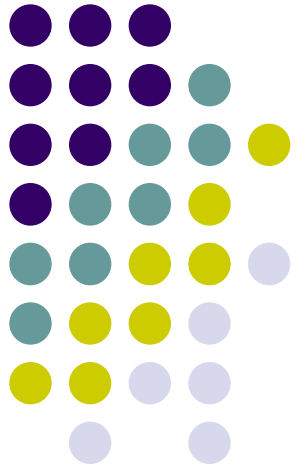
Improving Text Simplification Language Modeling Using Unsimplified Text Data

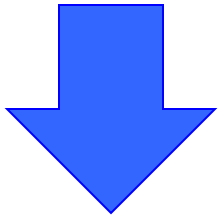
David Kauchak
Computer Science Department
Middlebury College
dkauchak@middlebury.edu



Text simplification

Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction.

- Albert Einstein ?



Simpler is better.

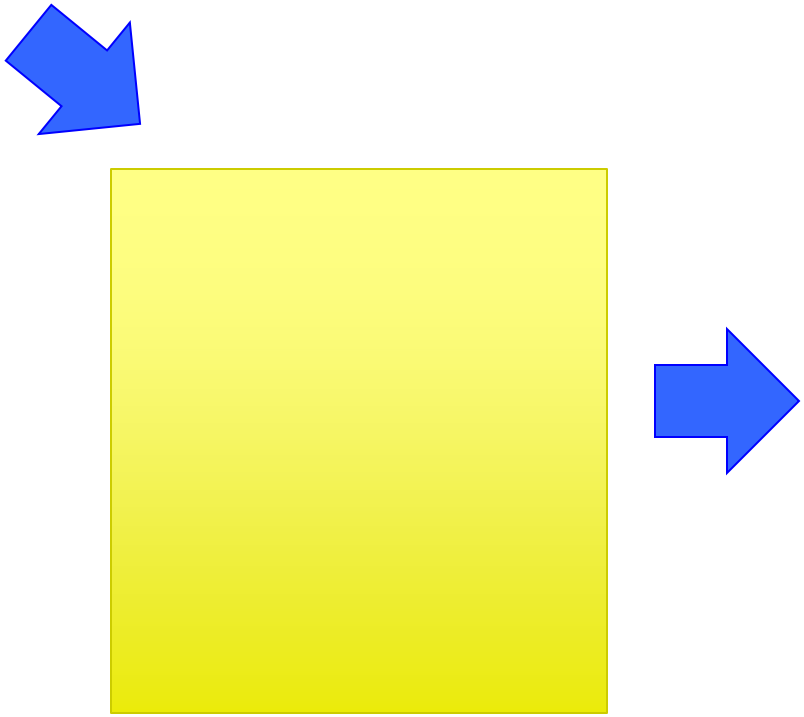
Simpler is better

Goal:

Reduce the reading complexity of text by incorporating more accessible vocabulary and structure while maintaining the content.

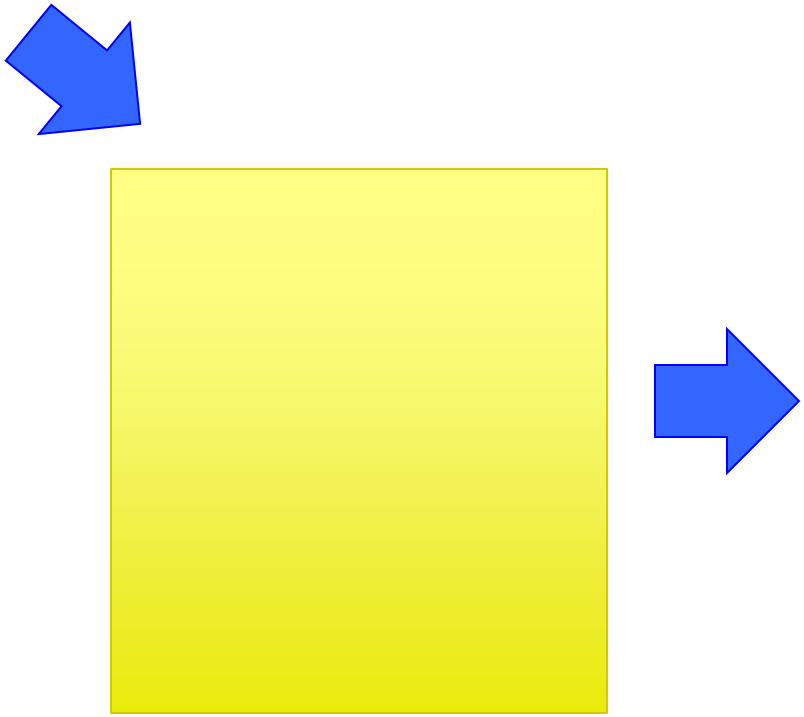
Text-to-text translation

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.



Text-to-text translation

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.



model

I do not like green eggs and ham.

Text-to-text translation

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.

translation model

language model

length model

…

model

I do not like green eggs and ham.

Text-to-text translation

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.

Models the likelihood of the output text

translation model

language model

length model

…

model

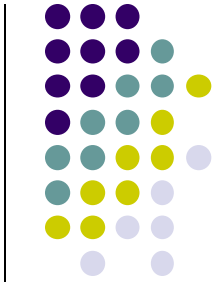
I do not like green eggs and ham.

Data availability

How much data is available to train a ***simple*** English language model?

~0.5 millions sentences

Data availability



How much data is available to train an English language model?

A lot more.

Idea: Just use unsimplified data

(normal) (simple)

Sentence aligned corpus
(137K sentence pairs)

simple n-grams found in normal Wikipedia

n-grams	simple→normal % overlap
1	96%
2	80%
3	68%
4	61%
5	55%

Idea: Just use unsimplified data

(normal) (simple)

Sentence aligned corpus
(137K sentence pairs)

simple n-grams found in normal Wikipedia

n-grams	simple→normal % overlap
1	**96%**
2	**80%**
3	68%
4	61%
5	55%

Good news:
- some reasonable overlap
- It's English

Idea: Just use unsimplified data

(normal) (simple)

Sentence aligned corpus
(137K sentence pairs)

simple n-grams found in normal Wikipedia

n-grams	simple→normal % overlap
1	96%
2	80%
3	68%
4	**61%**
5	**55%**

Possibly bad news: a lot of missing data!

Idea: Just use unsimplified data

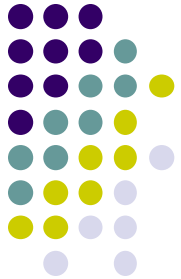
(normal) (simple)

Sentence aligned corpus
(137K sentence pairs)

n-grams	simple→normal	normal→simple
1	96%	87%
2	80%	68%
3	68%	58%
4	61%	51%
5	55%	46%

Bad news: different distributions over English!

Big questions



How do these distribution differences affect language modeling performance?

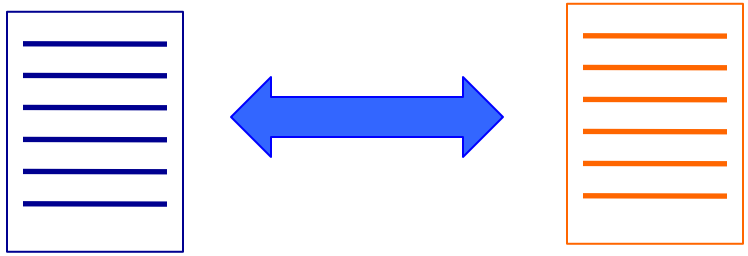
Is unsimplified data useful for simple language modeling?

What is the best way to utilize unsimplified data?

Document Aligned Corpus

en.wikipedia.org/wiki/**England**

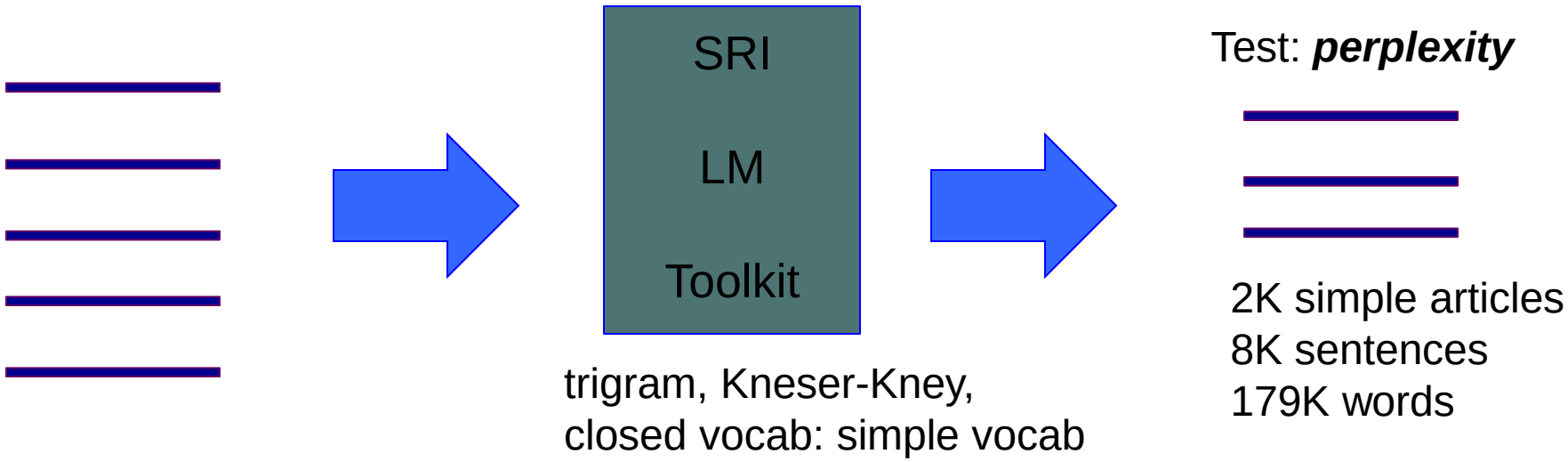
simple.wikipedia.org/wiki/**England**



Articles	60K	60K
Sentences	2540K	385K
Words	64.7M	7.15M
Vocab size	307K	78K

order of magnitude larger

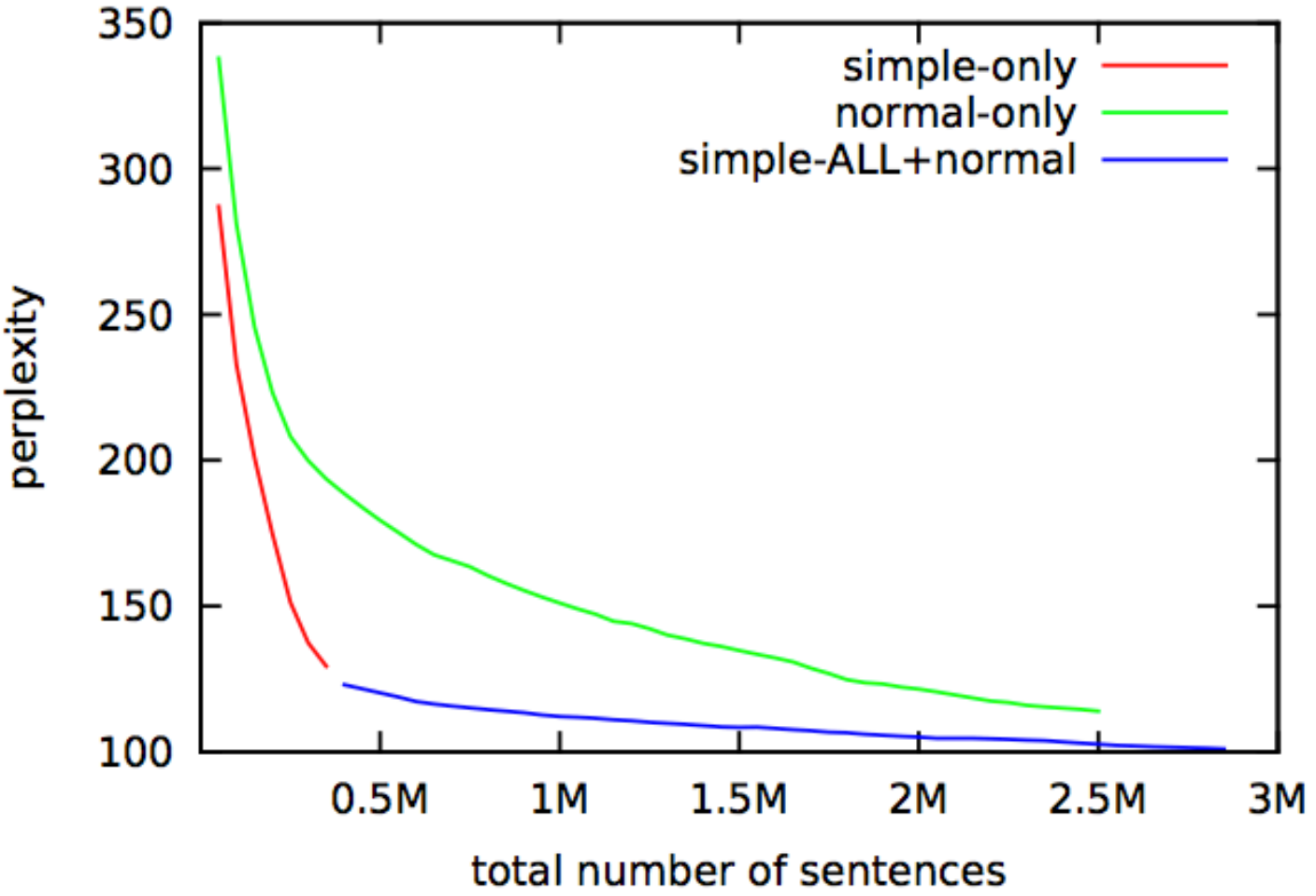
Task 1: LM perplexity



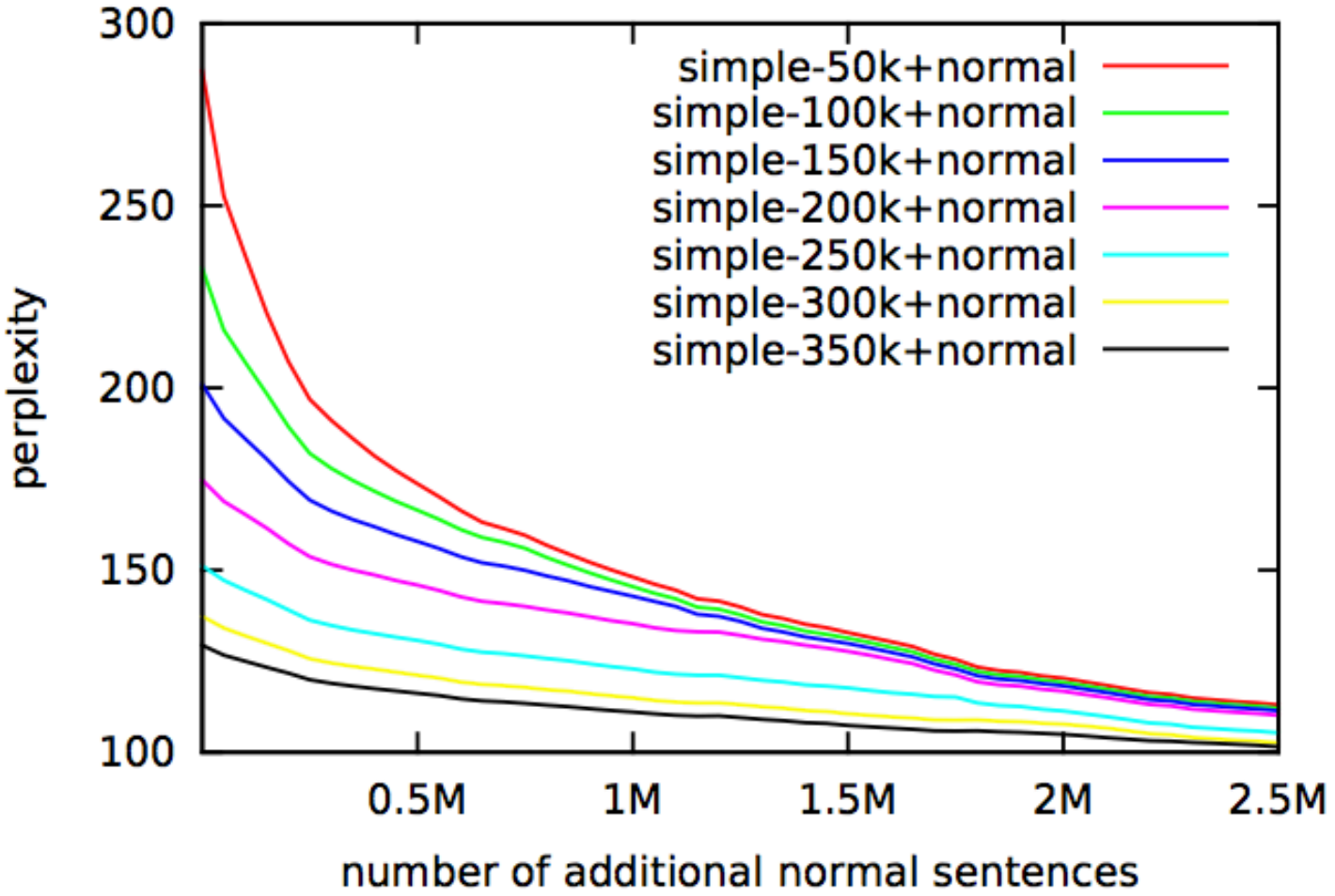
simple-only: simple English Wikipedia sentences

normal-only: English Wikipedia sentences

simple-*X*+normal: *X* simple sentences combined with varying amounts of normal sentences



23% improvement in perplexity by adding normal data to simple data



normal data helps even more if the simple data is limited

Language model adaptation

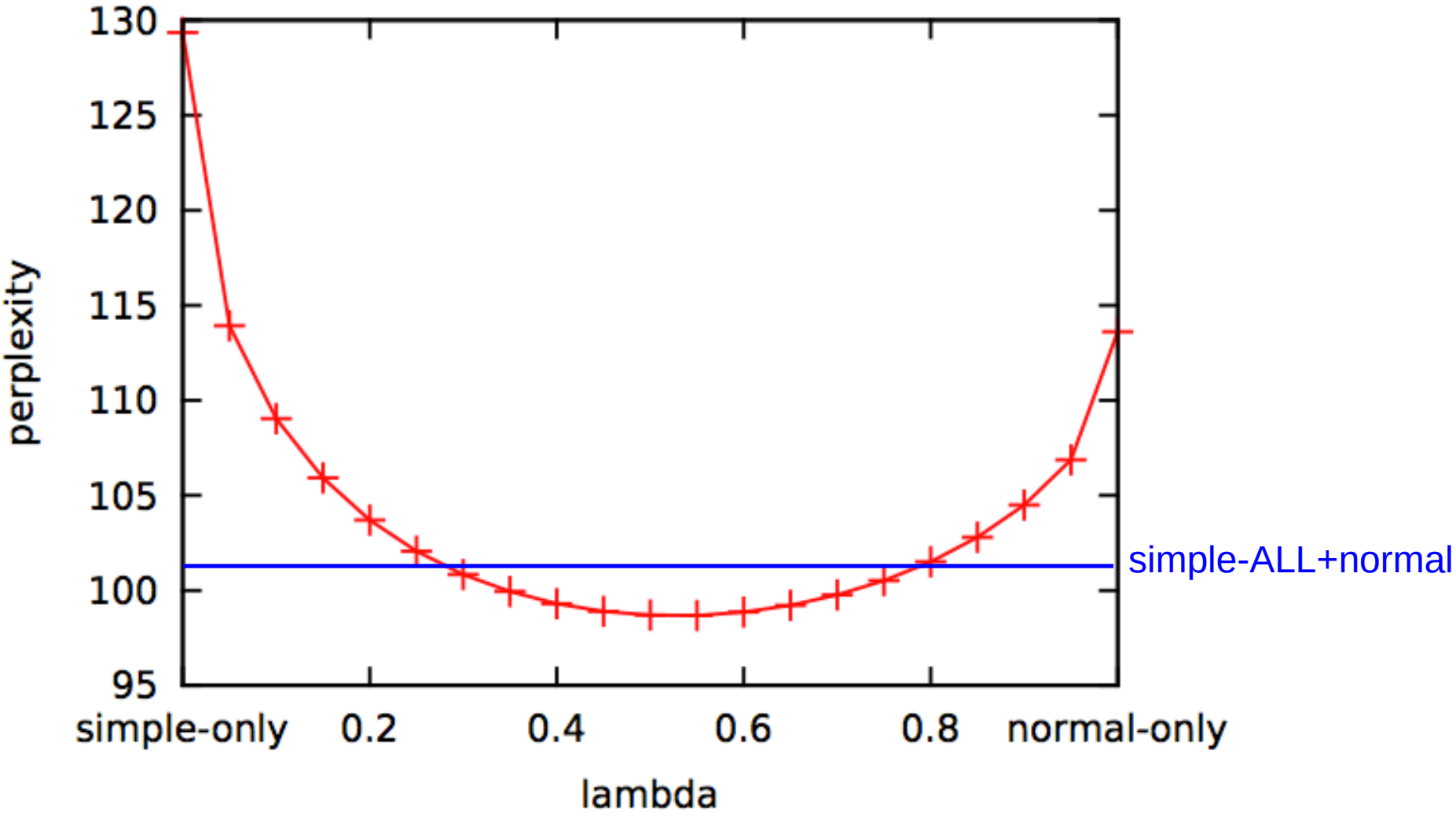
Linearly interpolated language model:

$$p_{\text{interpolated}}(w_i \mid w_{i-2} w_{i-1}) =$$

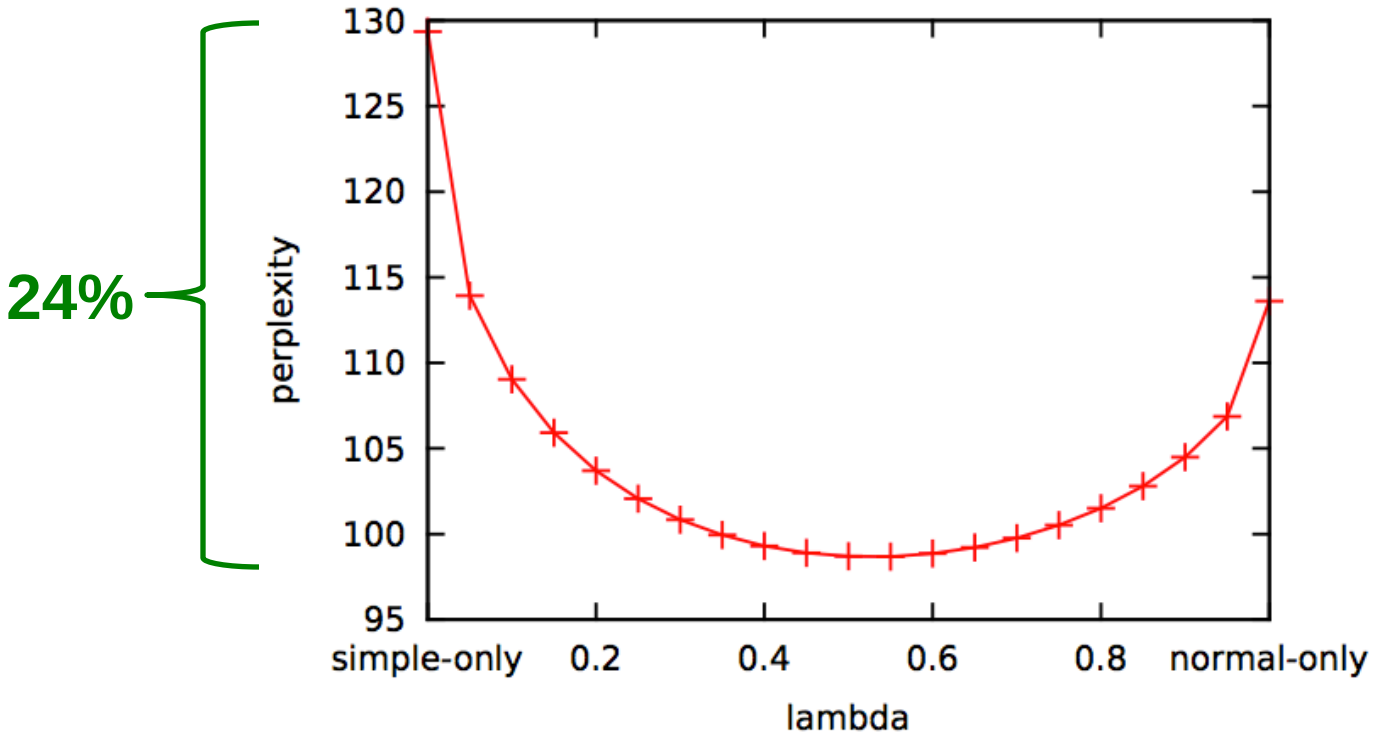
$$\lambda p_{normal}(w_i \mid w_{i-2} w_{i-1}) + (1-\lambda) p_{simple}(w_i \mid w_{i-2} w_{i-1})$$

normal-only **simple-only**

Language model adaptation



Task 1 summary



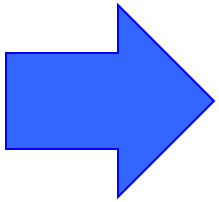
~24% improvement in perplexity over models trained with ALL available simple data by using normal data

Task 2: Lexical simplification

SemEval 2012 task:

With the physical market as ***tight*** as it has been in memory, silver could fly at any time.

Candidates
constricted
pressurised
low
high-strung
tight



Task: ranker

Human simplicity ranking
tight
low
constricted
pressurised
high-strung

Task 2: Lexical simplification

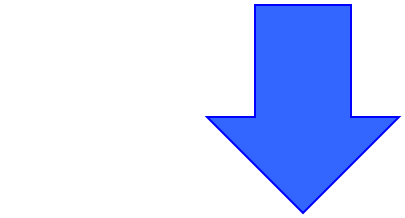
With the physical market as ***constricted*** as it has been …

With the physical market as ***pressurised*** as it has been ...

With the physical market as ***low*** as it has been …

With the physical market as ***high-strung*** as it has been ...

With the physical market as ***tight*** as it has been ...



Language Model

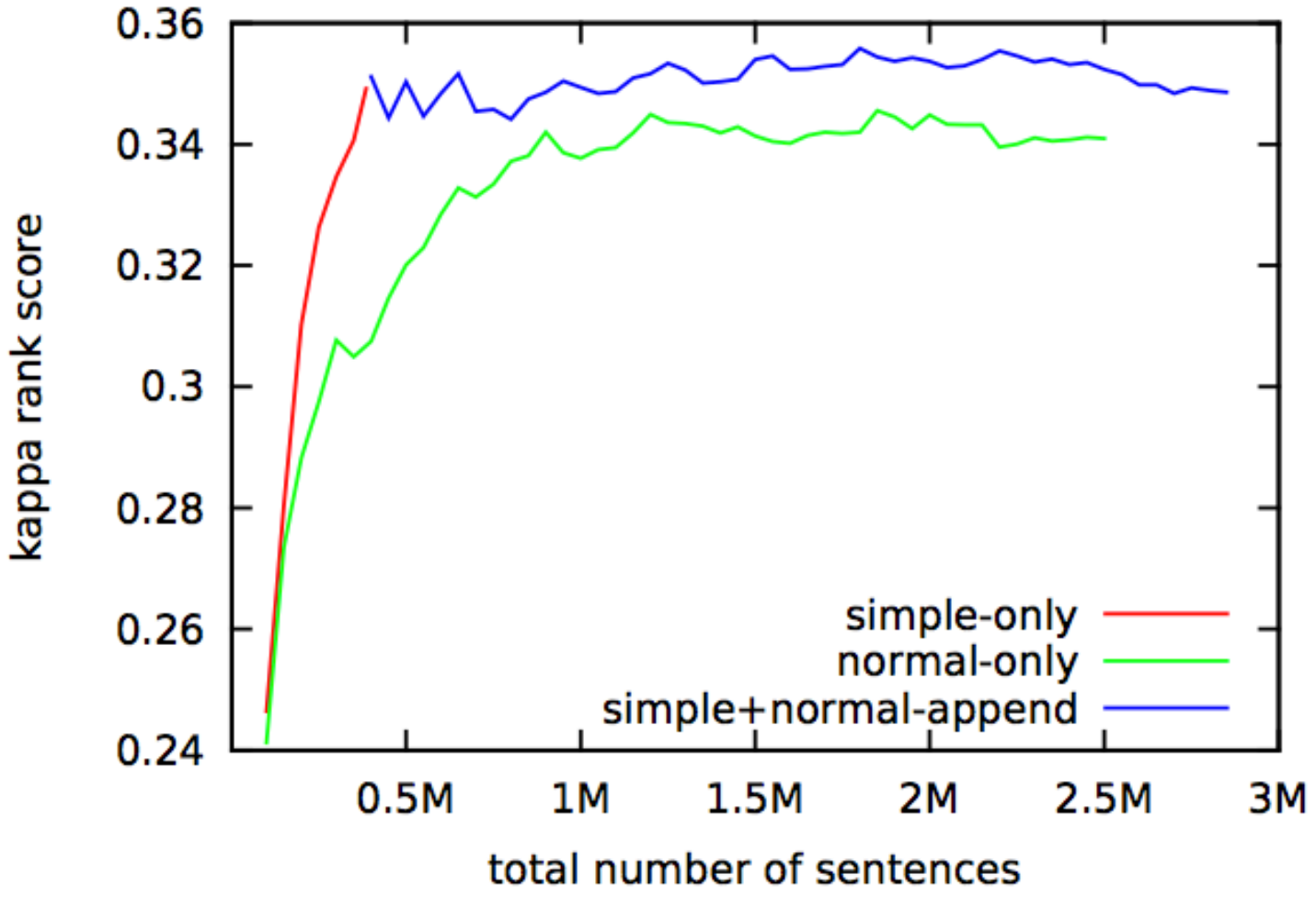
Rank by LM score

Task 2: Evaluation

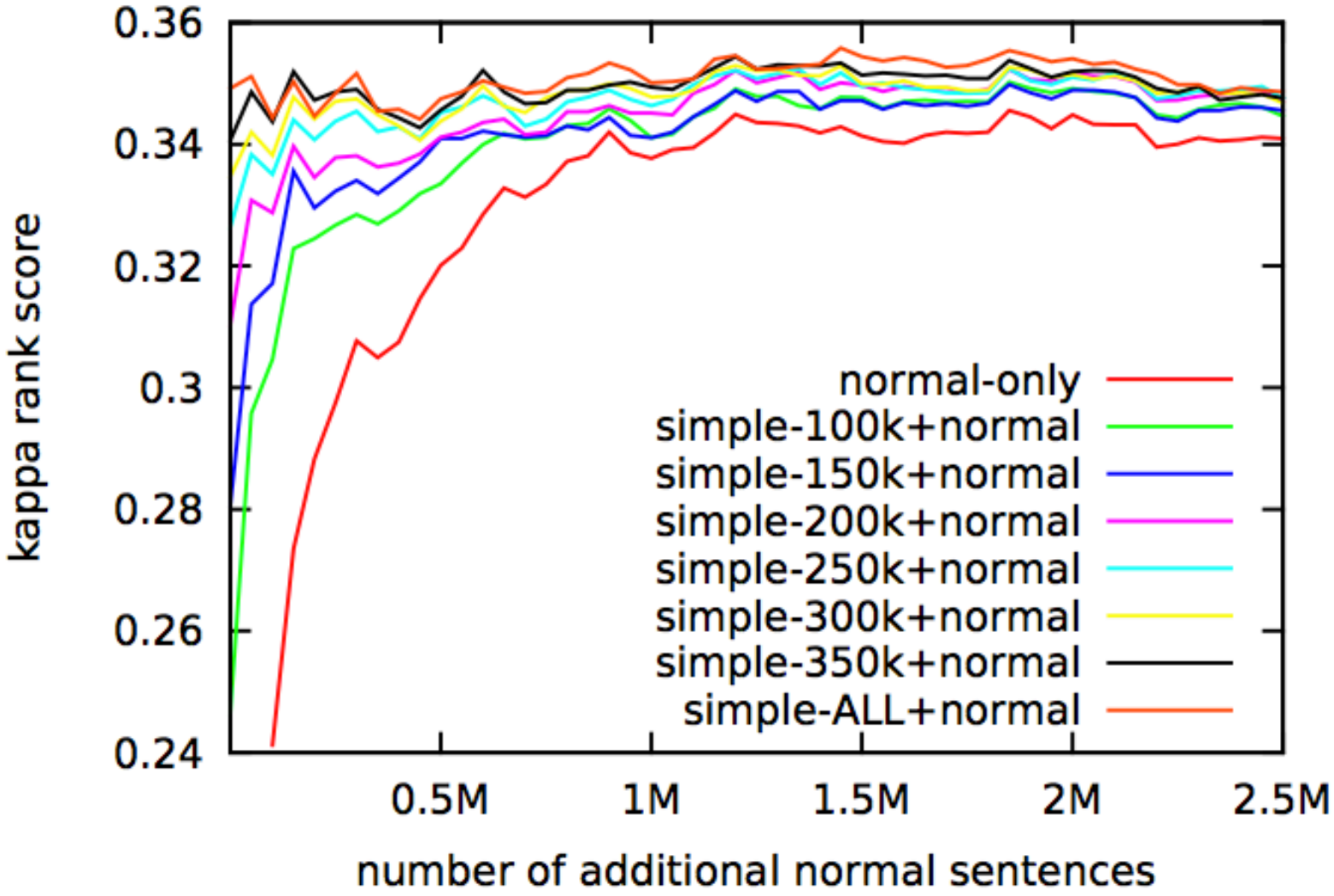
Human	System
tight	tight
low	low
constricted	constricted
pressurised	pressurised
high-strung	high-strung

kappa rank score: Cohen's kappa coefficient between the system ranking and the human ranking based on pairwise rank comparisons (evaluation used in the SemEval 2012 task)

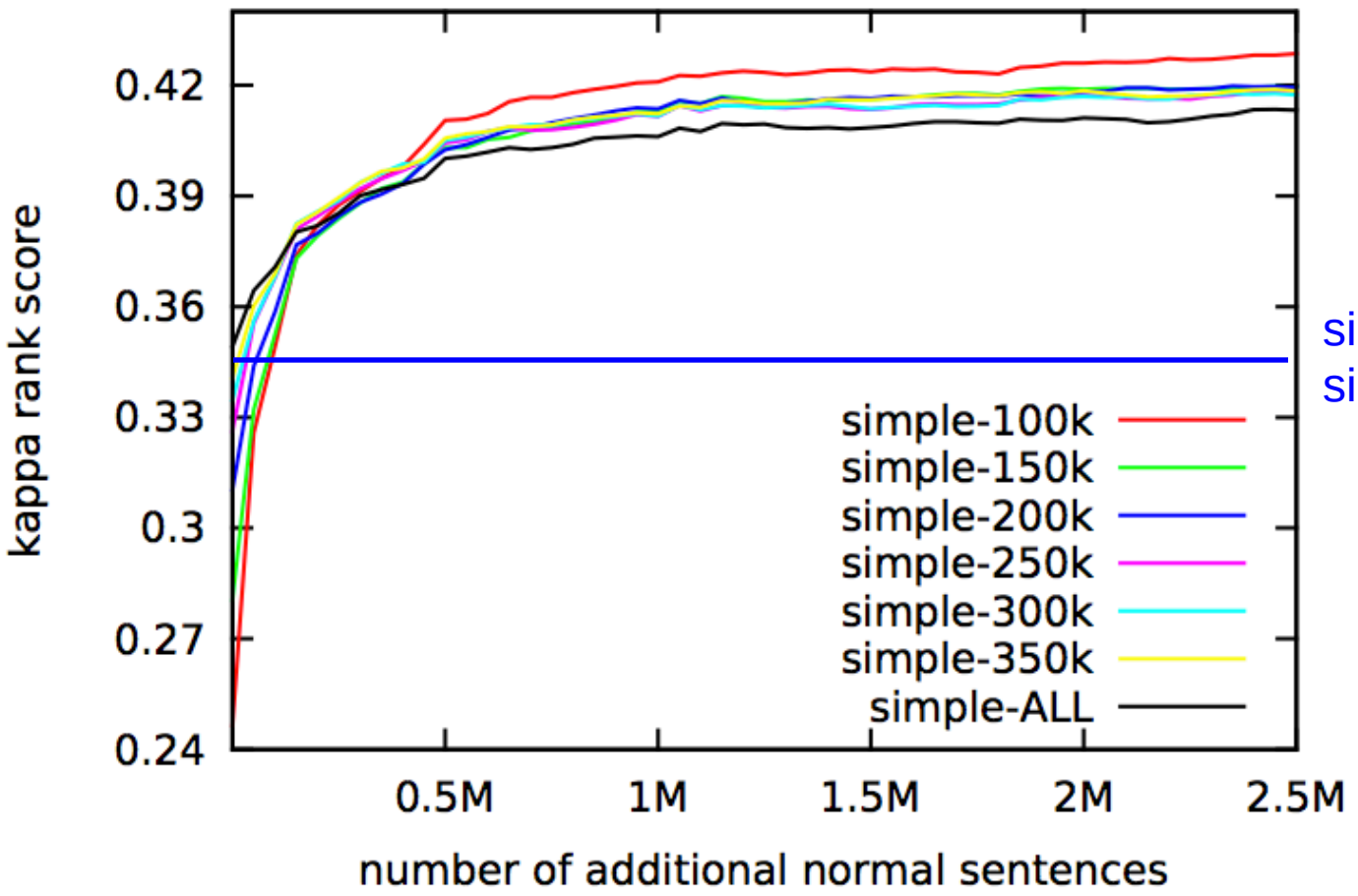
Lexical simplification results



Less simple data



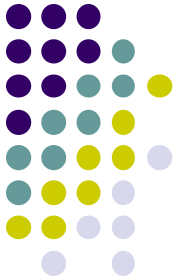
Language model adaptation



simple-only
simple-ALL+normal

23% improvement over simple-only model!

Why does normal data help?



Our guess: more *n*-grams

How many more *n*-grams are seen in normal data compared to simple?

Why does normal data help?

How many more *n*-grams are seen in normal data compared to simple?

	Perplexity test data	Lexical simplification data
	normal contains:	
unigrams	9.4% more	6.2% more
bigrams	24% more	56% more
trigrams	46% more	117% more

Application matters

Optimal λ (weighting between simple and normal) for linearly interpolated models:

Perplexity task	Lexical simplification task
λ = 0.5	**λ = 0.98**
An equal balance between simple and normal models	A very strong bias towards the simple model

Conclusions

Unsimplified data **is** useful for simple English language modeling

>23% improvement on both perplexity and lexical simplification tasks over model using *ALL simple data available*

LM domain adaption techniques are important, but are application specific

Data available:

http://www.cs.middlebury.edu/~dkauchak/simplification/

Open questions

How much unsimplified data can we utilize?

How does source/domain affect perfomance?

How does the LM quality affect other simplification applications (e.g. full sentence simplification)?

Better LM domain adaptation techniques.