What Kind of Language Is Hard to Language-Model?

ACL 2019

Sabrina J. Mielke *and* Ryan Cotterell, Kyle Gorman, Brian Roark, Jason Eisner

Johns Hopkins University // City University of New York Graduate Center // Google
`sjmielke@jhu.edu`

Twitter: `@sjmielke` – paper and thread pinned!

Questions and answers

0. Do current language models do equally well on all languages?

Questions and answers

0. Do current language models do equally well on all languages? *No.*

Questions and answers

0. Do current language models do equally well on all languages? *No.*

1. Which one do they struggle more with: German or English?

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese?

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese? *It depends.*

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese? *It depends.*
3. What makes a language harder to model?

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese? *It depends.*
3. What makes a language harder to model? *Actually, rather technical factors.*

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese? *It depends.*
3. What makes a language harder to model? *Actually, rather technical factors.*
4. Is Translationese easier?

Questions and answers

0. Do current language models do equally well on all languages? *No.*
1. Which one do they struggle more with: German or English? *German.*
2. What about non-Indo-European languages, say Chinese? *It depends.*
3. What makes a language harder to model? *Actually, rather technical factors.*
4. Is Translationese easier? *It's different, but not actually easier!*

Outline

"Difficulty"

Outline

“Difficulty”

Models and languages

Outline

“Difficulty”

Models and languages

What correlates with difficulty?

Outline

“Difficulty”

Models and languages

What correlates with difficulty?

And... is Translationese really easier?

How to measure "difficulty"?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	I love Florence!	0.03	$\Rightarrow$ 5 bits

How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	I love Florence!	0.03	$\Rightarrow$ 5 bits
de	Ich grüße meine Oma und die Familie dahein.	0.008	$\Rightarrow$ 7 bits

How to measure "difficulty"?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	I love Florence!	0.03	$\Rightarrow$ 5 bits
de	Ich grüße meine Oma und die Familie dahein.	0.008	$\Rightarrow$ 7 bits
nl	Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.	0.0004	$\Rightarrow$ 11 bits

How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	I love Florence!	0.03	$\Rightarrow$ 5 bits
de	Ich grüße meine Oma und die Familie dahein.	0.008	$\Rightarrow$ 7 bits
nl	Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.	0.0004	$\Rightarrow$ 11 bits

Issue 1: Different topics/styles/content

How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

Issue 1: Different topics/styles/content

Solution: train and test on translations!

Europarl: 21 languages share ~40M chars
Bibles: 62 languages share ~4M chars

How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

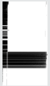
		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

Issue 1: Different topics/styles/content

Solution: train and test on translations!

Europarl: 21 languages share ∼40M chars
Bibles: 62 languages share ∼4M chars

↰ and this one takes a big ILP to solve, which is really fun

Gurobi()

How to measure “difficulty”?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

Issue 1: Different topics/styles/content

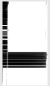
Solution: train and test on translations!

Europarl: 21 l… ars

Bibles … s

$\sum$ 69 languages
13 language families

… one takes
a big ILP to solve,
which is really fun

Gurobi()

How to measure "difficulty"?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

Issue 1: Different topics/styles/content

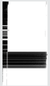
Solution: train and test on translations!

Europarl: ... ars

Bibles ... s

$\sum$ 69 languages
13 language families

... one takes
a big ILP to solve,
which is really fun

Gurobi()

Issue 2: Comparing scores

How to measure "difficulty"?

Language models measure surprisal/information content (NLL; $-\log p(\cdot)$):

		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

Issue 1: Different topics/styles/content

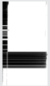
Solution: train and test on translations!

Europarl: 21 l… ars

Bibles … s

$\sum$ 69 languages
13 language families

… one takes
a big ILP to solve,
which is really fun

Gurobi $\left(\right.$  $\left. \right)$

Issue 2: Comparing scores

Use **total bits** of an **open-vocabulary model.**

Why?

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is "cheating" – morphologically rich languages have more UNKs, unfairly advantaging them.

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is "cheating" – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can't normalize per word or even per character in languages individually.**
   Example: if puč$_{\text{cz}}$ and Putsch$_{\text{de}}$ are equally likely, they should be equally "difficult."

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is “cheating” – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can’t normalize per word or even per character in languages individually.**
   Example: if puč$_{\text{cz}}$ and Putsch$_{\text{de}}$ are equally likely, they should be equally “difficult.”

$\Rightarrow$ **just use overall bits (i.e., surprisal/NLL) of an aligned sentence**

How to compare your language models across languages

1. **We need to be open-vocabulary – no UNKs.**
   Every UNK is "cheating" – morphologically rich languages have more UNKs, unfairly advantaging them.

2. **We can't normalize per word or even per character in languages individually.**
   Example: if puč$_{\text{cz}}$ and Putsch$_{\text{de}}$ are equally likely, they should be equally "difficult."

⇒ **just use overall bits (i.e., surprisal/NLL) of an aligned sentence**

[note: total easily obtainable from BPC or perplexity by multiplying with total chars/words]

How to aggregate multiple intents' surprisals into "difficulties"?

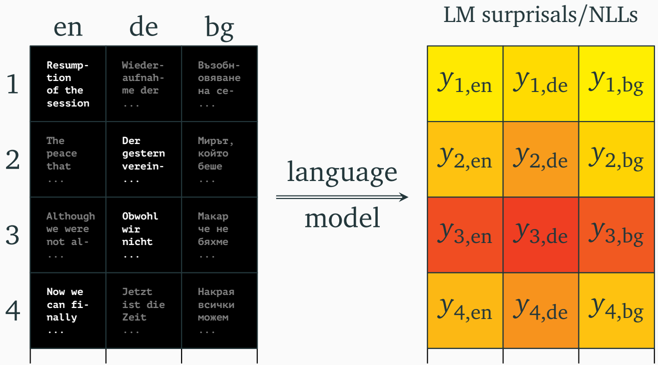
For fully parallel corpora...

	en	de	bg
1	Resump- tion of the session	Wieder- aufnah- me der ...	Възобн- овяване на се- ...
2	The peace that ...	Der gestern verein- ...	Мирът, който беше ...
3	Although we were not al- ...	Obwohl wir nicht ...	Макар че не бяхме ...
4	Now we can fi- nally ...	Jetzt ist die Zeit ...	Накрая всички можем ...

aligned multi-text

How to aggregate multiple intents' surprisals into "difficulties"?

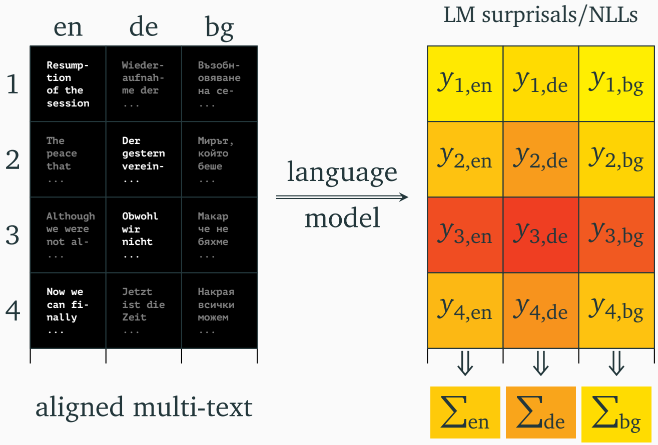
For fully parallel corpora...



aligned multi-text

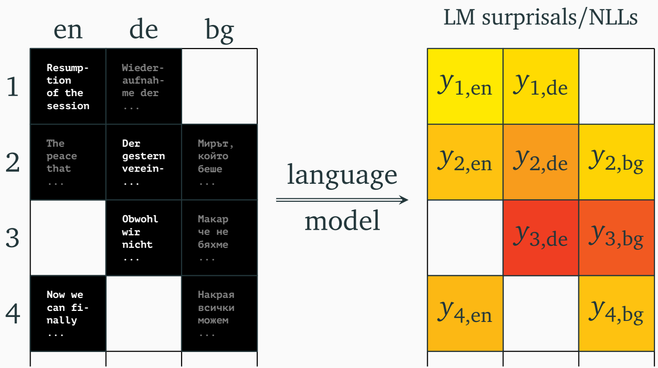
How to aggregate multiple intents’ surprisals into “difficulties”?

For fully parallel corpora… we can just sum everything up and compare – that is *fair*.



How to aggregate multiple intents' surprisals into "difficulties"?

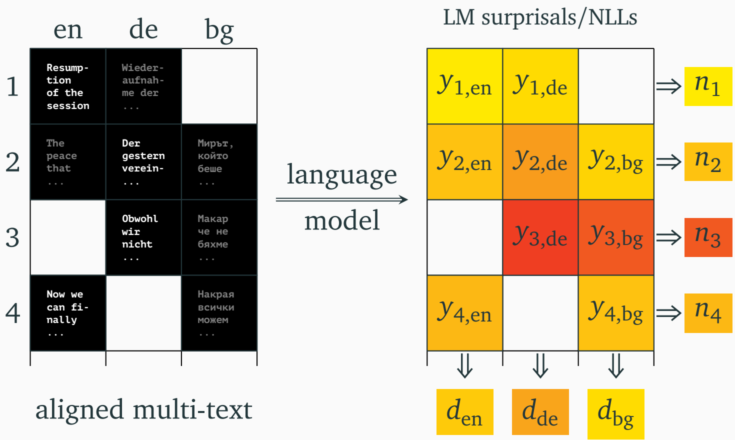
But what if there's missing data? Or we want robustness?



aligned multi-text

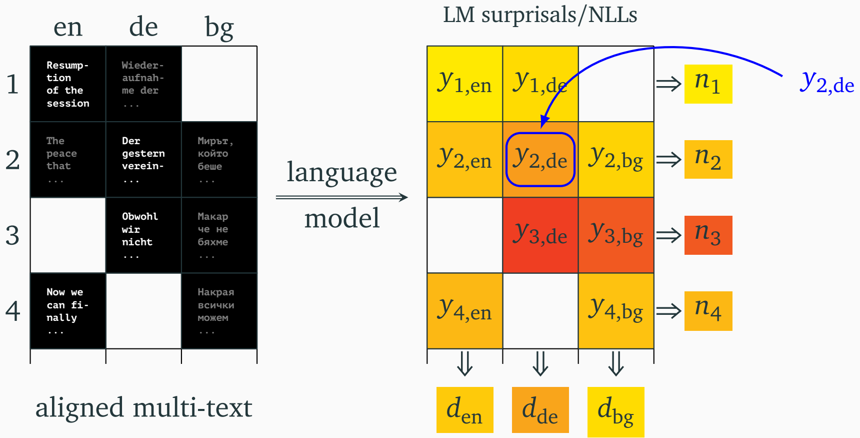
How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?



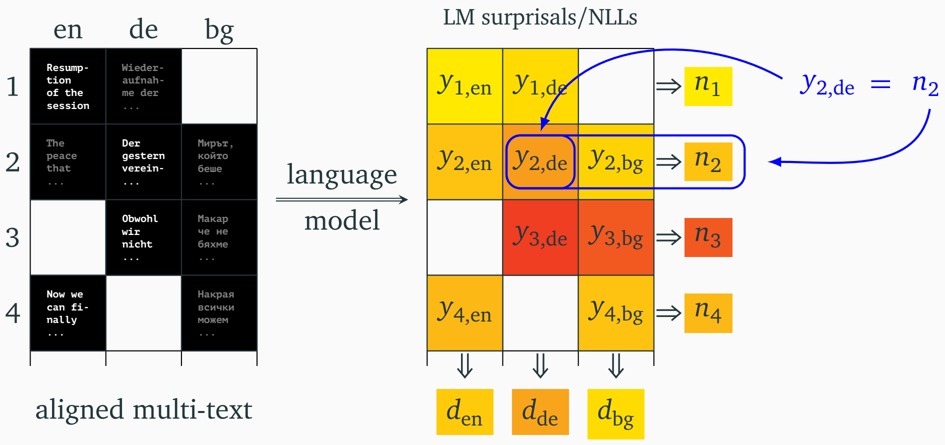
How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?



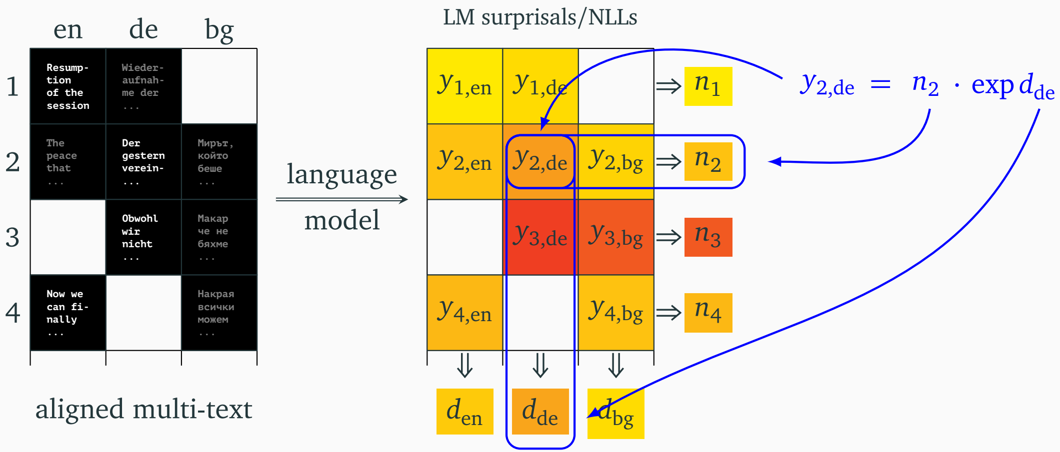
How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?



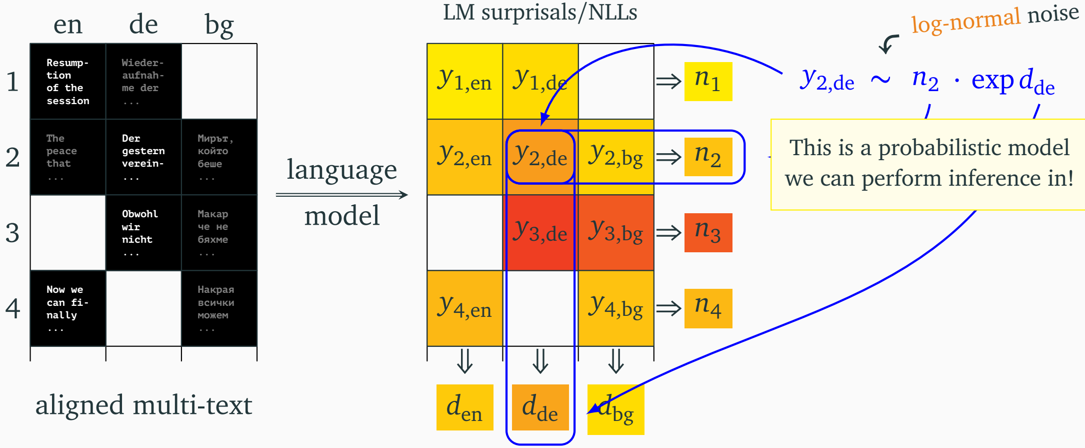
How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?



How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?



This is a probabilistic model we can perform inference in!

How to aggregate multiple intents' surprisals into "difficulties"?

But what if there's missing data? Or we want robustness?

	en	de	bg
1	Resumption of the session	Wiederaufnahme der ...	
2	The peace that ...	Der gestern verein- ...	Мирът, който беше ...
3		Obwohl wir nicht ...	Макар че не бяхме ...
4	Now we can finally ...		Накрая всички можем ...

aligned multi-text

$\xrightarrow{\text{language model}}$

LM surprisals/NLLs

y_1 n_1

log-normal noise

$$y_{2,\mathrm{de}} \sim n_2 \cdot \exp d_{\mathrm{de}}$$

d_{en} d_{de} d_{bg}

not quite, our actual model is

HETEROSCEDASTIC

$$y_{ij} = n_i \cdot \exp(d_j) \cdot \exp(\epsilon_{ij})$$
$$\sigma_i^2 = \ln\left(1 + \frac{\exp(\sigma^2)-1}{n_i}\right)$$
$$\epsilon_{ij} \sim \mathcal{N}\left(\frac{\sigma^2-\sigma_i^2}{2}, \sigma_i^2\right)$$

Outline

“Difficulty” ✓

Models and languages

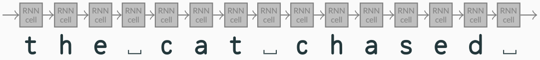
What correlates with difficulty?

And... is Translationese really easier?

Good open-vocabulary language models

Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

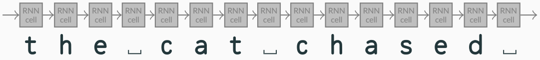
char-RNNLM:



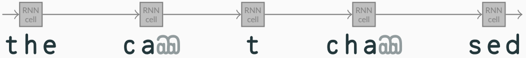
Good open-vocabulary language models (Mielke and Eisner, 2019)

Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

char-RNNLM:



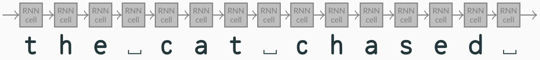
BPE-RNNLM, few merges:



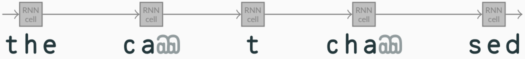
Good open-vocabulary language models (Mielke and Eisner, 2019)

Formerly state-of-the-art-ish AWD-LSTM (Merity et al., 2018) language models:

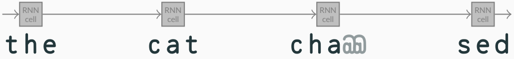
char-RNNLM:



BPE-RNNLM, few merges:

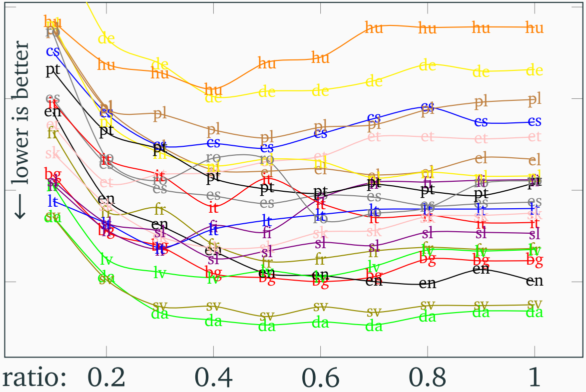


BPE-RNNLM, many merges:



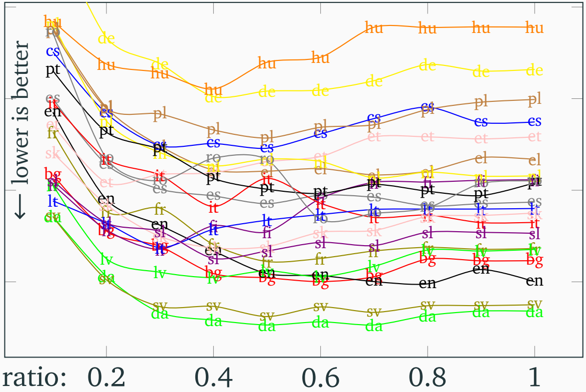
Choosing the number of BPE merges: how many is best?

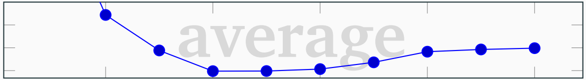
It depends on the language (total surprisal, given merges as a ratio of the vocabulary):



Choosing the number of BPE merges: how many is best?

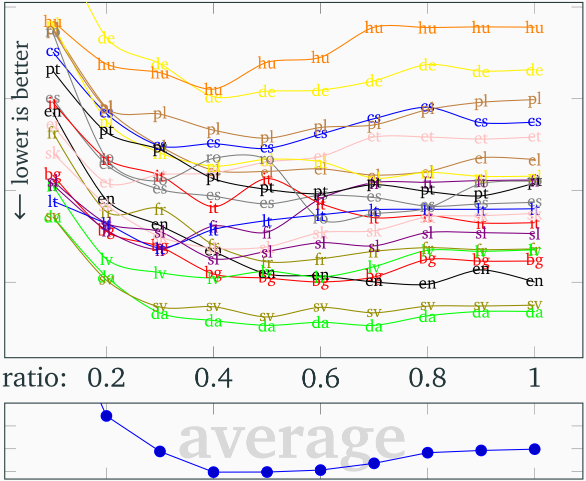
It depends on the language (total surprisal, given merges as a ratio of the vocabulary):





Choosing the number of BPE merges: how many is best?

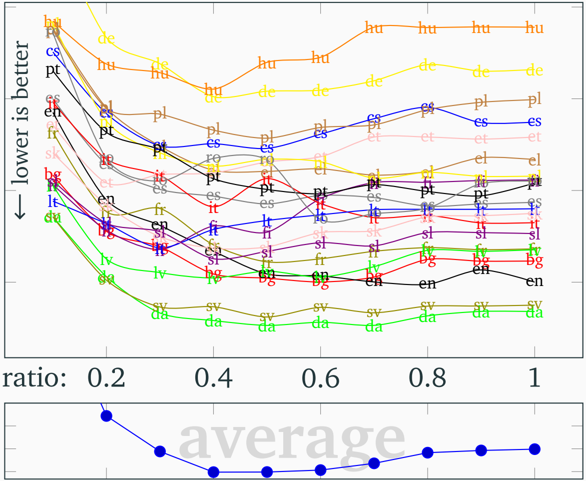
It depends on the language (total surprisal, given merges as a ratio of the vocabulary):



↰ is this one going to be fine?

Choosing the number of BPE merges: how many is best?

It depends on the language (total surprisal, given merges as a ratio of the vocabulary):



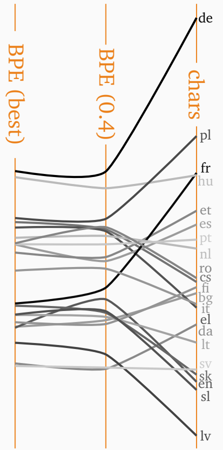
↰ is this one going to be fine?

Yeah:

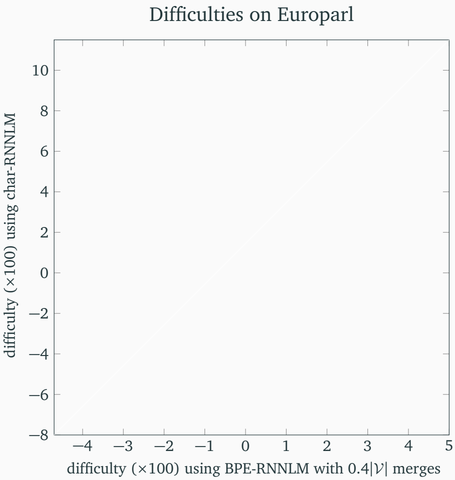
it doesn't

matter

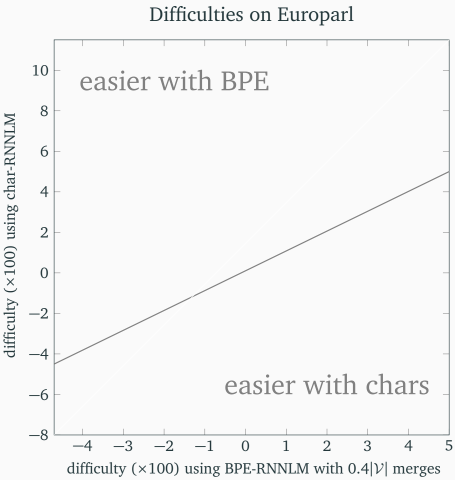
that much.



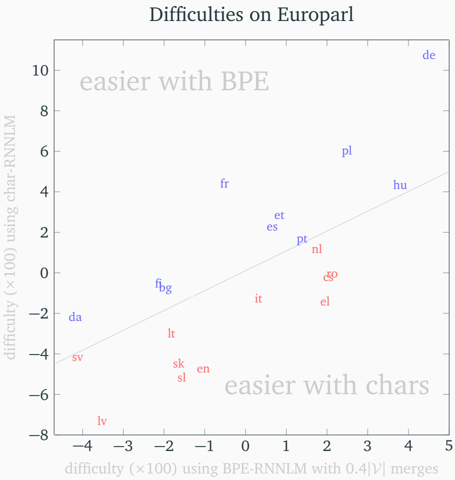
Difficulties for char-/BPE-RNNLM: 21 Europarl languages



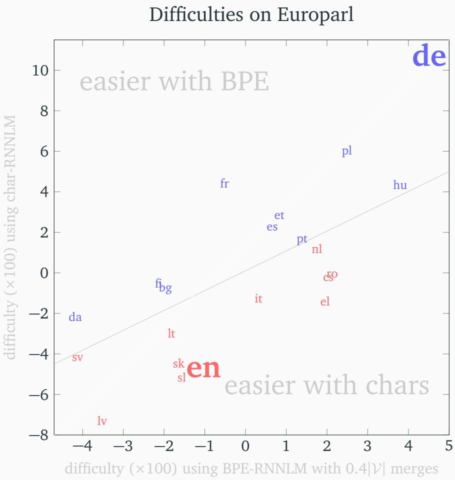
Difficulties for char-/BPE-RNNLM: 21 Europarl languages



Difficulties for char-/BPE-RNNLM: 21 Europarl languages

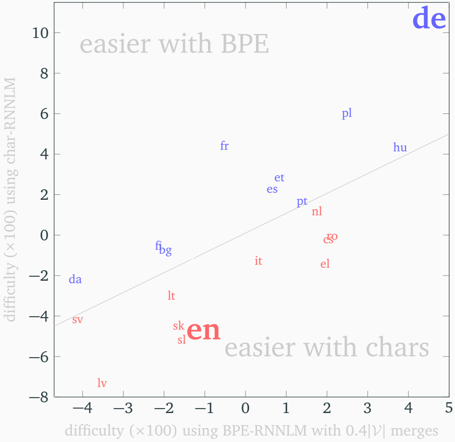


Difficulties for char-/BPE-RNNLM: 21 Europarl languages



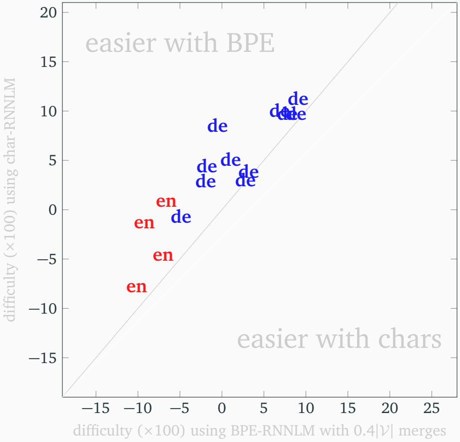
Difficulties for char-/BPE-RNNLM: 21 Europarl languages and Bibles

Difficulties on Europarl

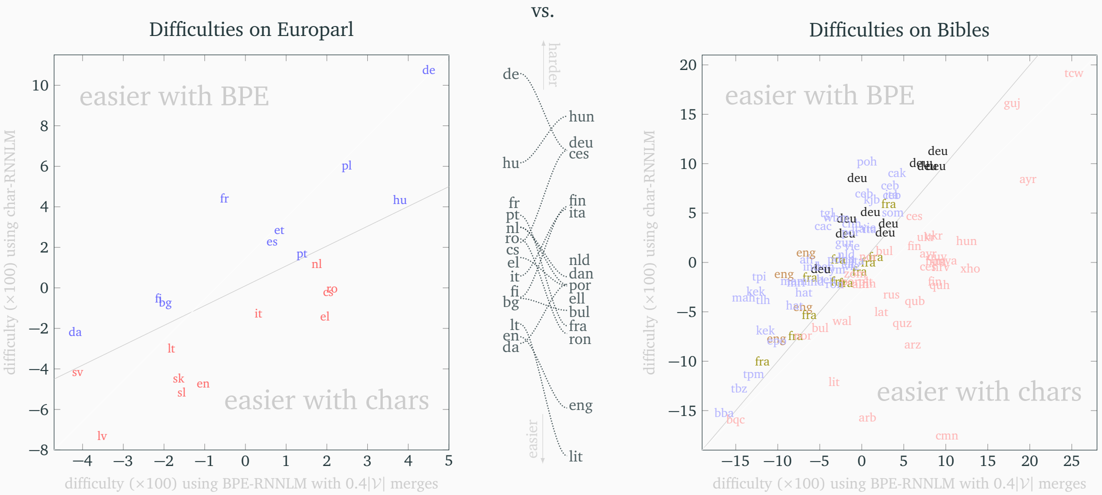


vs.

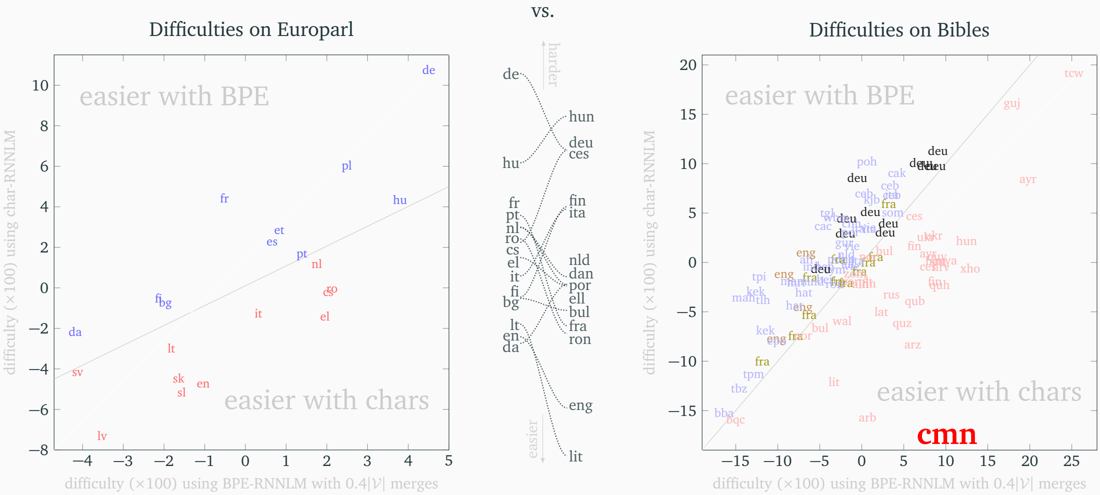
Difficulties on Bibles



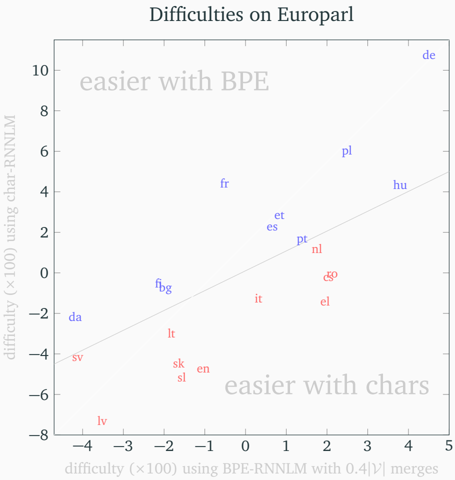
Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles

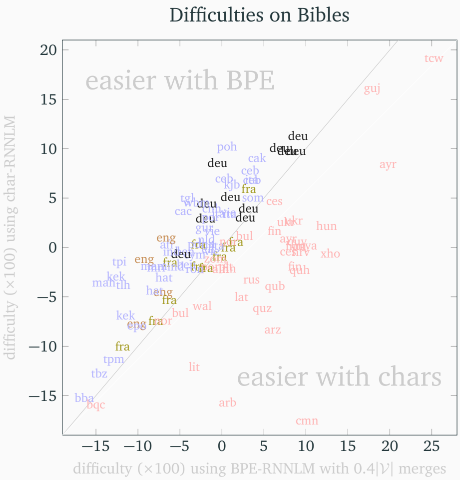


Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles



Difficulties for char-/BPE-RNNLM: 21 Europarl languages and 106 Bibles





Outline

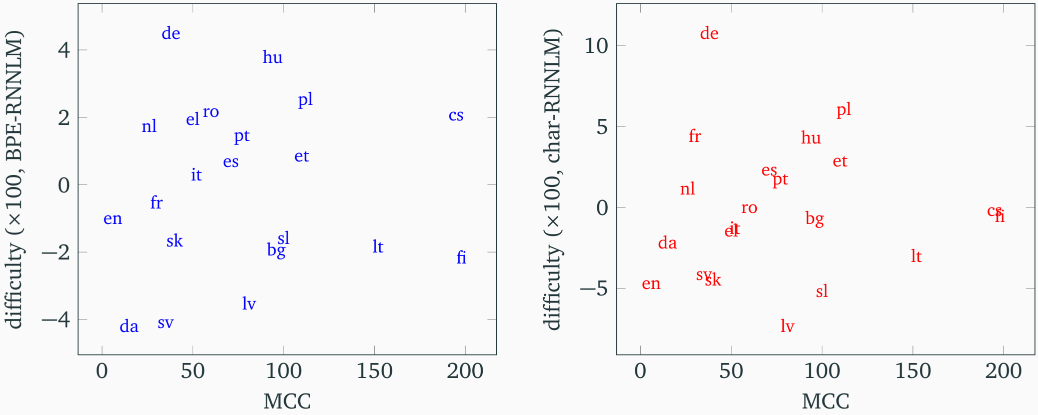
"Difficulty" ✓

Models and languages ✓

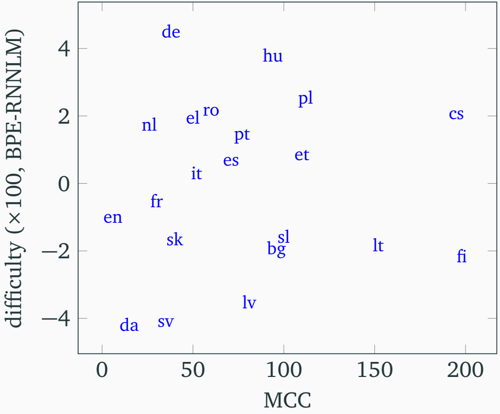
What correlates with difficulty?

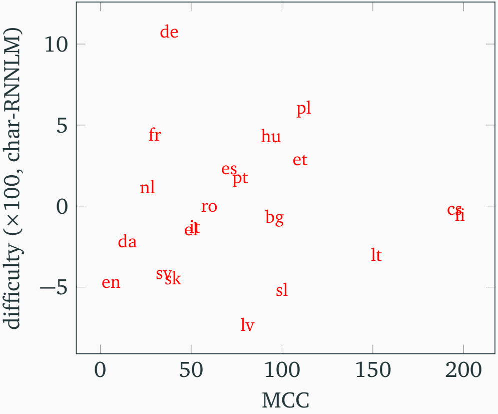
And... is Translationese really easier?

How about: morphological counting complexity (Sagot, 2013)



How about: morphological counting complexity (Sagot, 2013)





...not particularly striking. Perhaps Finnish was an outlier in Cotterell et al. (2018)?

Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing […] Morphology” (for languages where present)**?**

Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing […] Morphology” (for languages where present)**?**

…no visible differences.

Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)**?**

...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)**?**

Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing [...] Morphology" (for languages where present)**?**

...no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)**?**

...no visible differences.

Other linguistically motivated regressors

WALS: “Prefixing vs. Suffixing [...] Morphology” (for languages where present)**?**

...no visible differences.

WALS: “Order of Subject, Object and Verb” (for languages where present)**?**

...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)**?**

Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing [...] Morphology" (for languages where present)**?**

...no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)**?**

...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)**?**

...neither mean and skew show correlation.

Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing [...] Morphology" (for languages where present)**?**

...no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)**?**

...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)**?**

...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))**?**

Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing […] Morphology" (for languages where present)**?**

…no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)**?**

…no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)**?**

…neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))**?**

…correlation! But not significant after correcting for multiple hypotheses.

Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing [...] Morphology" (for languages where present)**?**

...no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)**?**

...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)**?**

...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))**?**

...correlation! But not significant after correcting for multiple hypotheses.

This is **disappointing**.

Very simple heuristics are very predictive

Raw sequence length / # predictions
$\rightarrow$ char-RNNLM difficulty

Significant on:

- Europarl at $p < .01$
- Bibles at $p < .001$

i.e., for the char-RNNLM
puč$_{cz}$ is easier than Putsch$_{de}$!

Very simple heuristics are very predictive

Raw sequence length / # predictions
→ char-RNNLM difficulty

Significant on:

- Europarl at $p < .01$
- Bibles at $p < .001$

i.e., for the char-RNNLM
$\text{pu\v{c}}_{\text{cz}}$ is easier than $\text{Putsch}_{\text{de}}$!

Raw vocabulary size
→ BPE-RNNLM difficulty

Significant on:

- not Europarl
- but Bibles at $p < .0000000001$

i.e., the BPE-RNNLM still suffers
if a language has high type-token-ratio!

Very simple heuristics are very predictive

Raw sequence length / # predictions
→ char-RNNLM difficulty

Significant on:

- Europarl at $p < .01$
- Bibles at $p < .001$

i.e., for the char-RNNLM
puč_{cz} is easier than $\text{Putsch}_{\text{de}}$!

Raw vocabulary size
→ BPE-RNNLM difficulty

Significant on:

- not Europarl
- but Bibles at $p < .0000000001$

i.e., the BPE-RNNLM still suffers
if a language has high type-token-ratio!

Wow! What is happening here? We have many conjectures...

Outline

"Difficulty" ✓

Models and languages ✓

What correlates with difficulty? ✓

And… is Translationese really easier?

Translationese: translations as a separate language?

Common assumption: Translationese is somehow simpler than "native" text.

Translationese: translations as a separate language?

Common assumption: Translationese is somehow simpler than "native" text.

We have partial parallel data that we can use to evaluate our models:

$\mathbf{en}_{\text{original}}$	$\mathbf{en}_{\text{translated}}$	$\mathbf{de}_{\text{original}}$	$\mathbf{de}_{\text{translated}}$	$\mathbf{nl}_{\text{original}}$	$\mathbf{nl}_{\text{translated}}$	...
Resumption...			Wiederauf-...		Hervatten...	...
The German...			Der deutsche...		De Duitse...	...
	Thank you...	Vielen Dank...			Hartelijk...	...
...	...	...	...	...	...	

Translationese: translations as a separate language?

Common assumption: Translationese is somehow simpler than “native” text.

We have partial parallel data that we can use to evaluate our models:

$\textbf{en}_{\textbf{original}}$	$\textbf{en}_{\textbf{translated}}$	$\textbf{de}_{\textbf{original}}$	$\textbf{de}_{\textbf{translated}}$	$\textbf{nl}_{\textbf{original}}$	$\textbf{nl}_{\textbf{translated}}$	…
Resumption...			Wiederauf-...		Hervatten...	…
The German...			Der deutsche...		De Duitse...	…
	Thank you...	Vielen Dank...			Hartelijk...	…
…	…	…	…	…	…	

...and indeed the original languages **seem** harder.

Translationese: translations as a separate language?

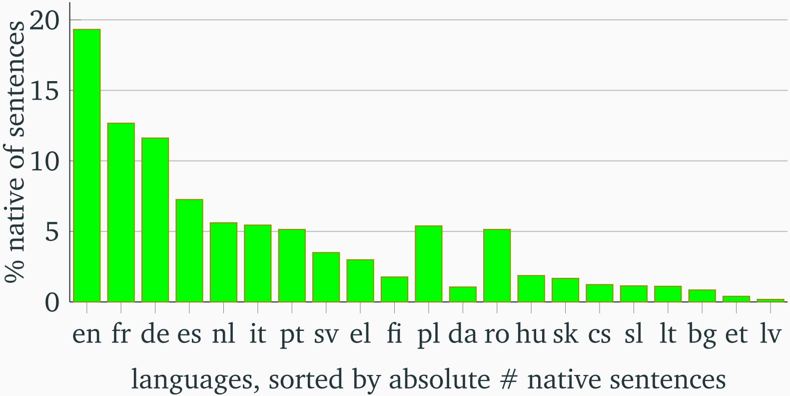
Common assumption: Translationese is somehow simpler than "native" text.

We have partial parallel data that we can use to evaluate our models:

$\mathbf{en}_{\text{original}}$	$\mathbf{en}_{\text{translated}}$	$\mathbf{de}_{\text{original}}$	$\mathbf{de}_{\text{translated}}$	$\mathbf{nl}_{\text{original}}$	$\mathbf{nl}_{\text{translated}}$	···
Resumption...			Wiederauf-...		Hervatten...	···
The German...			Der deutsche...		De Duitse...	···
	Thank you...	Vielen Dank...			Hartelijk...	···
···	···	···	···	···	···	

...and indeed the original languages seem harder. **But we missed something!**

We trained on mostly translationese!



Of course we will then find it easier...

Repeat the experiment with fairly balancing training data

Change the training sets!

We can **rebalance a single language**, leaving the others merged, i.e.:

en$_{\textbf{original}}$	**en**$_{\textbf{translated}}$	**de**	**nl**	...
Resumption...		Wiederauf-...	Hervatten...	...
~~The German...~~		~~Der deutsche...~~	~~De Duitse...~~	~~...~~
	Thank you...	Vielen Dank...	Hartelijk...	...
...	...	...	...	...

Repeat the experiment with fairly balancing training data

Change the training sets!

We can **rebalance a single language**, leaving the others merged, i.e.:

$\mathbf{en}_{\mathbf{original}}$	$\mathbf{en}_{\mathbf{translated}}$	**de**	**nl**	…
Resumption…		Wiederauf-…	Hervatten…	…
~~The German…~~		~~Der deutsche…~~	~~De Duitse…~~	~~…~~
	Thank you…	Vielen Dank…	Hartelijk…	…
…	…	…	…	…

And the result: the **difficulties are now the same**!

(more precisely, "native" is 0.004 ± 0.02 *easier*)

Conclusion: cross-linguistic comparisons are tricky (hope we didn't mess up!)

Conclusion: cross-linguistic comparisons are tricky (hope we didn’t mess up!)

1. Make sure your training data is comparable and fair.

Conclusion: cross-linguistic comparisons are tricky (hope we didn't mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.

Conclusion: cross-linguistic comparisons are tricky (hope we didn't mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.
3. Make sure your stats are fair (no p-hacking!).

Conclusion: cross-linguistic comparisons are tricky (hope we didn't mess up!)

1. Make sure your training data is comparable and fair.
2. Make sure your metrics are comparable and fair.
3. Make sure your stats are fair (no p-hacking!).
4. Work on more NLP resources for more languages!

What Kind of Language Is Hard to Language-Model?

ACL 2019

Sabrina J. Mielke *and* Ryan Cotterell, Kyle Gorman, Brian Roark, Jason Eisner

Johns Hopkins University // City University of New York Graduate Center // Google
`sjmielke@jhu.edu`

Twitter: `@sjmielke` – paper and thread pinned!