FastKASSIM: A Fast Tree Kernel-Based Syntactic Similarity Metric

Maximillian Chen, Caitlyn Chen, Xiao Yu, Zhou Yu

Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

Existing Approaches to Syntactic Similarity

Utterance 1: When we hate, we always move away from the grace of God. When we become resentful and unforgiving, the world around us seems spiteful and meaningless.

Utterance 2: How can you be skiing if you are already swimming?

FastKASSIM Score: 0.219 ✓

CASSIM Score: 0.838 ✗

LSM Score: 0.623 ✗

Utterance 1: I like swimming because it is cool.

Utterance 2: I love running because it is fun.

FastKASSIM Score: 0.928 ✓

CASSIM Score: 0.962 ✓

LSM Score: 1.0 ✓

- **Syntactic similarity** is an important evaluation for syntactic consistency.
- **Existing approaches** to document-level similarity **are too computational expensive or inaccurate** to be feasible.

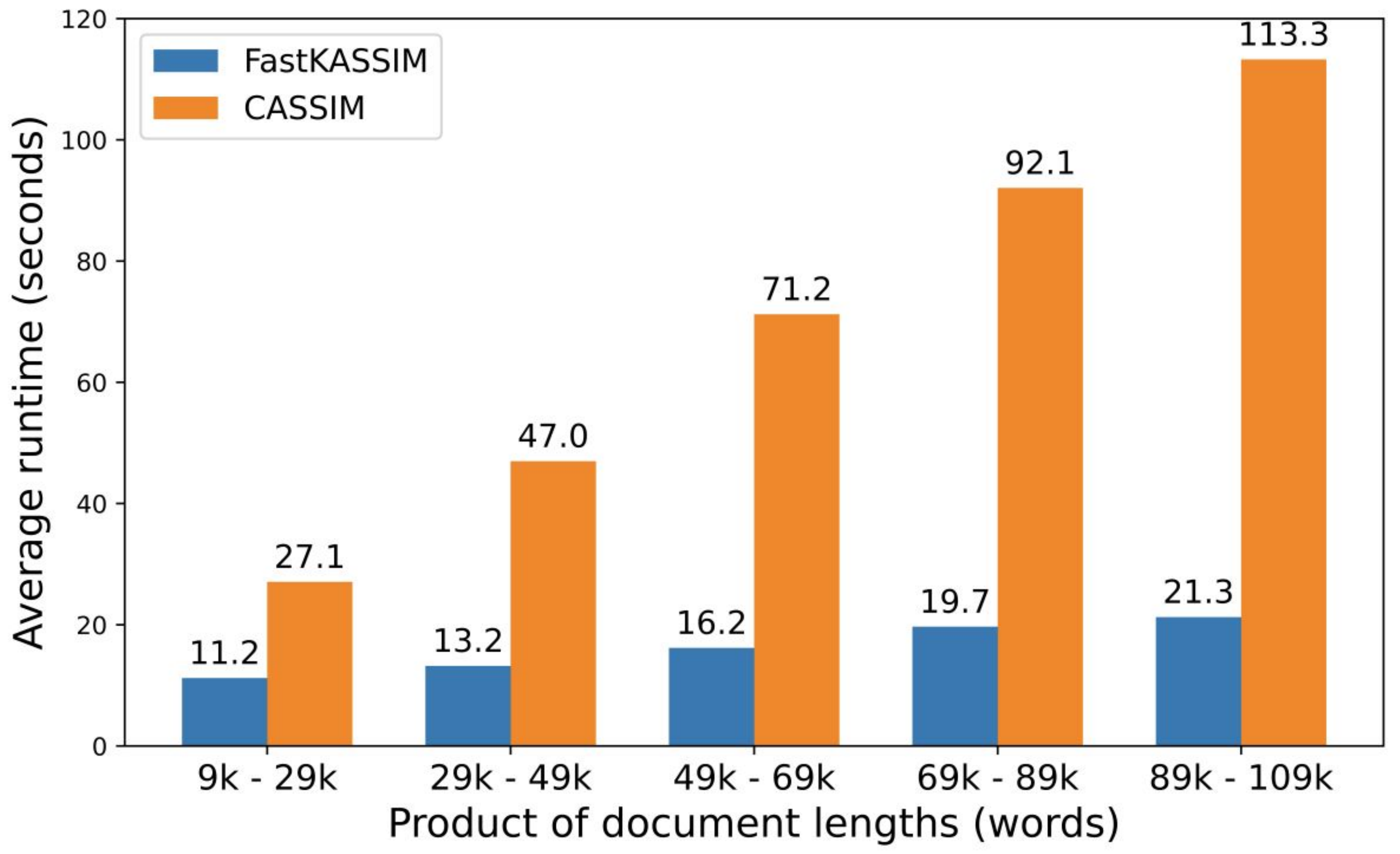
Identifying Similar and Dissimilar Documents

Metric	Acc.	SR	SP	DR	DP
LSM	46.2	92.5	30.8	30.7	92.5
LSM_a	65.6	81.1	40.6	60.4	90.6
CASSIM	25.1	**100.**	25.0	0.11	**100.**
CASSIM_a	48.8	47.7	23.8	49.2	73.8
BERTScore	25.0	100.	25.0	00.0	00.0
BERTScore_a	74.6	99.3	49.6	66.4	99.6
Sentence-BERT	18.9	19.8	74.0	2.70	0.20
Sentence-BERT_a	34.3	9.50	19.2	59.3	39.3
FastKASSIM	**88.3**	96.1	**69.1**	**98.5**	85.6

- **FastKASSIM is holistically better** at discerning between similarity and dissimiliarity than existing syntax metrics and embedding-based semantic similarity metrics

Paper **Code**

End-to-End Runtime Improvements



- FastKASSIM, powered by the **Label-based Tree Kernel**, runs up to **5.32x faster** than its predecessor and scales better with document size

Application: Authorship Attribution

Features	**Acc.**$_{(\sigma)}$	**F1**$_{(\sigma)}$
Majority Baseline	0.767	0.868
Bag of Words	$0.892_{(0.02)}$	$0.867_{(0.02)}$
Bag of Words + Syntax	$0.923_{(0.02)}$	$0.922_{(0.01)}$
RoBERTa	$0.939_{(0.01)}$	$0.935_{(0.00)}$
RoBERTa + Syntax	$\mathbf{0.945}_{(0.01)}$	$\mathbf{0.938}_{(0.01)}$

- **Using FastKASSIM to create syntax feature vectors** improves authorship attribution classification performance on the Australian High Court Judgment dataset

Application: Persuasion on ChangeMyView

- **Matching communication styles** creates familiarity, which **improves conversational outcomes**[1,2]
- **Hypothesis:** Arguments which are more syntactically similar to opinions will be more successful on r/ChangeMyView
- **Findings:**
  - Successful arguments **tend to be more syntactically similar** to viewpoints
  - Arguments which are similar to viewpoints **tend to be more successful**

[1] Jared R Curhan and Alex Pentland. 2007. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology, 92(3):802.

[2] Howard Giles. 2016. Communication accommodation theory. The international encyclopedia of communication theory and philosophy, pages 1–7.