.4 tokens

vs.

S2S length: 10.8 tokens



Results: Human Evaluation



Error Analysis

Error Analysis

Poor substitution due to word embedding proximity

Complex

In Beijing kite circles, Fei is widely known as the elder **<u>statesman</u>**.

Error Analysis

Poor substitution due to word embedding proximity

Complex	In Beijing kite circles, Fei is widely known as the elder **statesman**.
Our Model	In Beijing, Fei is considered a **doctor**.

Error Analysis

Long and complex sentences with multiple clauses

Complex

Wal-Mart, which imports more fruits and vegetables from Mexico than any other U.S. company, announced its effort to force improvements up and down its supply chain.

Error Analysis

Long and complex sentences with multiple clauses

Complex

Wal-Mart, which imports more fruits and vegetables from Mexico than any other U.S. company, announced its effort to force improvements up and down its supply chain.

Our Model

and vegetables from the company.

Current Evaluation Metrics

- BLEU (Papineni et al., 2002)
    - Counts the number of overlapping *n*-grams

Current Evaluation Metrics

- BLEU (Papineni et al., 2002)
    - Counts the number of overlapping *n*-grams

- SARI (Xu et al., 2016)
    - Counts how often the output correctly keeps, deletes, and adds *n*-grams

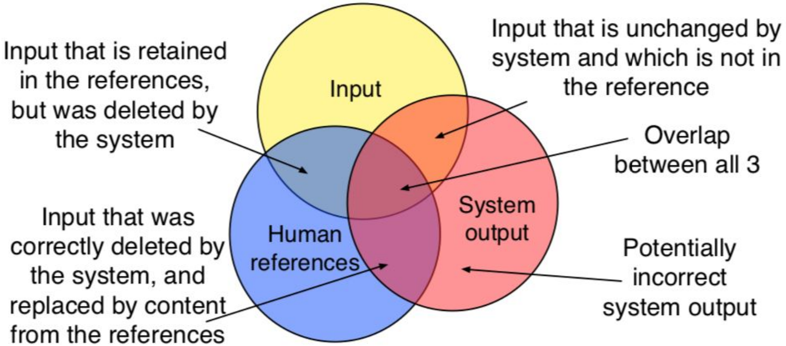


Illustration of SARI (Xu et al., 2016)

Simplification Evaluation Metrics: Limitations

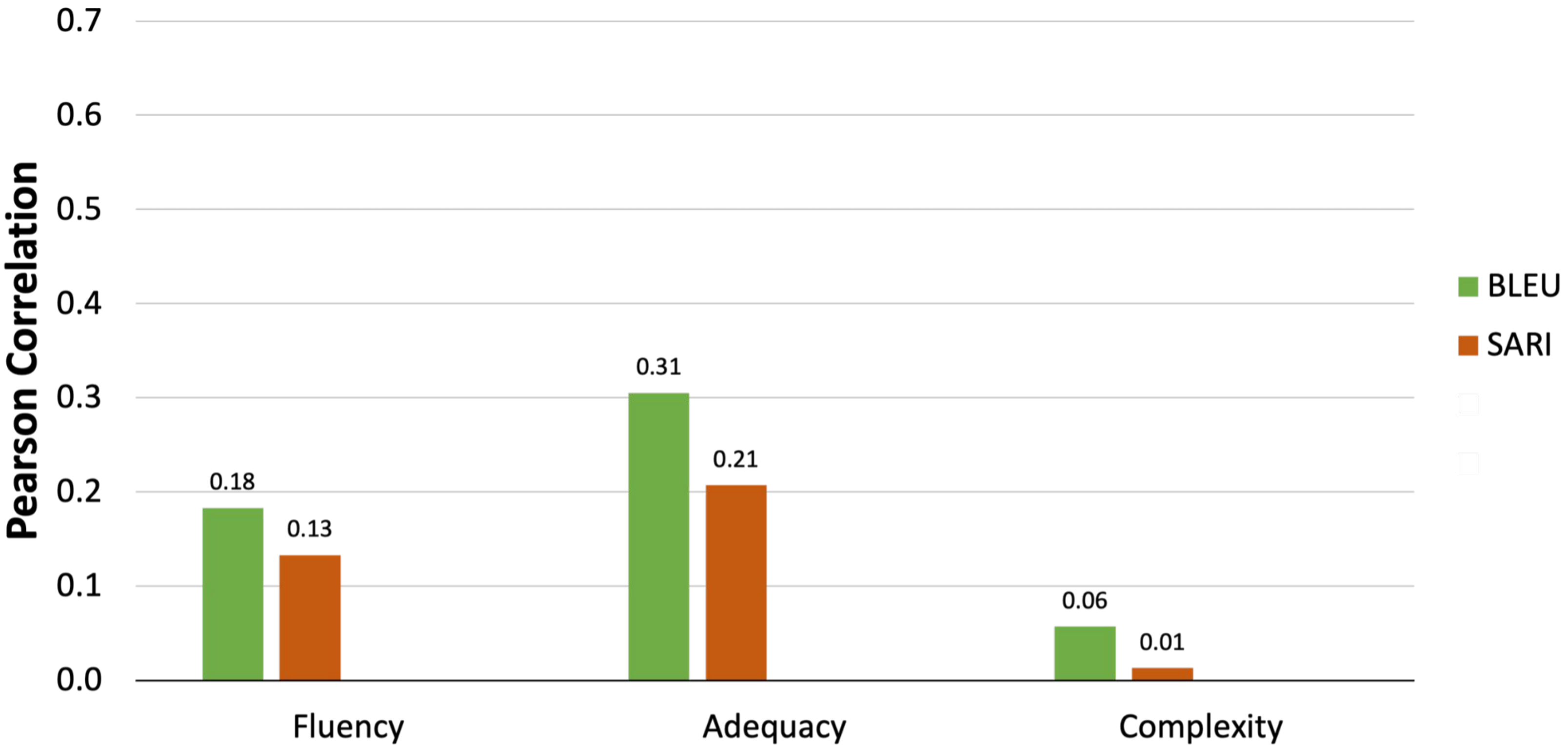
- Low correlation with human judgments
    - BLEU negatively correlates with deletion (Sulem et al., 2018)

Simplification Evaluation Metrics: Limitations

- Low correlation with human judgments
  - BLEU negatively correlates with deletion (Sulem et al., 2018)

- Require 1+ reference simple sentences

Simplification Evaluation Metrics: Results

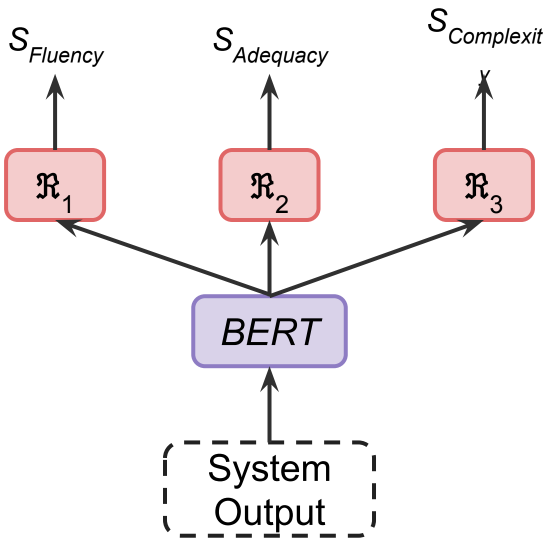


BERT-based Quality Estimation (Xenouleas et al., 2019)

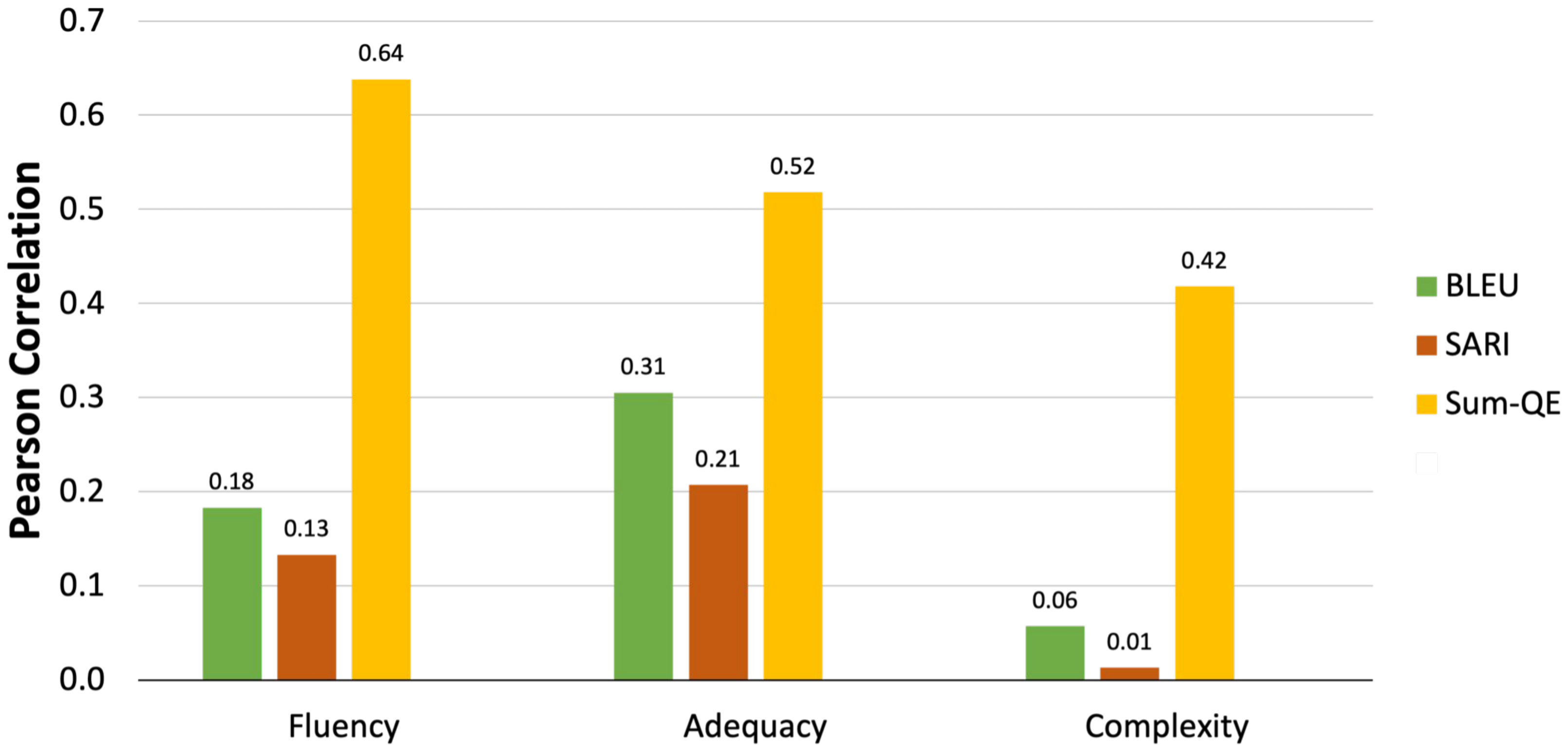
Sum-QE: Rates summaries with respect to five linguistic qualities

BERT-based Quality Estimation (Xenouleas et al., 2019)

<u>Sum-QE</u>: Rates summaries with respect to five linguistic qualities



Simplification Evaluation Metrics: Results

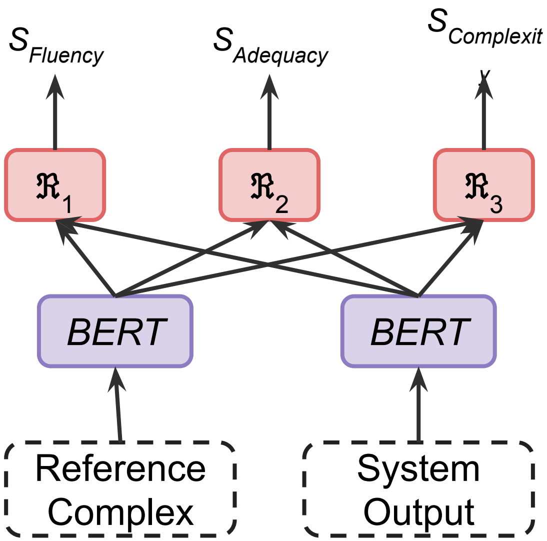


Simplification QE: Method

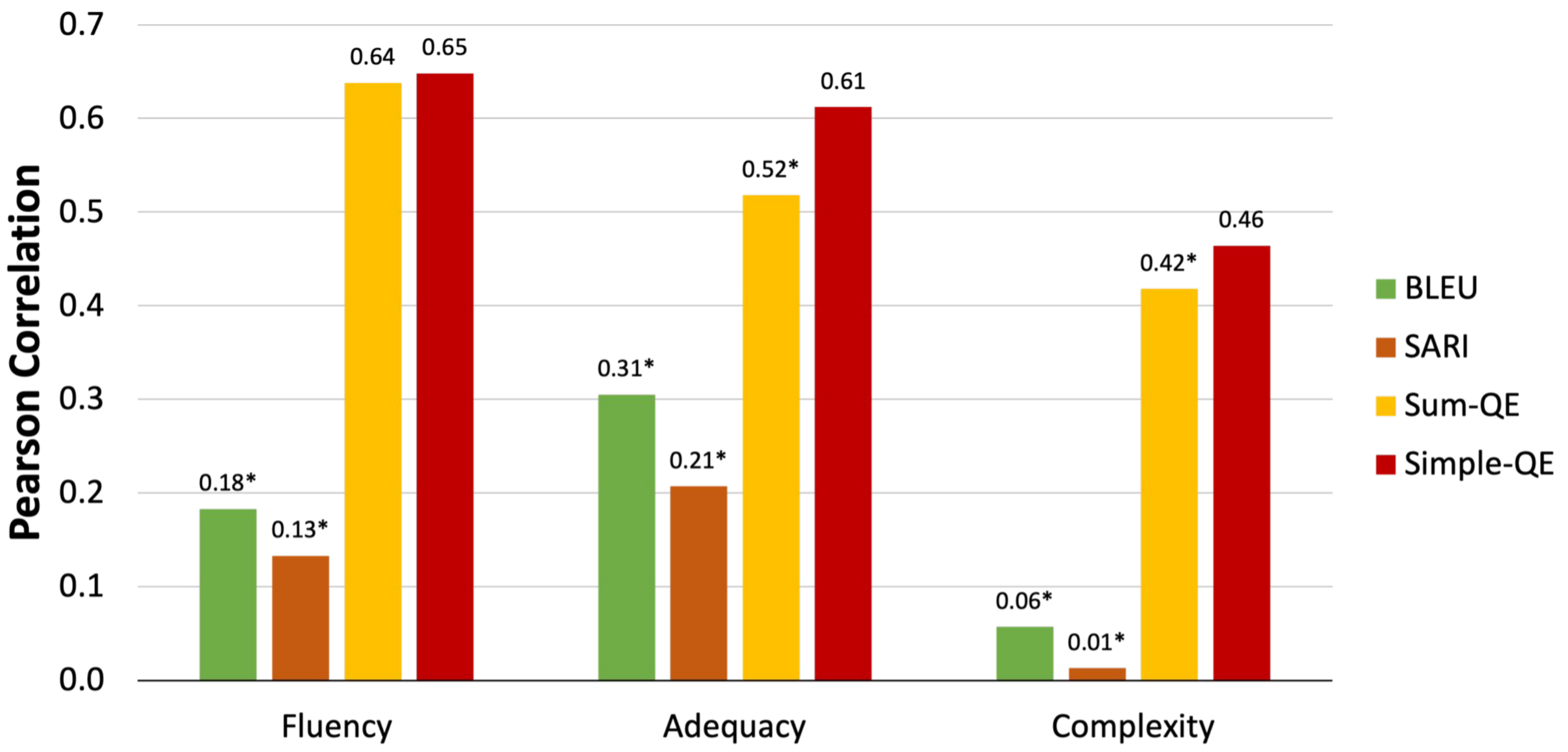
Simple-QE: Extension of Sum-QE that takes reference complex sentence into account

Simplification QE: Method

Simple-QE: Extension of Sum-QE that takes reference complex sentence into account



Simplification Evaluation Metrics: Results



Thesis Outline

1. Simplification using Paraphrases and Context-based Lexical Substitution
2. Modeling and Evaluation of Sentence Simplification
3. **Recasting Text Simplification as a Retrieval Task**

Motivating Example

Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).
"Statistical learning" redirects here. For statistical learning in linguistics, see statistical learning in language acquisition.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to carry out tasks without explicit instructions, such as by using pattern recognition and inference. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[3][4] In its application across business problems, machine learning is also referred to as predictive analytics.

Motivating Example

Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).
"Statistical learning" redirects here. For statistical learning in linguistics, see statistical learning in language acquisition.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to carry out tasks without explicit instructions, such as by using pattern recognition and inference. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[3][4] In its application across business problems, machine learning is also referred to as predictive analytics.

Motivating Example

Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).

"Statistical learning" redirects here. For statistical learning in linguistics, see statistical learning in language acquisition.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to carry out tasks without explicit instructions, such as by using pattern recognition and inference. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[3][4] In its application across business problems, machine learning is also referred to as predictive analytics.

Reformulating as Information Retrieval Problem

Reformulating as Information Retrieval Problem

- Critical concept identification

Reformulating as Information Retrieval Problem

- Critical concept identification
- Leveled document retrieval

Reformulating as Information Retrieval Problem

- Critical concept identification
- Leveled document retrieval
- Document filtering

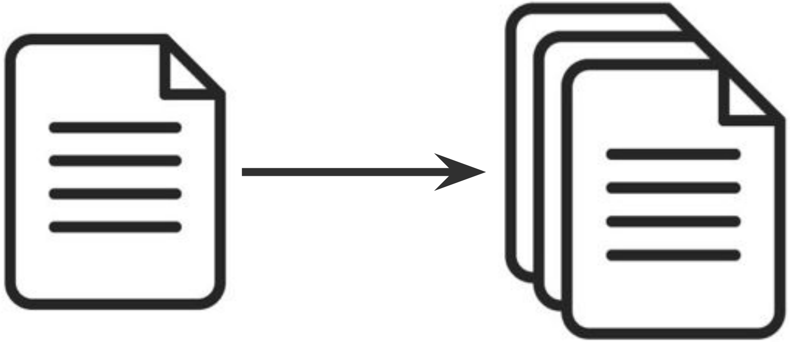
Reformulating as Information Retrieval Problem

- Critical concept identification
- **Leveled document retrieval**
- **Document filtering**

Complexity-Sensitive Retrieval: Problem Setup

Input: a phrase,
sentence, or document

Complexity-Sensitive Retrieval: Problem Setup



Input: a phrase, sentence, or document

k related texts at a variety of complexity levels

Complexity-Sensitive Retrieval: Problem Setup

<u>Input</u>: a phrase, sentence, or document

k related texts at a variety of complexity levels

A filtered list of texts based on appropriateness

Complexity-Sensitive Retrieval: Initial Experiment

We can again use Newsela!

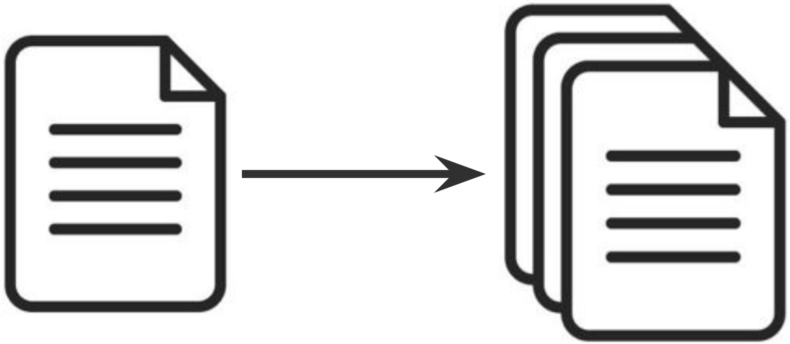
Complexity-Sensitive Retrieval: Initial Experiment

We can again use Newsela!

Input: each complex
Newsela article D_{i4}

Complexity-Sensitive Retrieval: Initial Experiment

We can again use Newsela!



Input: each complex Newsela article D_{i4}

All other articles, ranked by similarity to D_{i4}

Complexity-Sensitive Retrieval: Initial Experiment

We can again use Newsela!

Input: each complex Newsela article D_{i4}

All other articles, ranked by similarity to D_{i4}

The rank of the aligned Newsela articles D_{ij}, where $0 \leq j \leq 3$

Complexity-Sensitive Retrieval: Initial Experiment

We can again use Newsela!

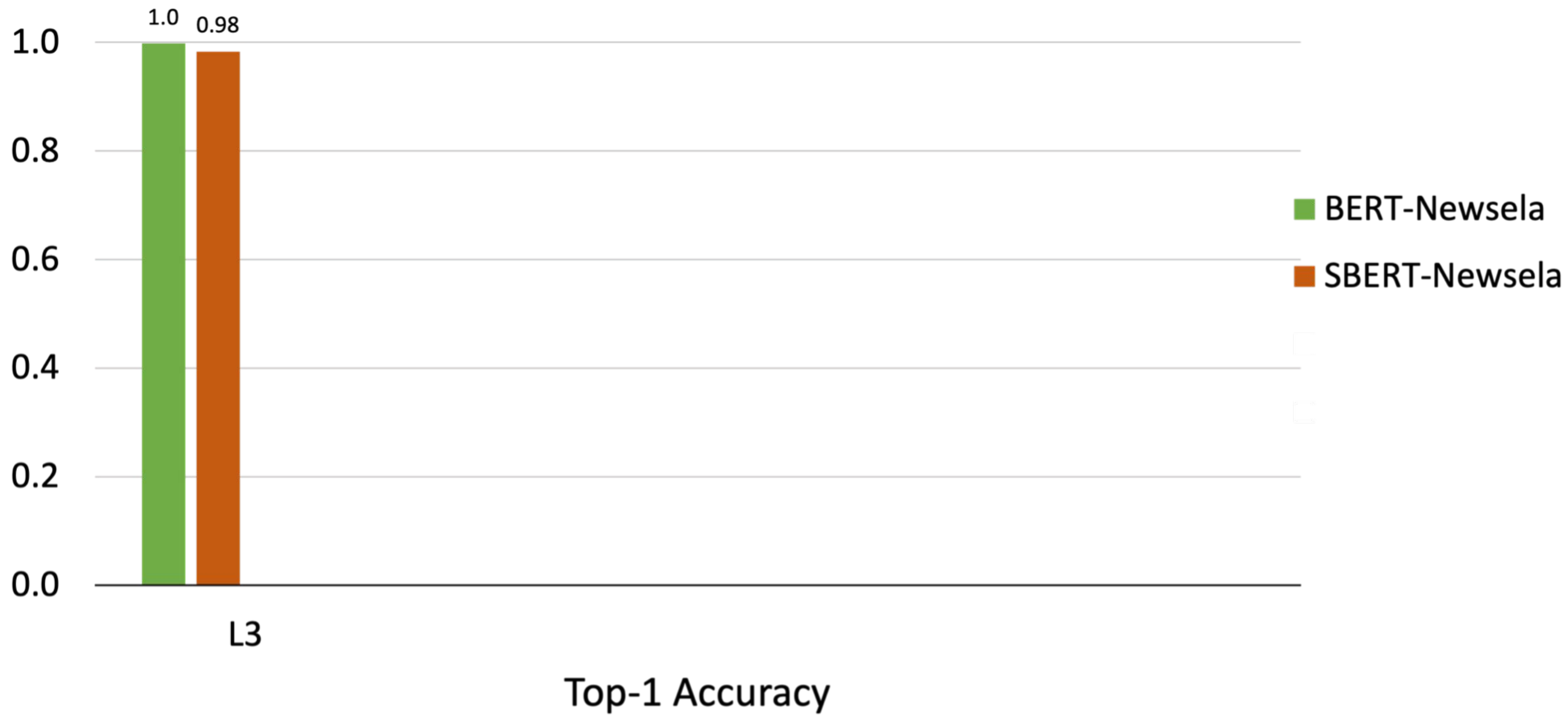
- Since this is a relatively small corpus, we also include ~1 million New York Times articles as additional distractor documents

Input: each complex Newsela article D_{i4}

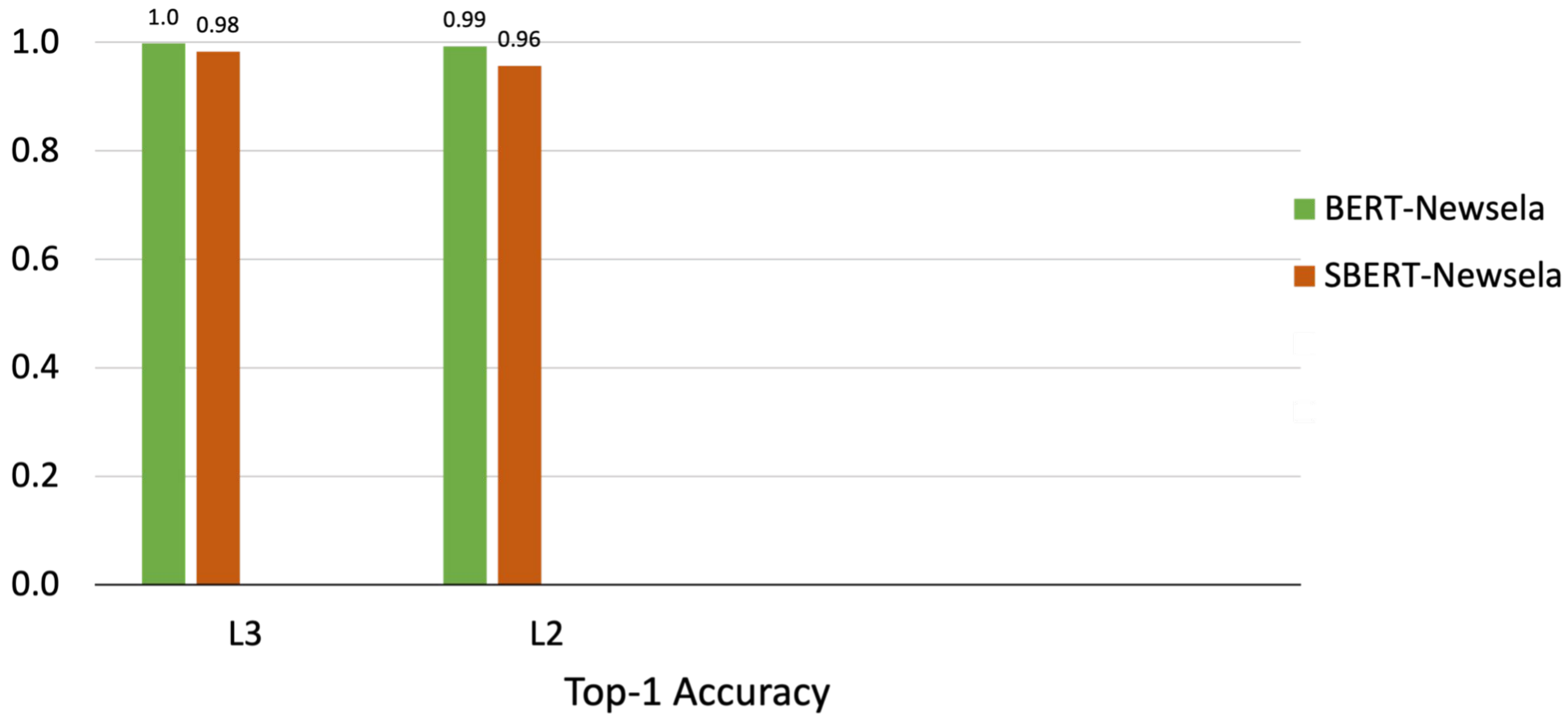
All other articles, ranked by similarity to D_{i4}

The rank of the aligned Newsela articles D_{ij}, where $0 \leq j \leq 3$

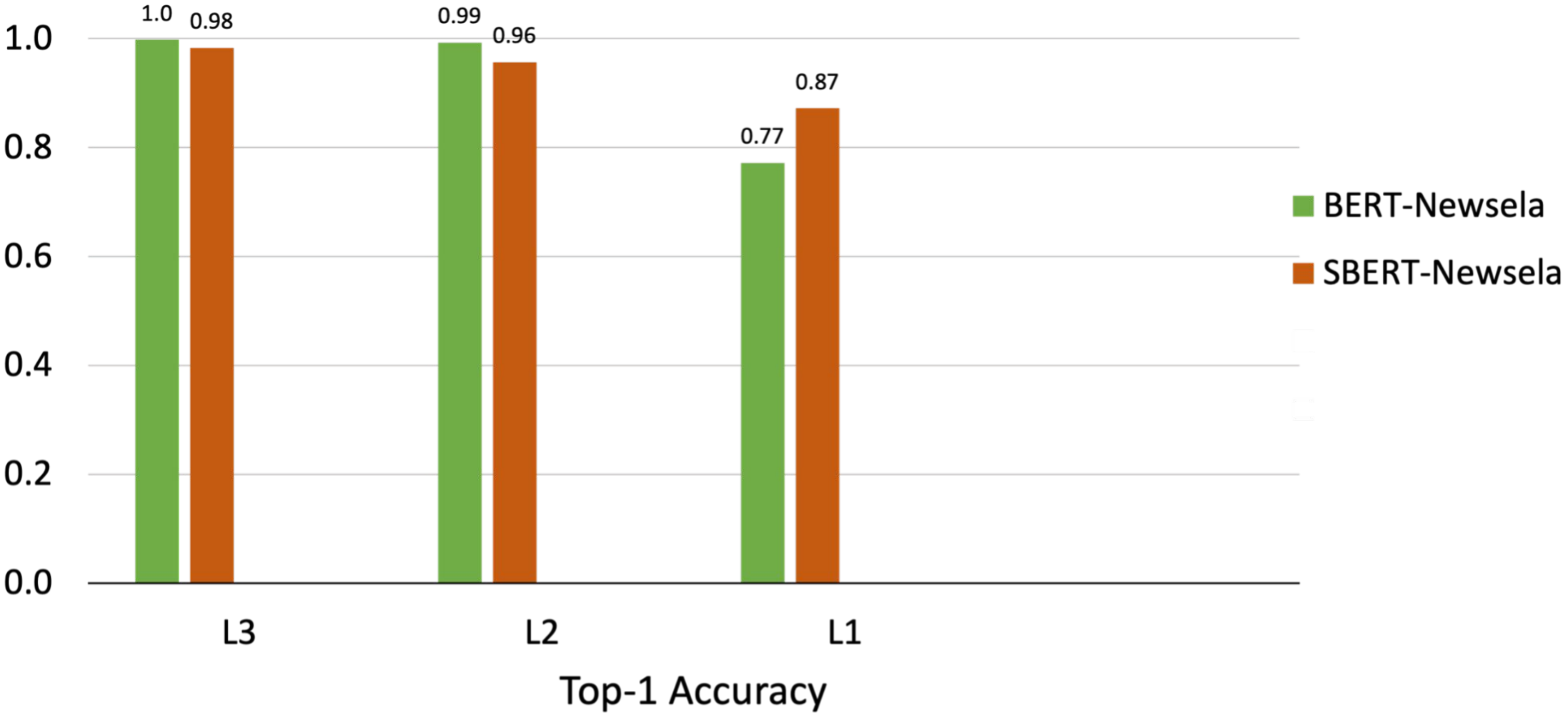
Complexity-Sensitive Retrieval: Results



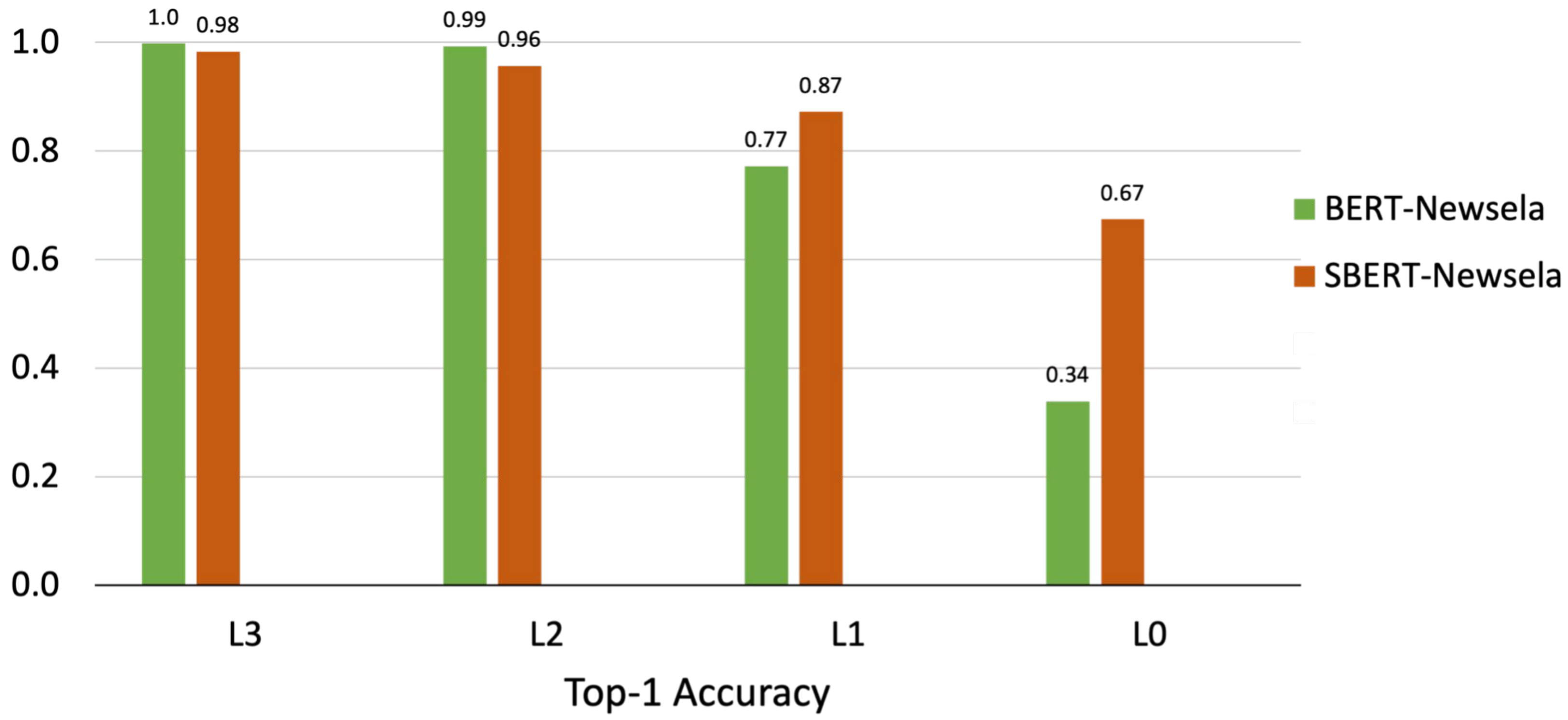
Complexity-Sensitive Retrieval: Results



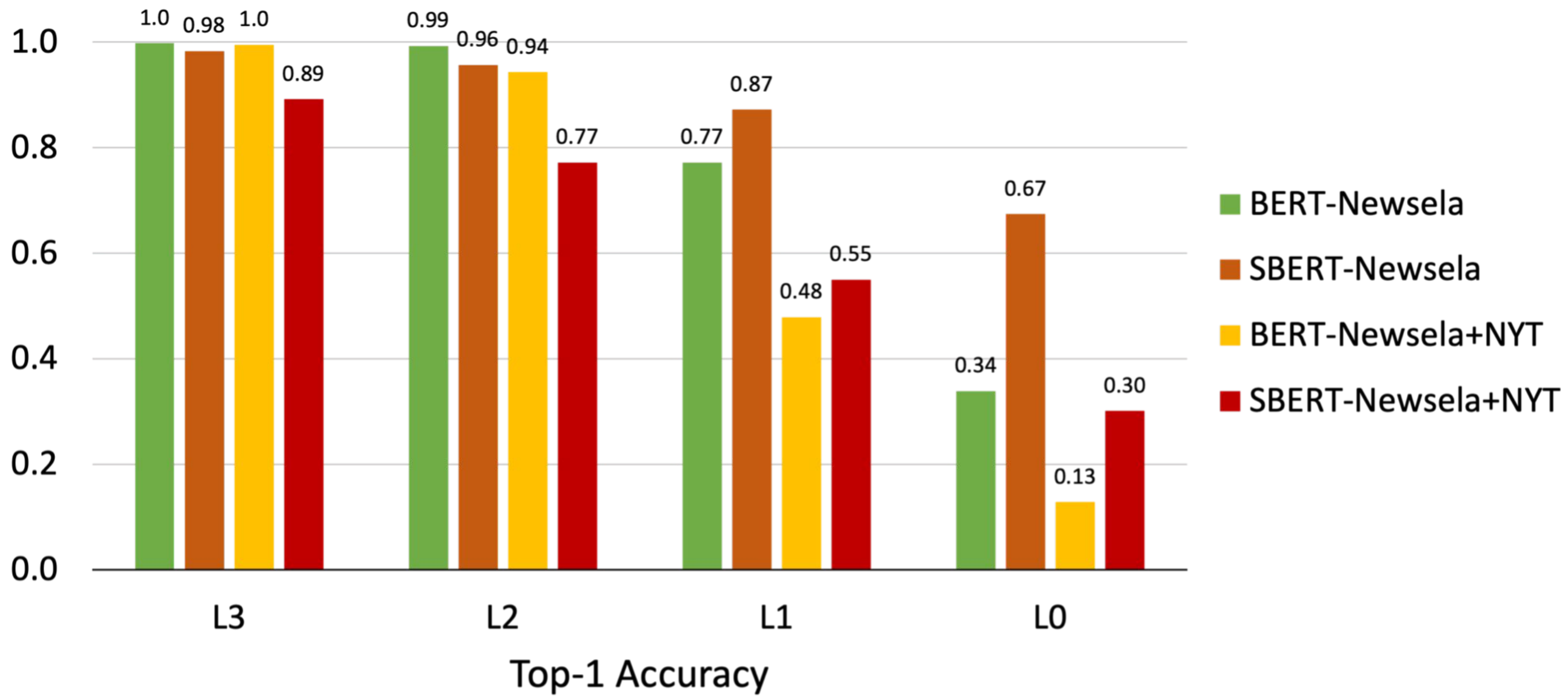
Complexity-Sensitive Retrieval: Results



Complexity-Sensitive Retrieval: Results



Complexity-Sensitive Retrieval: Results



Complexity-Sensitive Retrieval: Problem with Results

- In practice, there will not be only a single relevant document

Complexity-Sensitive Retrieval: Problem with Results

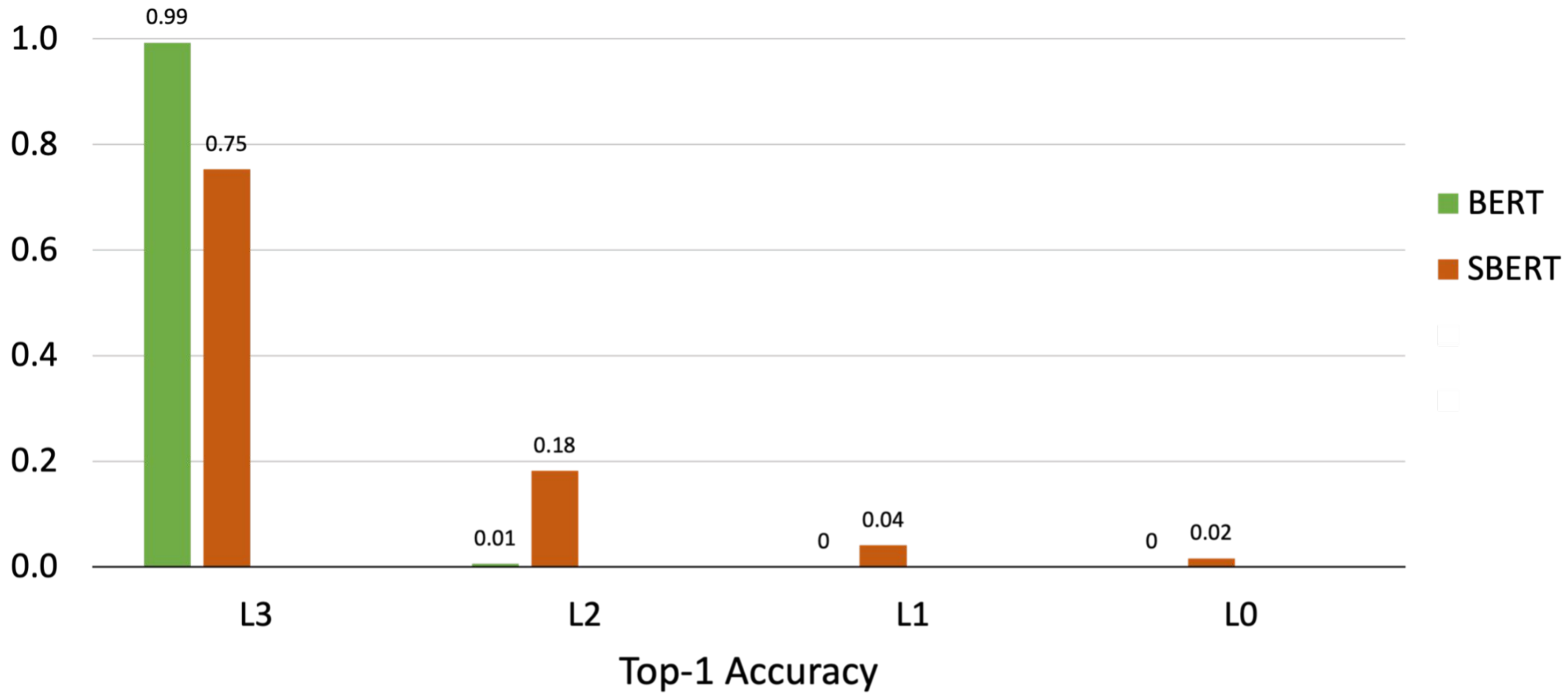
- In practice, there will not be only a single relevant document

- When considering all aligned documents at once, those written at lower complexity levels (D_{i2}, D_{i1}, D_{i0}) are rarely ranked highest
  - Known as the *Keep All Aligned Documents* evaluation approach, or KAAD

Complexity-Sensitive Retrieval: Problem with Results

- In practice, there will not be only a single relevant document
- When considering all aligned documents at once, those written at lower complexity levels (D_{i2}, D_{i1}, D_{i0}) are rarely ranked highest
  - Known as the *Keep All Aligned Documents* evaluation approach, or KAAD
- Need: A way to filter documents based on their complexity

Complexity-Sensitive Retrieval: KAAD Newsela Results



Document-Level Complexity Prediction: Method

Use a fine-tuned BERT model to predict document complexity

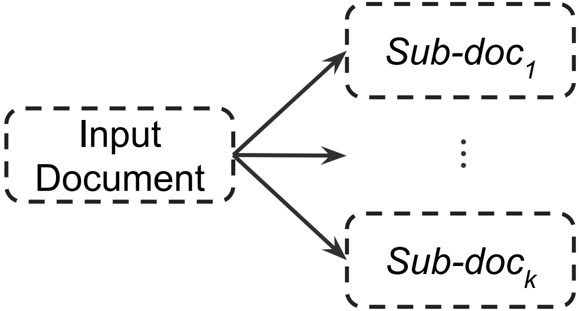
Document-Level Complexity Prediction: Method

Use a fine-tuned BERT model to predict document complexity

Input Document

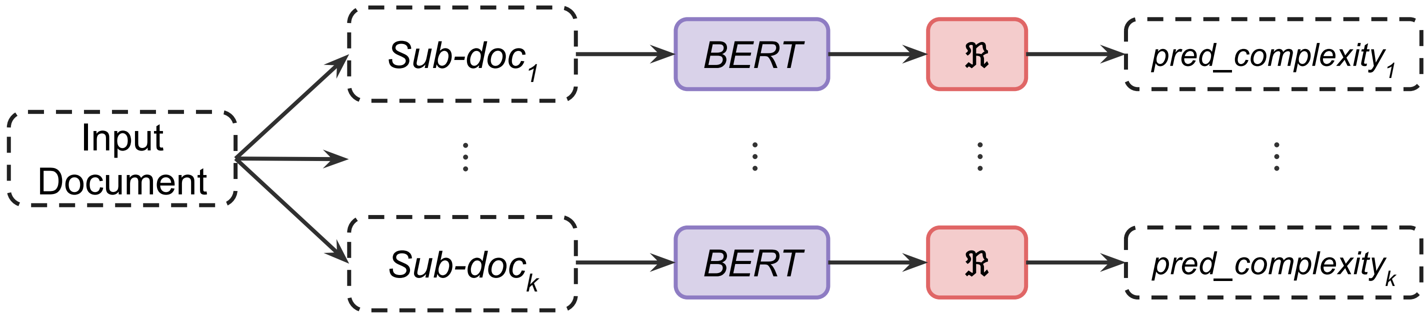
Document-Level Complexity Prediction: Method

Use a fine-tuned BERT model to predict document complexity



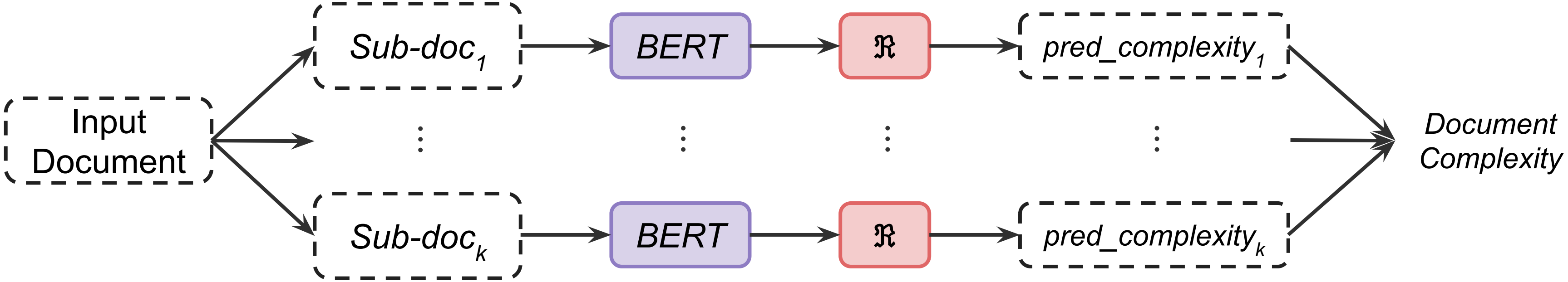
Document-Level Complexity Prediction: Method

Use a fine-tuned BERT model to predict document complexity

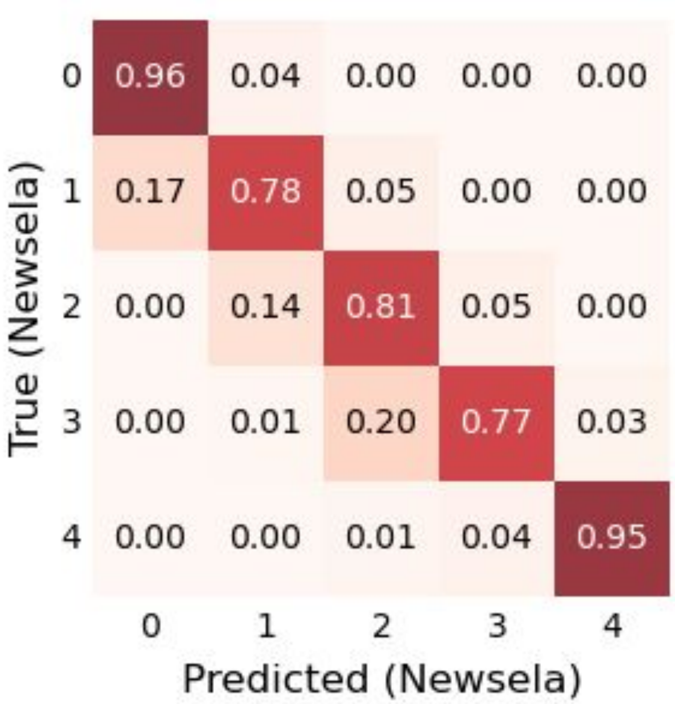


Document-Level Complexity Prediction: Method

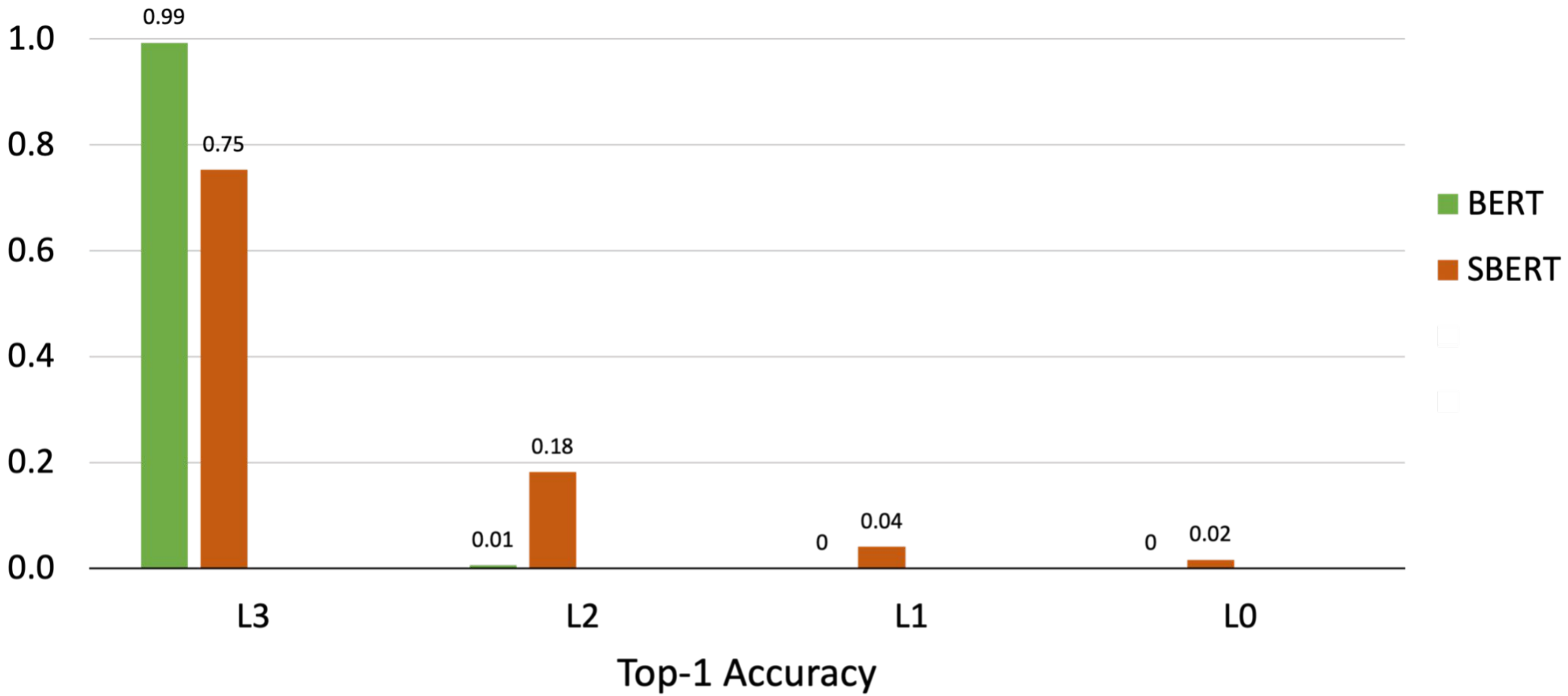
Use a fine-tuned BERT model to predict document complexity



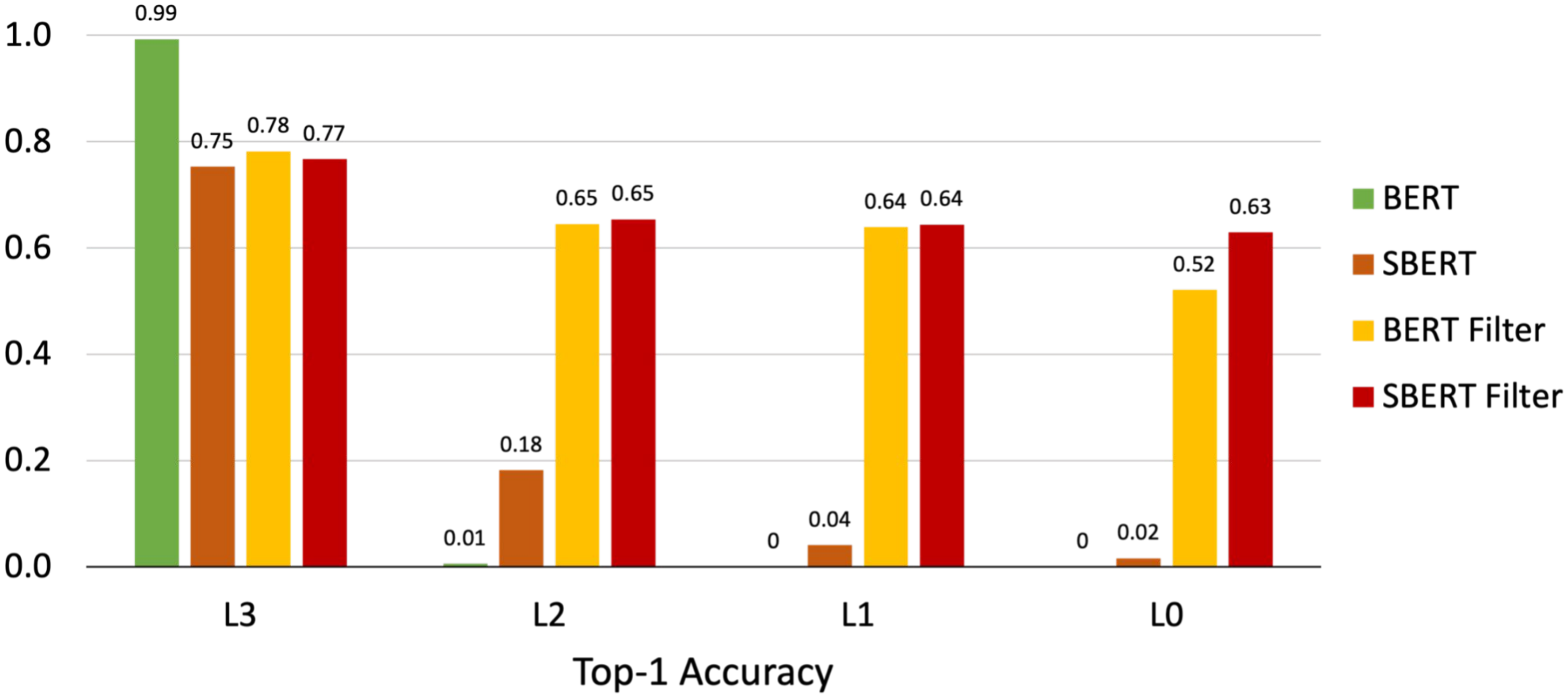
Document-Level Complexity Prediction: Results



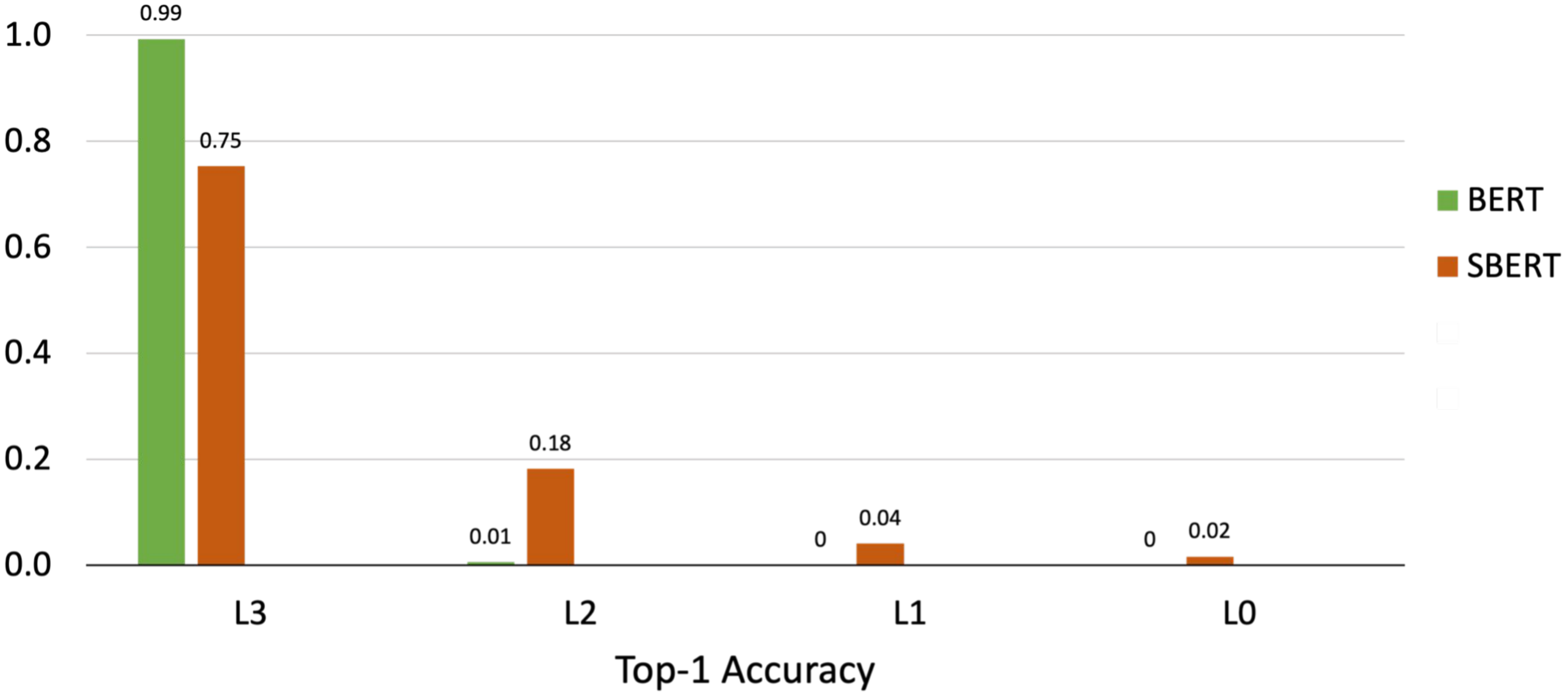
Complexity-Sensitive Retrieval: KAAD Newsela Results



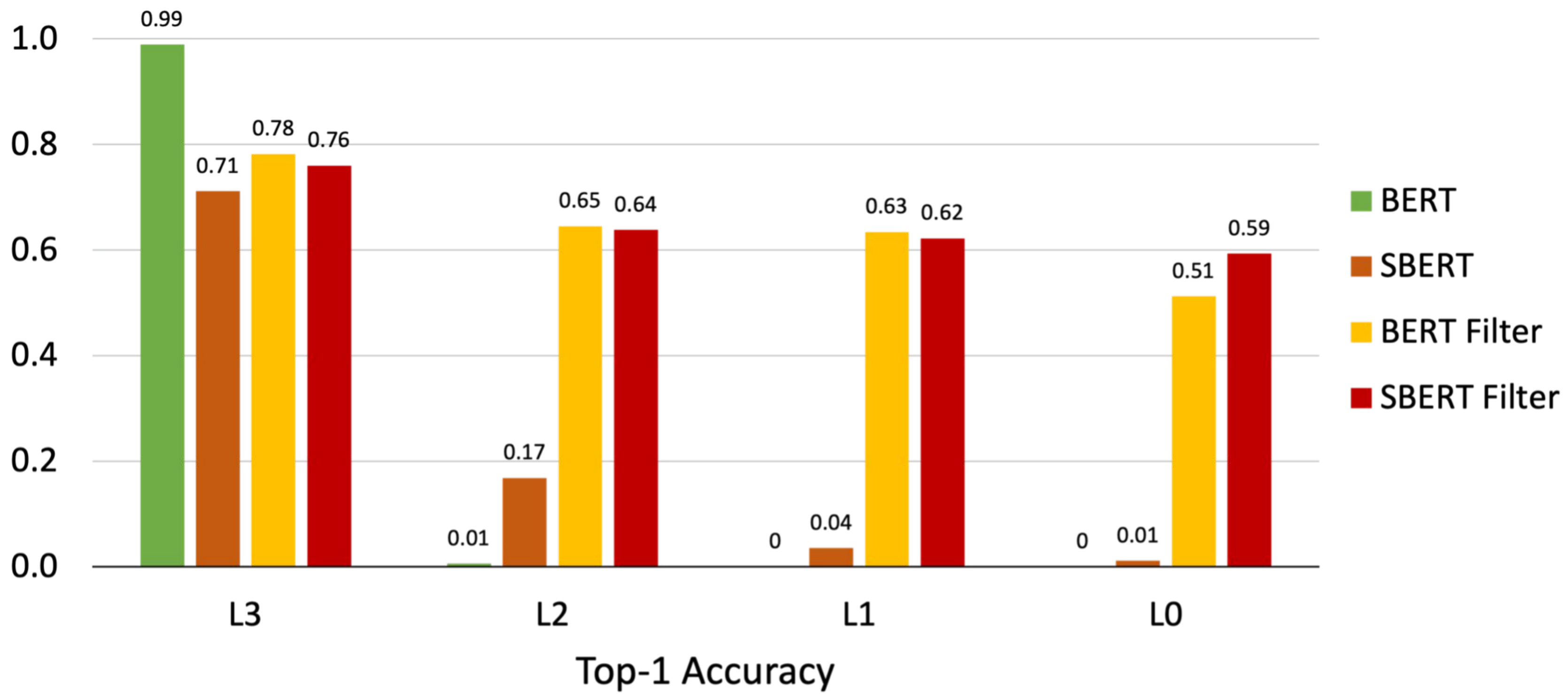
Complexity-Sensitive Retrieval: KAAD Newsela Results



Complexity-Sensitive Retrieval: KAAD Newsela+NYT Results



Complexity-Sensitive Retrieval: KAAD Newsela+NYT Results



Takeaways

- Utilizing additional context allows for improvement of lexical simplification systems

Takeaways

- Utilizing additional context allows for improvement of lexical simplification systems
- Incorporating diverse candidate generation and re-ranking increases performance of neural sentence simplification models

Takeaways

- Utilizing additional context allows for improvement of lexical simplification systems

- Incorporating diverse candidate generation and re-ranking increases performance of neural sentence simplification models

- Leveraging pre-trained language models to estimate the quality of sentence simplification outputs improves correlation with human judgments

Takeaways

- Utilizing additional context allows for improvement of lexical simplification systems
- Incorporating diverse candidate generation and re-ranking increases performance of neural sentence simplification models
- Leveraging pre-trained language models to estimate the quality of sentence simplification outputs improves correlation with human judgments
- Reformulating text simplification as a retrieval task has the potential to make the problem more practically useful

Relevant Publications

Reno Kriz, Eleni Miltsakaki, Marianna Apidianaki, and Chris Callison-Burch. *Simplification using paraphrases and context-based lexical substitution*. In NAACL 2018.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. *Complexity-weighted loss and diverse reranking for sentence simplification*. In NAACL 2019.

Daphne Ippolito*, **Reno Kriz***, Maria Kustikova, João Sedoc, and Chris Callison-Burch. *Comparison of Diverse Decoding Methods from Conditional Language Models*. In ACL 2019.

Reno Kriz, Marianna Apidianaki, and Chris Callison-Burch. *Simple-QE: Better Automatic Quality Estimation for Text Simplification*. arXiv 2020.

Reno Kriz, Eleni Miltsakaki, Jaime Rojas, Rebecca Iglesias-Flores, Megha Mishra, Marianna Apidianaki, and Chris Callison-Burch. *Recasting Text Simplification as a Document Retrieval Task*. arXiv 2021.

Thanks to my collaborators!

Thanks to my collaborators!