A Neural Attention Model for Abstractive Sentence Summarization

Alexander Rush Sumit Chopra Jason Weston

Facebook AI Research

Harvard SEAS

Sentence Summarization

Source

Russian Defense Minister Ivanov called Sunday for the creation of a joint front for combating global terrorism.

Target

Russia calls for joint front against terrorism.

Summarization Phenomena:

Sentence Summarization

Source

Russian Defense Minister Ivanov *called Sunday for the creation of a joint front for combating global terrorism.*

Target

Russia *calls for joint front against terrorism.*

Summarization Phenomena:

- **Generalization**

Sentence Summarization

Source

Russian Defense Minister Ivanov called **Sunday** *for the creation of a joint front for combating global terrorism.*

Target

Russia calls for joint front against terrorism.

Summarization Phenomena:

- Generalization
- **Deletion**

Sentence Summarization

Source

Russian Defense Minister Ivanov called Sunday for the creation of a joint front **for combating** *global terrorism.*

Target

Russia calls for joint front **against** *terrorism.*

Summarization Phenomena:

- Generalization
- Deletion
- **Paraphrase**

Types of Sentence Summary

[Not Standardized]

- **Compressive**: deletion-only

  Russian Defense Minister Ivanov called Sunday for the creation of a joint front for combating global terrorism.

- **Extractive**: deletion and reordering

- **Abstractive**: arbitrary transformation

  Russia calls for joint front against terrorism.

Elements of Human Summary

Jing 2002

	Phenomenon	Abstract	Compress	Extract
(1)	Sentence Reduction	✓	✓	✓
(2)	Sentence Combination	✓	✓	✓
(3)	Syntactic Transformation	✓		✓
(4)	Lexical Paraphrasing	✓		
(5)	Generalization or Specification	✓		
(6)	Reordering	✓		✓

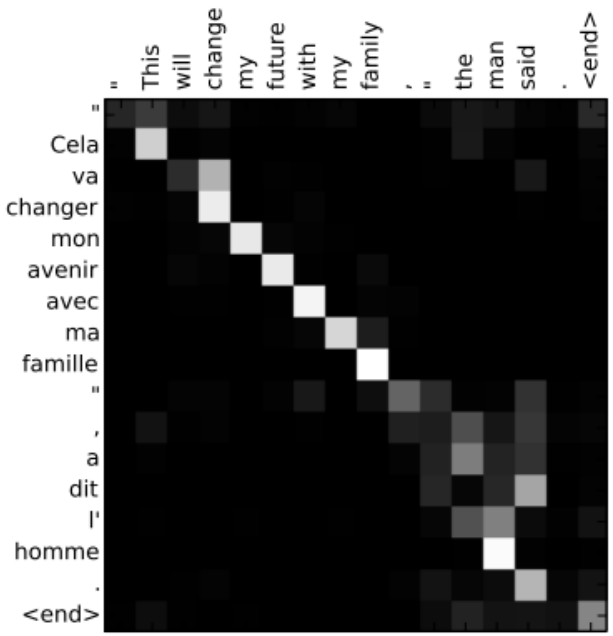
Related Work: Ext/Abs Sentence Summary

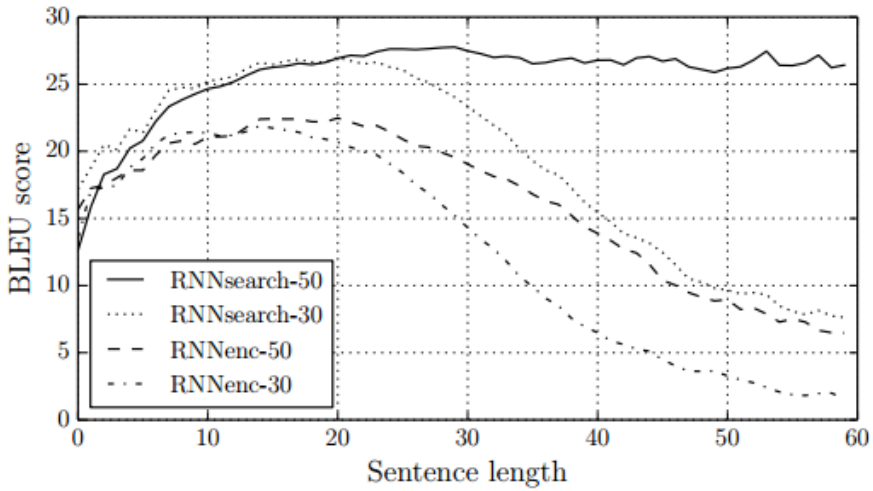
- **Syntax-Based** [Dorr, Zajic, and Schwartz 2003; Cohn and Lapata 2008; Woodsend, Feng, and Lapata 2010]
- **Topic-Based** [Zajic, Dorr, and Schwartz 2004]
- **Machine Translation-Based** [Banko, Mittal, and Witbrock 2000]
- **Semantics-Based** [Liu et al. 2015]

Related Work: Attention-Based Neural MT

Bahdanau, Cho, and Bengio 2014

- Use attention ("soft alignment") over source to determine next word.
- Robust to longer sentences versus encoder-decoder style models.
- No explicit alignment step, trained end-to-end.





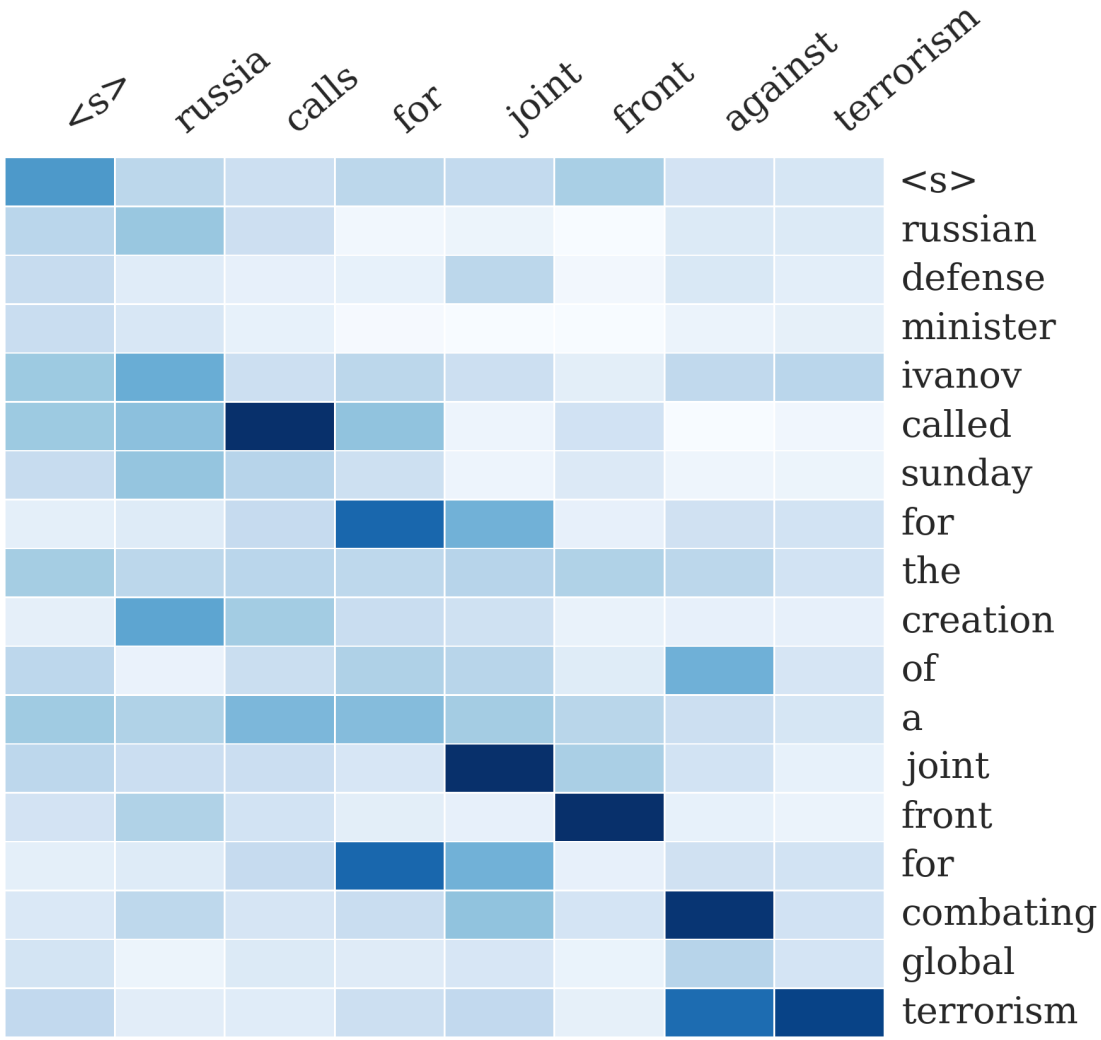
A Neural Attention Model for Summarization

Question: Can a data-driven model capture abstractive phenomenon necessary for summarization without explicit representations?

Properties:

- Utilizes a simple attention-based neural conditional language model.
- No syntax or other pipelining step, strictly data-driven.
- Generation is fully abstractive.

Attention-Based Summarization (ABS)



Model

Summarization Model

Notation:

- **x**; Source sentence of length M with $M >> N$
- **y**; Summarized sentence of length N (we assume N is given)

Summarization Model

Notation:

- $\mathbf{x}$; Source sentence of length M with $M >> N$
- $\mathbf{y}$; Summarized sentence of length N (we assume N is given)

Past work: Noisy-channel summary [Knight and Marcu 2002]

$$\arg\max_{\mathbf{y}} \log p(\mathbf{y}|\mathbf{x}) = \arg\max_{\mathbf{y}} \log p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$$

Summarization Model

Notation:

- $\mathbf{x}$; Source sentence of length M with $M >> N$
- $\mathbf{y}$; Summarized sentence of length N (we assume N is given)

Past work: Noisy-channel summary [Knight and Marcu 2002]

$$\arg\max_{\mathbf{y}} \log p(\mathbf{y}|\mathbf{x}) = \arg\max_{\mathbf{y}} \log p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$$

Neural machine translation: Direct neural-network parameteriziation

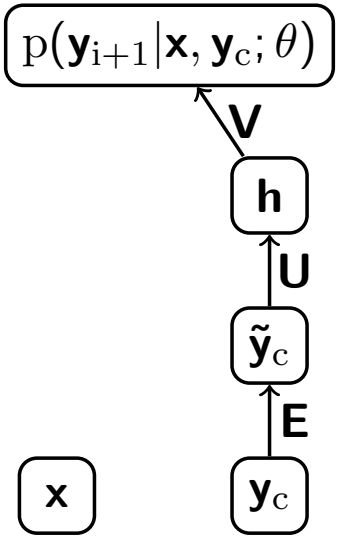
$$p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) \propto \exp(\mathrm{NN}(\mathbf{x}, \mathbf{y}_c; \theta))$$

where $\mathbf{y}_{i+1}$ is the current word and $\mathbf{y}_c$ is the context

Most neural MT is non-Markovian, i.e. $\mathbf{y}_c$ is full history (RNN, LSTM) [Kalchbrenner and Blunsom 2013; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2014]

Feed-Forward Neural Language Model

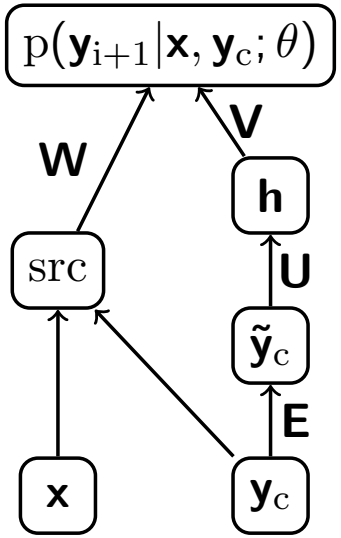
Bengio et al. 2003



$$
\begin{aligned}
\tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \ldots, \mathbf{E}\mathbf{y}_i], \\
\mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c), \\
p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h}).
\end{aligned}
$$

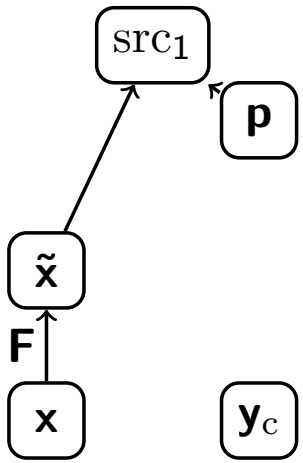
Feed-Forward Neural Language Model

Bengio et al. 2003



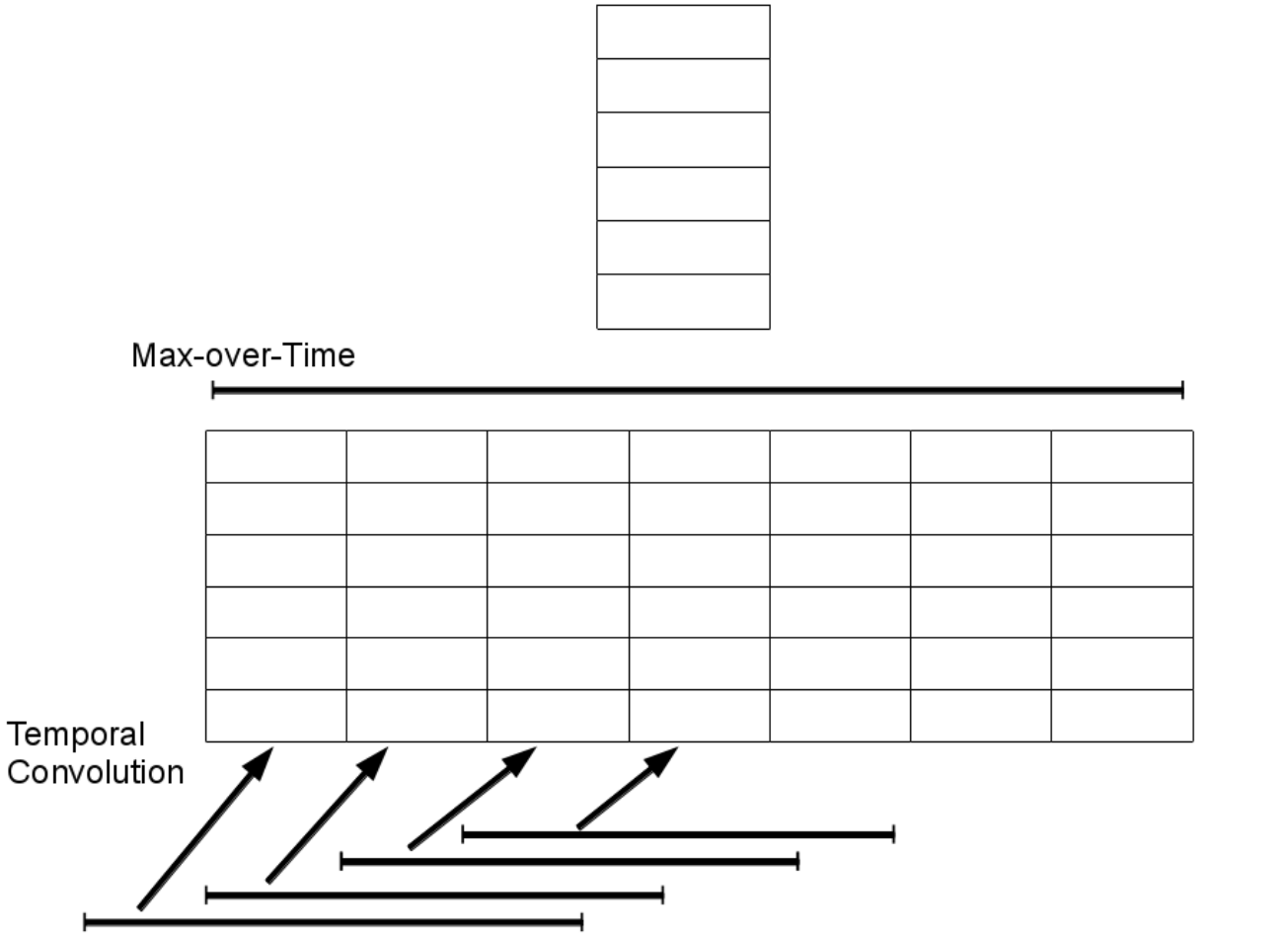
$$
\begin{aligned}
\tilde{\mathbf{y}}_c &= [\mathbf{E}\mathbf{y}_{i-C+1}, \ldots, \mathbf{E}\mathbf{y}_i], \\
\mathbf{h} &= \tanh(\mathbf{U}\tilde{\mathbf{y}}_c), \\
p(\mathbf{y}_{i+1}|\mathbf{y}_c, \mathbf{x}; \theta) &\propto \exp(\mathbf{V}\mathbf{h} + \mathbf{W}\mathrm{src}(\mathbf{x}, \mathbf{y}_c)).
\end{aligned}
$$

Source Model 1: Bag-of-Words Model



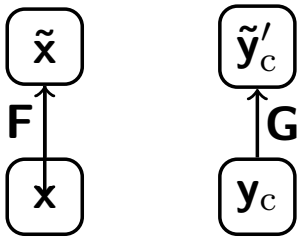
$$
\begin{aligned}
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\mathbf{p} &= [1/M, \ldots, 1/M], \text{[Uniform Distribution]} \\
\text{src}_1(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \tilde{\mathbf{x}}.
\end{aligned}
$$

Source Model 2: Convolutional Model



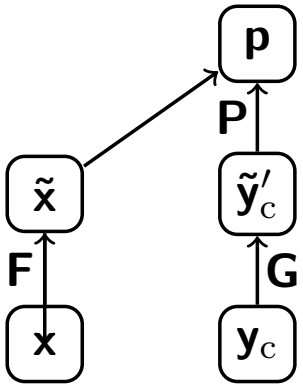
Russian Defense Minister Ivanov called Sunday for the creation of a joint front ...

Source Model 3: Attention-Based Model



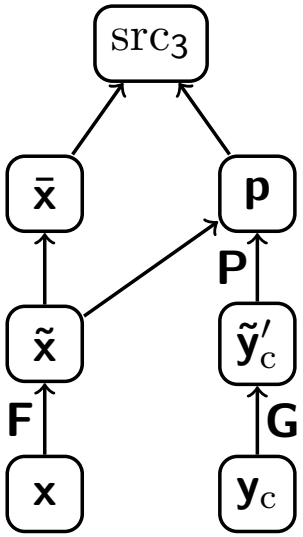
$$\begin{aligned}\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\ \tilde{\mathbf{y}}'_{\mathrm{C}} &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i],\end{aligned}$$

Source Model 3: Attention-Based Model



$$
\begin{aligned}
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{y}}'_c &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i], \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}}\mathbf{P}\tilde{\mathbf{y}}'_c), \quad \textbf{[Attention Distribution]}
\end{aligned}
$$

Source Model 3: Attention-Based Model

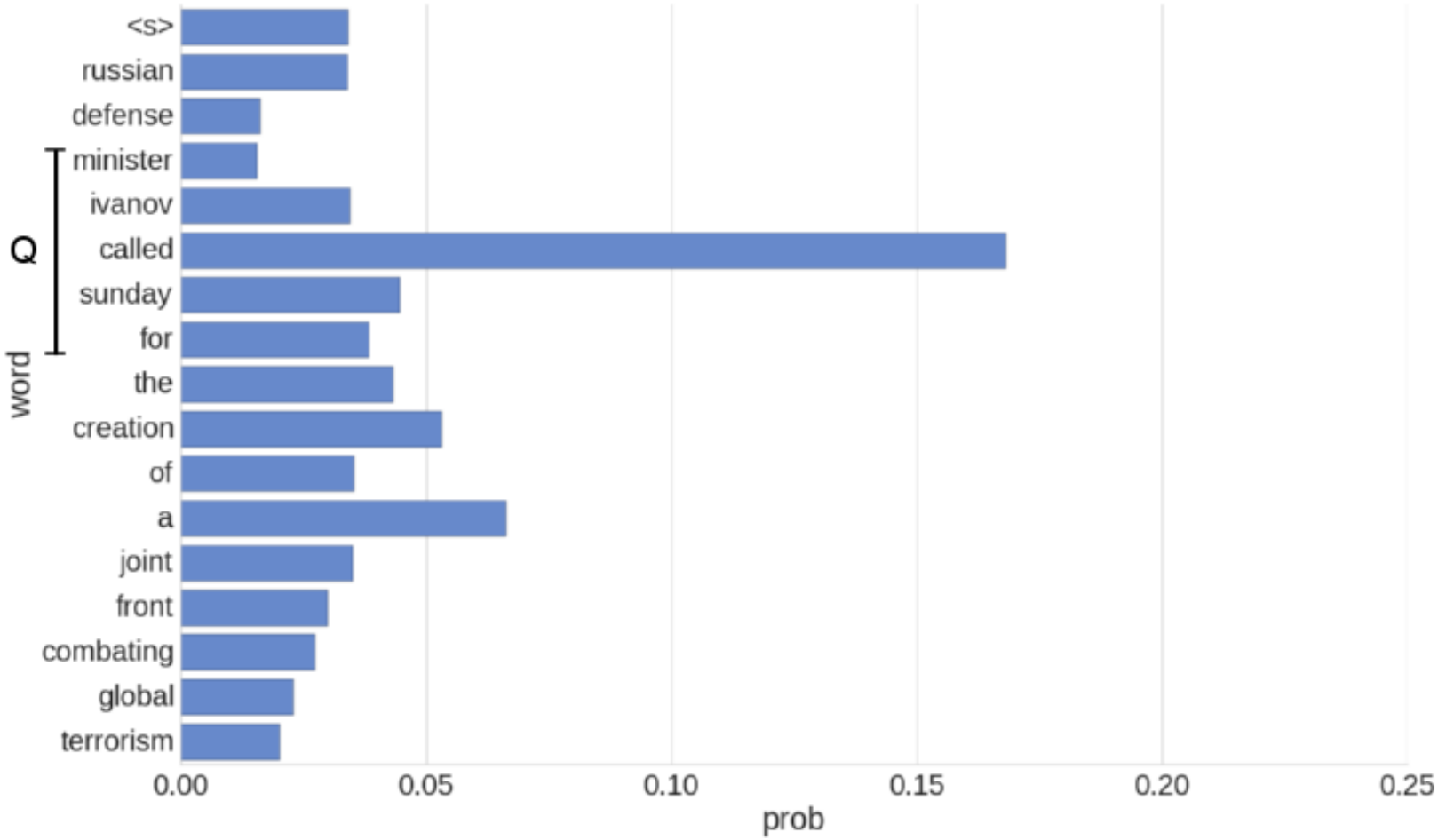


$$
\begin{aligned}
\tilde{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \ldots, \mathbf{F}\mathbf{x}_M], \\
\tilde{\mathbf{y}}'_c &= [\mathbf{G}\mathbf{y}_{i-C+1}, \ldots, \mathbf{G}\mathbf{y}_i], \\
\mathbf{p} &\propto \exp(\tilde{\mathbf{x}}\mathbf{P}\tilde{\mathbf{y}}'_c), \quad \textbf{[Attention Distribution]} \\
\forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-(Q-1)/2}^{i+(Q-1)/2} \tilde{\mathbf{x}}_i / Q, \quad \textbf{[Local Smoothing]} \\
\text{src}_3(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \bar{\mathbf{x}}.
\end{aligned}
$$

ABS Example

[⟨s⟩ Russia calls]	for
$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

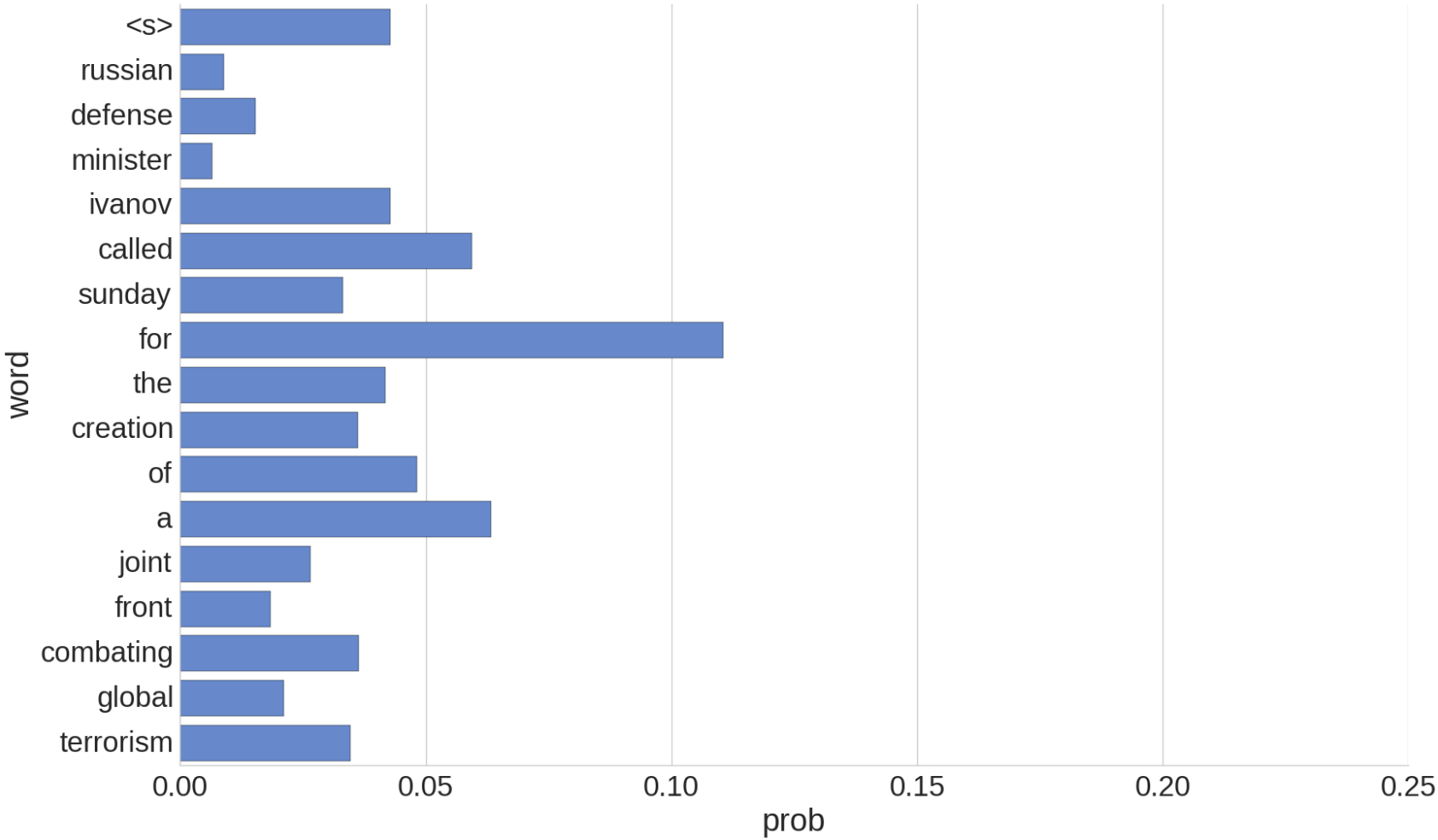
x



ABS Example

[⟨s⟩ Russia calls for]	**joint**
$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

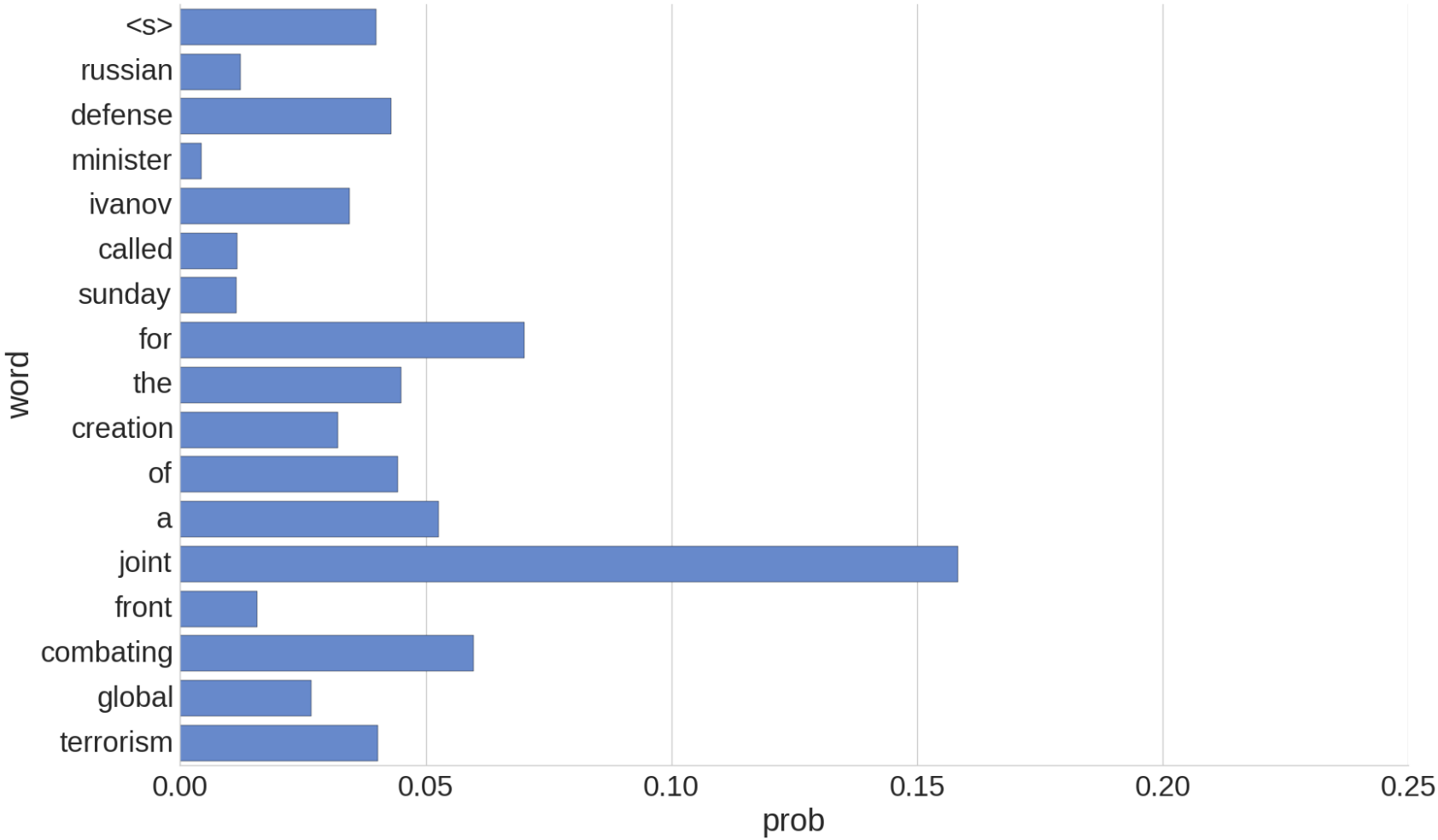
x



ABS Example

[⟨s⟩ Russia calls for joint]	**front**
$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

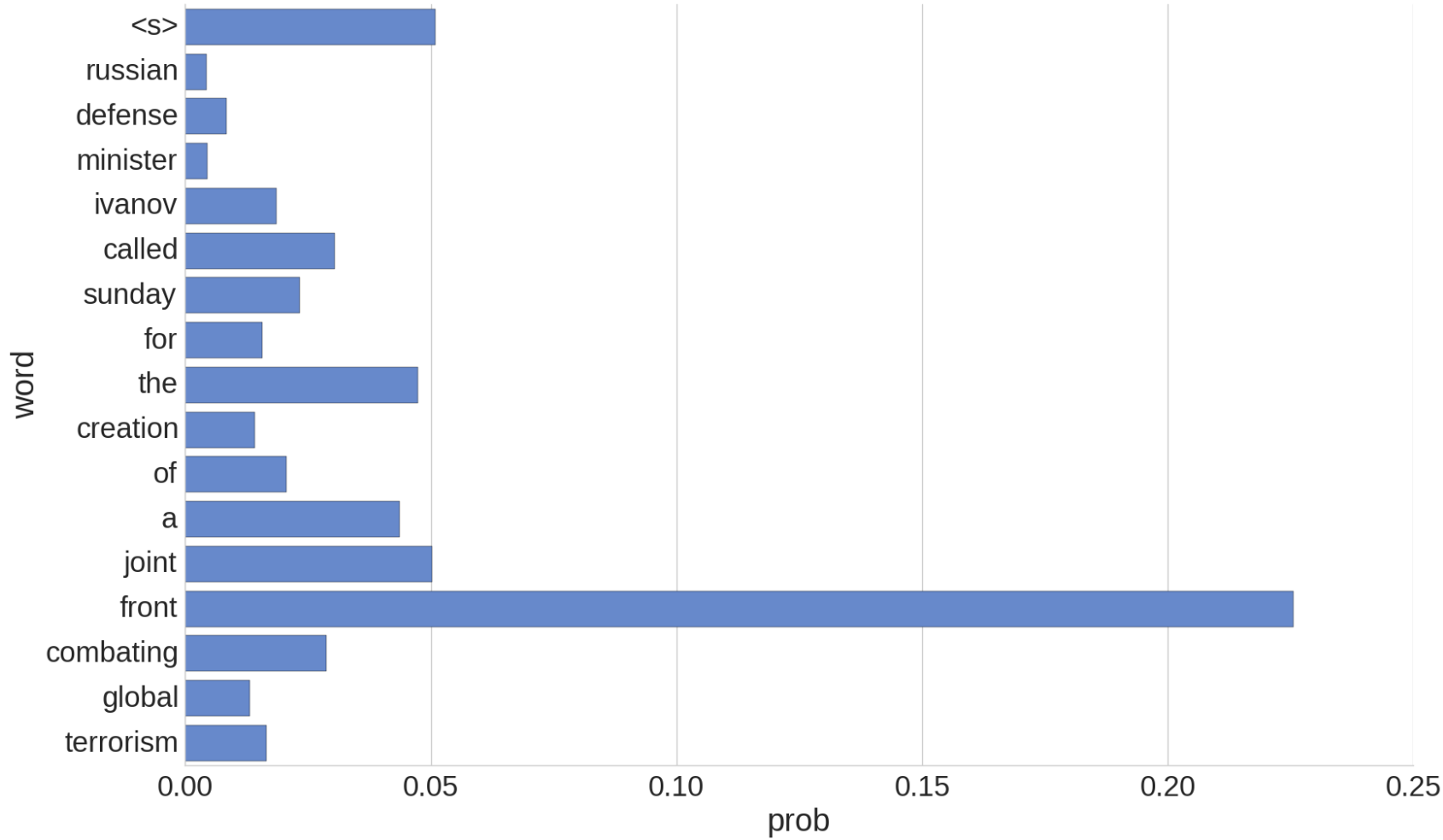
x



ABS Example

$\langle s \rangle$	[Russia calls for joint front]	**against**
	$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

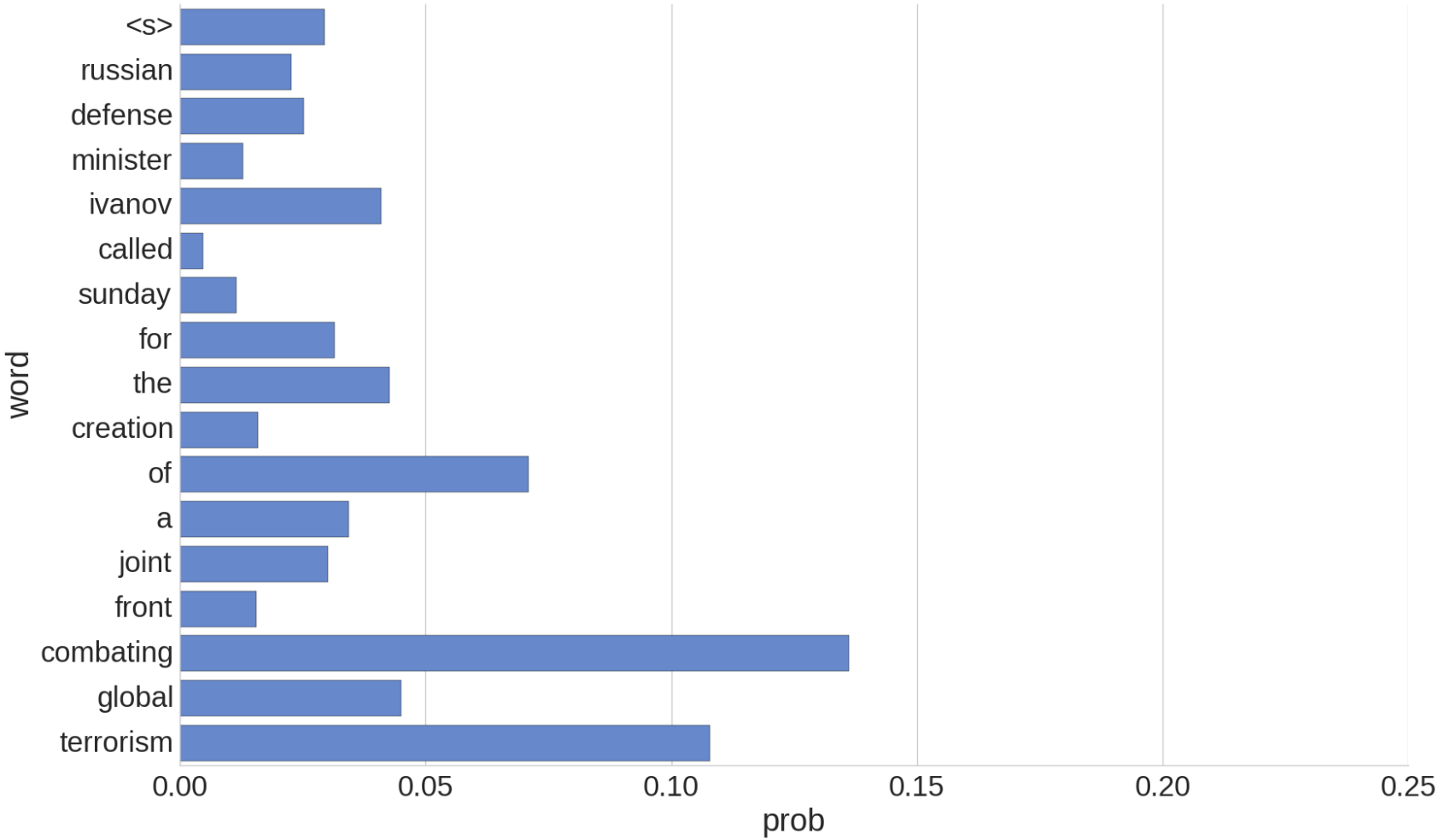
x



ABS Example

$\langle s \rangle$ Russia	[calls for joint front against]	**terrorism**
	$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

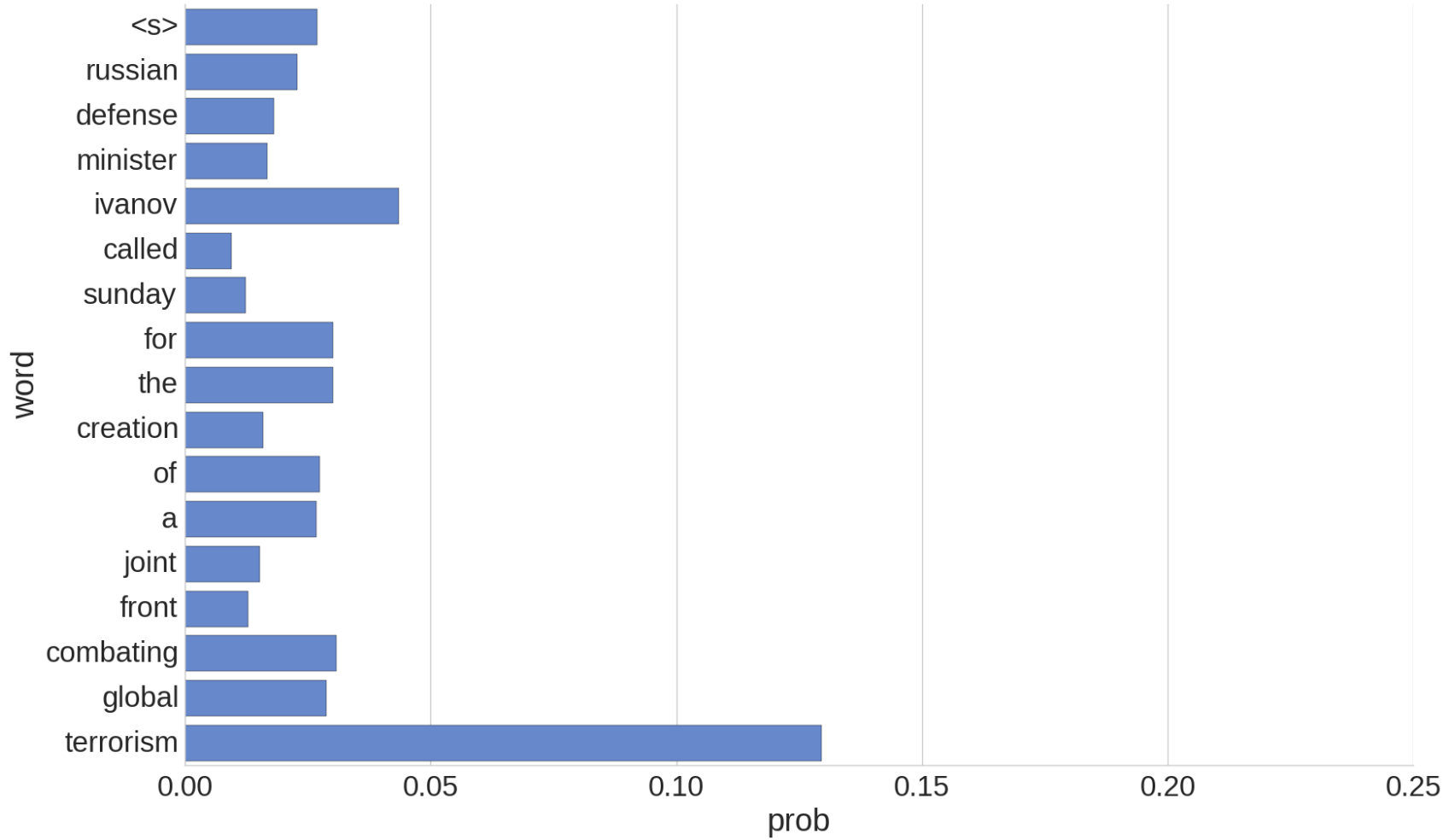
x



ABS Example

⟨s⟩ Russia calls	[for joint front against terrorism]	.
	$\mathbf{y}_c$	$\mathbf{y}_{i+1}$

x



Sep 13, 3:17 PM EDT

GERMANY IMPLEMENTS TEMPORARY BORDER CHECKS TO LIMIT MIGRANTS

BY GEIR MOULSON AND SHAWN POGATCHNIK
ASSOCIATED PRESS

BERLIN (AP) -- Germany introduced temporary border controls Sunday to stem the tide of thousands of refugees streaming across its frontier, sending a clear message to its European partners that it needs more help with an influx that is straining its ability to cope.

AP Photo/Kay Nietfeld

Germany is a preferred destination for many people fleeing Syria's civil war and other troubled nations in the migration crisis that has bitterly divided Europe. They have braved dangerous sea crossings in flimsy

Headline Generation Training Set

Graff et al. 2003; Napoles, Gormley, and Van Durme 2012

- Use Gigaword dataset.

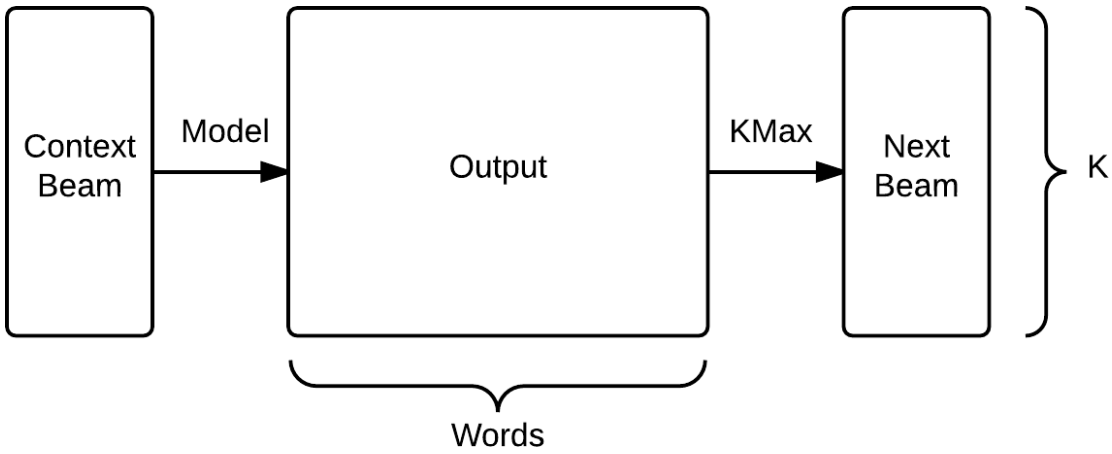
Total Sentences	3.8 M
Newswire Services	7
Source Word Tokens	119 M
Source Word Types	110 K
Average Source Length	31.3 tokens
Summary Word Tokens	31 M
Summary Word Types	69 K
Average Summary Length	8.3 tokens
Average Overlap	4.6 tokens
Average Overlap in first 75	2.6 tokens

Comp with [Filippova and Altun 2013] 250K compressive pairs (although Filippova et al. 2015 2 million)

Training done with mini-batch stochastic gradient descent.

Generation: Beam Search

russia	calls	for	joint
defense	minister	calls	joint
joint	front	calls	terrorism
russia	calls	for	terrorism
...			



- Markov assumption allows for hypothesis recombination.

Extension: Extractive Tuning

- Low-dim word embeddings unaware of exact matches.
- Log-linear parameterization:

$$p(\mathbf{y}|\mathbf{x};\theta,\alpha) \quad \propto \quad \exp(\alpha^\top \sum_{i=0}^{N-1} f(\mathbf{y}_{i+1}, \mathbf{x}, \mathbf{y}_c)).$$

Features f :

1. Model score (neural model)
2. Unigram overlap
3. Bigram overlap
4. Trigram overlap
5. Word out-of-order

- Similar to rare-word issue in neural MT [Luong et al. 2015]
- Use MERT for estimating α as post-processing (not end-to-end)

Results

Baselines

- Type: [A]bstractive, [C]ompressive, [E]xtractive
- Data: [S]ource, [T]arget, [B]oth, [N]one

Model	Dec.	Type	Data	Cite
PREFIX	N/A	C	N	
TOPIARY	HT	A	N	[Zajic, Dorr, and Schwartz 2004]
W&L	ILP	-	N	[Woodsend, Feng, and Lapata 2010]

Baselines

- Type: [A]bstractive, [C]ompressive, [E]xtractive
- Data: [S]ource, [T]arget, [B]oth, [N]one

Model	Dec.	Type	Data	Cite
PREFIX	N/A	C	N	
TOPIARY	HT	A	N	[Zajic, Dorr, and Schwartz 2004]
W&L	ILP	-	N	[Woodsend, Feng, and Lapata 2010]
IR	BM-25	A	B	
T3	Trans.	A	B	[Cohn and Lapata 2008]
Compress	ILP	C	T	[Clarke and Lapata 2008]
MOSES+	Beam	A	B	[Koehn et al. 2007]

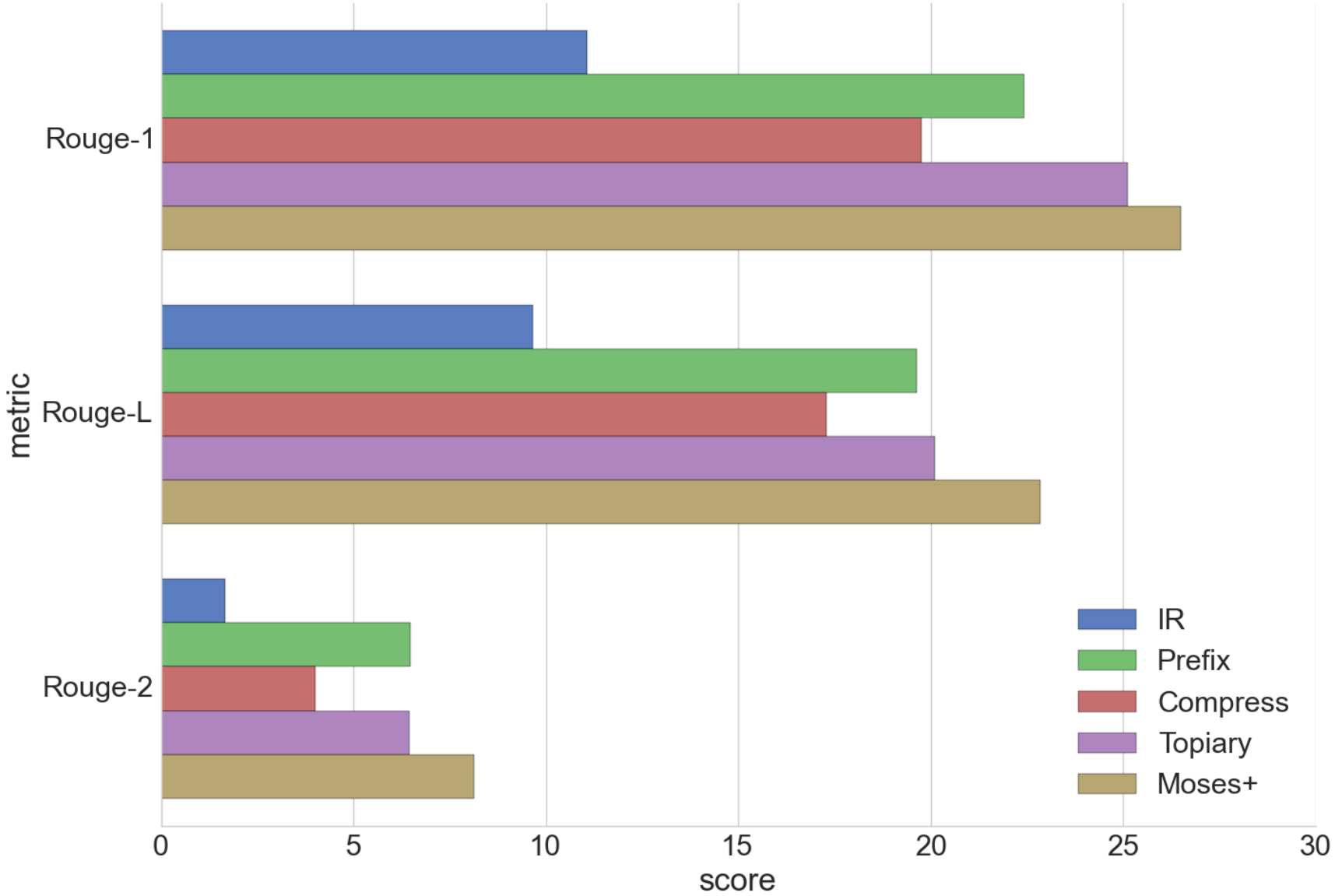
Baselines

- Type: [A]bstractive, [C]ompressive, [E]xtractive
- Data: [S]ource, [T]arget, [B]oth, [N]one

Model	Dec.	Type	Data	Cite
PREFIX	N/A	C	N	
TOPIARY	HT	A	N	[Zajic, Dorr, and Schwartz 2004]
W&L	ILP	-	N	[Woodsend, Feng, and Lapata 2010]
IR	BM-25	A	B	
T3	Trans.	A	B	[Cohn and Lapata 2008]
Compress	ILP	C	T	[Clarke and Lapata 2008]
MOSES+	Beam	A	B	[Koehn et al. 2007]
ABS	Beam	A	B	This Work
ABS+	Beam	A	B	This Work

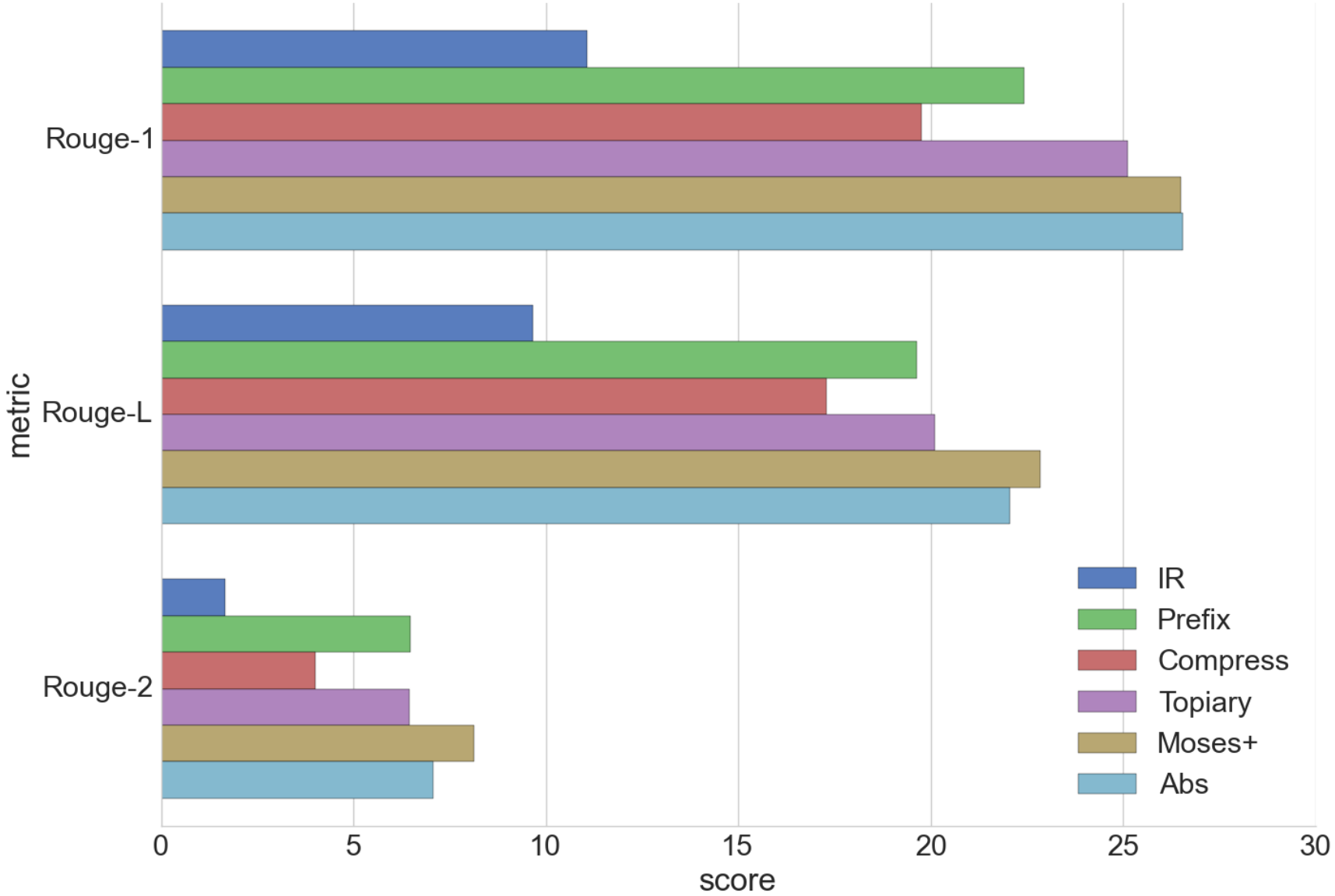
Summarization Results: DUC 2004

(500 pairs, 4 references, 75 characters)



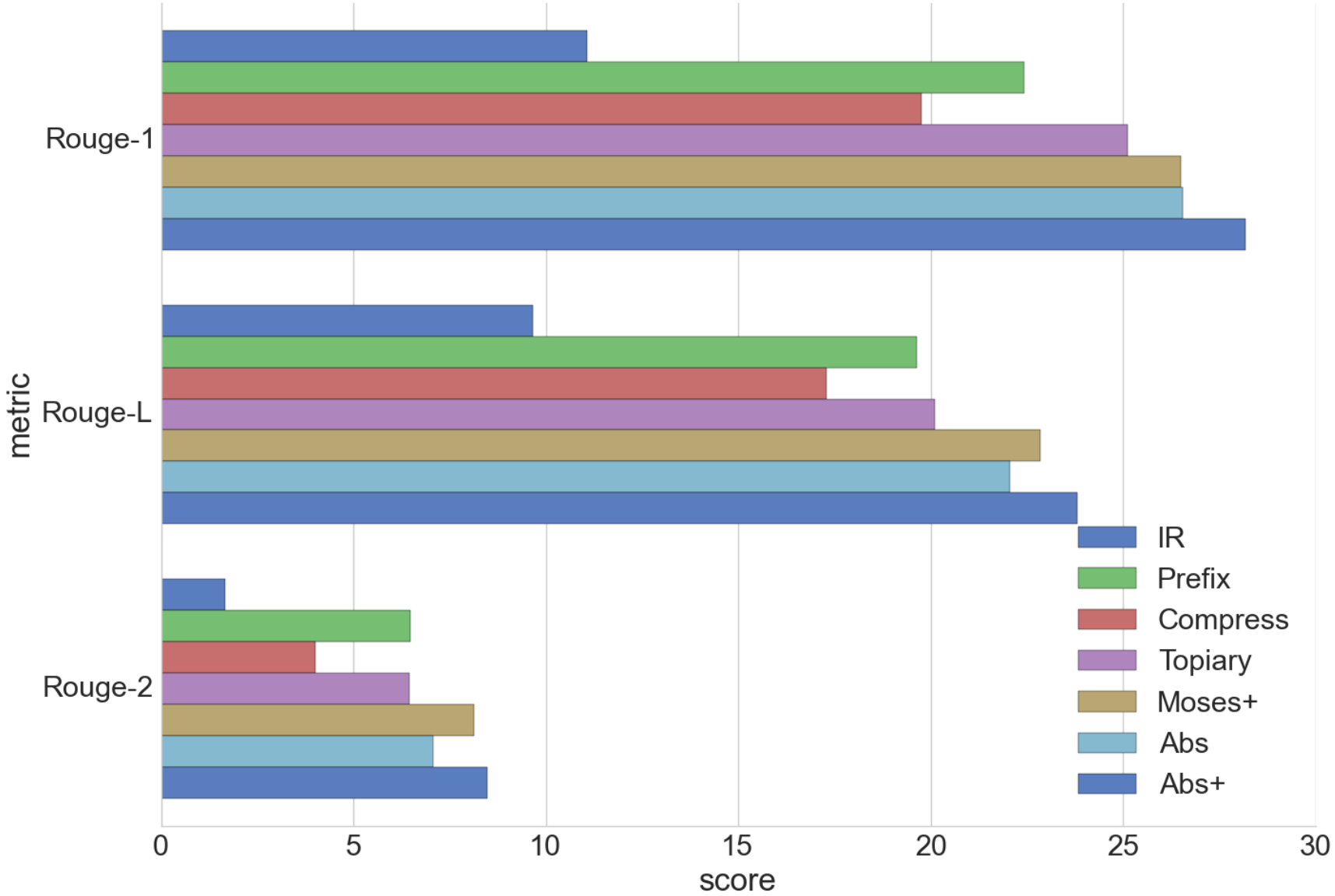
Summarization Results: DUC 2004

(500 pairs, 4 references, 75 characters)



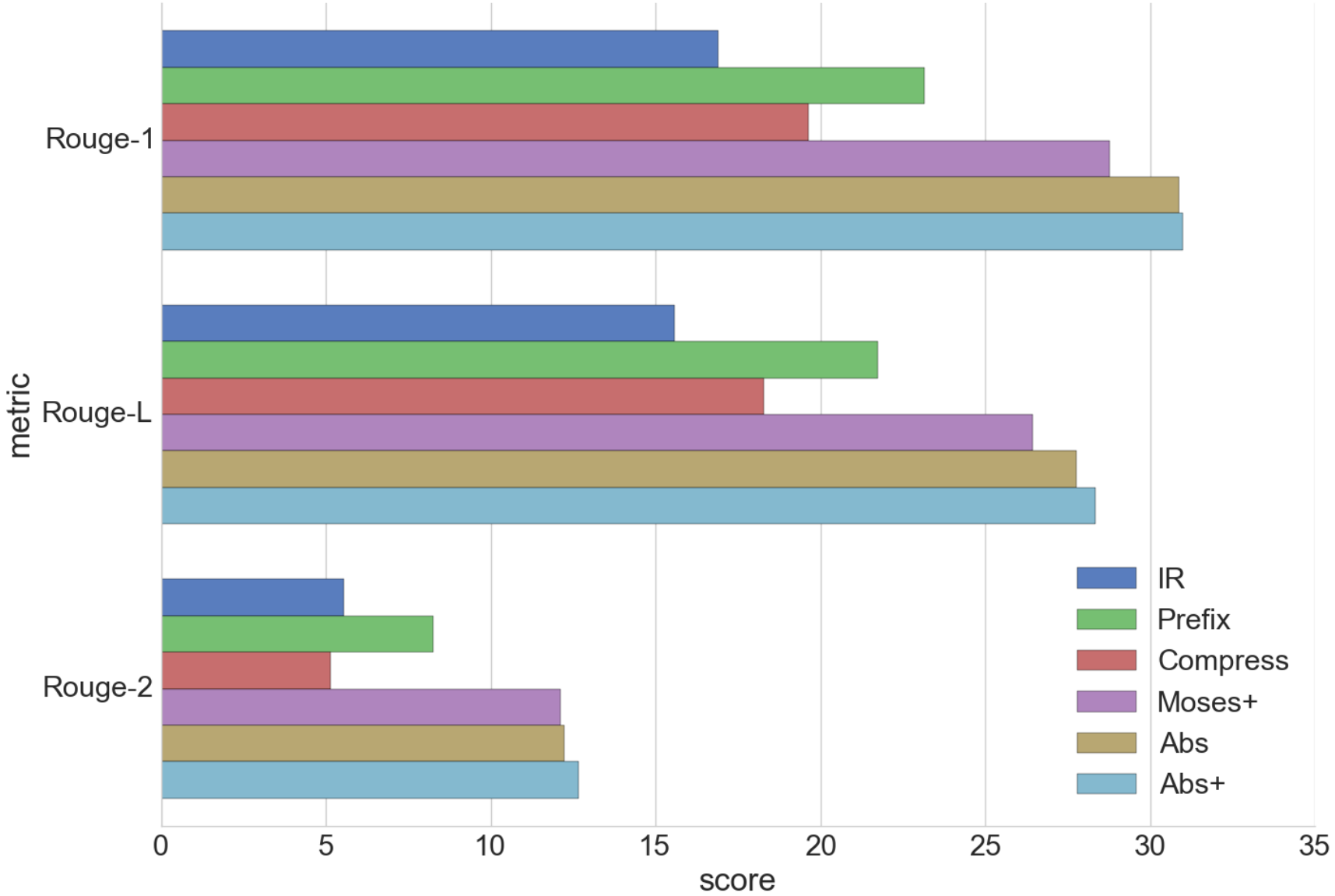
Summarization Results: DUC 2004

(500 pairs, 4 references, 75 characters)



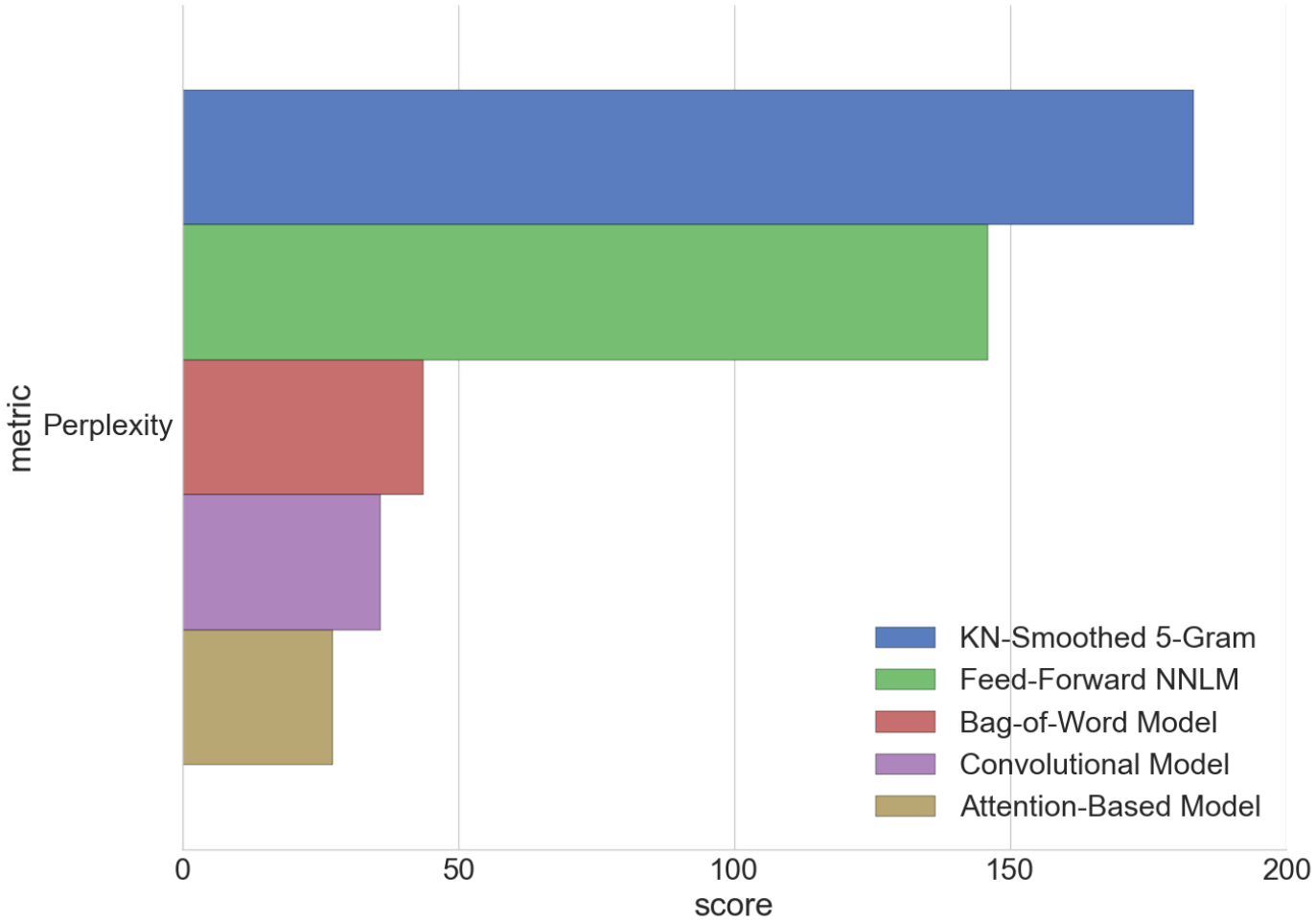
Summarization Results: Gigaword Test

(2000 pairs, 1 reference, 8 words)



Model Comparison

Perplexity Gigaword Development Set



Ablations

Decoder	Model	Cons.	R-1	R-2	R-L
Greedy	ABS+	Abs	26.67	6.72	21.70
Beam	BOW	Abs	22.15	4.60	18.23
Beam	ABS+	Ext	27.89	7.56	22.84
Beam	ABS+	Abs	28.48	8.91	23.97

Generated Sentences on Gigaword I

Source:

a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .

Ref: iranian-american academic held in tehran released on bail

Abs: detained iranian-american academic released from jail after posting bail

Abs+: detained iranian-american academic released from prison **after hefty bail**

Generated Sentences on Gigaword II

Source:

ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .

Ref: european mediterranean ministers gather for landmark conference by julie bradford

Abs: mediterranean neighbors gather for unprecedented conference **on heavy security**

Abs+: mediterranean neighbors gather under heavy security for unprecedented conference

Generated Sentences on Gigaword III

Source:

the death toll from a school collapse in a haitian shanty-town rose to ## after rescue workers uncovered a classroom with ## dead students and their teacher , officials said saturday .

Ref: toll rises to ## in haiti school unk : official

Abs: death toll in haiti school **accident** rises to ##

Abs+: death toll in haiti school to ## **dead** students

Generated Sentences on Gigaword IV

Source:

australian foreign minister stephen smith sunday congratulated new zealand 's new prime minister-elect john key as he praised ousted leader helen clark as a " gutsy " and respected politician .

Ref: time caught up with nz 's gutsy clark says australian fm

Abs: australian foreign minister congratulates **new nz pm after election**

Abs+: australian foreign minister congratulates **smith new zealand** as leader

Generated Sentences on Gigaword V

Source:

two drunken south african fans hurled racist abuse at the country 's rugby sevens coach after the team were eliminated from the weekend 's hong kong tournament , reports said tuesday .

Ref: rugby union : racist taunts mar hong kong sevens : report

Abs: south african fans hurl racist **taunts at rugby sevens**

Abs+: south african fans **racist** abuse at rugby sevens tournament

Generated Sentences on Gigaword VI

Source:

christian conservatives – kingmakers in the last two us presidential elections – may have less success in getting their pick elected in #### , political observers say .

Ref: christian conservatives power diminished ahead of #### vote

Abs: christian conservatives may have less success in #### election

Abs+: christian conservatives **in the last two** us presidential elections

Generated Sentences on Gigaword VII

Source:

the white house on thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions .

Ref: us warns iran of step backward on nuclear issue

Abs: **iran** warns of possible new sanctions on nuclear work

Abs+: **un nuclear watchdog** warns iran of possible new sanctions

Generated Sentences on Gigaword VIII

Source:

thousands of kashmiris chanting pro-pakistan slogans on sunday attended a rally to welcome back a hardline separatist leader who underwent cancer treatment in mumbai .

Ref: thousands attend rally for kashmir hardliner

Abs: thousands **rally in support** of hardline **kashmiri** separatist leader

Abs+: thousands of kashmiris rally to welcome back **cancer treatment**

Generated Sentences on Gigaword IX

Source:

an explosion in iraq 's restive northeastern province of diyala killed two us soldiers and wounded two more , the military reported monday .

Ref: two us soldiers killed in iraq blast december toll ###

Abs: # us two soldiers killed in restive northeast province

Abs+: explosion in restive northeastern province kills two us soldiers

Generated Sentences on Gigaword X

Source:

russian world no. # nikolay davydenko became the fifth withdrawal through injury or illness at the sydney international wednesday , retiring from his second round match with a foot injury .

Ref: tennis : davydenko pulls out of sydney with injury

Abs: davydenko **pulls out** of sydney international with foot injury

Abs+: russian world no. # davydenko **retires at sydney international**

Generated Sentences on Gigaword XI

Source:

russia 's gas and oil giant gazprom and us oil major chevron have set up a joint venture based in resource-rich northwestern siberia , the interfax news agency reported thursday quoting gazprom officials .

Ref: gazprom chevron set up joint venture

Abs: **russian oil giant chevron** set up **siberia joint venture**

Abs+: **russia 's gazprom** set up joint venture in siberia

Open-Source

- Torch/Lua
- Important optimizations (heavily CUDA/GPU dependent)
  - Source-length grouped for batching
  - Batch matrix multiply
  - GPU full soft max
- Code, dataset construction, tuning, and evaluation available: `http://www.github.com/facebook/NAMAS/`

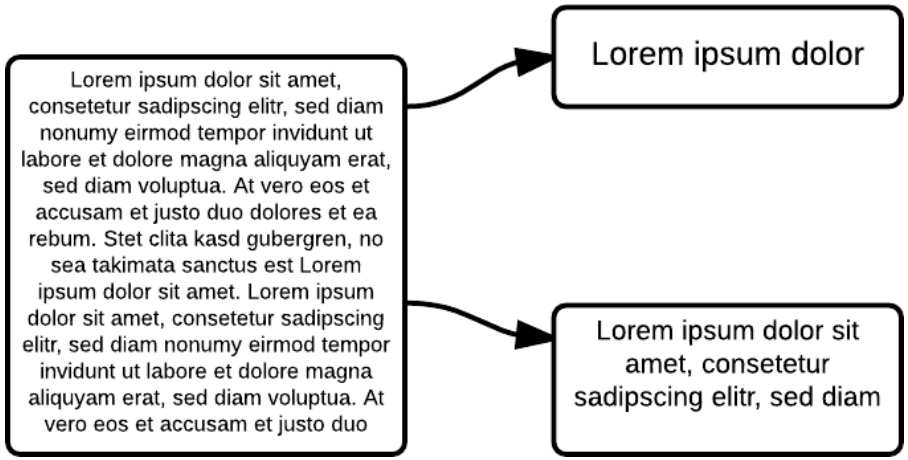
Conclusion

Qualitative Issues:

- Repeating semantic elements.
- Altering semantic roles.
- Improper generalization.

Future Work:

- Move from Feed-Forward NNLM to RNN-LM.
- Summarizing longer documents.
- Incorporating syntactic evaluation.



References I

Jing, Hongyan (2002). "Using hidden Markov modeling to decompose human-written summaries". In: *Computational linguistics* 28.4, pp. 527–543.

Dorr, Bonnie, David Zajic, and Richard Schwartz (2003). "Hedge trimmer: A parse-and-trim approach to headline generation". In: *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*. Association for Computational Linguistics, pp. 1–8.

Cohn, Trevor and Mirella Lapata (2008). "Sentence compression beyond word deletion". In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 137–144.

Woodsend, Kristian, Yansong Feng, and Mirella Lapata (2010). "Generation with quasi-synchronous grammar". In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 513–523.

References II

Zajic, David, Bonnie Dorr, and Richard Schwartz (2004). "Bbn/umd at duc-2004: Topiary". In: *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pp. 112–119.

Banko, Michele, Vibhu O Mittal, and Michael J Witbrock (2000). "Headline generation based on statistical translation". In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 318–325.

Liu, Fei et al. (2015). "Toward abstractive summarization using semantic representations". In:

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473. URL: http://arxiv.org/abs/1409.0473.

Knight, Kevin and Daniel Marcu (2002). "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". In: *Artificial Intelligence* 139.1, pp. 91–107.

References III

Kalchbrenner, Nal and Phil Blunsom (2013). "Recurrent Continuous Translation Models." In: *EMNLP*, pp. 1700–1709.

Sutskever, Ilya, Oriol Vinyals, and Quoc VV Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in Neural Information Processing Systems*, pp. 3104–3112.

Bengio, Yoshua et al. (2003). "A neural probabilistic language model". In: *The Journal of Machine Learning Research* 3, pp. 1137–1155.

Filippova, Katja and Yasemin Altun (2013). "Overcoming the Lack of Parallel Data in Sentence Compression." In: *EMNLP*, pp. 1481–1491.

Filippova, Katja et al. (2015). "Sentence Compression by Deletion with LSTMs". In:

Graff, David et al. (2003). "English gigaword". In: *Linguistic Data Consortium, Philadelphia*.

References IV

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme (2012). "Annotated gigaword". In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, pp. 95–100.

Luong, Thang et al. (2015). "Addressing the Rare Word Problem in Neural Machine Translation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 11–19. URL: http://aclweb.org/anthology/P/P15/P15-1002.pdf.

Clarke, James and Mirella Lapata (2008). "Global inference for sentence compression: An integer linear programming approach". In: *Journal of Artificial Intelligence Research*, pp. 399–429.

Koehn, Philipp et al. (2007). "Moses: Open source toolkit for statistical machine translation". In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 177–180.