Improving Text Simplification Language Modeling Using Unsimplified Text Data

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

I do not like green eggs and ham.
I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.

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

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

Sentence aligned corpus (137K sentence pairs)

<table>
<thead>
<tr>
<th>n-grams</th>
<th>simple ➔ normal % overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>68%</td>
</tr>
<tr>
<td>4</td>
<td>61%</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
</tr>
</tbody>
</table>

simple n-grams found in normal Wikipedia
Idea: Just use unsimplified data

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

Simple n-grams found in normal Wikipedia

<table>
<thead>
<tr>
<th>n-grams</th>
<th>Simple→Normal % overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>68%</td>
</tr>
<tr>
<td>4</td>
<td>61%</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
</tr>
</tbody>
</table>

Sentence aligned corpus (137K sentence pairs)
Idea: Just use unsimplified data

 simple n-grams found in normal Wikipedia

<table>
<thead>
<tr>
<th>n-grams</th>
<th>simple→normal % overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>68%</td>
</tr>
<tr>
<td>4</td>
<td>61%</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
</tr>
</tbody>
</table>

Possibly bad news: a lot of missing data!

Sentence aligned corpus (137K sentence pairs)
Idea: Just use unsimplified data

<table>
<thead>
<tr>
<th>n-grams</th>
<th>simple $\rightarrow$ normal</th>
<th>normal $\rightarrow$ simple</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96%</td>
<td>87%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
<td>68%</td>
</tr>
<tr>
<td>3</td>
<td>68%</td>
<td>58%</td>
</tr>
<tr>
<td>4</td>
<td>61%</td>
<td>51%</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Sentence aligned corpus (137K sentence pairs)

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

Test: perplexity

- 2K simple articles
- 8K sentences
- 179K words

SRI LM Toolkit

trigram, Kneser-Kney, closed vocab: simple vocab
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:

\[ p_{\text{interpolated}}(w_i | w_{i-2} w_{i-1}) = \lambda p_{\text{normal}}(w_i | w_{i-2} w_{i-1}) + (1 - \lambda) p_{\text{simple}}(w_i | w_{i-2} w_{i-1}) \]

normal-only \hspace{2cm} \text{simple-only}
Language model adaptation

![Graph showing the impact of lambda on perplexity for simple-ALL+normal adaptation. The graph illustrates a curve where perplexity decreases as lambda increases, reaching a minimum before increasing again.]
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

**Human simplicity ranking**

- tight
- low
- constricted
- pressurised
- high-strung

*Task: ranker*
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

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

![Graph showing kappa rank score vs. total number of sentences for different methods: simple-only, normal-only, and simple+normal-append. The graph demonstrates the performance of each method across various sentence counts.]
Less simple data
Language model adaptation

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?

<table>
<thead>
<tr>
<th>$n$-grams</th>
<th>Perplexity test data</th>
<th>Lexical simplification data</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigrams</td>
<td>9.4% more</td>
<td>6.2% more</td>
</tr>
<tr>
<td>bigrams</td>
<td>24% more</td>
<td>56% more</td>
</tr>
<tr>
<td>trigrams</td>
<td>46% more</td>
<td>117% more</td>
</tr>
</tbody>
</table>
Application matters

Optimal $\lambda$ (weighting between simple and normal) for linearly interpolated models:

<table>
<thead>
<tr>
<th>Perplexity task</th>
<th>Lexical simplification task</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0.5$</td>
<td>$\lambda = 0.98$</td>
</tr>
<tr>
<td>An equal balance between simple and normal models</td>
<td>A very strong bias towards the simple model</td>
</tr>
</tbody>
</table>
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 performance?

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

Better LM domain adaptation techniques.