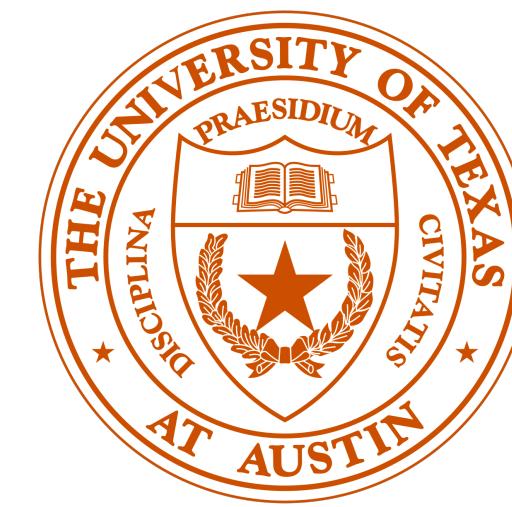


Decontextualization: Making Sentences Stand-Alone

**Eunsol Choi, Jennimaria Palomaki, Matthew Lamm,
Tom Kwiatkowski, Dipanjan Das, Michael Collins**



Google AI



Decontextualization: Task

Given a Wikipedia page and a sentence inside it, rewrite the sentence such that it can stand alone.

Decontextualization: Example

Croatia at the FIFA World Cup

From Wikipedia, the free encyclopedia

Croatia national football team have appeared in the [FIFA World Cup](#) on five occasions (in 1998, 2002, 2006, 2014 and 2018) since gaining independence in 1991. Before that, from 1930 to 1990 Croatia was part of [Yugoslavia](#). Their best result thus far was reaching the [2018 final](#), where they lost 4-2 to [France](#).

Original

Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.

Decontextualized

The Croatia national football team's best result *in the FIFA world cup* thus far was reaching the 2018 final, where they lost 4-2 to France.

Examples

Original

In 1850 , the first experimental electric telegraph line was started between Calcutta and Diamond Harbour.

Decontextualized

In 1850, the first experimental electric telegraph line *in India* was started between Calcutta and Diamond Harbour.

Original

It stars professional dancer Lauren Taft alongside Petricca .

Decontextualized

The music video for Shut Up and Dance stars professional dancer Lauren Taft alongside *Nicholas* Petricca .

Original

Christy 's daughter , Violet (Sadie Calvano) , who was born when Christy was 16, has also become a teen mother by her boyfriend , Luke (Spencer Daniels) .

Decontextualized

In Mom (TV series), Christy *Plunkett* (Anna Faris)'s daughter , Violet *Plunkett* (Sadie Calvano) , who was born when Christy was 16 , has also become a teen mother by her boyfriend , Luke (Spencer Daniels) .

Examples

Original

In 1850 , the first experimental electric telegraph line was started between Calcutta and Diamond Harbour.

Bridging

Decontextualized

In 1850, the first experimental electric telegraph line *in India* was started between Calcutta and Diamond Harbour.

Original

It stars professional dancer Lauren Taft alongside Nicholas Petricca .

Coreference Resolution

Decontextualized

The music video for Shut Up and Dance stars professional dancer Lauren Taft alongside Nicholas Petricca .

Original

Christy 's daughter , Violet Plunkett (Anna Faris) , who was born when Christy was 16, has also become a teen mother by her boyfriend , Luke (Spencer Daniels) .

Global Scoping

Decontextualized

In Mom (TV series), Christy Plunkett (Anna Faris)'s daughter , Violet Plunkett (Sadie Calvano) , who was born when Christy was 16 , has also become a teen mother by her boyfriend , Luke (Spencer Daniels) .

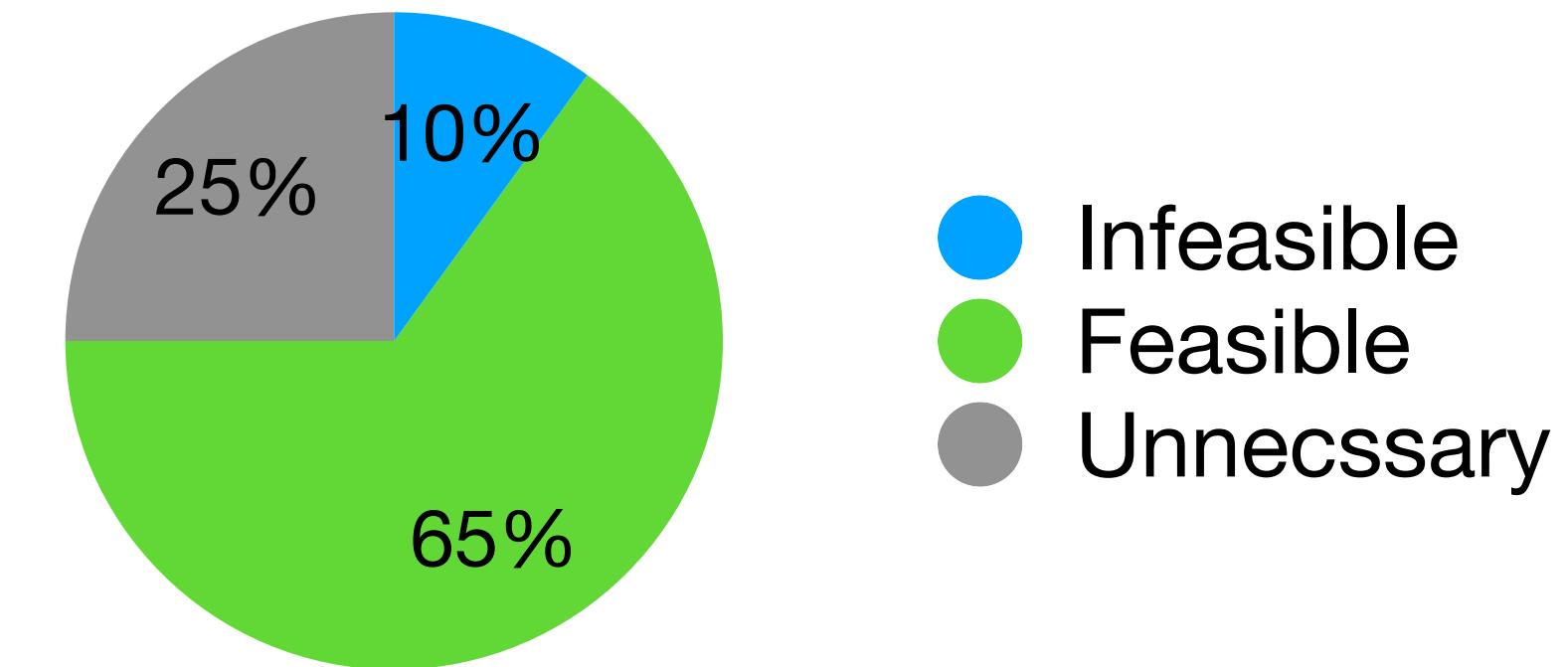
Data Analysis: Semantics of Edits

Phenomena	Example	% examples with phenomena
Pronoun / NP Swaps	he → Bernie Sanders	40%
Bridging	characters → characters of Toy Story 3	13%
Global Scoping	→ at the Cannes Film Festival	7%
Name / Acronym Expansion	Clinton → Hillary Clinton	11.5%
Add Information	Charles Darwin → Charles Darwin, an English naturalist and biologist	10%
Discourse Marker removal	However →	3.5%

- Coreference resolution alone cannot handle decontextualization!

Data Statistics

- We annotated about 10K examples (train set, one way), 4K examples (evaluation set, 5-way)
- Distribution of example types (agreement Fleiss' Kappa: 0.50)



- The length of sentence after the decontextualization was increased by about 30%.

Scope of Our Study

- Does not cover implicatures
- Extra-linguistic (spatial, temporal) contexts stay

Original

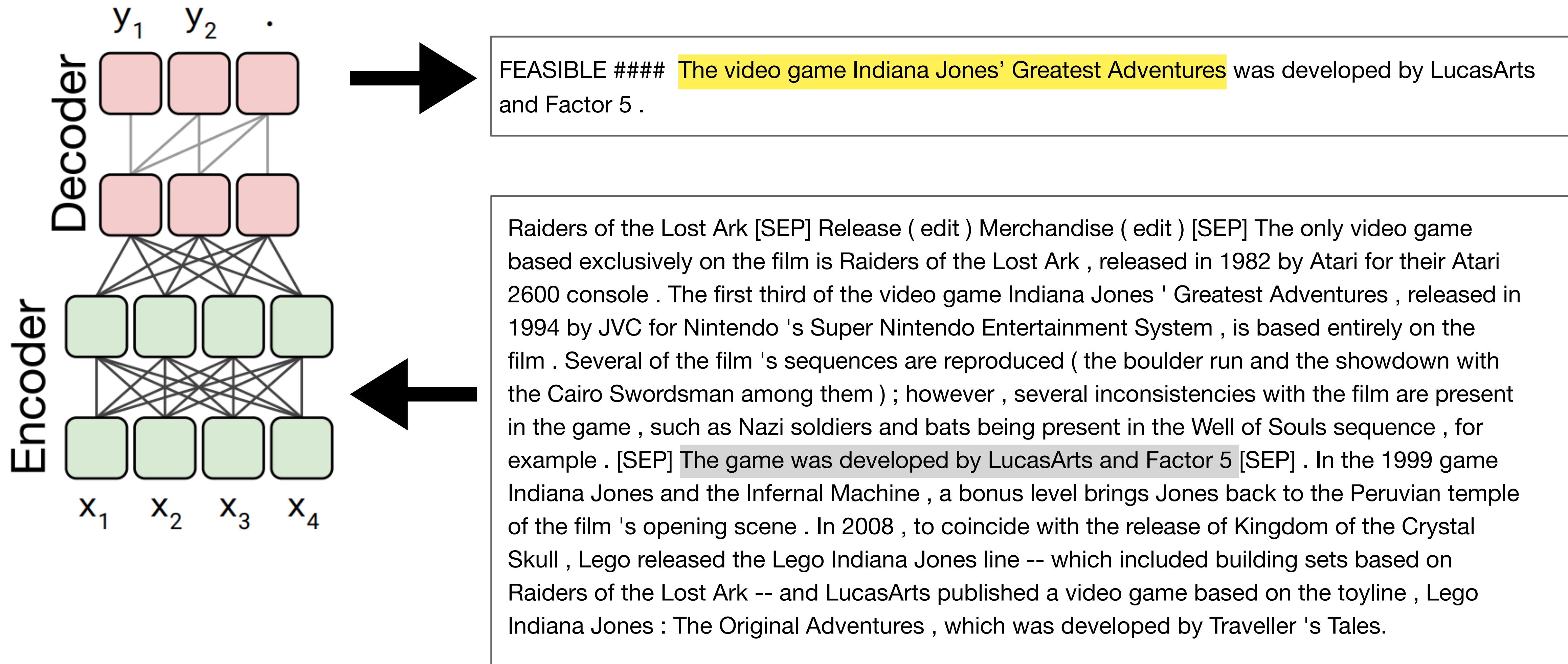
Their best result **thus far** was reaching the 2018 final, where they lost 4-2 to France.

Decontextualized

The Croatia national football team's best result in the FIFA world cup thus far was reaching the 2018 final, where they lost 4-2 to France.

- Does not cover diverse types of text – only English Wikipedia.

Supervised Baseline: Seq2Seq Model



Evaluation Metric

- SARI score [Xu et al, TACL 16]: compares original sentence, reference sentences, and output sentence, computes token-level F1 scores on **edits**
 - Add: overlap for added tokens in the output and those in the references
 - Del: overlap for deleted tokens in the output and those in the references

Original

Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.

References

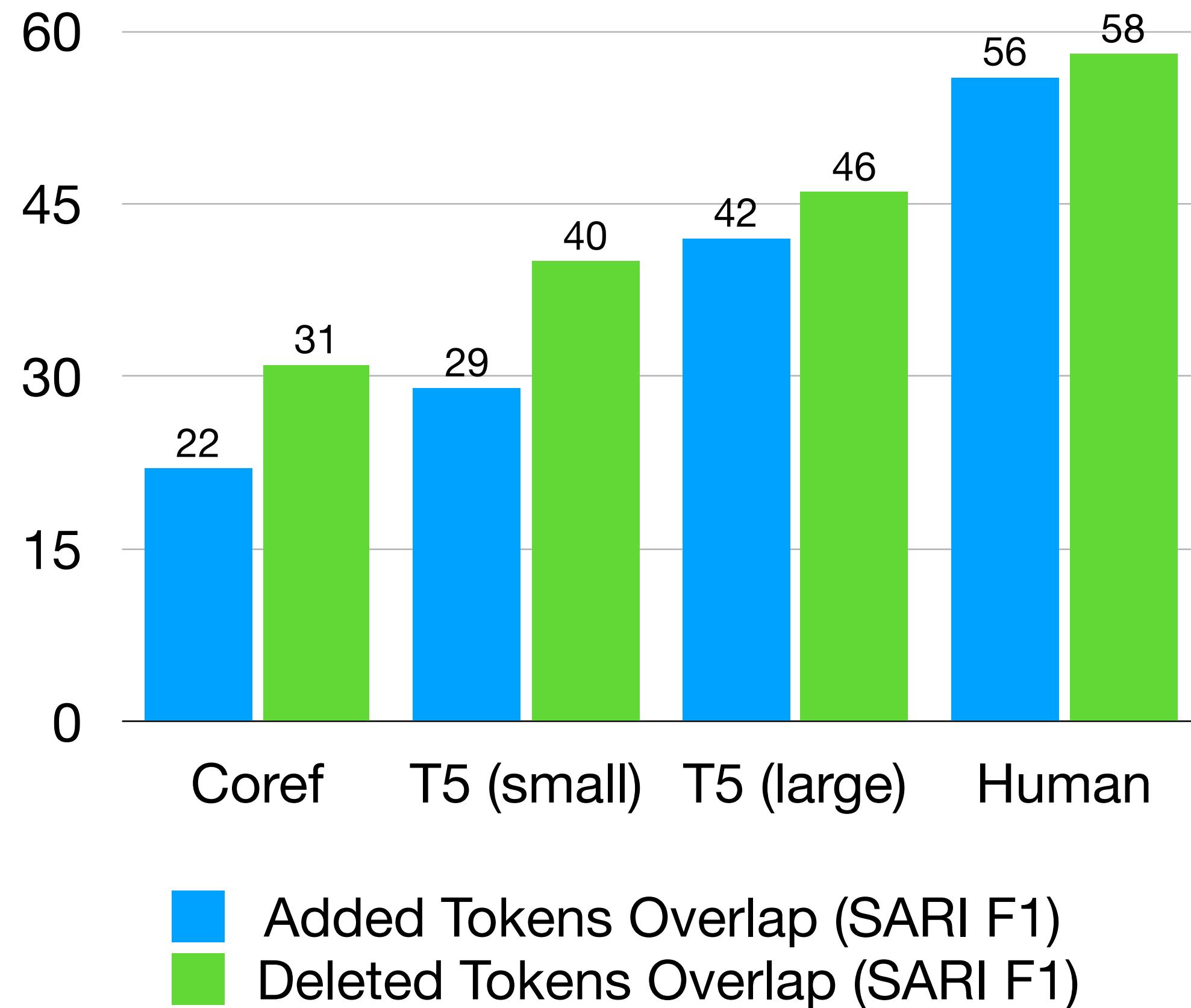
[Their / *The Croatia national football team's*] best result *in the FIFA world cup* thus far was reaching the 2018 final, where they lost 4-2 to France.

[Their / *Croatia's*] best result *in the FIFA world cup* thus far was reaching the 2018 final, where they lost 4-2 to France.

Model Output

[Their / *Croatia's national team's*] best result *in the world cup* thus far was reaching the 2018 final, where they lost 4-2 to France.

Automatic Evaluation Results:



- Fine-tuned seq2seq model performs competitively, especially for larger variants.
- Simpler metrics (i.e., exact sentence match) show a smaller gap between human and models.

Why do we want to make sentences stand alone?

Applications

- Decontextualization as pre-processing
- Decontextualization as is

Retrieval Corpus for Open Domain QA

Question: Which US president is known as pragmatic caretaker?

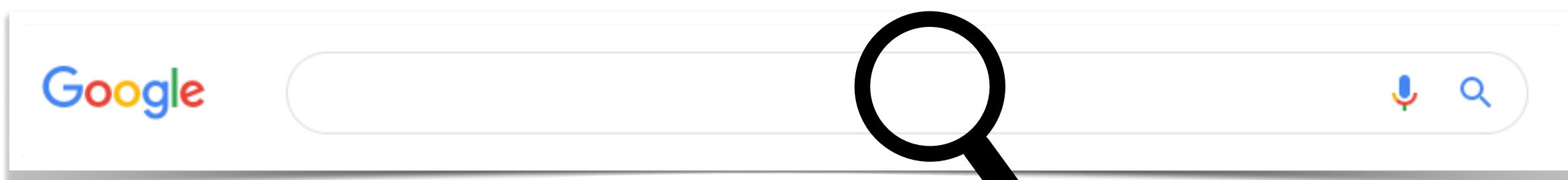


WIKIPEDIA
The Free Encyclopedia



Retrieval Corpus for Open Domain QA

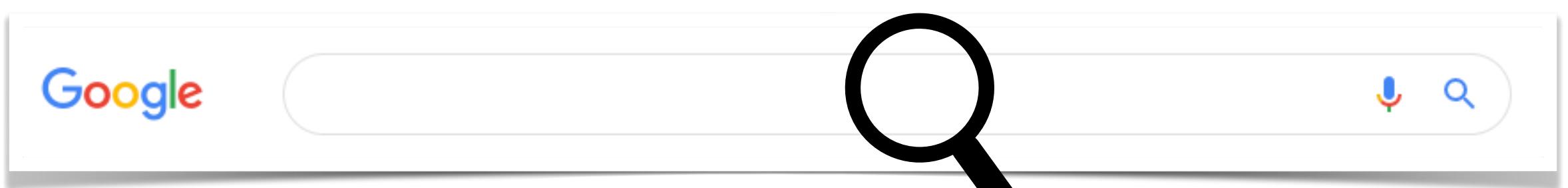
Bush was widely seen as a "pragmatic caretaker" president who lacked a unified and compelling long-term theme in his efforts. Indeed, Bush's sound bite where he refers to the issue of overarching purpose as "the vision thing" has become a metonym applied to other political figures accused of similar difficulties.



Starting in 2010, Clinton began serving as Assistant Vice-Provost for the Global Network University of New York University, working on international recruitment strategies. She is the co-founder of the Of Many Institute for Multifaith Leadership at NYU and serves as its co-chair.

Retrieval Corpus for Open Domain QA

George H.W. Bush was widely seen as a "pragmatic caretaker" president who lacked a unified and compelling long-term theme in his efforts. Indeed, Bush's sound bite where he refers to the issue of overarching purpose as "the vision thing" has become a metonym applied to other political figures accused of similar difficulties.



Starting in 2010, Chelsea Clinton began serving as Assistant Vice-Provost for the Global Network University of New York University, working on international recruitment strategies. She is the co-founder of the Of Many Institute for Multifaith Leadership at NYU and serves as its co-chair.

Applications

- Decontextualization as pre-processing
- Decontextualization as is

Original Sentence Answer vs. Decontextualized Sentence Answer

Q: what is the most viewed video on youtube in 24 hours?

Q: what's the best results from Croatia's football team in the World cup?

Original Sentence Answer vs. Decontextualized Sentence Answer

Original

Q: what is the most viewed video on youtube in 24 hours?

The most viewed music video **in this time period** is Taylor Swift 's Look What You Made Me Do.

Decontextualized

Q: what's the best results from Croatia's football team in the World cup?

Their best result thus far was silver position at the 2018 final , where they lost 4-2 to France .

The most viewed music video **on Youtube within 24 hours of its release** is Taylor Swift 's Look What You Made Me Do.

The Croatia national football team's best result thus far in the FIFA World Cup was reaching the 2018 final, where they lost 4-2 to France.

Original Sentence Answer vs. Decontextualized Sentence Answer

Q: what is the most viewed video on youtube in 24 hours?

Original

The most viewed music video **in this time period** is Taylor Swift 's Look What You Made Me Do.

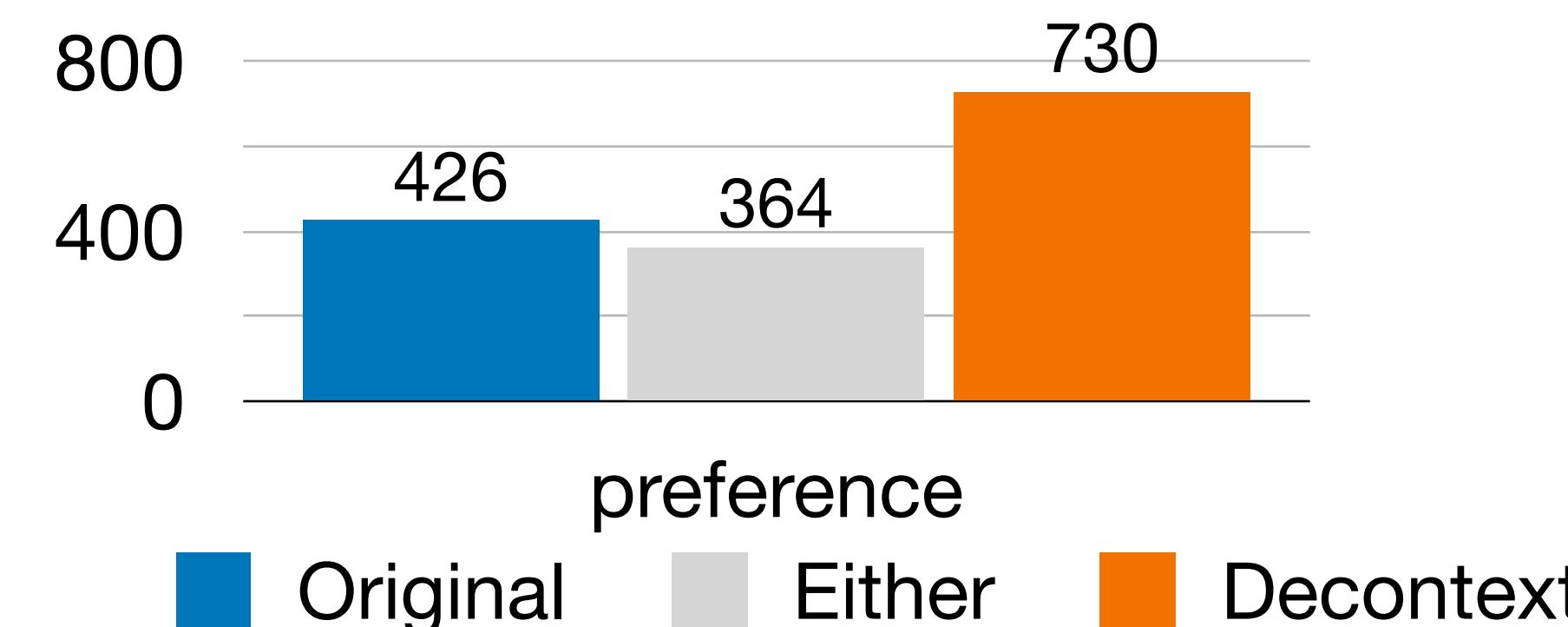
Q: what's the best results from Croatia's football team in the World cup?

Decontextualized

The most viewed music video **on Youtube within 24 hours of its release** is Taylor Swift 's Look What You Made Me Do.

Their best result thus far was silver position at the 2018 final , where they lost 4-2 to France .

The Croatia national football team's best result thus far **in the FIFA World Cup** was reaching the 2018 final, where they lost 4-2 to France.



Conclusion

- Decontextualization recovers implicit arguments and ambiguous phrases, making the sentence interpretable and clear without the full document context.
- We can train models to make sentences stand alone with modest amount of annotated data.
- The decontextualization can help downstream NLP systems as a subcomponent of a larger system and by generating useful stand-alone sentences for user facing tasks.
- Data & Colab released!

Paper Link: <https://arxiv.org/pdf/2102.05169.pdf>

